Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной Международной конференции «Диалог» (2011)

Выпуск 10

Computational Linguistics and Intellectual Technologies

Papers from the Annual International Conference “Dialogue” (2011)

Issue 10

УДК 80/81; 004
ББК 81.1
К63

Программный комитет конференции выражает
искреннюю благодарность Российскому фонду фундаментальных
исследований за финансовую поддержку,
грант № 11-06-06056-г

К63 **Компьютерная** лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 25–29 мая 2011 г.). Вып. 10 (17). — М.: РГГУ, 2011.

Сборник включает 73 доклада международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2011», представляющих широкий спектр теоретических и прикладных исследований в области описания естественного языка, моделирования языковых процессов, создания практически применимых компьютерных лингвистических технологий.

Для специалистов в области теоретической и прикладной лингвистики и интеллектуальных технологий.

УДК 80/81; 004
ББК 81.1

Предисловие

10-ый выпуск ежегодника «Компьютерная лингвистика и интеллектуальные технологии» содержит материалы 17-й Международной конференции «Диалог». Для сборника было отобрано 73 доклада, охватывающих наиболее актуальные направления теоретической и прикладной лингвистики, связанные с компьютерным анализом естественного языка. В настоящем сборнике представлены:

- Лингвистическая семантика и семантический анализ
- Формальные модели языка и их применение
- Теоретическая и компьютерная лексикография
- Создание и применение универсальных компьютерных лексических ресурсов
- Методы оценки (evaluation) систем и методов анализа текстов
- Корпусная лингвистика; создание, применение, оценка корпусов
- Интернет как лингвистический ресурс; лингвистические технологии в Интернете
- Извлечение знаний из текстов
- Компьютерный анализ документов: реферирование, классификация, поиск
- Машинный перевод
- Модели общения; коммуникация, диалог и речевой акт
- Анализ и синтез речи.

«Диалог» является наиболее крупной российской конференцией по компьютерной лингвистике.

Принципиальной особенностью конференции является особое внимание к лингвистически ориентированным подходам к решению задач автоматического анализа языка. Именно этим объясняется и состав участников, и программа конференции, в которой соседствуют доклады теоретического и прикладного характера.

Традиционно важное место в программе «Диалога» занимают исследования звучащей речи, коммуникативных стратегий, невербальных компонентов процесса общения.

Каждый год Программный комитет выбирает отдельные темы или направления в качестве доминант очередной конференции. Им посвящаются специальные заседания, круглые столы, обзорные выступления приглашенных докладчиков. В этом году в центре обсуждений проблемы корпусометрии. Можно сказать, что тема анализа текстовых корпусов не была выбрана Программным Комитетом, а оказалась в центре внимания естественным образом. Практически всякое лингвистическое исследование ведется сегодня с привлечением корпусных данных. При этом далеко не всегда четко формулируется, какими

свойствами должен обладать корпус и методы работы с ним, чтобы полученные результаты заслуживали доверие.

Начиная с этого года сборник становится принципиально двуязычным. Это отражает одинаковую важность для «Диалога» двух взаимосвязанных задач:

- Создание ресурсов, моделей и технологий для поддержки анализа русского языка. Это становится особенно важным в связи с тем, что сегодня русский язык де-факто не входит в список языков, которым уделяется существенное внимание в мировой компьютерной лингвистике;
- Преодоление того методического и технологического отставания, которое, к сожалению, характерно для российской компьютерной лингвистики, несмотря на ее достижения.

Для успешного решения второй задачи Программный комитет «Диалога» пытается внедрить международные стандарты оценки присылаемых работ, этой же цели служит и выбор английского языка в качестве рабочего для тех направлений «Диалога», которые относятся к мировому технологическому мейнстриму. Это дает, в частности, важную возможность привлекать к отбору докладов и иностранных экспертов.

Тематика «Диалога» существенно шире, чем может продемонстрировать данный выпуск. Более цельную картину можно получить на сайте конференции www.dialog-21.ru , где представлены обширные электронные архивы прошлых лет и форумы по основным направлениям «Диалога».

Программный комитет конференции «Диалог»
Редколлегия ежегодника «Компьютерная лингвистика и интеллектуальные технологии»

Организаторы

Ежегодная конференция «Диалог» проводится под патронажем Российского Фонда Фундаментальных Исследований при организационной поддержке компании ABBYY.

Основными учредителями конференции являются:

- Институт лингвистики РГГУ
- Институт проблем информатики РАН
- Институт проблем передачи информации РАН
- Компания ABBYY
- Компания Яndex
- Филологический факультет МГУ

Конференция проводится при поддержке Российской ассоциации искусственного интеллекта.

Международный программный комитет

Буате Кристиан	Гренобльский университет
Богуславский Игорь Михайлович	Политехнический университет Мадрида
Гельбух Александр Феликсович	Национальный политехнический институт, Мехико
Иомдин Леонид Лейбович	Институт проблем передачи информации РАН
Кибрик Александр Евгеньевич	Филологический факультет МГУ
Кобозева Ирина Михайловна	Филологический факультет МГУ
Козеренко Елена Борисовна	Институт проблем информатики РАН
Кронгауз Максим Анисимович	Институт лингвистики РГГУ
Лукашевич Наталья Валентиновна	НИВЦ МГУ
Мельчук Игорь Александрович	Монреальский университет
Нивре Йоаким	Уппсальский университет
Ниренбург Сергей	Университет Нью-Мексико
Осипов Геннадий Семёнович	Институт программных систем РАН
Попов Эдуард Викторович	РосНИИ информационной техники и САПР
Сегалович Илья Валентинович	Компания Яndex
Селегей Владимир Павлович	Компания ABBYY
Флюр-Семёнова Вера	Компания SCIPER
Шаров Сергей	University of Leeds, UK
Ыйм Халдур	Тартуский университет

Организационный комитет и Редсовет

Селегей Владимир Павлович, *председатель*	Компания ABBYY
Азарова Ирина Владимировна	Санкт-Петербургский государственный университет
Беликов Владимир Иванович	Институт русского языка им. В. В. Виноградова РАН
Добров Борис Викторович	НИВЦ МГУ
Иомдин Леонид Лейбович	Институт проблем передачи информации РАН
Лауфер Наталия Исаевна	ООО «проФан Продакшн»
Ляшевская Ольга Николаевна	Universitetet i Tromsø, Norway
Соколова Елена Григорьевна	РосНИИ искусственного интеллекта
Толдова Светлана Юрьевна	Филологический факультет МГУ

Секретариат

Талис Валентина Львовна, *секретарь оргкомитета, редактор сайта*	Компания ABBYY
Мытникова Татьяна Александровна, *координатор*	Компания ABBYY

Рецензенты

Августинова Таня
Азарова Ирина Владимировна
Апресян Валентина Юрьевна
Баранов Анатолий Николаевич
Беликов Владимир Иванович
Богданов Алексей Владимирович
Богданова Наталья Викторовна
Богуславский Игорь Михайлович
Борщев Владимир Борисович
Браславский Павел Исаакович
Губин Максим Вадимович
Добров Борис Викторович
Добровольский Дмитрий Олегович
Зарецкая Елена Наумовна
Захаров Леонид Михайлович
Зуев Константин Алексеевич
Иомдин Борис Леонидович
Иомдин Леонид Лейбович
Кибрик Андрей Александрович
Кобозева Ирина Михайловна
Козеренко Елена Борисовна
Крейдлин Григорий Ефимович
Кронгауз Максим Анисимович
Левонтина Ирина Борисовна
Лобанов Борис Мефодьевич
Лукашевич Наталья Валентиновна
Ляшевская Ольга Николаевна
Пазельская Анна Германовна
Подлесская Вера Исааковна
Ронжин Андрей Леонидович
Савельев Василий Евгеньевич
Сегалович Илья Валентинович
Селегей Владимир Павлович
Сокирко Алексей Викторович
Соколова Елена Григорьевна
Старостин Анатолий Сергеевич
Тестелец Яков Георгиевич
Тихомиров Илья Александрович
Толдова Светлана Юрьевна
Урысон Елена Владимировна
Филиппова Екатерина Александровна
Циммерлинг Антон Владимирович
Шаров Сергей Александрович
Янко Татьяна Евгеньевна

Contents*

Section I. Guest reports

Corbett Greville G.
Lexical Splits and Morphological Complexity 1

Hovy Eduard
A New Semantics: Merging Propositional and Distributional Information 3

Kibrik A. E.
The Basis of Natural Human Language and its Main Parameters 4

McCarthy Diana
Exploiting Distributional Similarity for Lexical Acquisition 19

Section II. Main program of the conference

Alekseev A. A., Loukachevitch N. V.
Automatic Detection of Near-Synonyms in News Clusters 32

Avgustinova T.
Parallel Construction of Slavic Grammatical Resources 42

Baranov A. N., Dobrovol'skii D. O.
Semantic Relations in Phraseology 53

Belikov V. I.
What are Sociolinguists and Lexicographers Lacking in a Digitized World? 63

Benigni V., Cotta Ramusino P.
Italian Constructions with Support Verb Fare in Comparison with Russian 72

Berdichevskii A.
E-mail vs. Chat: the Influence of the Communication Channel on the Language 89

Bergel'son M. B.
Modern Russian Public Discourse: Do Changes in Information Technology Lead to New Discourse Strategies, or to New Worldview? 99

Bocharov V., Bichineva S., Granovskii D., Ostapuk N., Stepanova M.
Quality Assurance Tools in the OpenCorpora Project 107

* The reports of each section are ordered by the surname of the first author in compliance with the English alphabet.

Bogdanova N. V., Os'mak N. A.
Some Lexical "Discoveries" on the Material of Russian Spontaneous Speech, a Corpus Study 116

Bol'shakov I. A., Gel'bukh A. F.
A Large Electronic Dictionary as a Polythematic Guide and Shaper of Queries to the Web 131

Boriskina O. O.
A Corpus-based Study of Noun Cryptotypes in English 142

Borisova E. G., Ovchinnikova T. E.
Parameter of Nearness in the Metaphorical Space 153

Braslavskii P., Kiselev Iu.
To Find Out or to Buy? Product Review vs. Web Shop Classifier 160

Bylinina E. G.
"Functional"Standard in Russian and English Degree Constructions 169

Chetverkin I. I., Loukachevitch N. V.
Three-way Movie Review Classification 177

Davydov A. G., Kiselev V. V., Kochetkov D. S.
Voice Emotion Classification: Problems and Solutions 187

Erekhinskaia T. N., Titova A. S., Okat'ev V. V.
Syntax Parsing for Texts with Misspellings in Dictascope Syntax 196

Fedorova O. V., Uspenskaia A. M.
Experimental Analysis of Discourse: the Impact of a Potential Referential Conflict on the Choice of the Referring Expression (on the Material of Russian) 207

Frolova T., Podlesskaia O.
Tagging Lexical Functions in Russian Texts of SynTagRus 219

Giliarova K. A.
Characteristics of Student-Professor E-mail Communication 232

Grashchenkov P., Ionov M., Maliutina S.
Semi-tagged Corpora Method Exemplified with a Study of Ossetic Nominalization 250

Grishina E. A.
Multimodal Clusters in Spoken Russian 258

Iagunova E. V., Pivovarova L. M.
A Study of the News Text Structure as a Consequence of Connected Segments 274

Ianko T. E.
Accent Placement Principles in Russian 289

Iomdin B. L., Piperski A. Ch., Russo M. M., Somin A. A.
How Different Languages Categorize Everyday Items 303

Iomdin L. L., Lobanov B. M., Getsevich Iu. S.
The Talking ETAP. Using the ETAP Parser in Russian Speech Synthesis 315

Karpenko M. P., Protasov S. V.
Some Methods for Language Model Pruning 327

Karpova O. S., Rakhilina E. V., Reznikova T. I., Ryzhova D. A.
Meaning of Estimation in Semantic Shifts of Rebranding Type in Adjectives and Adverbs (on the Material of the Database of Semantic Shifts in Russian Adjectives and Adverbs) 340

Kiseleva K. L.
Antonyms in Phraseology: Formal Similarity as a Condition of a Semantic Oppositeness 354

Kotov A. A.
Types of Simulated Emotional Expressive States in the Russian Emotional Corpus 364

Kozerenko A. D.
Gesture Idioms and Gestures: Types of Correspondence 375

Kozerenko E. B.
Linguistic Motivation for Statistical Translation Models 384

Kreidlin G. E.
Nonverbal Dialogue in the History of Kinesics 401

Kriuchkova O. Iu., Gol'din V. E.
A Corpus of Russian Dialectal Speech: the Concept and Parameters of Evaluation 412

Kudinov A. S., Voropaev A. A., Kalinin A. L.
A High Precision Method for the Recognition of Sentence Boundaries 422

Kustova G. I.
Constructions with Abstract Nouns in an Electronic Database 434

Kuznetsov I. P.
Identifying Role Functions of People on the Basis of Knowledge Structures 447

Letuchii A. B.
Pronominalization of Sentential Arguments in Russian 460

Levontina I. B.
On Some Non-Assertive Verbs 472

Litvinenko A. O.
Speech Reporting Strategies in Russian Comic-Based Stories 484

Lobanov B. M., Getsevich Iu. S.
Statistical Characteristics of Syntagmatic Segmentation of Utterances from the Viewpoint of Expressive Text-To-Speech Synthesis 494

Logacheva V. K., Klyshinskii E. S.
Non-stochastic Learning of Cross-language Transliteration Rules from a Small Dataset 509

Loukachevitch N. V., Dobrov G. B., Kibrik A. A., Khudiakova M. V., Linnik A. S.
Factors of Referential Choice: Computational Modeling 519

Lukashevich N. Iu., Kobozeva I. M.
Character Nominations in Ontological Perspective 530

Liudovyk T. V., Pylypenko V. V., Robeiko V. V.
Automatic Recognition of Spontaneous Ukrainian Speech Based on the Ukrainian Broadcast Speech Corpus 541

Nikolaeva Iu.
Illustrative Gestures as Markers for Discourse Macrostructure 553

Paducheva E. V.
Meanings, Diatheses and Ontological Categories of the Russian Word Vpechatlenie 'Impression' 560

Pazel'skaia A. G., Solov'ev A. N.
A method of Sentiment Analysis in Russian Texts 575

Piperski A. Ch.
Generic Terms in Everyday Vocabulary as a Sphere of Subtle Differences Between Serbian and Croatian 588

Podlesskaia V. I.
Relative Clauses in Spoken Russian and Elsewhere: a Corpus Approach 594

Potemkin S. B., Kedrova G. E.
Exploring Semantic Orientation of Adverbs 603

Renkovskaia E. A.
Some Peculiarities of the Syntactic Structure of Russian Proverbs: a Study of One-Predicate Sentences 610

Romanov A. S., Meshcheriakov R. V.
Gender Identification of the Author of a Short Message 621

Savchuk S. O.
A Corpus-based Study of Morphological Variability: Variation of Gender Forms of Russian Nouns 628

Seryi A. S., Sidorova E. A.
Object Identification in Problem of Automatic Document Processing 646

Sharov S., Nivre J.
The Proper Place of Men and Machines in Language Technology. Processing Russian without any Linguistic Knowledge 658

Sizov V. G., Podlesskaia O. Iu.
Reflecting Accentuation in the Russian Morphological Dictionary of the Multifunctional Linguistic Processor ETAP-3 672

Skatov D., Liverko S.
Anaphora Resolution of the Third-person Pronoun in Texts from Narrow Subject Domains with Grammatical Errors and Mistypings 686

Smirnova N. S., Chistikov P. G.
Software for Automated Statistical Analysis of Phonetic Units Frequency in Russian texts and its Application for Speech Technology Tasks 699

Sokolova E. G., Semenova S. Iu., Zagorul'ko Iu. A., Zakharov V. P., Kononenko I. S., Krivnova O. F.
Selection and Preparation of Terms for the Russian-English Thesaurus of Computational Linguistics 712

Testelets Ia. G.
Case as a Characteristic of Identity under Ellipsis in Russian 726

Trub V. M.
On the Dynamic Semantics of the Word Mif 'Myth' 738

Uryson E. V.
Concessive Conjunction Khotia 'Though' and "Cancelled Expectation" 747

Voznesenskaia M. M.
Enantiosemy in Russian Phraseology 760

Zalizniak A. A., Mikaelian I. L.
On one Use of Simple Imperfectives in Russian 769

Zevakhina N. A.
Exclamatives in Russian: a Corpus Study 782

Zimmerling A.
Scrambling Types in the Slavic Languages 796

Section III. Other areas of the "Dialogue"

Mikheev M. Iu.
Multiple Narrators in Varlam Shalamov's Texts 813

Shmeleva E. Ia., Shmelev A. D.
Interlingual Puns in Russian Jokes 829

Abstracts 837

Index 856

Section I.
Guest reports

LEXICAL SPLITS AND MORPHOLOGICAL COMPLEXITY

Greville G. Corbett (G.Corbett@surrey.ac.uk)

Surrey Morphology Group
University of Surrey,
Guildford, Surrey, United Kingdom

Key words: lexical splits, possible word, possible lexical splits, typology

A key notion in understanding and modelling language is 'possible word'. While some words (lexemes) are internally homogeneous and externally consistent, we find others with splits in their internal structure (morphology) and inconsistencies in their external behaviour (syntactic requirements). I begin with the characteristics of the simplest lexemes, adopting the approach of Canonical Typology. In this approach, we push our definitions to the logical limit, in order to establish a point in the theoretical space from which we can calibrate the real examples we find. Defining canonical inflection, allows us to schematize the interesting phenomena which deviate from this idealization. These include suppletion, syncretism, deponency and defectiveness. I then look at the different ways in which lexemes are 'split' by these phenomena. Consider the French verb *aller* 'go', which is split by suppletion. Some of its forms are based on the stem *all-* (as in *allons*), some on *v-* (as in *vont*) and some on *ir-* (as in *irons*). This example demonstrates that a lexeme's forms need not have any phonology in common. From this point of view, the split is as radical as it could be; in certain other respects there are more remarkable examples. I therefore set out a typology of possible lexical splits, along four dimensions:

1) form versus composition/structure of the paradigm: in the French suppletion example the split concerns forms only and does not affect the structure of the paradigm; contrast this with the deeper split in the Russian verb, where different segments of the paradigm are sensitive to different featural requirements (gender is marked in the past: but not in the present).
2) motivated versus morphology-internal (morphomic): the Russian split follows a boundary which is motivated from outside the paradigm (it follows tense), while the French split is purely morphology-internal.

3) regular versus irregular: splits may be fully regular, extending across the lexicon (all Russian verbs share the featural split), or they may be lexically specified, as we find in Archi (Daghestanian), where particular cells of individual personal pronouns must be specified as taking agreement (while the remaining cells do not).
4) externally relevant versus irrelevant: we would expect such splits to be internal to the lexeme, as with English *go~went*, but some have external relevance, in that they lead to different syntactic requirements. Instances include the different alignments found with certain tense-aspect-mood forms in Georgian, and lexemes whose splits bring with them different gender values, as in Czech, Scots Gaelic and the Tromsø dialect of Norwegian.

Our typology specifies these four dimensions independently. They are orthogonal to each other, so that the unexpected patterns of behaviour may cooccur in particular lexemes, giving rise to some remarkable examples, especially where periphrasis is involved. These examples show that the notion 'possible word' is challenging for theoretical and for applied work. And, given how unlikely some some of the theoretical combinations appeared, the typology proves remarkably complete.

A NEW SEMANTICS: MERGING PROPOSITIONAL AND DISTRIBUTIONAL INFORMATION

Eduard Hovy (hovy@isi.edu)

Information Sciences Institute, University of Southern California
Marina del Rey, Southern California

Key words: semantic content, NLP, distributional semantics, distribution information.

Despite hundreds of years of study on semantics, theories and representations of semantic *content* — the actual meaning of the symbols used in semantic propositions — remain impoverished. The traditional extensional and intensional models of semantics are difficult to actually flesh out in practice, and no large-scale models of this kind exist. Recently, researchers in Natural Language Processing (NLP) have increasingly treated *topic signature word distributions* (also called 'context vectors', 'topic models', 'language models', etc.) as a de facto placeholder for semantics at various levels of granularity. This talk argues for a new kind of semantics that combines traditional symbolic logic-based proposition-style semantics (of the kind used in older NLP) with (computation-based) statistical word distribution information (what is being called Distributional Semantics in modern NLP). The core resource is a single lexico-semantic 'lexicon' that can be used for a variety of tasks. I show how to define such a lexicon, how to build and format it using tensors, and how to use it for various tasks. I discuss some of the recent work on composing vectors and tensors in attempts to produce statistically-based compositional semantics. Combining the two views of semantics opens many fascinating questions that beg study, including the operation of logical operators such as negation and modalities over word(sense) distributions, the nature of ontological facets required to define concepts, and the action of compositionality over statistical concepts.

БАЗА ЕСТЕСТВЕННОГО ЧЕЛОВЕЧЕСКОГО ЯЗЫКА И ЕЕ ОСНОВНЫЕ ПАРАМЕТРЫ

А. Е. Кибрик (aekibrik@gmail.com)

Филологический факультет МГУ (ОТиПЛ), Москва, Россия

В статье затрагиваются некоторые, но заведомо не все базовые свойства языка. Речь будет идти о языке и знаковых системах, о базовых (обязательных) и вторичных (возможных) языковых функциях, об основных социальных регистрах языка, об умирающих языках и истории изменения языков, о причинах множественности языков.

Ключевые слова: знаковые системы, языковые функции, регистры языка, изменение языка, множественность языков

THE BASIS OF NATURAL HUMAN LANGUAGE AND ITS MAIN PARAMETERS

A. E. Kibrik (aekibrik@gmail.com)

Lomonosov Moscow State University, Moscow, Russian Federation

The paper discusses some, although not all, basic properties of language. I discuss language and sign systems (symbolic signs, indexical signs, iconic signs), as well as the functions of language, including the primary (epistemic, cognitive, and communicative) and the secondary ones (the functions of: social solidarity, individuation, support of social comfort, getting in contact [phatic]; the aesthetic function, the fascination function, the emotional function, and the metalinguistic function). I also treat the main social registers of language (idiolect, subdialect, dialect, language, literary language) and the issues of language death, language change, and linguistic diversity.

Key words: sign systems, functions of language, registers of language, language change, linguistic diversity.

1. Язык и знаковые системы

Теория языка — ядро лингвистического знания о природе Языка, принципах его устройства, объективных внутренних ограничениях на возможный естественный язык, о его отличиях от всех прочих знаковых систем. «Наука о знаках называется семиотикой, она занимается общими принципами, лежащими в основе структуры всех знаков» [Якобсон 1985: 320]. «Подразделение знаков на индексные, иконические и символические, которое Пирс предложил в знаменитой работе 1867 года, на самом деле основывается на двух дихотомиях. Одна из них — это противопоставление смежности и сходства. Индексное отношение между signans и signatum (= означающим и означаемым) зиждется на их фактической, существующей в действительности смежности. Типичный пример индекса — это указание пальцем на определенный предмет. Иконическое отношение между signans и signatum — это, по словам Пирса,'простая общность по некоторому свойству' [Peirce 1965], то есть относительное сходство, ощущаемое тем, кто интерпретирует знак, например, картина, в которой зритель узнает знакомый ему пейзаж. В знаке-символе signans и signatum соотнесены безотносительно к какой бы то ни было фактической связи. Смежность между двумя составляющими компонентами символа можно назвать приписанным свойством» [Якобсон 1985:322].

Для ясности приведем минимальные пары (может быть, тройки?) знаков.

Таблица 1[1]. Виды знаков

Символы	**Индексы**	**Иконы**
нос [орган обоняния]	*не суй нос не в свои дела*	*нос корабля,* *носик чайника*
голова [часть тела]	*пять голов* [единица счета животных]; *городской голова* [начальник]	*головка чеснока, лука*
школа [учебное заведение, место]	*школа* [здание]; *пражская лингвистическая школа* [группа людей, объединенных взглядами]	*школка* [место, где выращивают рассаду]
дерево [разновидность многолетнего растения]	*древо познания*; *дерево* = ветки и корни вместе с центральным стволом	*генеалогическое дерево;* *дерево зависимостей,* *синтаксическое дерево*
сеть [предмет с ячейками]	*сетка* [расписание]; *сеть Интернета*	*сетка* [сумка для переноски вещей]; *сеть* для рыбной ловли; *семантическая сеть*

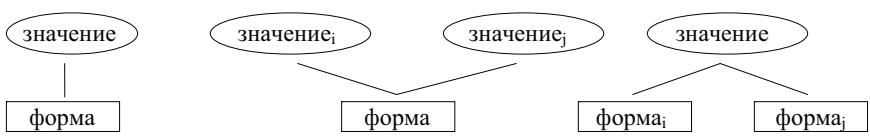
[1] Индексы и иконы имеют традиционные терминологические корреляты: иконы соотносятся с метафорой (отношение сходства), символы соотносятся с метонимией (отношение смежности).

Якобсон говорит о том, что не всегда у знака имеется интерпретатор: «Мнемонический узел на носовом платке, служивший для русских напоминанием о важном деле, является типичным примером внутренней коммуникации между прошлым и последующим состоянием одного и того же человека» [Якобсон 1985:324]. «Широко распространены и непреднамеренные иконические знаки; так, Фрейд отмечает,что некоторые грибы легко вызывают фаллический образ. Возможно, в некоторых случаях подобное сравнение можно определить, в терминах Пирса, как символическо-иконические знаки, порожденные или по меньшей мере подкрепленные в воображении индивида метафорическими ассоциациями…» [Якобсон 1985:325].

Соотношение формы и значения

Рассмотрим различные способы взаимодействия формы и значения (см. Схему 1, 2).

Имеется три традиционные схемы: однозначное соответствие; многозначное соответствие (полисемия, омонимия), когда одна форма имеет более одного значения; синонимия, когда одно значение имеет более одной формы.



(**1**) однозначность (**2**) полисемия, омонимия (**3**) синонимия

Схема 1. Традиционные схемы соответствия формы и значения языкового знака

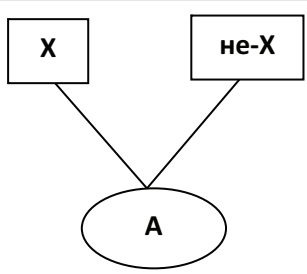
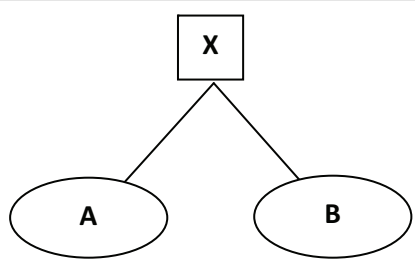
(а) некая языковая форма «как бы» обозначает противоположные сущности:	б) разные языковые формы «как бы» обозначают одну и ту же сущность:
	

Схема 2. Аномальные корреляции формы и значения

На Схеме 2 языковая форма А (означающее) обозначает некоторое значение (означаемое) Х и прямо противоположное антонимичное значение НЕ-Х, или значение Х, выражаемое формами А и В. Встает вопрос, почему одна языковая форма А выражает противоположные значения, а то же значение Х выражается разными языковыми формами А и В?

Случай (б) давно известен как синонимия языковой формы. Этот случай в своей абсолютной формулировке аномален функциональной избыточностью. Если иметь в виду синонимию в строгом смысле, то она означает полную вариативность формы (абсолютную взаимозаменимость синонимичных форм в любом высказывании). Однако полная вариативность, при ближайшем рассмотрении, никогда не подтверждается. Всегда обнаруживаются контексты, когда смысл *X* выразить при помощи формы *A* можно, а при помощи *B* нельзя, и наоборот.

Особенно парадоксален случай (а), так как противоположные (антонимичные) значения *X* и *не-X* выражаются одной формой *A*: **что же на самом деле означает *A*?**

Во флективных языках некоторые флексии имеют более одного значения, и восстановить значение морфологического маркера не представляет труда: он ведет себя как обычная языковая форма, просто значением флективной морфемы является конъюнкция параметрических значений, с сохранением принципа семантической аддитивности[2].

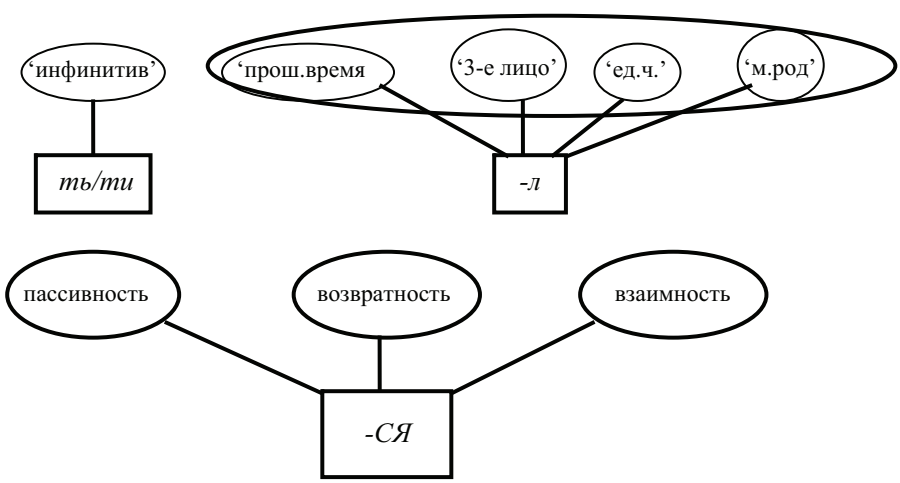


Схема 3. Комплексное значение флективного показателя

Показатель *-ть/ти* имеет значение инфинитива, и здесь аддитивность вырожденная. Глагольный показатель *–л* является флективной служебной морфемой, значение которой есть конъюнкция ‘прошедшее время’ & ‘3-е лицо’ &

[2] Семантическая аддитивность — сохранение смысла компонентов языкового выражения при сочетании параметрических значений.

'ед. число' & 'муж. род' Наконец, возвратная частица *–ся* традиционно имеет значения 'пассивность', 'возвратность', 'взаимность', например, *Чины людьми даются* (пассивность)*, Маруся отравилась* взаимность)*, Парень с девицей целуются* (взаимность).

Означает ли схема 3 многозначность морфемы *–ся* и как определить количество ее значений[3]? Почему адресат обычно не замечает смены значения, и для него это одна частица?

Если перейти на уровень генерализации, эти три значения суть конкретные реализации одного общего значения. Все эти значения объединяет потеря прямого дополнения, см. Схему 4.

Под языком как объектом изучения лингвистики прежде всего имеют в виду (в оппозиции к искусственным языкам и языку животных) естественный человеческий язык, возникновение и существование которого связано с возникновением и существованием человека разумного, *homo sapiens*. Язык является видообразующим свойством: человек является человеком постольку, поскольку он говорит на естественном человеческом языке.

Термин «язык» имеет по крайней мере два взаимосвязанных значения:

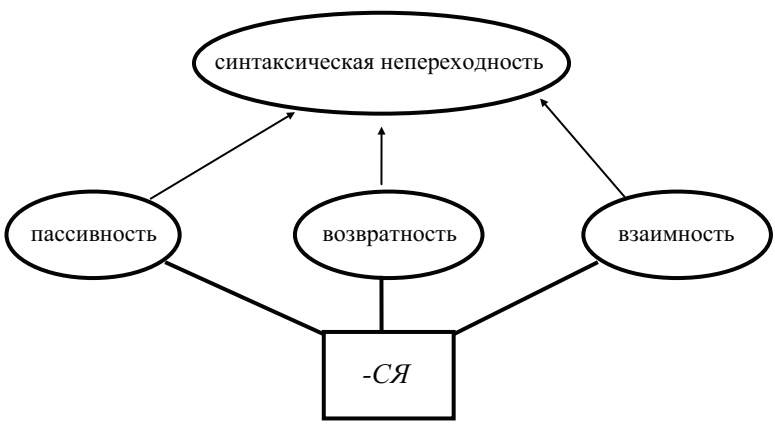


Схема 4. Генерализация частных значений частицы *–ся*

2. Язык вообще и идиоэтнический язык

Язык вообще, язык как определенный класс знаковых систем; Язык в первом значении — это абстрактное представление о едином человеческом языке, средоточии универсальных свойств всех конкретных языков. Язык вообще есть естественно (на определенной стадии развития человеческого общества) возникшая и закономерно развивающаяся семиотическая (знаковая) система,

[3] Оно зависит от схемы описания, а не вытекает из объективного положения вещей.

обладающая свойством социальной предназначенности, и естественно (на определенной стадии развития человеческого общества), существующая прежде всего не для отдельного индивидуума, а для определенного социума. Кроме того, на эту знаковую систему наложены ограничения, связанные с ее функциями и используемым субстанциальным (звуковым) материалом. Язык в первом значении — это абстрактное представление о едином человеческом языке, средоточии универсальных свойств всех конкретных языков. Конкретные языки — это многочисленные реализации свойств языка вообще.

Конкретный, так называемый **этнический (идиоэтнический) язык** — некоторая реально существующая знаковая система, используемая в некотором социуме, в некоторое время и в некотором месте. Конкретные языки — это многочисленные реализации свойств языка вообще.

2.1. Функции языка

Существенно, что язык, обладая внутренней целостностью и единством, является полифункциональной системой (имеет множество функций). Рассмотрим базовые (=первичные) функции Среди всех функций важнейшими можно считать те, которые связаны с основными операциями над информацией. Язык является устройством, опосредованным мышлением. Будучи солидарно разделяемым всем языковым сообществом, а не только некоторыми его носителями, он есть форма отражения и обработки реальных и гипотетических знаний, полученных и получаемых человеком различными способами. Язык используется индивидуумом как средство участия в процессе создания, хранения и передачи информации.

Во-первых, выделяется так называемая эпистемическая (= эписемическая) функция, иногда называемая накопительной (= аккумулятивной). Благодаря этой функции в единицах языка в виде гносеологических образов закрепляются и хранятся элементы как реального, так и виртуальных миров, выделенные, отображенные и обработанные сознанием человека.

Во-вторых, большое значение имеет познавательная (= когнитивная) функция. Она реализует фундаментальную связь языка и мышления и процесс получения нового знания. Она ответственна за то, что в единицах языка и их свойствах материализуются структура и динамика мысли, то есть языковые единицы приспособлены как для номинации элементов действительности (и, далее, хранения знаний), так и для обеспечения потребностей мыслительного процесса.

В то же время язык, в-третьих, является основным средством человеческого общения (коммуникативная функция) в коммуникативной среде, средством передачи информации от говорящего к слушающему (= адресату). В силу этого базовые свойства языка естественным образом согласованы с потребностями и условиями протекания коммуникативной деятельности человека, составляющей важнейший аспект его социального поведения, так как общественная, в том числе трудовая деятельность человека, невозможна без обмена информацией

Благодаря языку человек может хранить информацию об окружающей действительности неограниченно долго, в течение многих поколений. Так, большие по величине тексты «Илиады» и «Одиссеи» были созданы еще задолго до появления письменности и сохранились лишь благодаря языку. С возникновением письменности хранить и передавать знания от поколения к поколению стало многократно проще, но коммуникативная функция осталась той же. Развитие этой функции, как и раньше, состояло в приспособлении языка к фундаментальной потребности человека, превращающей его в существо, способное использовать накопленный опыт применительно к всевозможным жизненным ситуациям.

Чтобы хранить информацию, надо ею владеть. Человек не просто регистрирует элементы опыта, он анализирует и обобщает их. С этой способностью связана познавательная функция языка.

Человек в своей эволюции сформировался как существо социальное, с чем связана необходимость передавать информацию (знания) друг другу. С этой потребностью связана коммуникативная функция языка.

Эти три функции являются базовыми, универсальными, представленными в любом человеческом языке. Встает вопрос о том, какая из функций является самой главной. Часто утверждают, что это несомненно коммуникативная функция. Однако многими авторитетными лингвистами это мнение оспаривается[4] Можно привести немало аргументов в пользу приоритета познавательной и эпистемической функций над коммуникативной. Это значит, что статус первичности нельзя приписать ни одной из рассмотренных функций. Они образуют неразрывное единство и противопоставлены всем прочим, безусловно важным функциям.

Рассмотрим теперь вторичные функции.

К широкому классу вторичных функций относится функция социальной солидарности всех говорящих, являющаяся важным фактором социализации.

Функции социализации противопоставлена функция развития индивидуализации, сознательная или бессознательная демонстрация отличности индивида от всех прочих говорящих на данном языке. К бессознательной индивидуализации относятся «характерные качества голоса, фонетическая организация речи, быстрота и относительная четкость произношения, длина и строение предложений, характер и объем словаря, употребительность наукообразной лексики, способность слов откликаться на потребности социальной среды, и в частности, ориентация речи на языковые привычки своих собеседников» см. [Сепир 1993:233], а также просодия и тембр В итоге язык можно считать средством идентификации личности, схожим с отпечатками пальцев.

К функции социальной солидарности близка функция поддержания социального комфорта (нормального бесконфликтного пребывания в коммуникативной

[4] См. [Сепир 1993:231], где в качестве доказательства анализируется аутическая речь — речь детей, больных аутизмом, пользующихся языком, но не вступающих в коммуникацию.

среде). Сюда относится, в частности, удовлетворение принципа поддержания принятого в обществе ритма общения, часто сводящегося к пустой «светской» болтовне, например, разговоры о здоровье, погоде, (не)урожае, курсе валют.

Также социальной является контактоустанавливающая (фатическая) функция (использование языка для установления контакта), ср. *Алё-алё* в начале телефонного разговора, приветствия при встречах (*привет, салям алейкум*[5]*, не скучаешь ли*[6]*,* разговоры о погоде и здоровье.

С творческим аспектом языка связана эстетическая (= поэтическая) функция (создание художественных текстов, построение текстов, в которых мастерски используются языковые ресурсы — достаточно вспомнить часто отмечаемую непереводимость поэзии..

С базисной коммуникативной функцией связана фасцинативная функция привлечения и поддержания эмоционального подъема слушающего благодаря силе воздействия, личному шарму, необычному, особому поведению говорящего, его харизме.

С модальной сферой взаимодействует эмоциональная функция (выражение чувств и эмоций).

Особой, уникальной является метаязыковая функция (описание и познание языка в терминах самого языка), присущая только человеку. Примечательно, что с этой функцией неразрывно связан тот факт, что лингвистические штудии, анализ языковых данных, любые ментальные операции с языком осуществляются благодаря существованию продуктов языковой деятельности.

Следует иметь в виду, что это далеко не полный перечень вторичных функций.

Наряду с полифункциональными ограничениями субстанциальный материал — звуковая (акустическая) природа языка — также накладывает значительные ограничения на общие свойства языка, в частности, предопределяет наличие незнаковых единиц (фонем — звуков) и линейную организацию знаковых единиц (морфем, слов, словосочетаний, предложений).

3. Социальные формы существования языка

С точки зрения социального функционирования имеется несколько форм, различающихся числом говорящих, территориальным распространением, функциональными особенностями. В русской лингвистической традиции принято различать такие формы, как говор, диалект, литературный язык. Во второй половине XX века закрепился также термин идиолект.

1. Идиолект. Это индивидуальный язык одного конкретного носителя языка. В некотором смысле у каждого индивида свой уникальный идиолект.

---

[5] Обязательный элемент культуры в мусульманском мире.

[6] Приветствие в арчинском этносе.

2. Говор. Говор есть множество структурно очень близких идиолектов, обслуживающее одну небольшую, территориально замкнутую группу людей, внутри которой не обнаруживается никаких заметных (территориально характеризуемых) языковых различий. Говор может характеризовать также не территорию обитания, а семью, членов одного родственного клана, улицу в небольшом населенном пункте или часть этого пункта: верхняя / нижняя его часть, или населенный пункт в целом. Это может также быть некоторая профессионально или семейно объединенная целостная группа людей, сообщество людей одного возраста или окказионально возникающая группа из 3-х-4-х индивидумов, объединенных общим предшествующим опытом и собравшихся на рыбалку, вечеринку, игру в покер, культурное мероприятие.

3. Диалект. Диалект — группа говоров (в частном случае — единичная), в которой сохраняется значительное внутриструктурное единство. Территориальная непрерывность распространения диалекта не является его обязательным признаком. В частности, в диалект может входить несколько территориально не контактирующих, удаленных друг от друга идентичных говоров. Например, африканский западно-атлантический язык пулар-фульфульде — язык «с сильным диалектным дроблением»,который «не имеет ни единого центра, ни наддиалектной формы. Рассеянные на огромном пространстве (от Атлантики до долины Голубого Нила) диалектные варианты, долговременно лишенные коммуникативных связей, естественно, не могут избежать большей или меньшей языковой дивергенции» [Коваль 2010:213–214].

Термин *говор* принят в русской традиции, в Европе он неизвестен. Иными словами, в англоязычной литературе говор и диалект не противопоставлены друг другу.

4. Язык. Язык является очень важным термином с социальной точки зрения, так как он противопоставлен диалекту. Язык, как правило, — это множество диалектов, допустимые различия между которыми могут в значительной мере варьировать и зависеть не только от чисто языковых факторов, но и, что особенно важно, от социальных параметров: языкового самосознания носителей (носителями какого языка они себя считают), наличия или отсутствия единой письменности, религиозного единства (ср. сербский и хорватский, отличающиеся кириллицей VS латиницей, ортодоксальной VS католической конфессией, социальной престижности диалектов, численности носителей отдельных диалектов, традиции и т. д.. В этом отношении интересные факты преподносит нам история греческого языка, в начале классического периода состоящего из множества диалектов. Среди этих диалектов выделялся аттический диалект, прежде всего своей социальной престижностью. Требования государственного устройства в период политического объединения всех диалектов оказались причиной возникновения общегреческого диалекта, так называемого койне (эллинистический период ,1–4 вв. до Р.Х., классический период, с 4–7 вв. до 4в. до Р.Х.

Койне поглотил прочие диалекты, и они были утрачены. Однако на месте этого койне возникло несколько новых диалектов и языковое единство

греческого языка распалось. На протяжении почти полутора тысяч лет такого рода смены происходили волнообразно, и носители более поздних языковых состояний уже не понимали речи на первичном койне. Нынешний новогреческий язык имеет единую литературную норму и множество сестринских диалектов.

В настоящее время на земном шаре продолжает существовать достаточно большое количество языков. Точную цифру никто не может указать, но по разным подсчетам количество современных языков колеблется от 6000 до 7000. Будем исходить из меньшей цифры 6000.

Языки очень неравномерно распределены по числу говорящих (см. Таблицу 2).

Таблица 2. Распределение языков по числу говорящих

Численность говорящих	Число языков	Доля от всех языков
> более 150 миллионов	7	0,1%
> более 50 миллионов	20	0,3%
> более 1 миллиона	138	2,3%
> более 100 тысяч	258	4%
> более 10 тысяч	597	10%
< менее 10 тысяч	5400	90%

Языков, на которых говорит более 150 миллионов, всего семь. Если добавить к ним языки с числом говорящих более 50 миллионов, число таких языков увеличится до двадцати. Всего у 138 языков число говорящих более миллиона. Далее численность говорящих быстро падает. 258 языков имеют численность говорящих более 100 тысяч, 597 языков имеют численность говорящих более 10 тысяч. Это всего 10% от общего числа языков. Подавляющее большинство языков (около 5400) имеют ничтожное количество говорящих — менее десяти тысяч человек.

Итак, доля «крупных» языков (более 100 тысяч говорящих) составляет всего 4%, а доля «малых» языков (менее десяти тысяч говорящих) — 90% (доля языков промежуточной зоны — от 10 до 100 тысяч — 6%). Малые языки составляют на земном шаре подавляющее большинство.

Очевидно, что численность говорящих является одним из важнейших факторов, влияющих на сохранение языка в обозримом будущем, см. [Kibrik 1991, Кибрик 1992]. Очевидно также, что малые языки находятся в смертельно опасной зоне с точки зрения угрозы их исчезновения. Прогнозы лингвистов на ближайшее столетие выглядят устрашающе. По пессимистическому сценарию Мишеля Крауса [Krauss 1992], через сто лет исчезнет 95% ныне существующих языков. По оптимистическому сценарию, сформулированному лингвистами, сотрудничающими с Volkswagen Foundation, см. [Noonan 2006:351], к концу XXI века умрет 60–70% языков. В обоих случаях к здоровым языкам, существование которых в течение ближайшего столетия гарантировано, может быть причислено сравнительно малое количество языков, а подавляющее большинство языков относится к «больным» или «умирающим» языкам.

Проблема надвигающегося массового вымирания языков усугубляется крайне неравномерной их изученностью. Из всего мирового лингвистического сообщества 90–98% исследователей занимается горсткой наиболее престижных языков с национальной традицией изучения. Это, как правило, языки, имеющие статус государственного или регионального языка и поддерживаемые на государственном и образовательном уровне. Так, в Дагестане имеется несколько региональных языков, имеющих письменность и изучающихся в школе как родной язык (аварский, лезгинский, даргинский, лакский, кумыкский).

Количество лингвистов, занимающихся тысячами бесписьменных и младописьменных языков, составляет очень малую величину. А количество тех, кто занимается умирающими языками, — вообще исчезающе малая величина. Иными словами, почти все научные исследовательские ресурсы сконцентрированы на тех языках, существованию которых ничто не грозит, а теми языками, которые могут в ближайшее время умереть, лингвисты практически не занимаются. Почти полная неизученность многих языков, сведения о которых зачастую добываются из косвенных источников, неполнота инвентаризации конкретных языков и различия в принципах их разграничения приводят, в частности, к расхождениям в оценке числа языков. В авторитетном издании Ethnologue от издания к изданию число языков постепенно изменяется. Около 7000 языков зафиксировано в 1996 году, см. [Grimes et al. 1996], а в 1988 году в нем значилось 6170 языков. В последней, компьютерной версии 2005 года (www.ethnologue.com) — 6912 языков. По личным, возможно слишком оптимистичным, оценкам Нунэна [Noonan 2006:352] только 500 языков имеют полное грамматическое описание, словари, большое количество текстов, то есть являются прилично документированными, порядка 2000 языков имеют краткие грамматические очерки и словари, как правило неудовлетворительного качества, а документация всех прочих языков весьма рудиментарна или вовсе отсутствует.

5. Литературный язык. На определенном этапе национального и социального развития некоторые стихийно существующие и развивающиеся языки вступают в высшую форму своего существования — форму литературного языка, характеризующегося социально регламентированной нормированностью и наличием более или менее широкого диапазона функциональных стилей: стиль повседневного разговорного бытового общения, газетно-политический, официально-деловой, научный и многие другие. Литературный язык является основным и почти единственным объектом лингвистических штудий. Описание конкретных языков сводится к изучению различных форм литературного языка и различных его компонентов. Под литературным языком понимается не только его письменная форма (ср. весьма дробное членение специализаций в русистике[7]) и не только язык художественной литературы.

[7] Уровневое изучение русского языка: фонетика, морфология, синтаксис, семантика, лингвистика текста, .синхрония *vs* диахрония (история русского языка) этимология, стилистика, культура речи, орфоэпия (правильное произношение), орфография (правила письма), лексикология (структура словарного состава), лексикография (теория

Это может быть и язык, не имеющий письменности. Таков, например, язык Ветхого завета, возникший задолго до возникновения письменности, в отличие от языка Нового завета, появившегося в эпоху уже существовавшей письменности. Литературным языком является также язык эпического фольклора, различные сакральные языки и тайные языки.

4. Живые, «больные», «умирающие» и мертвые языки: каковы научные приоритеты. История существования языков

При отсутствии специальной нормирующей деятельности, направленной на консервацию языкового состояния (ср. классический арабский язык), языки постоянно претерпевают изменения во всех звеньях своей структуры, происходит их непрерывное историческое развитие. Конкретные причины этого процесса не вполне выявлены, но, несомненно, что они заложены, во-первых, в принципах самого устройства языка и, во-вторых, в функциональном механизме его использования. Нередко процесс исторического развития приводит к умиранию этих языков. Наблюдается давно известное явление исчезновения некоторых малых языков, не имеющих письменности и достаточного уровня социального престижа.

В фиксированный момент времени число индивидуальных реализаций языка — идиолектов не меньше (а учитывая двуязычие, больше) числа говорящих на земном шаре людей (исчисляется миллиардами), а живых языков в социально признанном смысле насчитывается от трех до семи тысяч. Важный предмет размышлений лингвиста составляет как феномен множественности человеческих языков, так и характерная для них тенденция к изменению.

За время существования *homo loquens* для науки безвозвратно утеряно огромное количество материальных свидетельств о существовавших ранее языках. Во-первых, это умершие языки, не оставившие потомства, во-вторых, это предшествующие состояния («предки») современных языков.

Не дожившие до настоящего времени бесписьменные языки никогда уже не будут доступны лингвистическому анализу. Предки современных языков (праязыки) могут быть до известной степени реконструированы по данным современных языков и / или по сохранившимся письменным памятникам. Реконструкция предшествующих состояний языков является главной целью сравнительно-исторического метода.

Такое положение вещей, оставленное современной описательной лингвистике в наследство от предшествующего лингвистического опыта, не является

и практика составления словарей), фразеология (изучение семантически несвободных слов и предложений), ономастика (изучение собственных имен), топонимика (географические названия), антропонимика (антропонимы — собственные названия людей, отчества, псевдонимы), зоонимика (зоонимы — клички животных), диалектология (изучение говоров и диалектов во всех аспектах) и др.

случайным. В Европе развитие лингвистики шло от греко-латинской грамматической традиции, погруженной в классические древности, к расширению эмпирической базы и вовлечению живых современных национальных языков. До начала XX века лингвистика не осознавалась как единая мировая научная дисциплина, она развивалась в рамках отдельных национальных традиций, объединяемых своими национальными языками.

Вследствие этого в европейских странах доминировала почти исключительная концентрация исследований в области своих национальных языков. Для России это был, естественно, русский язык, для Франции — французский, для Англии — английский и т. д.

Развитие общей лингвистической теории (что принято называть общим языкознанием) также базировалось, прежде всего, на языковых данных, извлеченных из родного языка исследователей.

Поскольку национальные языки основных государств Европы, которые могли себе позволить роскошь развивать лингвистическое знание, являются родственными и входят в одну большую семью индоевропейских языков, общая лингвистическая теория, обобщающая эмпирическое языковое разнообразие на ограниченном материале достаточно близких по базовой структуре и лексикону языков, также может быть характеризована как страдающая европоцентризмом. Многим лингвистам казалось, а зачастую и сегодня кажется, что все существенные свойства человеческого языка можно обнаружить в своем родном языке плюс, при особой образованности, те иностранные языки, что входят в стандартный образовательный набор.

Несмотря на огромные изменения, потрясавшие лингвистику в XX веке, мало что изменилось в понимании фундаментального значения эмпирической языковой базы во всем ее разнообразии. Достаточно сказать, что основные презумпции наиболее популярной в США теории — теории порождающей грамматики — унаследовали все тот же англоцентризм, и лишь под внешним давлением стали вовлекаться данные прочих языков, часто поверхностно и неполно, лишь в той степени, в какой они не колеблют основных догматов этой теории. Что касается русской лингвистической традиции, то в ней интерес к языкам России был довольно устойчивым, хотя даже беглое ознакомление с традиционными грамматиками и словарями этих языков показывает, что за точку отсчета принимается, как правило, русский язык и, более конкретно, традиционная нормативная грамматика русского языка.

Такова вкратце беглая характеристика основного направления лингвистической мысли, расставляющая приоритеты в описательной и теоретической лингвистике. Малым языкам в этой парадигме действительно нет места. Их изучают энтузиасты-одиночки, сознательно выбирая это непрестижное экзотическое занятие. И в основном благодаря их усилиям мы имеем то немногое, что все-таки имеем.

В России в XIX веке такова была деятельность барона П. К. Услара, посвятившего себя изучению языков Кавказа и значительно опередившего существовавшую в его время теорию, и политического ссыльного В. Г. Богораза, оставившего нам блестящее описание и словарь чукотского языка. В XX веке

изучение языков СССР было поставлено на академическую основу, и это способствовало значительному прогрессу в инвентаризации и исследовании языков народов Севера, Сибири, Памира и Кавказа. Можно назвать немало достойных упоминания имен, но это увело бы нас слишком далеко в сторону от целей данной работы. Важно, что языковых «белых пятен» на нынешнем постсоветском пространстве практически не осталось (чего нельзя сказать о многих регионах земного шара, заселенных тысячами языков, куда не ступала нога лингвиста), хотя степень изученности большинства языков имеет скорее ознакомительный, а не документирующий характер. И даже этот эмпирический материал почти не востребован лингвистическим сообществом, с ним знакомы, по тем или иным причинам, также лишь лингвисты-одиночки.

Прогнозируемый процесс исчезновения языков не является чем-то неслыханным, он сопровождал всю историю существования человека говорящего. Как правило, языки исчезали вместе с исчезновением говорящих на них этносов в результате истребительных войн или естественных причин — эпидемий и природных катаклизмов. Еще свежа память о великих географических открытиях, повлекших за собой бесчеловечную индустрию уничтожения коренного населения новых для европейцев земель. Так, в начальный период захвата европейцами американского и австралийского континентов многие племена аборигенов были полностью истреблены или вытеснены в места, мало пригодные для выживания. Удивительно не то, как много языков было тогда уничтожено, а то, что несколько сотен языков выжило и сохранилось до наших дней. Нередко языки уничтожались насильственно, когда завоеватели принуждали побежденных переходить на язык завоевателей. Миллионы африканцев, проданные в рабство в Америку, утратили свои родные языки. Никто не интересовался, какие этнические группы были полностью переселены на американский континент и на каких языках они тогда говорили.

5. Причины множественности языков

Множество человеческих языков нельзя считать случайным. Независимо от решения проблемы происхождения языка, требует объяснения непреложная тенденция языка к изменению (в особенности при отсутствии специальной нормирующей деятельности, нацеленной на консервацию языкового состояния).

Причины изменений в конкретных языках могут иметь различную мотивацию, но, как было сказано выше, они предопределяются тем, каковы системные отношения между элементами грамматики и лексикона данного языка, а также каков механизм его использования в окружающей языковой среде.

С точки зрения неискушенного человека, говорящего на престижном «большом» языке, проблема, рассматриваемая в данной работе, не представляет для него особого интереса. Существование на земном шаре большого количества языков обычно кажется ему излишней роскошью, если не практической помехой для беспрепятственного распространения информации

в современном мире. Это положение вещей, кажется ему, находится в конфликте с жизненными интересами человечества, и его необходимо изменить. И действительно, такой процесс идет быстрыми темпами, и он вызывает серьезную обеспокоенность у лингвистического сообщества.

Возникновение новых языков, наряду с их исчезновением, является постоянным процессом. Основной источник этого процесса — переход говоров в диалекты, а диалектов — в языки по мере количественного и качественного накопления различий между ними. Это сопровождается также постоянным переходом языков в диалекты, обычно отличающиеся от предыдущей диалектальной стратификации, ср. ситуацию в языке пулар-фульфульде [Коваль 2010].

Многократное членение русского языка на множество диалектов не останавливает процесса этого членения. Изменению подвергается как русский язык в целом, так и первая волна диалектов. Каждый из диалектов в свою очередь членится, и результат может быть различным. Русский язык благодаря своей престижности не может создавать новые языки, но видно, что каждый язык в целом является центром «клубка» бесконечно движущихся и изменяющихся диалектов разной степени устойчивости и изменяемости. Такая картина подтверждает выше высказанное утверждение, что причины языковых изменений лежат в самом языке, а именно в его структуре. Существенно, что процесс изменений никогда не может быть остановлен, это удивительная, но очевидная данность. Поэтому неудивительно, что для крупных языков возникает потребность нормировать все элементы языковой структуры — фонетику, морфологию, синтаксис, семантику и даже стилистику.

References

1. *Crystal David*. 1987. Language .The Cambridge Encyclopedia of Language.
2. *Grimes* (ed.). 1996, 1998, 2005. Ethnologue. Languages of the World.
3. *Jakobson Roman*. 1970. Language in Relation to Other Communication Systems. Linguaggi nella Societa e nella Tecnika : 3–16.
4. *Kibrik A. E*. 1991. The Problem of Endangered Languages in the USSR : 257–273.
5. *Koval' A.I*. 2010. Concord in the Noun Classe of Pular-Fulfude [Soglasovanie v Imennoi Gruppe Pular-ful'fude]. Osnovy Afrikanskogo Iazykoznaniia. Sintaksis imennykh I glagol'nykh grupp : 211–354.
6. *Krauss M*. 1992. The World's Languages in Crisis. Language, 68 (1) : 4–10.
7. *Noonan M. M.* 2006. Grammar Writing for a Grammar-reading Audience. Perspectives on Grammar Writing. Special Issue of Studies in Language, 30 (2) : 351–365.
8. *Peirce Charles*. 1965. Collected Papers I IV.

EXPLOITING DISTRIBUTIONAL SIMILARITY FOR LEXICAL ACQUISITION

McCarthy Diana (diana@dianamccarthy.co.uk)

Lexical Computing Ltd., Brighton, East Sussex

Lexical acquisition has been dubbed the bottleneck of large scale robust natural language processing applications for at least two decades. There is now a substantial body of research dedicated to this important subfield of computational linguistics. Since the 1990s, researchers have turned to corpora for automatic lexical acquisition, rather than rely on extraction from existing online lexical resources. This allows for coverage of new domains, genres and languages without existing resources and where available resources do not provide sufficient coverage or require tailoring to the specific text type. A large body of lexical acquisition from corpora uses distributional similarity whereby the similarity between two words is calculated from the extent that the words have similar contexts of occurrence. Distributional similarity approaches are used for smoothing unseen events using data from seen events. They are also used as an approximation of semantic similarity since there is a strong tendency for words that exhibit similar distributional behaviour to share in their underlying semantics. This paper provides a summary of research that I, along with various collaborators, have conducted using distributional similarity to automatically acquire sense frequency information, selectional preferences and estimates of semantic non-compositionality of putative multiwords.

Key words: lexical acquisition, distributional similarity, NLP, semantic similarity

1. Introduction

Automatic lexical acquisition has received considerable interest for the past twenty years and more since without it computational linguistic systems simply will not scale and due to the emphasis on the lexicon as the appropriate repository for the majority of linguistic information (Gazdar, 1996). The focus quickly shifted from acquisition from electronic resources to acquisition from corpora since it was felt that this would avoid the errors and lack of coverage that beset man made resources. Extraction from corpora furthermore allows acquisition to languages and tailoring to domains which are not covered, or are poorly served by pre existing resources. Corpora also provide the much needed frequency information that is the backbone of computational linguistics systems, which since the 1990s are invariably statistical. That said, corpus approaches suffer from errors that arise in automatic processing of naturally occurring data and are dependent on sufficient language data of the appropriate type

being available in electronic form. Neither approach provides a panacea and many solutions are found in hybrid approaches (Klavans and Resnik, 1996).

One major area of research in automatic acquisition from corpora has been the use of distributional similarity. In distributional similarity approaches, words are represented by the contexts that they occur in and the frequency of occurrence in these contexts. A vector capturing this information can be used directly for representation and a measure of distributional similarity is used to compare the representation of one word with that of another. Automatic distributional "thesauruses" can be produced from this data. In these thesauruses, a word entry is listed with other words that have the most distributional contexts in common with the target word. Distributional similarity can be applied to linguistic phrases beyond the lexical level (Mitchell and Lapata, 2008) however in this paper, we focus on the application to lexical acquisition.

Lexical acquisition encompasses a wide variety of different areas of linguistics: phonology, morphology, syntax and semantics. Certain aspects that relate to pragmatics are also beginning to receive attention, such as the widespread interest in sentiment. Topics that have been particularly prevalent have been the acquisition of word senses and information associated with specific senses such as collocations, subcategorisation (predicate argument structure), selectional restrictions or preferences (for parsing and semantic role labelling), and multiwords. In this paper I provide an overview of some of my research in lexical acquisition in the last decade focusing particularly on work exploiting distributional thesauruses. I will focus the paper on acquisition of word sense frequency information and non-compositionality detection of putative multiwords.

1.1. Distributional similarity

Distributional similarity is an approach which uses statistics concerning the contexts of occurrence of words and determines the similarity between two words given this information. A word is represented by a vector of values, usually frequency values, from a corpus and each dimension of the vector represents a particular context. The definition of context varies considerably. It can be a document, a specified grammatical relation or within a window of words around the target. Distributional similarity uses these vectors and calculates a similarity score designed to measure the similarity between the vectors. The distributional similarity score can be used for smoothing statistical models. In such an approach, seen information occurring with a word is used for a rarer or unseen word that is related to the more frequent word by distributional similarity. It can also used as an approximation of semantic similarity since there is a strong tendency for words that exhibit similar distributional behaviour to share in their underlying semantics. To this end the vectors have been used for semantic representation in vector space models (Schütze, 1998). The similarity score can also be used to produce a "distributional thesaurus" by ranking other words in terms of their similarity to the target word and the top K (where K is a threshold such as 10 or 50) words are provided in rank order as the nearest neighbours to the target word along with the distributional similarity score used to rank them.

There are many different distributional similarity measures (see Weeds (2003) for a survey). Though we have used various measures in our work (Weeds et al., 2004; McCarthy and Navigli, 2009), we have predominantly used the measure proposed by Lin (1998) and found it to perform well on our tasks. As a rule, we have used the grammatical relations output from RASP (Briscoe and Carroll, 2002) as our contexts, though we have also observed good results with proximity relations (McCarthy et. al., 2007) which bodes well for applying the methods to data without a suitable parser.

2. Word sense frequency acquisition

Words have different meanings and we expect our computational models to reflect this. Naturally, we therefore expect to represent different meanings in the lexicon somehow, and in doing so it is necessary to have an automatic method of associating the word forms in natural language data with the senses in the lexicon. Such automatic methods fall under the rubric of word sense disambiguation. Word sense frequency information is arguably the most important information for this enterprise.

Word sense disambiguation is performed using clues such as collocations and domain information which can be automatically acquired from training data where the target senses have been marked up by human annotators, or from existing resources, or automatically from corpora. The best performing word sense disambiguation methods however rely on a very simple heuristic to supplement information from the context. This is known as the first (or most frequent) sense heuristic. The first sense heuristic is particularly powerful (Navigli, 2009) and particularly so when the contextual evidence is weak and when the entropy is low, that is the sense frequency distribution for a given word is particularly skewed. Of course contextual evidence is required to disambiguate words effectively, nevertheless, in many typical texts there is a strong tendency for the same sense to occur throughout a discourse (Gale et al., 1992) . McCarthy et al. (2004) proposed a method to automatically determine the most likely sense given a particular corpus as training data and a predefined inventory of senses. Researchers had been using predominant sense information for many years but what was new in this work is that sense predominance could be estimated from corpus data that had not been annotated by hand. Manually tagging a corpus with word senses is a laborious and costly process (Ng, 1997). The use of an "unsupervised" system that did not require manually labelled training data meant that not only was the technique applicable to a language without a handtagged corpus (Iida et al., 2008), but also that the method can be applied to corpus data from a given domain which will give more appropriate sense frequency information compared to using a general purpose resource, at least for words that are salient to that domain (Koeling et al., 2005).

In this paper, we give a brief overview of the method and some of our main findings. For a full account of the method and results, please see McCarthy et al. (2007) and the various references in this paper.

2.1. Method

The approach first reported in McCarthy et al. (2004) works as follows. Given a listing of word senses from an inventory such as WordNet (1998), we calculate a ranking score over those senses. As an example, take the noun *tie*. In WordNet (version 3.0) there are in fact 9 senses, but if we use just the first three for this example we have

1. **necktie**, tie — (neckwear consisting of a long narrow piece of material worn (mostly by men) under a collar and tied in knot at the front; "he stood in front of the mirror tightening his necktie"; "he wore a vest and tie")
2. **affiliation**, association, tie, tie-up — (a social or business relationship; "a valuable financial affiliation"; "he was sorry he had to sever his ties with other members of the team"; "many close associations with England")
3. tie — (**equality of score** in a contest)

When we applied the Lin (1998) distributional similarity score to data from the British National Corpus (Leech, 1992) parsed with RASP, we observe the following top 10 neighbours with their corresponding distributional similarity scores used for ranking shown in parenthesis:

BNC:
links (0.165) shirt (0.162) scarf (0.152) jacket (0.142) bond (0.130) match (0.128) trousers (0.126) link (0.125) collar (0.125) dress (0.121)[1]

We can intuitively see that while the neighbours reflect different senses of *tie*, there are more that relate to the **necktie** sense. To calculate the ranking score for each sense, we take each distributional similarity score of each neighbour and allocate a proportion of it to each of the three senses. We do this such that the proportion is reflected in the semantic similarity between the sense and that neighbour. Neighbours are words and so may have multiple senses. To calculate the semantic similarity between a sense and a neighbour the algorithm picks whichever sense of the neighbour maximises the semantic similarity to the target word. The calculation for semantic similarity depends on what sense inventory we have. For our work with WordNet, we tried various measures from the WordNet Similarity Package (Patwardhan and Pedersen, 2003). The JCN (Jiang and Conrath, 1997) and Lesk (Lesk, 1986) proved to perform well. The JCN uses the hypernym structure of WordNet to estimate semantic similarity while Lesk uses the overlap of dictionary definitions. Lesk is therefore useful in many cases where a sense inventory is like a standard dictionary with definitions but without the semantic relationships encoded in WordNet. The method produces a score for each sense by summing the distributional similarity scores (0.165 0.162 etc.) each multiplied by a weight for that sense and that neighbour where the weight is the maximum semantic similarity (JCN for example) between that sense and any of the senses of that neighbour. Thus for a sense (s) of a word (w) the calculation is as follows:

[1] We use only the top 10 neighbours here for the sake of brevity.

$$ranking\ score(s \in senses(w)) = \sum_{i=1}^{k} distsim(w, n_i) \times maximum(jcn(ns \in senses(n_i)), s)$$

Where distsim represents the distributional similarity between w and the neighbour of w at rank i. jcn is the semantic similarity measure that weights the contribution from this neighbour according to its semantic similarity with s.

Thus, while there are neighbours obtained from the BNC related to different senses, the majority here are most strongly related (intuitively and by measuring with JCN) to the first **necktie** sense of *tie*.

Although we can get this information from sense tagged texts such as SemCor (Miller et al. 1993), the sense distributions will naturally vary in different domains. We can see this by looking at domain specific data we (Koeling et al., 2005) collected from the Reuters Corpus (Rose et al., 2002) in finance and sport. The top 10 neighbours of *tie* are:

Finance:
relation (0.329) links (0.247) relationship (0.232) cooperation (0.228) contact (0.142) partnership (0.141) trade (0.137) role (0.133) integration (0.133) finances (0.132)

Sport:
qualifier (0.191) match (0.174) clash (0.150) round (0.135) semifinal (0.132) series (0.129) fixture (0.125) matchup (0.120) encounter (0.120) win (0.116)

The majority of neighbours in Finance are most strongly associated with the **affiliation** sense of *tie*, whereas those from Sport are most strongly associated with the third **equality of score** sense.

2.2. Further work

In addition to the various studies with corpus data that has been classified for domain manually, we have also demonstrated that we can apply this method successfully where the corpus data needed for training our models has been marked up for domain automatically and also where the input data itself is likewise annotated automatically (Koeling et al., 2007) .

As well as adaptation to different domains, we (Iida et al., 2008) have also applied our method to another language, Japanese, and show that where a dictionary does not have the structure that WordNet does, then we can use the Lesk score. We also propose an adapted Lesk score which uses distributional similarity to refine the overlap measure between the definition of a sense to be ranked and the senses of the neighbours. Rather than summing the exact matches between any of the words occurring in the two definitions, we use the sum of the distributional similarity scores of the words in the paired definitions where words that are present in both get the maximum distributional similarity score of 1. This has the effect of coping with sparse data to give a more productive overlap method.

Another aspect of our more recent work is to use the sense ranking for word sense disambiguation, i. e. taking account the context rather than simply applying the top ranking sense irrespective of context. As well as automatically detecting the domain (Koeling et al., 2007; Koeling and McCarthy, 2007), we have used the ranking score to estimate the entropy of the sense distribution to better gauge when the predominant sense heuristic will be more powerful, because the distribution is skewed, or when the distribution is flatter and it is more important to look for contextual evidence (Jin et al., 2009). We have obtained modest improvements using the grammatical relation in the target sentence to help determine which neighbours, and therefore sense, is more relevant in the context (Koeling and McCarthy, 2008).We have also recently used the sense ranking information to help in initialising domain specific graphical methods for word sense disambiguation (Reddy et al., 2010). Accuracy improves by 11 percentage points when domain specific sense ranking information is used.

3. Non-compositionality detection of putative multiwords

A crucial aspect of lexical acquisition is to determine exactly which entries should be stored in the lexicon. Nevertheless, I and various collaborators have been developing acquisition methods that aim to detect cases of semantic non-compositionality because ultimately such techniques could be used to determine the boundaries of what entries go in the lexicon and what stays out.

Multiwords have received considerable attention in computational linguistics over the past decade and particularly since the seminal paper by Sag et al. (2002). There has been a series of ten workshops run at the main international computational linguistics conferences in the last decade focusing on various aspects of computational representation, handling and application of multiwords. One important and reoccurring issue is the difficulty of a precise definition to make a clear boundary between what is and what it not a multiword. Coverage of multiwords in man made lexicons varies considerably for this very reason and also because of their abundance and the fact that multiword neoglisms are coined all the time. There are many reasons why the boundaries vary, but for many purposes there is some level of idiosyncratic behaviour. This might be syntactic, for example *wine and dine,* or pragmatic, for example *good morning*[2]*,* but in most cases we care about semantic non-compositionality which may give rise to other types of idiosyncratic behaviour. In addition to organisation of some of the multiword expression workshops and a journal special issue, my involvement in this area has been in automatic methods for detecting compositionality, or the lack of it, on the grounds that this will help determine the boundaries of what should be stored in the lexicon.

My research has focused on English and has emphasised the fact that compositionality is on a continuum. In McCarthy et al. (2003) we conducted experiments contrasting distributional similarity of the phrasals and the constituent verbs

[2] I am indebted to Timothy Baldwin for these examples.

to determine the extent that putative phrasal verbs (such as *blow up* and *eat up*) are compositional. In McCarthy et al. (2007) we conducted experiments on verb-object combinations, such as (*draw breath* and *light cigarette*). We again used distributional similarity, but this time rather than comparing the distributional profile of constituents to that of the whole phrase we used the nearest neighbours for modelling the selectional preference of the verb and then determine if the object was prototypical as an argument or not. If the object is not semantically related, using distributional similarity as a proxy for semantic similarity, to the typical objects seen with that verb then this is an indication of non-compositionality. We use preference strength directly to measure this. The next two subsections give a little more detail on these two works but we refer the interested reader to the papers cited for further details.

3.1. Detecting compositionality of phrasal verbs

For these experiments, we were interested in estimating the semantic compositionality of phrasal verbs which had been found by the RASP parser. We investigate various measures which compare the nearest neighbours of the verb constituent (e. g. *eat*) with the phrasal verb (e. g. *eat up*) or which scrutinize the list of nearest neighbours of the phrasal for occurrence of the constituents, or which combine both these approaches. More specifically the methods were:

- overlap: overlap of the top K neighbours of the phrasal and the constituent verb.[3]
- Sameparticle: the number of neighbours in the top 500 of the phrasal containing the same particle, for example *nibble up* has the same particle as *eat up*.
- Sampleparticle-simplex : as for Sameparticle but where we deduct the number containing the same particle which occurred in the simplex (constituent) verb's neighbours (the neighbours of *eat*).
- Simplexasneighbour: whether the simplex verb (*eat*) occurs in the top 50 nearest neighbours of the phrasal.
- Rankofsimplex: the rank of the simplex in the top 500 neighbours
- Scoreofsimple: the distributional similarity score of the simplex in the top 500 neighbours of the phrasal
- OverlapS: the overlap in the top K neighbours of the phrasal with those of the constituent verb's neighbours but where we remove all particles from the phrasal's neighbours (so for example, *nibble up* would become *nibble*).

We experimented with data from the BNC using Lin's measure of distributional similarity. We evaluated our methods by ranking a list of candidate phrasals according to these measures and correlating them using Spearman's rho with a gold-standard. We created the gold-standard by asking a set of three human annotators how compositional the candidate phrase was on a scale of 0–10 (idiomatic — fully compositional). Correlation was highest and highly significant for sameparticle, sameparticle-simplex and for the overlapS when using 30 or 50 neighbours. An interesting finding was that

[3] We experimented with different values of K, 30, 50, 100, 500.

statistics often used for multiword detection, such as Chi-squared, the log-likelihood ratio (Dunning, 1993) and pointwise mutual information (Church and Hanks, 1991) gave much lower, though significant, correlations. Phrasal frequency was not even significantly correlated.

3.2. Using Distributional Similarity for detecting compositionality of verb-object pairs

In this work, we used the dataset of verb-direct object pairs provided by Venkatapathy and Joshi (2005) which contained compositionality judgments on a scale following McCarthy et al (2003). This time, rather than use the distributional similarity neighbours for comparison of the constituents to the whole, we used them to build selectional preference models to estimate the preference strength of the verb for the given object. It is assumed that a weak preference for a particular direct object would indicate that the particular verb and object combination does not exhibit the normal semantic behaviour of the verb, and that this combination is non-compositional. For example, in the expressions *I'll eat my hat*, the direct object *hat* is not prototypical of the types of object we usually see with *eat* and our model should indicate this. We use a measure of selectional preference strength as an estimate of compositionality.

One issue for selectional preference acquisition is that it is acquired from automatically parsed data and multiwords are present in such data. Indeed selectional preference acquisition was one of our main motivations for detecting compositionality of multiwords in the first place (McCarthy et al. 2003). To avoid this problem we contrasted standard WordNet models (Li and Abe, 1998) which use direct object token instances from the training data to determine the classes, with type based models which use word types rather than tokens. We proposed both WordNet and distributional similarity type based models and contrasted these with the traditional token based models. Traditional token based models, such as (Resnik, 1993; Li and Abe, 1998) use direct object data for a given verb to populate the WordNet noun hierarchy with frequencies and obtain a probability distribution over WordNet classes. Our WordNet type based models use word types to determine the classes used for representation, rather than tokens, before then calculating the probability distribution using the tokens. Only if there are several types of a semantic class does the model include that class. For example, though *eat hat* might be reasonably frequent in a corpus, the type based models would not retain the probability under a **clothing** class simply because there are no other word types to support the use of that class in the model, whereas for *wear hat* that would not be the case due to the occurrence of words such as *coat*, *scarf* and *dress* which are semantically related. Furthermore, in these type based models we disambiguate an object that occurs at (directly or by virtue of hypernymy) several WordNet classes by assigning it to the class with the maximum number of types in the object data.

We contrasted the type based WordNet models with type based distributional models that use distributional similarity to group the objects in the training data into "classes". In these distributional similarity models the classes are a subset of the objects selected

automatically so as to maximise[4] the inclusion of the object types from the training data for this verb in the top K neighbours. The training data for each verb was obtained from the direct objects detected for that verb from RASP parses of the BNC. The probability distribution associated with each class is then estimated using the frequency of the objects occurring as distributional neighbours (in the top K) to these words representing the classes. Where an object occurs as a neighbour of multiple words selected as classes, the class selected will be that which has the maximum number of object types as neighbours. A portion of the model acquired for the direct object slot of *park* is shown in table 1.

We compared these three types of model on the subset of the Venkatapathy and Joshi dataset that contained common nouns as objects (rather than adjectives, pronouns and complements). All models produced significant results. The type based WordNet models outperformed the token based models but were themselves outperformed by the models that used distributional similarity for the

Table 1. A portion of the distributional similarity selectional preference for the direct object of the verb 'park'

Class (probability)	Disambiguated objects (frequency)
van (0.86)	car (174) van (11) vehicle (8) . . .
mile (0.05)	street (5) distance (4) mile (1) . . .
yard (0.03)	corner (4) lane (3) door (1)

Classes without requiring a manually constructed resource like WordNet. This is an encouraging result as it means that the method can be applied to a language without such a resource.

The methods also outperformed the individual features that Venkatapathy had used on the same portion of the data. These features included vector space models (Baldwin et al., 2003), pointwise mutual information and an existing method for detecting compositionality using distributional similarity to find non productive combinations (Lin, 1999). The best results were obtained when using the distributional similarity selectional preferences combined with some of these other features. This demonstrates that while the selectional preferences are useful features for non-compositionality detection of verb-object multiwords, no one approach is a panacea.

3.3. Further work

We (Reddy et al., 2011) are currently engaged in further work to examine compositionality judgments of humans in more detail by considering not only judgments for the phrase as a whole, but also for the individual constituents. We are developing distributional similarity methods that likewise compare the distributional profile (the vector containing contexts of occurrence) of the constituent words with the vector for

[4] We use a greedy algorithm.

the whole phrase combined also with the distributional similarity between models of a composition of the constituent vectors and the vector for the whole phrase. The composition vector representations use both additional and multiplication composition functions over the constituent vectors (Mitchell and Lapata, 2008). Furthermore we refine the constituent vectors, inspired by Erk and Pado (2010) by considering only the contexts that are shared by both constituents but not including the contexts occurring with the candidate multiword.

4. Conclusions and future directions

In this paper, I have given a summary of research I have conducted, along with various collaborators, in exploiting distributional similarity for lexical acquisition. I have focused this paper on work on sense ranking and on compositionality detection for multiwords. The compositionality detection itself involved use of distributional similarity models for acquiring selectional preferences. There are others using distributional similarity for selectional preference acquisition (Erk, 2007) and we look forward to trying out these models for new purposes, such as automatic detection of diathesis alternations where previously we used token based WordNet models (McCarthy, 2000).

Another direction for research has been the representation of sense using distributional similarity. There are several strands of such research (Panel and Lin, 2002; Erk and McCarthy 2009). I am particularly interested in alternative ways of annotating and evaluating distributional models of semantics using paraphrases (McCarthy and Navigli, 2009; McCarthy et al., 2010), translations (Mihalcea et al. 2010) and usage similarity judgments (Erk et al., 2009). I have been examining the relationships between these different types of annotations (McCarthy, 2011).

Acknowledgments

I would like to thank all my collaborators on these projects. In alphabetical order these are John Carroll, Katrin Erk, Abhilash Inumella, Nick Gaylord, Spandana Gella, Ryu Iida, Bill Keller, Rob Koeling, Peng Jin, Aravind Joshi, Suresh Manadhar, Rada Mihalcea, Roberto Navigli, Mark Stevenson, Siva Reddy, Ravi Sinha, Sriram Venkatapathy, Julie Weeds and David Weir. I am indebted to Siva Reddy for a careful reading of an earlier draft. All the faults that remain are my own.

References

1. *Baldwin T., Bannard C., Tanaka T., Widdows D.* 2003. An Empirical Model of Multiword Expression Decomposability. Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment : 86–96
2. *Briscoe E., Carroll J.* 2002. Robust Accurate Statistical Annotation of General Text. Proceedings of the Third International Conference on Language Resources and Evaluation (LREC). : 1499–1504
3. *Church K.,Hanks P.* 1991. Word Association Norms, Mutual Information and Lexicography. Computational Linguistics, 16 (1): 22–29
4. *Dunning T.* 1993. Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics, 19 (1): 61–74
5. *Erk K.* 2007. A Simple, Similarity-based Model for Selectional Preferences. Proceedings of ACL 2007.
6. *Erk K., McCarthy D.* 2009. Graded Word Sense Assignment. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2009).
7. *Erk K., McCarthy D., Gaylord N.* 2009. Investigations on Word Senses and Word Usages. Proceedings of the Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing ACL-IJCNLP.
8. *Erk K., Pado S.* 2010. Exemplar-Based Models for Word Meaning In Context.Proceedings of ACL 2010.
9. *Fellbaum C.* (editor).1998. WordNet, An Electronic Lexical Database.
10. *Gazdar G.* 1996. Paradigm merger in Natural Language Processing. Computing Tomorrow: Future Research Directions in Computer Science : 88–109
11. *Gale W., Church K., Iarovski D.* 1992. One Sense Per Discourse. Proceedings of the 4th DARPA Speech and Natural Language Workshop : 233–237
12. *Iida R., McCarthy D. and Koeling R.* 2008. Gloss-Based Semantic Similarity Metrics for Predominant Sense Acquisition. Proceedings of the Third International Joint Conference on Natural Language Processing : 561–568
13. *Jiang J., Conrath D.* Semantic similarity Based on Corpus Statistics and Lexical Taxonomy. 10th International Conference on Research in Computational Linguistics : 19–33
14. *Jin, P. McCarthy, D. Koeling R. , Carroll J.* 2009. Estimating and Exploiting the Entropy of Sense Distributions. Proceedings of the North American Chapter of the Association for Computational Linguistics — Human Language Technologies (NAACL HLT) 2009 Conference
15. *Klavans J., Reznik P.* (editors.). 1996. The Balancing Act: Combining Symbolic and Statistical Approaches to Language.
16. *Koeling R., McCarthy D.* 2007. Sussx: WSD using Automatically Acquired Predominant Senses. Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007) : 314–317

17. *Koeling R., McCarthy D.* 2008. From Predicting Predominant Senses to Using Local Context for Word Sense Disambiguation. Semantics in Text Processing. STEP 2008 Conference Proceedings :129–138
18. *Koeling, R., McCarthy D., Carroll J.* 2005. Domain-Specific Sense Distributions and Predominant Sense Acquisition. Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing : 419–426.
19. *Koeling R., McCarthy D. , Carroll J.* 2007. Text Categorization for Improved Priors of Word Meaning. Proceedings of the Eighth International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2007)
20. *Leech G.* 1992. 100 million Words of English: the British National Corpus. Language Research, 28(1):1–13
21. *Lesk M.* 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone From an Ice Cream Cone. Proceedings of the ACM SIGDOC Conference : 24–26
22. *Li H., Abe N.* 1998. Generalizing Case Frames Using a Thesaurus and the MDL Principle. Computational Linguistics, 24(2) : 217–244
23. *Lin D.* 1998. An Information-Theoretic Definition of Similarity. Proceedings of the 15th International Conference on Machine Learning.
24. *Lin D.* 1999. Automatic Identication of Noncompositional Phrases. Proceedings of ACL-1999 : 317–324.
25. *Mihalcea R., Sinha R., McCarthy D.* 2010. SemEval-2010 Task 2: Cross-Lingual Lexical Substitution. Proceedings of SemEval-2010: 5th International Workshop on Semantic Evaluations ACL 2010
26. *McCarthy D.* 2000. Using Semantic Preferences to Identify Verbal Participation in Role Switching Alternations. Proceedings of the first Conference of the North American Chapter of the Association for Computational Linguistics.
27. *McCarthy D.* 2011. Measuring Similarity of Word Meaning in Context with Lexical Substitutes and Translations. Computational Linguistics and Intelligent Text Processing 12th International Conference, CICLing 2011 : 238–252.
28. *McCarthy D., Keller B., Carroll J.* 2003. Detecting a Continuum of Compositionality in Phrasal Verbs. Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment.
29. *McCarthy D., Keller, B., Navigli, R.* 2010. Getting Synonym Candidates from Raw Data in the English Lexical Substitution Task. Proceedings of the 14th EURALEX International Congress. Leeuwarden.
30. *McCarthy D., Koeling R., Weeds J., Carroll J.* 2004. Finding Predominant Senses in Untagged Text. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics: 280–287.
31. *McCarthy D., Koeling R., Weeds J. ,Carroll J.* Unsupervised Acquisition of Predominant Word Senses. Computational Linguistics, 33 (4) : 553–590.
32. *McCarthy, D., R. Navigli.* 2009. The English Lexical Substitution Task. Language Resources and Evaluation 43 (2) Special Issue on Computational Semantic Analysis of Language: SemEval-2007 and Beyond : 139–159.

33. *McCarthy D., Venkatapathy S., Joshi A. K.* 2007. Detecting Compositionality of VerbObject Combinations using Selectional Preferences. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2007) : 369–379.
34. *Miller G. A., Leacock C., Tengi R., Bunker R. T.* 1993. A Semantic Concordance. Proceedings of the ARPA Workshop on Human Language Technology : 303–308.
35. *Mitchell J. , Lapata M.* 2008. Vector-based Models of Semantic Composition. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies : 236–244.
36. *Navigli R.* 2009. Word Sense Disambiguation: a Survey. ACM Computing Surveys, 41(2) : 1–69.
37. *Ng H. T.* 1997. Getting Serious about Word Sense Disambiguation. Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How? :1–7.
38. *Pantel P., Lin D.* 2002. Discovering Word Senses from Text. Proceedings of ACM Conference on Knowledge Discovery and Data Mining (KDD-02) : 613–619.
39. *Patwardhan S., Pedersen T.* 2003. The CPAN WordNet::Similarity Package.// http://search.cpan.org/~sid/WordNet-Similarity-0.05/ 2003
40. *Reddy S., Inumella I., McCarthy D., Stevenson M.* 2010. IIITH: Domain Specific Word Sense Disambiguation. Proceedings of SemEval-2010: 5th International Workshop on Semantic Evaluations.
41. *Reddy S., McCarthy D., Manandhar S., Gella, S.* 2011. Exemplar-based Word-Space Model for Compositionality Detection.
42. *Reznik P.* 1993. Selection and Information: A Class-Based Approach to Lexical Relationships.
43. *Rose T. G., Stevenson M., Whitehead M.* 2002. The Reuters Corpus Volume 1 — From Yesterday's News to Tomorrow's Language Resources. Proceedings of the Third International Conference on Language Resources and Evaluation : 827–833.
44. *Sag I., Baldwin T., Bond F., Copestake A., Flickinger D.* 2002. Multiword Expressions: A Pain in the Neck for NL. Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics : 1–15.
45. *Venkatapathy S., Joshi A. K.* 2005. Measuring the Relative Compositionality of Verb-Noun (V-N) Collocations by Integrating Features. Proceedings of the joint conference on Human Language Technology and Empirical methods in Natural Language Processing : 899–906.
46. *Weeds J.* 2003. Measures and Applications of Lexical Distributional Similarity.
47. *Weeds J., Weir D., McCarthy D.* 2004. Characterising Measures of Lexical Distributional Similarity. Proceedings of the 20th International Conference of Computational Linguistics : 1015–1021.

Section II.
Main program of the Conference

АВТОМАТИЧЕСКОЕ ОБНАРУЖЕНИЕ КВАЗИСИНОНИМОВ В НОВОСТНЫХ КЛАСТЕРАХ

А. Алексеев (a.a.alekseevv@gmail.com)

Н. Лукашевич (louk_nat@mail.ru)

Московский Государственный Университет, Москва, Россия

В данной работе рассматривается метод извлечения квазисинонимов — вариантов наименования одной и той же сущности в новостном кластере. Метод основан на тематической структуре новостного кластера и использует как сравнение разного рода контекстов употребления выражений, так и сопоставление употребления выражений в одних и тех же и соседних предложениях.

Ключевые слова: квазисинонимы, кластеры, новостные кластеры, контекст, метод, метод обнаружения, обнаружение.

AUTOMATIC DETECTION OF NEAR-SYNONYMS IN NEWS CLUSTERS

A. Alekseev (a.a.alekseevv@gmail.com)

N. Loukachevitch (louk_nat@mail.ru)

Lomonosov Moscow State University, Moscow, Russian Federation

The paper presents a method for extraction of alternative names of a concept or a named entity mentioned in a news cluster. The method is based on the structural organization of news clusters and exploits comparison

of various contexts of words. Word contexts are used as basis for multiword expression extraction and detection of alternative names. As a result of cluster processing we obtain groups of near-synonyms, in which the central synonym of each group is determined.

Key words: near-synonims, clusters, news clusters, context, method, detection method, detection

1. Introduction

An important step in news processing is thematic clustering of news articles describing the same event. Such news clusters are the basic units of information presentation in news services.

After a news cluster is formed, it undergoes various kinds of automatic processing:

— Duplicates are removed from the cluster. Duplicate is a message that almost completely repeats the content of an initial document,

— A cluster is categorized to a thematic category,

— A summary of a cluster is created, usually containing the sentences from different documents of the cluster (multi-document summary) etc.

The formation of a cluster can represent a serious problem. It is especially difficult to form clusters correctly for complex hierarchical events having some duration in time and distributed geographic location (world championships, elections) (Dobrov, Pavlov, 2010).

A part of news cluster forming and processing problems is due to the fact that in cluster documents, the same concepts or entities may be named differently. Lexical chain approaches could partly overcome this problem using thesaurus information (Li et. al., 2007; Loukachevitch, Dobrov, 2009). However in a pre-created resource, it is impossible to fix all possible alternatives for entities naming in various clusters. For example, the U.S. air base in Kyrgyzstan may be called in documents of the same news cluster as *Manas base, Manas airbase, Manas, base at Manas International Airport, U.S. base, U.S. air base* and etc.

The problem of alternative names for named entities is partly solved by coreference resolution techniques (*Russian President Dmitry Medvedev, President Medvedev, Dmitry Medvedev*) (Ermakov, 2007; Ng, 2005), but the variability of entity names in news clusters refers not only to concrete entities but also to concepts.

In this paper we consider a method for extraction of alternative names of a concept or a named entity mentioned in a news cluster. The method is based on the structural organization of news clusters and exploits comparison of various contexts of words. The word contexts are used as basis for multiword expression extraction and alternative names detection. At the end of cluster processing we obtain groups of near-synonyms, in which the main synonym of a group is determined. Such synonym groups include both single words and multiword expressions.

2. Principles of cluster processing

Processing of cluster texts is based on the structure of coherent texts, which have such properties as the topical structure and cohesion.

Van Dijk (Van Dijk, 1985) describes the topical structure of a text, the macrostructure, as a hierarchical structure in a sense that the theme of a whole text can be identified and summed up to a single proposition. The theme of the whole text can usually be described in terms of less general themes which in turn can be characterized in terms of even more specific themes. Every sentence of a text corresponds to a subtheme of the text.

The macrostructure of a connected text defines its global coherence: "Without such a global coherence, there would be no overall control upon the local connections and continuations" (Van Dijk, 1985). Sentences must be connected appropriately according to the given local coherence criteria, but the sequence would go simply astray without some constraint on what it should be about globally.

Cohesion, that is surface connectivity between text sentences, is often expressed through anaphoric references (i. e. pronouns) or by means of lexical or semantic repetitions. Lexical cohesion is modeled on the basis of lexical chains (Hirst, St-Onge, 1998).

The proposition of the main theme, that is interaction between theme participants, should be represented in specific text sentences, which should refine and elaborate the main theme. This means that if a text is devoted to description of relations between thematic elements $C_1 \ldots C_n$, then references to these participants should be met in different roles to the same verb in text sentences.

Thus if even very semantically close entities C_1 and C_2 often co-occur in the same sentences of a text, it means that the text is devoted to consideration of relations between these entities and they represent different elements of the text theme (Hasan, 1984; Loukachevitch, 2009). At the same time, if two lexical expressions C_1 and C_2 are rarely met in the same sentences but occur very frequently in neighbor sentences then we can suppose that they are elements of lexical cohesion, and there is a semantic relation between them.

A news cluster is not a coherent text but cluster documents are devoted to the same theme. Therefore statistical features of the topical structure are considerably enhanced in a thematic cluster, and on such a basis we try to extract unknown information from a cluster.

3. Stages of cluster processing

Cluster processing consists of three main stages. At the first stage noun and adjective contexts are accumulated. The second stage is devoted to multiword expression recognition. At the third stage the search of near-synonyms is performed.

In next sections we consider processing stages in more detail. As an example we use the news cluster, which is devoted to Kyrgyzstan and the United States agreement denunciation over U.S. air base located at the Manas International Airport (19.02.2009). This news cluster contains 195 news documents and is assembled on the basis of the algorithm described in (Dobrov, Pavlov, 2010).

3.1. Extraction of word contexts

Sentences are divided into segments between punctuation marks. Contexts of word W include nouns and adjectives situated in the same sentence segments as W. The following types of contexts are extracted:

— Neighboring words: neighboring adjectives or nouns situated directly to the right or left from W (*Near*),
— Across verb words: adjectives and nouns occurring in sentence segments with a verb, and the verb is located between W and these adjectives or nouns (*AcrossVerb*),
— Not near words: adjectives and nouns that are not separated with a verb from W and are not direct neighbors to W (*NotNear*).

In addition, adjective and noun words that occur in neighboring sentences are memorized (Ns). For this context extraction only sentence fragments from the beginning up to a segment with a verb are taken into consideration. It allows us to extract the most significant words from neighboring sentences.

To illustrate how these contexts can help in extraction of near-synonyms we ran the following experiment.

Documents of the example cluster were matched with RuThes thesaurus entries (Loukachevitch, 2011); pairs of synonyms and directly related expressions were extracted (*USA — American, Kyrgyzstan — Kyrgyz Republic, base — airbase* etc.). We took pairs of such expressions with the frequencies more than half of the number of documents in the cluster (98). Then we calculated co-occurrence of the expressions in the same sentences *(Near+NotNear+AcrossVerb)* and in neighbor sentences *(Ns)*. For thesaurus-related expressions the ratio between the values was:

(1) (Near+NotNear+AcrossVerb) / Ns= 0,56

If to take all other (not-related) pairs of thesaurus expressions found in the example cluster (with the same restriction on frequencies) and to calculate the same values and the ratio between them then we obtain ***2.09***. This confirms our idea that near-synonyms tend to occur more often in neighbor sentences than in the same sentences of a document.

3.2. Extraction of multiword expressions

We consider recognition of multiword expressions as a necessary step before near-synonym extraction. An important basis for multiword expression recognition is the frequency of word sequences (Witten et. al., 1999). However, a news cluster is a structure where various word sequences are repeated a lot of times. We supposed that the main criterion for multiword expression extraction from clusters is the significant excess in co-occurrence frequency of neighbor words in comparison with their separate occurrence frequency in segments of sentences (see (2), cf. Dobrov et. al., 2003):

(2) Near > 2 * (AcrossVerb + NotNear)

In addition, the restrictions on frequencies of potential component words are imposed.

Search for candidate pairs is performed in order of the value "*Near — (AcrossVerb + NotNear)*" reducing. In case that a suitable pair has been found, its component words are joined together into a single object and all contextual relationships are recalculated. The procedure starts again and repeats until at least one join is performed.

As a result, such expressions as *Parliament of Kyrgyzstan, the U.S. military, denunciation of agreement with the U.S., Kyrgyz President Kurmanbek Bakiyev* are extracted from the example cluster.

3.3. Detection of near-synonyms

At the third stage, search for near-synonyms is produced. For assuming a semantic relationship between expressions U_1 and U_2, the following factors are used:

— U_1 and U_2 have formal resemblance (for example, words with the same beginning),
— U_1 and U_2 occur more often in neighboring sentences than within segments of the same sentence,
— U_1 and U_2 have similar contexts based on Near, AcrossVerb, NotNear and Ns features, which are determined by calculating scalar products of corresponding vectors (NearScalProd, AVerbScalProd, NotNearScalProd, NsentScalProd),
— U_1 and U_2 should be enough frequent in a cluster to be evident statistically.

Note that if comparison of word contexts is a well known procedure for synonym detection and taxonomy construction (Yang, Callan, 2009), but generation of contexts from neighboring sentences has not been described in the literature.

Near-synonyms detection consists of several steps. A different set of criteria is applied at each step. The lookup is performed in order of frequency decreasing: for every expression U_1 all expressions U_2 having a lower frequency than U_1, are considered. If all conditions are satisfied, then less frequent expression U_2 is postulated as a synonym of U_1 expression, all U_2 contexts are transferred to U_1 contexts, the expressions U_1 and U_2 become joined together. As a result the sets of near-synonyms (synonym groups) are produced, i. e. linguistic expressions that are equivalent with respect to the content of the cluster.

We assume that U_1 and U_2 expressions, when they are enclosed in such a synonym group, are closely related in sense, or their referents in current cluster are closely related to each other, so that U_2 does not represent separate thematic significance with respect to U_1. For example, such words as *parliament* and *parliamentarian* have a close semantic relationship between them in general context, but they are not synonyms. But within a particular cluster, e. g., in which decision-making process in a parliament is discussed, these words may be classified as near-synonyms.

At the first step (3.1) semantic similarity between expressions consisting of similar words is sought, e. g. *Kyrgyzstan — Kyrgyz, Parliament of Kyrgyzstan — Kyrgyz Parliament*. We used simple similarity measure — the same beginning of words.

To connect words with the same beginning in synonym groups, the following conditions are required: the co-occurrence frequency in neighboring sentences is significantly higher than co-occurrence frequency in the same sentences (3, 4) (see section 3.1); both expressions should have sufficient frequencies in the cluster. The procedure is iterative:

(3) Ns > 2 * (AcrossVerb + Near + NotNear)

(4) Ns > 1

If expressions are rarely located in neighboring sentences (Ns <2), then the scalar product similarity of contexts is required:

(5) NearScalProd + NotNearScalProd + AVerbScalProd + NSentScalProd > 0.4

At the second step (3.2) semantic similarity between expressions, one of which is included into another, is sought, for instance, *Parliament — Parliament of Kyrgyzstan, airbase — Manas airbase*. The meaning of this step lies in the fact that a cluster might not mention any other parliaments, except of the Kyrgyz Parliament, i. e. in both cases the same object is mentioned. Similarity of neighbor contexts is required here:

(6) NearScalProd > 0.1

At the third step (3.3) we are looking for semantic similarity between the expressions with equal length and including at least one the same word, for example, *Manas Base — Manas Airbase, the U.S. military — the U.S. side* (7). High frequency of co-occurrence in neighboring sentences is required (8, 9):

(7) NS > 2 * (AcrossVerb + Near + NotNear)

(8) NS > 1

Finally, at the last step (3.4) semantic similarity between arbitrary linguistic expressions, mentioned in cluster documents, is searched, e. g. *USA — American, Kyrgyzstan — Bishkek*.

An assumption on semantic similarity between arbitrary expressions requires the maximum number of conditions: high frequency of co-occurrence in neighboring sentences (9, 10); restrictions on occurrence frequencies of candidates, context similarity:

(9) NS > 2 * (AcrossVerb + Near + NotNear)

(10) NS > 0.1 * MaxAcrossVerb

The following synonym groups were automatically assembled for the example cluster as a result of described stages (the main synonym of a group, which was automatically determined, is highlighted with bold font):

— ***Manas base:*** *base, Manas Air Base, Air Base, Manas;*
— ***USA:*** *American, America;*
— ***Kyrgyzstan:*** *Kirghizia, Kyrgyz, Kyrgyz-American, Bishkek;*
— ***Parliament of Kyrgyzstan:*** *Kyrgyz parliament, parliament, parliamentary, parliamentarian;*
— ***Manas International Airport:*** *airport, Manas airport;*
— ***Bill:*** *law, legislation, legislative, legal* and etc.

4. Evaluation of method

To test the introduced method we took 10 news clusters on various topics with more than 40 documents in each cluster.

Two measures of quality were tested for multiword expression extraction. Firstly, we evaluated the percentage of syntactically correct groups among all extracted expressions. Secondly, we have attracted a professional linguist and asked her to select the most significant multiword expressions (5–10) for each cluster, and to arrange them in descending order of importance.

So for the example cluster, the following expressions were considered significant by the linguist:

— *Manas Airbase,*
— *Parliament of Kyrgyzstan,*
— *Manas base,*
— *Kyrgyz Parliament,*
— *Denunciation of agreement,*
— *Government's decision.*

Note that such an evaluation task differs from evaluation of automatic keyword extraction from texts (Su Nam Kim et. al., 2010), when experts are asked to identify the most important thematic words and phrases of a text. In our case we tested exactly multiword expression extraction. In addition, a list created by the linguist could contain repetitions (*Parliament of Kyrgyzstan — Kyrgyz Parliament*).

364 multiword expressions were automatically extracted from test clusters, 312 (87.9%) of which were correct syntactic groups. With account of phrase frequencies, correct syntactic expressions achieved 91.4% precision. The linguist chose 70 most important multiword expressions for clusters and 72.6% of them were automatically extracted by the system.

We tested extracted synonym groups evaluating semantic relatedness of every synonym in a group to its main synonym. Every occurrence of supposed synonyms was tested. If more than a half of all occurrences of such a synonym in a cluster were related to the main synonym in the group, the synonymic relation was considered as correct.

Table 1 contains information about the quality of generated synonym groups calculated in number of expressions and in their frequencies.

Table 1. Test results for automatic detection of synonym groups in news clusters

Step	Number of joins	Total join frequency	Percent of correct joins	Percent of correct joins by frequency
3.1. The same beginning expressions joining	155	4383	87.9%	91.4%
3.2. Embedded expressions joining	99	9131	91.4%	92.9%
3.3. Intersecting expressions joining	8	677	85.7%	80.8%
3.4. Arbitrary expressions joining	38	4822	62.5%	62.4%

To assess the contribution of co-occurrence in neighboring sentences, we conducted detailed testing of the same beginning expression joining (step 3.1) for the example cluster (Table 2). Table 2 shows that adding Ns factor, as it is done in step 3.1, improves precision and recall of near-synonym recognition.

Table 2. Test results for the different methods of the same beginning synonym joining

Method	Number of joined expressions	Total joining frequency	Correct joining frequency	Precision by frequency (%)	Recall byfrequency (%)
Expressions with the same beginning (BasicLine)	383	2266	1472	65%	100%
Expressions with the same beginning + scalar products (threshold 0.1)	38	996	834	83.7%	56.7%
Expressions with the same beginning + scalar products (threshold 0.4)	36	976	814	83.4%	55.3%
Step 3.1 conditions	36	965	873	**90.5%**	**59.3%**

Conclusion

In this paper we have described two experiments on news clusters: multiword expression extraction and near-synonyms detection. In addition to known methods of contexts comparison, we exploited co-occurrence frequency in neighboring sentences for synonym detection. We conducted the testing procedure for the introduced method.

In future we are going to use extracted near-synonyms in such operations as cluster boundaries correction, automatic summarization, novelty detection, formation of subclusters and etc. We also intend to study methods of combination automatically extracted near-synonyms and thesaurus relations.

References

1. *Dii k van T.* 1985. Semantic Discourse Analysis. Handbook of Discourse Analysis : 103–136.
2. *Dobrov B., Loukachevitch N., Syromyatnikov S.* 2003. Automatic Detection of Text Entries for Information Retrieval Thesaurus. Proceedings of the fifth Russian Scientific Conference "Digital Libraries: Advanced Methods and Technologies" : 201–210.
3. *Dobrov B., Pavlov A.* 2010. Basic Line for News Clusterization Methods Evaluation. Proceedings of the fifth Russian Scientific Conference "Digital Libraries: Advanced Methods and Technologies".
4. *Dobrov B., Loukachevitch N., Shternov S.* 2005. News Processing Based on Large Linguistic Resource. Internet Mathematics, available at: http://download.yandex.ru/company/grant/2005/10_Loukachevitch_103030.pdf
5. *Dobrov B., Loukachevitch N.* 2009. Summarization of News Clusters Based on Thematic Representation. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2009" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2009") : 299–305.
6. *Ermakov A.* 2007. Automatical Extraction of Facts from Texts of Personal Files: Experience in Anaphora Resolution. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2007" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2007").
7. *Hasan R.* 1984. Coherence and Cohesive Harmony. Understanding Reading Comprehension :181–219.
8. *Hirst G., St-Onge D.* 1998. Lexical Chains as Representation of Context for the Detection and Correction Malapropisms. WordNet: An Electronic Lexical Database and Some of its Applications.
9. *Li J., Sun L., Kit C., Webster J.* 2007. A Query-Focused Multi-Document Summarizer Based on Lexical Chains. Proc. of the Document Understanding Conference DUC-2007.
10. *Loukachevitch N.* 2009. Multigraph Representation for Lexical Chaining. Proc. of SENSE workshop :67–76.

11. *Loukachevitch N.* 2011. Thesauri for Information Retrieval Tasks.
12. *Ng V.* 2005. Machine Learning for Coreference Resolution: From Local Classification to Global Ranking. Proc. of ACL-2005.
13. *Su Nam Kim, Medelyan O., Min-Yen Kan, Baldwin T.* 2010. Automatic Keyphrase Extraction from Scientific Articles. Proc. of the 5-th International Workshop on Semantic Evaluation, ACL -2010: 21–26.
14. *Witten I., Paynter G., Frank E., Gutwin C., Newill-Manning C.* 1999. KEA.: Practical Automatic Keyphrase Extraction. Proc. of the fourth ACM Conference on Digital Libraries.
15. *Yang H., Callan J.* 2009. A Metric-Based Framefork for Automatic Taxonomy Induction. Proc. of ACL-2009.

ПАРАЛЛЕЛЬНОЕ СОЗДАНИЕ СЛАВЯНСКИХ ГРАММАТИЧЕСКИХ РЕСУРСОВ

Таня Августинова (avgustinova@coli.uni-saarland.de)

DFKI GmbH & Saarland University

Saarbruecken

Представлена идея параллельного создания компьютерных грамматик для славянских языков на основе формализма HPSG с использованием общеславянского модуля и совместимых с ним расширений для отдельных языков. Важным требованием к проекту является динамичная связь между создаваемыми грамматиками и синтаксически размеченными корпусами.

Ключевые слова: компьютерная грамматика, HPSG, славянские языки, грамматический ресурс, славянская грамматика.

PARALLEL CONSTRUCTION OF SLAVIC GRAMMATICAL RESOURCES

Tania Avgustinova (avgustinova@coli.uni-saarland.de)

DFKI GmbH & Saarland University

Saarbruecken

We present the idea of parallel construction of HPSG-based grammatical resources for Slavic languages using a common Slavic core module in combination with language specifi c extensions and corpus-based grammar elaboration at all stages of the project.

Key words: grammatical resource, HPSG, Slavic languages, computational grammar, Slavic grammar

1. Introduction

Our long-term goal is to develop grammatical resources for Slavic languages and to make them freely available for the purposes of research, teaching and natural language applications. We build upon our previous research in language-family oriented grammar design and systematic specification of shared and non-shared grammar for modeling Slavic morphosyntax. An imperative objective in this context concerns what we see as methodical corpus-based grammar elaboration. To this effect, we envisage exploiting freely available linguistically interpreted corpora at all stages of the project and especially in discovering structured knowledge to be reflected in our grammars. Interfacing a morphological analyzer is a crucial prerequisite for any grammar development activity involving Slavic languages. For research purposes, such systems are by and large freely available nowadays, and the grammar engineering environment we plan to use provides the required interface for integrating a morphological pre-processor. An important desideratum for the individual resource grammars is to eventually couple them with syntactically interpreted text corpora (treebanks) which either pre-exist or will be constructed in parallel. The development of Slavic resource grammars is part of an on-going international collaborative effort which became popular under the name DELPH-IN.[1] It is based on a shared commitment to re-usable, multi-purpose resources and active exchange. Our project utilizes DELPH-IN software (linguistic knowledge builder[2] with integrated evaluation and benchmarking tools[3]) as a grammar development platform, and has strong affinity to the LinGO[4] Grammar Matrix. We envision a core Slavic grammar whose components can be commonly shared among the set of languages, and facilitate individual resource grammar development.

The Slavic core grammar is intended to encode mutually interoperable analyses of a wide variety of linguistic phenomena, taking into account eminent typological commonalities and systematic differences. Determining what counts as a worthwhile linguistic phenomenon is a challenge in its own right. for a corresponding operational notion, however, it would suffice to conclusively reflect the fact that what grammatical representations have in common, independently of their theoretical origin or purpose, is that they (i) identify linguistic items of different motivation and complexity, (ii) encode their properties, and (iii) specify explicit or implicit relationships between

---

1 Deep Linguistic Processing with HPSG Initiative, URL: http://www.delph-in.net/

2 The LKB (Linguistic Knowledge Builder) system is a grammar and lexicon development environment for use with unification-based linguistic formalisms. While not restricted to HPSG, the LKB implements the DELPH-IN reference formalism of typed feature structures (jointly with other DELPH-IN software using the same formalism).
URL: http://wiki.delph-in.net/moin/LkbTop

3 [incr tsdb()] — URL: http://www.delph-in.net/itsdb/

4 The Linguistic Grammars Online (LinGO) team is committed to the development of linguistically precise grammars based on the HPSG framework, and general-purpose tools for use in grammar engineering, profiling, parsing and generation. URL: http://lingo.stanford.edu/

them. As the interconnectedness of grammatical phenomena is at the heart of research in theoretical syntax, one of our objectives is to contribute a language-family oriented perspective to the data-driven cross-linguistic exploration of that interconnection. Our concept of Slavic core grammar will shape up and crystallize through rigorous testing in parallel grammar engineering for a closed set of languages for which a variety of linguistic resources is already available. All individual grammars will be designed to support an innovative implementation of a Slavic core module that consolidates strategies for constructing a cross-linguistic resource.

2. Background

Rule-based precision grammars are linguistic resources designed to model human languages as accurately as possible. Unlike statistical grammars, they are hand-built and take into account the respective grammarian's theory and analysis of how to best represent various syntactic and semantic phenomena in the language of interest. A side effect of this is that such grammars tend to substantially differ from each other, with no established best practices or common representations.[5] As implementations evolved for several languages within the formalism of Head-driven Phrase Structure Grammar (Pollard and Sag 1994), it became clear that homogeneity among existing grammars could be increased and development cost for new grammars greatly reduced by compiling an inventory of cross-linguistically valid (or at least useful) types and constructions. Hence the LinGO Grammar Matrix has been set up as a multilingual grammar engineering project (Bender et al. 2002) in an attempt to distil the wisdom of already existing broad coverage grammars and document it in a form that can be used as the basis for new grammars. The generalizations observed across linguistic objects and across languages result in a cross-linguistic type hierarchy[6] coming with a collection of phenomenon-specific libraries, which would optimally represent salient dimensions of cross-linguistic variation.

The original Grammar Matrix consisted of types defining the basic feature geometry (Copestake et al. 2001), types for lexical and syntactic rules encoding the ways that heads combine with arguments and adjuncts, and configuration files for the LKB grammar development environment (Copestake 2002) and the PET system (Callmeier 2000). Subsequent releases have refined the original types and developed a lexical hierarchy, including linking types for relating syntactic to semantic

---

[5] Exceptions do exist, of course: ParGram (Parallel Grammar) project is one example of multiple grammars developed using a common standard. It aims at producing wide coverage grammars for a wide variety of languages. These are written collaboratively within the linguistic framework of Lexical Functional Grammar (LFG) and with a commonly-agreed-upon set of grammatical features. URL: http://www2.parc.com/isl/groups/nltt/pargram/

[6] In a lexicalized constraint-based framework, the grammars are expressed as a collection of typed feature structures which are arranged into a hierarchy such that information shared across multiple lexical entries or construction types is represented only on a single super-type.

arguments, and the constraints required to compositionally build up semantic representations in the format of Minimal Recursion Semantics (Copestake et al. 2005; Flickinger and Bender 2003; Flickinger et al. 2003). These constraints are intended to be language-independent and monotonically extensible in any given grammar. In its recent development, the Grammar Matrix project aims at employing typologically motivated, customizable extensions to a language-independent core grammar (Bender and Flickinger 2005) to handle cross-linguistically variable but still recurring patterns. A web-based configuration tool eliciting typological information from users-linguists through a questionnaire is currently under active construction. While users specify phenomena relevant to their particular language, the resulting selections are compiled from libraries of available analyses into starter grammars which can be immediately loaded into the LKB environment in order to parse sentences using the rules and constraints defined therein. The regression testing facilities of [incr tsdb()] allow for rapid experimentation with alternative analyses as new phenomena are brought into the grammars (Oepen et al. 2002). The ultimate ambition is thus to allow the linguist to revise decisions in the face of new information or improved linguistic analyses. Apart from the shared 'core' in the Grammar Matrix the customization script treats the individual languages as separate instances, which is rather insufficient for our purposes. Because it is driven purely by the specific phenomena in the target language, this strategy is consistent with "bottom-up" data driven investigation of linguistic universals and constraints on cross-linguistic variation. Obviously, the fact that we have to do with a group of systematically related languages cannot be taken into account in the original setting. With grammars being created individually, the treatment of shared phenomena would work to the degree that satisfies but does not guarantee cross-linguistic compatibility. It is a legitimate expectation, though, that the constraint definitions supplied to grammar developers can be extended to also capture generalizations holding only for subsets of languages. It is essential therefore to augment the approach with a "top-down" perspective introducing intermediate levels of typological variation.

3. Shared grammar

Successful multilingual natural language processing systems employ generic linguistic resources that are adaptable to specific language and application requirements. If parallel grammars for more than one language are needed for an application like machine translation or computer-assisted language learning, it pays off to define and implement shared grammars. The reuse of portions of grammars for the description of additional languages speeds up grammar development which is a demanding and time consuming task. A shared grammar approach not only facilitates the difficult task of maintaining consistency within and across the individual parallel grammars, but it also strongly supports for the area of natural language processing the prospects of what in programming language research is called modularity. In applied computational linguistics the need for employing operational notions of shared grammar stems from multilingual grammar engineering — cf. projects like (DiET 1997–1999; LinGO 2002; LS-GRAM 1994–1996;

ParGram 1995–2002; TSNLP 1993–1995; XTAG 2002). Computational linguists engaged in multilingual grammar development have always tried to reduce their labour by importing existing grammar components in a simple copy-paste-modify fashion. But there were also a number of systematic attempts to create and describe shared grammars that are convincingly documented in publications. (Kameyama 1988) demonstrates the concept for a relatively restricted domain, the grammatical description of simple nominal expressions in five languages. (Bemová et al. 1988) were able to exploit the grammatical overlap of two Slavic languages, for the design of a lean transfer process in Russian to Czech machine translation. In multilingual application development within Microsoft research, grammar sharing has extensively been exploited (Gamon et al. 1997; Pinkham 1996). Current international collaborative efforts within the DELPH-IN partnership (Uszkoreit et al. 2001; Uszkoreit 2002a; b) exploit the notion of shared grammar both for the rapid development of grammars for new languages and for the systematic adaptation of grammars to variants of languages. The leading idea is to combine linguistic and statistical processing methods for getting at the meaning of texts and utterances. Based on contributions from several members and joint development over many years, an open-source repository of software and linguistic resources has been created that already enjoys wide usage in education, research, and application building.

The construction of shared-grammar fragments proposed in (Avgustinova 2007) presupposes a common core module which is abstract enough to be shared by all Slavic languages modulo the appropriate further specification. For a language family, this module is expected to be relatively large and to cover the major phenomena areas. Certainly, there are groups and sub-groups of languages exhibiting particular properties and phenomena which are not attested in other members of the family. Yet, these phenomena constitute natural extensions of the common core module. So, for instance, one could distinguish a South-Slavic extension or an East-Slavic extension, and possibly extensions of any further granularity. Nevertheless, there are language-specific traits that identify specific languages and dialects. While the common core module is expected to be relatively large and to cover all major phenomena areas, it has still to be abstract enough in order to be shared by all Slavic languages modulo the appropriate further specification. Intuitively, the core incorporates what is interpretable as typical Slavic. The extensions can be of different granularity in order to encode properties and phenomena that are characteristic of respective subgroups, but need not be attested in other members of the family. Yet, all these phenomena have to be consistent with the common core module, constituting natural extensions. For example, a modular Bulgarian grammar in such a setting would include the common core, the South-Slavic extension, and the Bulgarian extension.

Methodologically, the adopted shared-grammar perspective reveals two different aspects of structuring the grammatical knowledge. One is cross-linguistic and can by and large be viewed as consisting of an under-specified core and a competence-driven specification "switching on" various parameters. The other aspect of structuring the grammatical knowledge can be called intra-linguistic and concerns the interesting interaction with specialized domain ontologies that model the expert knowledge in particular subject domains like electrical engineering or microbiology. It is never the case that the entire power of a natural language grammar is employed

in restricted specialized areas. A domain-specific grammar extraction relies, therefore, on specialized text corpora. A rather naive approach would be to assume an over-specified full-fledged grammar of a given language in combination with a performance-driven relaxation "switching off" various parameters. More attractive, however, is the idea that specific domain ontologies interact with grammatical ontologies to derive restricted grammars of, e. g., Russian or Bulgarian as used in, e. g., electrical engineering or microbiology. A theoretically rewarding result is the straightforward model of the linguistic knowledge of a domain expert. For example, an electrical engineer or a microbiologist hardly needs to be fluent in all languages he uses in order to be effectively multilingual in his restricted subject domain.

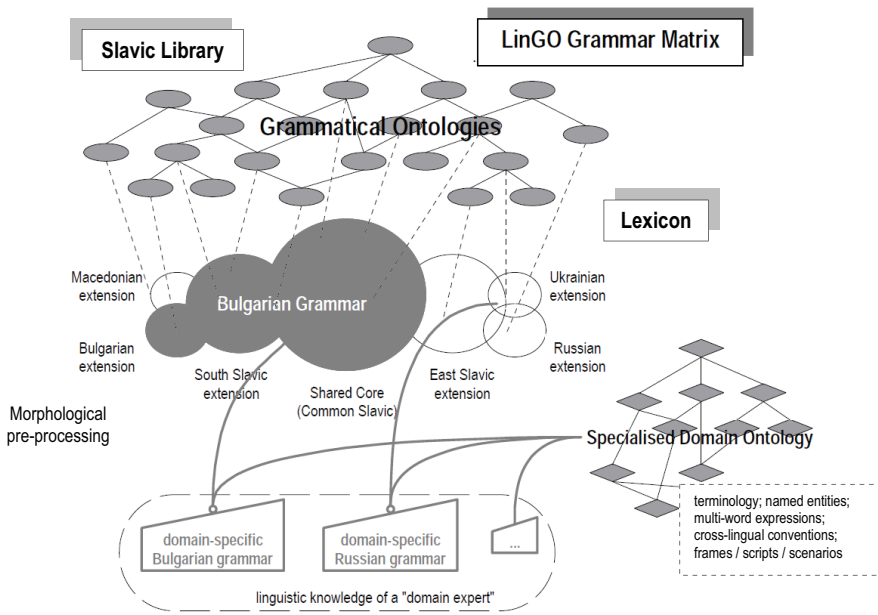


Figure 1. Resource architecture

4. Grammar resource development

Any approach to computational grammar design which maintains the notion of grammar sharing lends itself to a formal linguistic description of individual languages as well as groups of languages motivated by genetic origin or areal contact. The strategy we adopt is meant by design to be compatible with the current Grammar Matrix program: we use the customization system to quickly build small grammars for individual languages; shared analyses are put into a Slavic core; when the next language is added, the Slavic core helps to more efficiently build the new grammar, simultaneously receiving a cross-Slavic validation. Yet, a distinctive feature of our

approach to Slavic grammatical resources is that grammar engineering for each individual language takes place in a common Slavic setting. This in particular means that if, for example, two possibilities are conceivable of how to model a particular phenomenon observed in a certain Slavic language we strongly prefer the option that would potentially be consistent with what is found in the other grammars. The reason is that related languages share a much wider range of linguistic information than typically assumed in standard multilingual grammar architectures. We can, as a result, directly and effectively work with what has traditionally been regarded as "prototypically Slavic".

Currently we focus on the Russian resource grammar as a showcase on how its development can be assisted by interfacing with existing corpora and processing tools for the language (Avgustinova and Zhang 2009a; b; c; d; e). Applying the innovative corpus-oriented grammar development approach proposed by (Miyao et al. 2005) to the syntactically annotated and manually disambiguated part of the Russian National Corpus (Boguslavsky et al. 2000; Boguslavsky et al. 2002) we can obtain a unique Russian HPSG-style treebank (Avgustinova and Zhang 2010). We have also been looking into data-driven approaches of dependency parsing with the SynTagRus treebank. Specifically, we have rebuilt the transition-based dependency parsing models and cross-compare results with those reported in (Nivre et al. 2008). Subsequently we shall concentrate on resource grammars for Bulgarian and Polish, thus including representatives of the three main subgroups in the Slavic language family: East Slavic (Russian), South Slavic (Bulgarian); West Slavic (Polish).

For an illustration let us consider the Slavic case system, because it provides abundant scope for complex categorisation with many cross-linguistic tendencies. From the morphological perspective, there exists a spectrum of case marking possibilities. The most common and typical are the synthetic means like suffixation and inflexion, possibly in combination with supra-segmental distinctions. Apart from that, the case marking can involve analytical adpositional means like prepositions or postpositions. A further case marking possibility is the suppletion of forms (e. g., in pronominal paradigms), where the ultimate union of stem and case can be observed. Syntactically, there are two general ways in which case is acquired by the respective case-marked category: in concord (due to case-matching between a governor and a dependent) or under government (via non-congruent, non-agreeing case selection). This naturally results in distinguishing *concordial case* and *relational case*. Even though the term "case marking" traditionally refers to inflectional marking, it could successfully be extended to cover adpositions.

The majority of Slavic languages exhibits a rich inflectional case marking system and is traditionally classed among the *synthetic* languages. Characteristic of the synthetic language type is that prepositions, like verbs, *govern* cases. Moreover, relational cases can be expressed by the combination of a preposition and case inflection. Consider, for example, Russian prepositions such as *v* ('in'), *na* ('on'), *pod* ('under'), etc. Their combination with the locative/prepositional case inflection encodes location, while their combination with the accusative case inflection expresses direction. In the *analytic* language type, the situation is rather different.

Adpositions bear the sole burden of marking the relations and, thus, of expressing relational cases. For example, Bulgarian prepositions *combine* with the oblique form of the substantive, whereby any suffix or inflection, if available, is redundant with respect to case marking. Also, recall in this context that the above-mentioned opposition direction vs. location is altogether lost in this Slavic language.

To respond to the need of morphosyntactic abstraction over regular case variation and language-specific constraints with respect to case marking, the notion of *functional case* is employed. A shared Slavic case taxonomy encoded as a multiple inheritance hierarchy is sketched below — for detailed motivation cf. (Avgustinova 2007) p. 25–34.

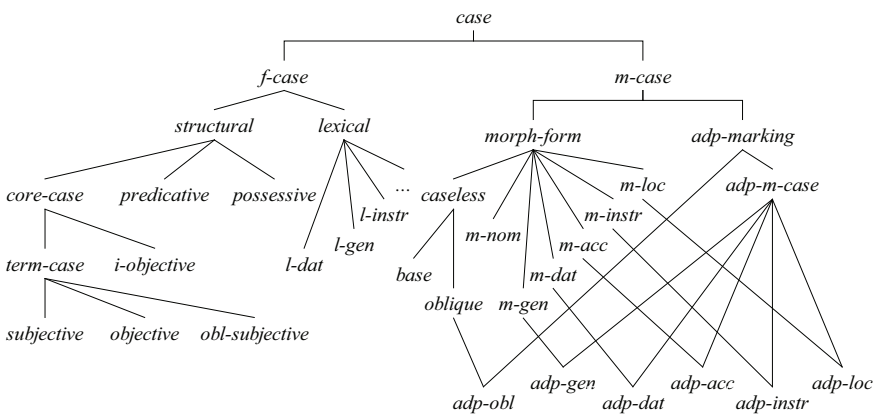


Figure 2. Slavic case system

The type *case* is classified along two dimensions: functional (*f-case*) and marking (*m-case*). Two types of relational (subcategorised) cases are assumed: *lexical* (inherent) and *structural* (syntactic, grammatical). These disjoint types partition the type *case* along the functional dimension, hence, are subtypes of *f-case* in the type hierarchy sketched below. This in particular means that in the lexical entry of a verb, the case value of subcategorised nominal categories is either specific, lexically predetermined or systematically under-specified, i. e. to be resolved by grammatical constraints / principles. A case value specified as *structural* is to be further instantiated to a particular (more specific) instance of *functional case*, namely, *subjective*, *objective* and *obl-subjective* (that is, the case specification of passivised subjects). These abstract case values will be expanded to their concrete instances on the basis of lexical and contextual constraints, taking into consideration the relevant (language-specific) morphological and adpositional case marking. An important aspect of such an approach is that the distinction between lexical and structural cases is applied not only to morphological, but also to adpositional case marking. In particular, the marking dimension of the case hierarchy is refined by allowing classification of *m-case* according to morphological form (*morph-form*) and adpositional marking (*adp-marking*). The type *morph-form* extends further either to concrete case inflection, i. e. to morphological nominative

(*m-nom*), morphological genitive (*m-gen*), morphological dative (*m-dat*), and so on, or just to *caseless*, i. e. to *base* or *oblique* morphological form — in the case of Bulgarian nouns. The type *adp-marking* encodes the adpositional marking on the noun, i. e. the respective PP; in particular, it interacts with *morph-form* in specifying the types *adp-oblique* (for Bulgarian) and *adp-m-case* (for other Slavic languages).

5. Outlook

The formal specification of shared grammar is also extremely important for developing stringent models of language change. Historical linguistics and sociolinguistics need formal models of grammar in which possible and factual shared grammars can be specified. It is a justified expectation that a formal notion of shared grammar should also be useful for theoretical and applied work on second-language acquisition. The precise specification of shared grammar could explain preferences of the second-language learner as well as contamination and interference phenomena. Designing specialized methodologies for second language learning that take into account the properties of the learner's first language could likewise benefit from a good description of shared grammar.

References

1. *Avgustinova T., Zhang Y.* 2009. Parallel Grammar Engineering for Slavic Languages. Workshop on Grammar Engineering Across Frameworks at the ACL/ IJCNLP 2009 Conference.
2. *Avgustinova T., Zhang Y.* 2009. Developing a Russian HPSG based on the Russian National Corpus. DELPH-IN Summit.
3. *Avgustinova T., Zhang Y.* 2009. Exploiting the Russian National Corpus in the Development of a Russian Resource Grammar. Workshop on Adaptation of Language Resources and Technology to New Domains at the RANLP 2009 Conference.
4. *Avgustinova T., Zhang Y.* 2009. Exploiting the Russian National Corpus in the Development of a Russian Resource Grammar. Proceedings of the RANLP-2009 Workshop on Adaptation of Language Resources and Technology to New Domains.
5. *Avgustinova T., Zhang Y.* 2009. Parallel Grammar Engineering for Slavic Languages. Workshop on Grammar Engineering Across Frameworks at the ACL/ IJCNLP.
6. *Avgustinova T., Zhang Y.* 2010. Conversion of a Russian Dependency Treebank into HPSG Derivations. Proceedings of the 9 th International Workshop on Treebanks and Linguistic Theories (TLT'9).
7. *Bemová A., Oliva K., Panevová J.* 1988. Some Problems of Machine Translation between Closely Related Languages. COLING'88.

8. *Bender E. M., Flickinger D., Oepen S.* 2002. The Grammar Matrix: An Open-Source-Kit for the Rapid Development of Cross-Linguistically Consistent BroadCoverage Precision Grammars. Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics : 8–14.
9. *Bender E. M., Flickinger D.* 2005. Rapid Prototyping of Scalable Grammars: Towards Modularity in Extensions to a Language-Independent Core. 2nd International Joint Conference on Natural Language Processing.
10. *Boguslavskii I., Grigor'eva S., Grigor'ev N., Kreidlin L., Frid N.* 2000. Dependency Treebank for Russian: Concept, Tools, Types of Information. COLING : 987–991.
11. *Boguslavskii I., Chardin I., Grigor'eva S., Grigor'ev N., Iomdin L., Kreidlin L., Frid. N.* 2002. Development of a Dependency Treebank for Russian and its Possible Applications in NLP. Third International Conference on Language Resources and Evaluation (LREC-2002) : 852–856.
12. *Callmeier U.* 2000. PET — a Platform for Experimentation with Efficient HPSG Processing Techniques. Natural Language Engineering, 6 : 99–107
13. *Copestake A., Lascarides A., Flickinger D.* 2001. An Algebra for Semantic Construction in Constraint-based Grammars. The 39th Meeting of the Association for Computational Linguistics.
14. *Copestake A.* 2002. Implementing Typed Feature Structure Grammars.
15. *Copestak, A., Flickinger D., Sag I. A., Pollard C.* 2005. Minimal Recursion Semantics: An Introduction. Journal of Research on Language and Computation, 3 (4) : 281–332.
16. *DiET,* 1997–1999. Diagnostic and Evaluation Tools for Natural Languages Applications, available at: http://diet.dfki.de/.
17. *Flickinger D., Bender E. M.* 2003. Compositional Semantics in a Multilingual Grammar Resource. ESSLLI Workshop on Ideas and Strategies for Multilingual Grammar Development : 33–42.
18. *Flickinger D., Bender E. M., Oepen S.* 2003. MRS in the LinGO Grammar Matrix: A Practical User's Guide.
19. *Gamon M., Lozano C., Pinkham J., Reutter T.* 1997. Practical Experiencer with Grammar Sharing in Multilingual NLP.
20. *Kameyama M.* 1988. Atomization in Grammar Sharing. 26th Annual Meeting of ACL
21. *LinGO.* 2002. Linguistic Grammars Online, available at: http://lingo.stanford.edu/.
22. *LS-GRAM.* 1994–1996. Large-Scale Grammatical Resources, available at: http://clwww.essex.ac.uk/group/projects/lsgram/.
23. *Miyao Y., Ninomiya T., Tsujii J.* 2005. Corpus-oriented Grammar Development for Acquiring a Head-driven Phrase Structure Grammar from the Penn Treebank. Natural Language Processing — IJCNLP 2004, LNAI3248 : 684–693.
24. *Nivre J., Boguslavskii I., Iomdin L.* 2008. Parsing the SynTagRus Treebank. COLING : 641–648.
25. *Oepen S., Tutanova K., Shieber S., Manning C., Flickinger D., Brants T.* 2002. The LinGO Redwoods treebank. Motivation and Preliminary Applications. 19th International Conference on Computational Linguistics

26. *ParGram.* 1995–2002. Parallel Grammar Project, available at: http://www.parc.xerox.com/stl/groups/nltt/pargram/.
27. *Pinkham J.* 1996. Grammar Sharing in French and English. IANLP'96.
28. *Pollard C., Sag I.* 1994. Head-Driven Phrase Structure Grammar TSNLP. 1993–1995. Test Suites for Natural Language Processing, available at: http://cl-www.dfki.uni-sb.de/tsnlp/.
29. *Uszkoreit H., Flickinger D., Oepen S.* 2001. Proposal of Themes and Modalities for International Collaboration on Deep Linguistic Processing with HPSG.
30. *Uszkoreit H.* 2002. DELPHIN: Deep Linguistic Processing with HPSG — an International Collaboration.
31. *Uszkoreit H.* 2002. New Chances for Deep Linguistic Processing. The 19th International Conference on Computational Linguistics COLING'02
32. *XTAG.* 2002. Lexicalised Tree Adjoining Grammar, available at: http://www.cis.upenn.

СЕМАНТИЧЕСКИЕ ОТНОШЕНИЯ ВО ФРАЗЕОЛОГИИ

А. Н. Баранов (baranov_anatoly@hotmail.com)

Д. О. Добровольский (dm-dbrv@yandex.ru)

Институт русского языка им. В. В. Виноградова,
Москва, Россия

Обычно выделяемые семантические отношения — синонимия, антонимия, полисемия, отношения включения (гиперо-гипонимии), конверсии и каузативности — проявляются во фразеологии специфическим образом. В докладе подробно рассматриваются некоторые из указанных типов отношений применительно к фразеологизмам.

Ключевые слова: фразеология, фразеологизмы, семантические отношения, типы семантических отношений.

SEMANTIC RELATIONS IN PHRASEOLOGY

A. N. Baranov (baranov_anatoly@hotmail.com)

D. O. Dobrovol'skii (dm-dbrv@yandex.ru)

Vinogradov Russian Language Institute,
Moscow, Russian Federation

Traditionally, the following types of semantic relations in the lexical system are distinguished: synonymy, antonymy, polysemy, hyponymy, conversion, and causativity. In the field of phraseology, these phenomena display some specific properties. The focus of our paper is on revealing and discussing some of these properties. The starting point of the discussion is the category of semantic fi eld. It provides the theoretical framework for considering semantic relationships between idioms. The semantic field is defined as a set of lexical units which are connected with each other by some salient semantic features. The totality of semantic fi elds along with the conceptual links between them constructs the thesaurus of a given language, which can be represented in the form of a semantic network. The most important type of semantic relations within the semantic fi eld is synonymy. Full synonymy is a rare phenomenon in phraseology, because the meaning of an idiom contains additional semantic features, namely the so-called image component. Idioms with identical actual meanings often reveal differences in their image components, and are perceived as near-synonyms, rather than full

synonyms. Antonymy is not typical of phraseology because in most cases it is impossible to single out the central semantic feature that could be considered responsible for meaning contrasts. Although traditionally idioms were mainly regarded as monosemous units of the lexicon, the results of our recent research prove that idioms' polysemy is a quite typical phenomenon.

Key words: phraseology, phraseologisms, semantic relations, types of semantic relations

Традиционные типы семантических отношений в лексике — это синонимия, антонимия, полисемия, отношения включения (гиперо-гипонимии), конверсии и каузативности. Во фразеологии эти отношения проявляются весьма специфическим образом. Особенности реализации указанных отношений во фразеологической системе обсуждались в литературе, однако явно недостаточно. Рассмотрим последовательно указанные типы отношений применительно к фразеологизмам, начав с понятия семантического поля, в рамках которого и формируются семантические связи между фразеологическими единицами.

1. Семантическое поле

В лексической семантике под семантическим полем понимается совокупность лексических единиц, объединенных общим нетривиальным семантическим признаком. Слишком абстрактные — «тривиальные» — семантические признаки порождают необозримые поля, не обладающие психологической реальностью для носителей языка. Так, множество одушевленных существительных вряд ли может рассматриваться как семантическое поле. Аналогично глаголы с семантическим компонентом 'сделать так, чтобы' также не образуют семантического поля, а, скорее, задают лексико-грамматическую категорию каузативности. С другой стороны, такие смыслы, как 'строение', 'транспортное средство', 'ложь/обман', 'цвет' формируют интуитивно приемлемые семантические поля.

Поскольку фразеологизмы представляют собой единицы словаря, то и в сфере фразеологии также выделяются семантические поля. Это, например, такие объединения фразеологизмов, как

ГНЕВ (*рвать и метать; пойти вразнос; заводиться с пол-оборота; с резьбы сорваться; выйти из себя; вожжа под хвост попала*);

СМЕХ (*животики надорвать; скалить зубы; лежать в лёжку; смешинка в рот попала*);

РАДОСТЬ (*прыгать до потолка; на седьмом небе; именины сердца; маслом по сердцу; ловить балду*);

СТРАХ (*душа в пятки ушла; в штаны наложил; дрожмя дрожать; слаб в коленках; мороз по коже; поджилки трясутся*);

ТРУСОСТЬ (*заячья душонка; страусиная политика; не сметь рта раскрыть; поджать хвост; спасать [свою] шкуру; прятаться за мамину юбку*);

ГОРЕ (*все глаза выплакал; рвать душу [на части]; сердце разрывается; хоть волком вой; хоть ложись да/и помирай*);

ПЬЯНСТВО (*принять на грудь; бутылку раздавить; под градусом; под мухой; глаза залить; хватить лишку; спится с круга; заложить за воротник*);

ДОБРОТА (*ангел во плоти; воды не замутит; тише воды, ниже травы; с человеческим лицом; и мухи не обидит*).

Организация словника по семантическим полям характерна для словарей-тезаурусов (см. [Тезаурус]). Традиционно считается, что структура полей словаря-тезауруса иерархична и представляет собой упорядоченное дерево. Исследования последних десятилетий показали, что лексика организована не по древесному принципу. Более того: для предикатной лексики (прежде всего для глаголов) данное отношение, скорее, является исключением, чем правилом. Сходным образом организована и фразеология. Во фразеологической системе представлены отношения причины-следствия, фазовости состояния или действия и ряд других неиерархических отношений. Например, существует группа фразеологизмов, которые описывают различные этапы развития конфликта и различные виды конфликта. Так, идиома *бросить перчатку* указывает на начало конфликта, *подливать масло в огонь* — на его углубление. Конец конфликта обозначается идиомами *идти на попятную, спустить на тормозах, закурить трубку мира*. Среди видов конфликта выделяются: спор (*копья ломать, с пеной у рта*), ссора/скандал (*катить бочку, переть бугром, сцена у фонтана*), конкуренция (*дышать в затылок, наступать на пятки*) и др. Таким образом, распределение лексики (и фразеологии) по семантическим полям должно ориентироваться не только на родо-видовые отношения.

Тезаурус можно представить в виде семантической сети. Семантическая сеть — это множество узлов, связанных между собой дугами, отражающими семантические отношения между узлами. При таком представлении узлы соответствуют семантическим полям. Семантическое поле — это и феномен, обладающий психологической реальностью, и удачный теоретический конструкт. Рассмотрим основные отношения, которые могут быть представлены в семантическом поле, подробнее.

2. Отношение синонимии

Важнейший тип отношений, связывающих лексические единицы внутри семантического поля, — отношение синонимии. По понятию синонимии существует обширная литература; см. по этому поводу, например, [Апресян 1995],

[Шмелев 2006: 114 и далее]. Синонимы в системе языка — это лексические единицы, имеющие одинаковую категориальную принадлежность, значения которых в основном совпадают, при этом формально определить меру совпадения довольно трудно. По большей части, это определяется интуитивно. Синонимы в речи — это лексические единицы, различия между которыми несущественны в конкретной ситуации общения.

Известно, что абсолютная синонимия — явление крайне редкое. Во фразеологии абсолютная синонимия еще более редкий феномен. Дело в том, что в плане содержания фразеологизмов, кроме актуального значения, выделяется еще и внутренняя форма, которая существенным образом влияет на семантику и употребление фразеологизма [Баранов, Добровольский 2008]. Внутренняя форма присутствует и в обычных словах, однако по большей части она стерта — конвенционализирована. Во фразеологизмах внутренняя форма, как правило, жива и реально ощущается носителями языка. Идиомы с семантикой смерти, например, *испустить дух* и *отправиться к праотцам*, с одной стороны, и *выставить кеды* и *склеить ласты* — с другой, традиционно должны описываться как синонимы, поскольку они описывают одно и то же событие. Однако очевидно, что в конкретной ситуации общения идиому *испустить дух* нельзя заменить на *выставить кеды* и *склеить ласты*. Аналогично *выставить кеды* нельзя заменить на *испустить дух* и *отправиться к праотцам*. Актуальные значения у них близки — 'умереть', однако внутренние формы радикально различаются.

Различия между квазисинонимами во фразеологии не всегда сводятся к внутренней форме. Так, идиома *протянуть ноги* описывает ситуацию, когда человек умирает из-за отсутствия жизненно необходимых ресурсов (как правило, из-за недоедания), а идиома *хватила кондрашка* уместна только в ситуации, когда человек неожиданно умирает естественной смертью, связанной с сильным душевным потрясением или приступом болезни [ФОС]. Во внутренней форме этих идиом нет почти ничего, что указывало бы на эти различия в значениях.

Отдельная — и весьма сложная — проблема состоит в разграничении между синонимами и вариантами. Высокая степень варьирования компонентов является типичным свойством фразеологической системы. В лексикографии решение о синонимии или вариантности фразеологизмов часто определяется прагматическими факторами — например, удобством алфавитизации. Если же говорить о содержательных критериях, то важнейший из них — идентичность внутренней формы. Если при модификации компонентов идиомы внутренняя форма не меняется, то решение принимается в пользу вариантов (ср. *замолвить словечко/словцо, положить/уложить на лопатки*), если же образы, лежащие в основе идиомы, различны, то естественно говорить о синонимии (ср. *дать по башке* и *дать по мозгам*, *нужен как рыбе зонтик* и *нужен как собаке пятая нога*).

Возвращаясь к синонимии можно констатировать, что это отношение в области фразеологии почти никогда не бывает полным. Однако в редких случаях абсолютная синонимия встречается и среди идиом. Так, русские идиомы *с головы до ног* (с вариантом *с ног до головы*), *с головы до пят* и *от ушей до пяток* обнаруживают одинаковые значения. Дело в том, что различия во внутренней

форме в данном случае оказываются несущественными: и ноги, и пятки выступают как символы низа, а голова и уши — как символы верха. Иными словами, образная составляющая при варьировании формы существенно не меняется. Следовательно, такие случаи можно рассматривать как пограничные между вариантностью и собственно синонимией.

3. Отношение антонимии

Как отношение синонимии, антонимия также различается в системе языка и в речи. Антонимы в системе языка — это лексические единицы, значения которых противопоставлены друг другу по какому-то смысловому признаку, который составляет ядро их значения, при этом формально определить степень «ядерности» признака трудно. Обычно это определяется интуитивно. Типичные антонимы — прилагательные, обозначающие концы некоторой шкалы признака: *холодный — горячий; хороший — плохой; ранний — поздний*.[1] В речи антонимия задается эксплицитным противопоставлением в контексте. Ср. известные пушкинские строки об Онегине и Ленском: *Волна и камень, / Стихи и проза, лед и пламень / Не столь различны меж собой*. Слова *волна* и *камень*, *стихи* и *проза*, *лед* и *пламень* не являются антонимами в системе языка. В контексте речевая антонимия может возникать из-за контраста, способного высвечивать и сопоставлять самые разные смыслы.

Антонимия представлена и во фразеологической системе. Как антонимы могут рассматриваться, например, идиомы *попасть в яблочко* и *попасть в молоко*. Поскольку антонимия выражается в противопоставлении по какому-то одному семантическому признаку, например, ‘низкий’ — ‘высокий’ как значения признака высота, то большая часть антонимов сосредоточена внутри одного семантического поля. Так, в поле бедность-богатство обнаруживаются следующие пары антонимичных идиом: *денег куры не клюют* и *ни гроша*; *сыт, пьян и нос в табаке* и *положить зубы на полку, дом полная чаша* и *ни кола ни двора*. Заметим, что идиомы *сыт, пьян и нос в табаке* и *положить зубы на полку* различны по категориальной принадлежности, что не позволяет рассматривать их как точные антонимы.

Точная антонимия в лексике явление нечастое. Для фразеологии это тем более верно, поскольку фразеологические единицы противопоставляются по множеству разнообразных параметров. Действительно, системное противопоставление фразеологизмов по смыслу сопровождается и другими различиями, не сводимыми к антонимии. Так, идиома *полная чаша* относится к достатку, который проявляется в комфортности быта, добротности одежды, показном обилии мебели, еды и т. д., что осмысляется как мещанский идеал существования: *Эка, сняв платье, примеряет крепдешиновый пеньюар мадам Зули: — Забавно: котлеты я чуть-чуть разогрела, они были еще теплые. — Дом — **полная чаша**, надо же, как живут еще некоторые… — удивляется свекровь*. [О. Иоселиани.

[1] О типах антонимии см., например, [Новиков 2001: 59–124; Кобозева 2000: 104–105].

Разбойники]. Фразеологизму *ни кола ни двора* (с вариантами *без кола без двора, без кола и двора*) присуща более общая семантика. Он используется для описания крайней бедности, проявляющейся в отсутствии своего жилья: *В одиночку Станислав Семенович жить не мог, а жена от него почему-то отказалась. Даже в дом его не пускала. Остался он* ***без кола без двора****, иди куда хочешь, хоть с Крымского моста в воду, и Левкина мать его пожалела.* [Ю. Трифонов. Время и место].

Важно иметь в виду, что идиомы, внешне противопоставленные друг другу по лексическому составу, совсем не обязательно оказываются антонимами. Так, фразеологизмы *поставить на ноги* и *сбить с ног*, вроде бы противоположные по значению из-за семантики глаголов *поставить* и *сбить*, на самом деле не являются антонимичными ни по внутренней форме (по образам), ни по актуальному значению. Действительно, они различны — но не противоположны — по актуальным значениям (*поставить на ноги* — 'помочь достигнуть самостоятельности' или 'вылечить', а *сбить с ног* — 'оказать сильное психологическое воздействие'), а также не противоположны и по образам: в образе идиомы *поставить на ноги* фиксирована идея 'придания человеку устойчивого вертикального положения, рассматриваемого как норма', что переосмысляется как приобретение самостоятельности (<…> *это еще беспомощное существо, хорошо, если дед успеет* ***поставить*** *его* ***на ноги*** [Ч. Айтматов. Белый пароход]) или выздоровление (*Думаю, недельки через три мы сумеем* ***поставить*** *больного* ***на ноги*** [Н. Носов. Незнайка на Луне]). Во внутренней форме идиомы *сбить с ног* реализуется идея резкого физического воздействия — удара, который переосмысляется как воздействие психологическое, ср. <…> *жалко девушке его,* <…> *сильного мужика,* ***сбитого с ног*** *неожиданным подлым горем* [Л. Корнешов. Газета].

Наряду с антонимией в точном смысле, во фразеологии выделяется также явление энантиосемии, то есть антонимии между значениями одной фразеологической единицы. Примером энантиосемии может служить идиома *вертится на языке (что-л. у кого-л.)*. В одном из значений это выражение употребляется в ситуации, когда кто-то хочет что-то сказать, но не может это сделать по тем или иным причинам, а также когда что-то — часто против воли человека — привлекло его внимание и он постоянно повторяет это «про себя»: *Ах, как я его уважаю… сказала бы… слово* ***вертится на языке****, — но не смею… Почему не сметь? Да, я его люблю, нет, боготворю!* [И. А. Гончаров. Обрыв]; *Кстати, а что такое «паскуда»?* ***Вертится на языке*** *с самого ранья…* [О. Дивов. Молодые и сильные выживут].

Другое значение описывает почти противоположную ситуацию — когда человек не может правильно сформулировать свою мысль, он не в состоянии подобрать правильное слово или вспомнить точную формулировку какой-то идеи: — *Нет, я боюсь, что придется выдумать за неимением слова; вы знаете:* ***вертится на языке****; и выходит не то* <…>. [А. Белый. Москва]. Отметим, что и в приведенных антонимичных значениях идиомы *вертится на языке (что-л. у кого-л.)* есть что-то общее, причем этот общий смысл нельзя отнести к числу «тривиальных». И в первом, и во втором значении эта идиома описывает ситуацию, в которой соответствующее слово (выражение) еще не произнесено. Эта семантическая общность семантики антонимичных

значений обеспечивается внутренней формой идиомы, указывающей на то, что «слова еще не сошли с языка».

Нечто вроде неполной энантиосемии обнаруживается в значениях идиомы *на днях*. В одних употреблениях она относится к будущему, в других — к прошлому. Ср. <…> ***на днях*** *он перетащил в холле телевизор из одного угла в другой* [Л. Улицкая. Пиковая дама] и <…> *старуха* ***на днях*** *непременно умрёт* [И. Грекова. Скрипка Ротшильда].

В целом, однако, отношение антонимии слабо выражено во фразеологической системе. Это, объясняет отсутствие значительных по объему фразеологических словарей антонимов.

4. Полисемия

Фразеологизмы, как и слова, обнаруживают те же типы многозначности, что и обычная лексика. Следует отметить, что хотя традиционно полисемия считалась нетипичной для фразеологии, в целом идиомы и коллокации обнаруживают весьма развитую многозначность. Принято различать цепочечную, радиальную и смешанную (радиально-цепочечную) полисемию. При **радиальной полисемии** «все значения слова мотивированы одним и тем же — центральным — значением», в случае же **цепочечной полисемии** «каждое новое значение слова мотивировано другим — ближайшим к нему — значением, но крайние значения могут и не иметь общих семантических компонентов» [Апресян 1974: 182].

Типичным примером радиальной полисемии могут служить три значения идиомы *положить на [обе] лопатки*:

1. Победить кого-л. в спортивном состязании, *что осмысляется как соответствие критерию победы в классической или вольной борьбе.*[2] (*Футболисты Челси* ***положили на обе лопатки*** *команду Португалии. Первый гол забит уже на десятой минуте*.)
2. Победить кого-л. в конфликте, предусматривающем применение силы, *что осмысляется как соответствие критерию победы в классической или вольной борьбе*. (*Германия в 1939 г.* ***положила*** *Польшу* ***на лопатки*** *в три недели*.)
3. Победить кого-л. в споре, дискуссии и т. п., *что осмысляется как соответствие критерию победы в классической или вольной борьбе.* (*Он высмеял Топоркова за доклад, а затем не спеша, с профессиональной сноровкой* ***уложил*** *отца психоанализа* ***на обе лопатки***. [Викт. Ерофеев. Трехглавое детище]).

В толковании компонент *что осмысляется как соответствие критерию победы в классической или вольной борьбе* не только эксплицирует внутреннюю

[2] Курсивом дана та часть толкования, которая соответствует внутренней форме.

форму идиомы, но и указывает направление семантической деривации от первого значения ко второму и третьему.

Цепочечная полисемия для идиом нехарактерна, что сближает идиоматику с обычной лексикой, в которой данный тип многозначности также встречается довольно редко. Фактически о цепочечной полисемии в идиоматике можно говорить только в том случае, если считать, что первым значением является буквальное значение выражения, образующего идиому. Поясним, что имеется в виду.

Идиома *из-под полы* обнаруживает в современном языке два значения:

1. В нарушение законов продавать в не предназначенных для этого местах какие-л. товары или предлагать какие-л. услуги так, чтобы проверяющие инстанции этого не обнаружили, *что описывается как сокрытие предлагаемого товара под одеждой*. (*Продавать контрабандные сигареты* ***из-под полы***)

2. Делать что-л. тайно, не имея официального разрешения, *что описывается по аналогии с ситуацией незаконной торговли*. (*Работа была, может быть, не строго научна, но, пожалуй, талантлива и написана хорошо по-русски, <…> но главное, что и поразило <…> была внутренне свободна. Мы видели ее однажды на кафедре, уже желтую, с потрепанными ушами… <…> Ею гордились не перечитывая, и кое-кому,* ***из-под полы****, показывали*. [А. Битов. Пушкинский дом]).

В данной идиоме первое значение 'незаконная торговля' мотивировано внутренней формой — '*что описывается как сокрытие предлагаемого товара под одеждой*'. Второе значение производно от первого на основе переосмысления 'незаконной торговли' как 'совершения чего-л. тайного, не имеющего официального разрешения'. Таким образом, формируется следующая цепочка семантической деривации: сокрытие предлагаемого товара под одеждой → тайная незаконная торговля → тайная деятельность.

Во фразеологии представлена также и смешанная полисемия.

Еще один аспект многозначности в сфере фразеологии — это регулярная полисемия. Регулярная полисемия — это «такая комбинация значений многозначного слова, которая повторяется во многих или во всех словах определенного семантического класса» [Апресян 1993: 10]. На материале обычной лексики это явление изучено довольно хорошо (ср. для русского языка работы Ю. Д. Апресяна, Е. В. Падучевой, Р. И. Розиной, Г. И. Кустовой и др.).

В идиоматике регулярная полисемия характерна прежде всего для семантических полей, связанных между собой переходом «физического» в «нефизическое». Так, многие идиомы со значением 'физического воздействия' употребляются и в значении 'порицания/наказания', 'превосходства/победы' и т. п. По этому принципу устроена многозначность таких идиом, как *дать/надавать по ушам, дать/надавать по соплям, всыпать по первое число, дать сдачи, сделать отбивную.* Ср. характерные пары контекстов:

*Генка **дал сдачи**, зеваки оживились… — Невозможно **дать сдачи** литературному злодею, не вывалявшись вместе с ним в болоте сквернословия.*

*— Вася, **навешай** ему **по первое число**, но без увечий, а потом отпусти. — Начальство вызывает только в трех случаях: чтобы дать задание, похвалить или **навешать по первое число**.*

По принципу порождения «нефизических» значений на основе «физических» устроены также и идиомы из поля смерть, переосмысляемые в значении ‘прекратить функционировать и/или перестать существовать’. Ср. фразы *компьютер **дышит на ладан**; если закроют завод, город **испустит дух**; холодная война **приказала долго жить**; идея коммунистического субботника **почила в бозе***.

Несмотря на относительную регулярность порождения вторичных значений у идиом таких типов, в этой сфере обнаруживается довольно много исключений. Так, те идиомы семантического поля физическое воздействие, физическое насилие, которые появились недавно, (еще) не развили «нефизического» значения. Ср. жаргонные идиомы *дать в табло, дать…*[3] *в бубен, дать… по щам*, которые имеют только одно — «физическое» — значение, в отличие, например, от выражения *дать сдачи*, развившего значение ‘нефизического сопротивления’. Встречаются и другие типы асимметрии. Например, внешне очень похожие идиомы *дать… по шапке* и *дать.. по башке* обнаруживают неидентичный набор значений. Выражение *дать по шапке*, не обладающее значением ‘физического воздействия’, входит в поля наказание (*За любую инициативу у нас **дают по шапке***) и увольнение (*А если тебе **по шапке дадут** с завода?*)[4]. А идиома *дать по башке* представлена только в поле наказание (ср. **А если тебе **по башке дадут** с завода?*).

5. Заключение

Часто высказывалась мысль о том, что с теоретической точки зрения фразеологизмы — это такие же единицы лексикона как обычные слова. Исследование стандартных типов семантических отношений в сфере фразеологии показывает, что их действие распространяется и на фразеологизмы. Однако особенности семантики и структуры фразеологизмов влияют на разнообразные факторы реализации этих отношений: на частоту (распространенность) того или иного отношения, а также на его регулярность и продуктивность. Основной причиной специфичности фразеологии в этой области следует считать наличие живой внутренней формы, расширяющей или, наоборот, ограничивающей круг возможных вариантов реализации того или иного семантического отношения.

---

[3] Многоточие указывает на возможность замены глагольного компонента на близкие по значению глагольные формы.

[4] Это значение данной идиомы воспринимается сегодня как устаревающее.

References

1. *Apresian Iu. D.* 1974. Lexical Semantics [Leksicheskaia semantika].
2. *Apresian Iu. D.* 1993. Lexicographic Concept of the New Large English-Russian Dictionary [Leksicographicheskaia Kontseptsiia Novogo Bol'shogo Anlgo-Russkogo Slovaria]. Novyi Bol'shoi Anglo-Russkii Slovar' :6–17.
3. *Apresian Iu. D.* 1995. New Explicative Synonimes Dictionary: Concept and Types of Information [Novyi Ob"iasnitel'nyi Slovar' Sinonimov: Kontseptsiia I Tipy Informatsii]. Novyi Ob"iasnitel'nyi Slovar' Sinonimov Russkogo Iazyka
4. *Baranov A. N., Dobrovol'skii D.O.* 2008. Aspects of the Theory of Phraseology [Aspekty Teorii Frazeologii].
5. *Kobozeva I. M.* 2000. Linguistic Semantic [Lingvisticheskaia Semantika].
6. *Novikov L. A.* 2001. Proceedings I: Problems of Linguistic Meanings [Izbrannye trudy I: Problemy Iazykovogo Znacheniia].
7. *Russian* Explicative Dictionary of Phraseology [Frazeologicheskii Ob"iasnitel'nyi Slovar' Russkogo Iazyka].2009.
8. *Shmelev D. N.* 2006. Problems of the Semantic Lexical Analysis [Problemy Semanticheskogo Analiza Leksiki] .

ЧЕГО НЕ ХВАТАЕТ В «ОЦИФРОВАННОМ МИРЕ» ЛЕКСИКОГРАФУ И СОЦИОЛИНГВИСТУ?

В. И. Беликов (vibelikov@gmail.com)

Институт русского языка им. В. В. Виноградова,
Москва, Россия

Ключевые слова: социолингвистика, лексикография, обработка информации, сегментная статистика.

WHAT ARE SOCIOLINGUISTS AND LEXICOGRAPHERS LACKING IN A DIGITIZED WORLD?

V. I. Belikov (vibelikov@gmail.com)

Vinogradov Russian Language Institute, Moscow,
Russian Federation

It is a common belief that text corpora provide the best testing ground for solving any kind of linguistic problems. As far as grammar is concerned, this may be true, but if we focus on investigating the lexicon the results often appear to be rather superfi cial. WWW contains some relatively homogeneous arrays of texts formed independently of linguists, in some cases emerging quite spontaneously. Text arrays with the most prominent social characteristics of their authors are regarded as independent Internet segments (digitized classical literature and 2010 teenager blogs are the most contrasting examples). Frequencies of the same lexical items differ greatly from one segment to another, and this statistics is very signifi cant for sociolinguistics. The main problem in applying the method of segmental statistics is the lack of a suitable instrument for automatic data processing Several case studies are presented, and the results of segmental statistics seem to be more indicative than those obtained from the Russian National Corpus

Key words: sociolinguistics, lexicography, segment statistics, data processing.

Ещё у меня есть где-то гора ссылок по корпусной лингвистике ‹…›
dimkaguarani, 29 декабря 2002 // Блогосфера

Примечание к эпиграфу: В НКРЯ словосочетание *корпусная лингвистика* не фиксируется, хотя четверть текстов с упоминанием *лингвистики* датировано там 2003 г. и позднее.

Начну с оптимистичной цитаты. «В русском языке есть глагол несовершенного вида реагировать. Его коррелятами совершенного вида могут быть несколько разных приставочных глаголов: прореагировать, отреагировать, среагировать (явление нередкое, особенно среди заимствований). Какой из этих приставочных коррелятов употребляется чаще? К каким контекстам тяготеет каждый из этих приставочных коррелятов (например, какой из них охотнее сочетается с наречием быстро)? Наконец, в какой последовательности они появляются в современном языке — одновременно или по очереди? Различается ли частота их употребления в разные периоды?»[*][1] [Плунгян 2005: 302]. На эти вопросы «лингвист может ответить с помощью Корпуса буквально за считанные минуты» [Там же].

Картина, выявляемая НКРЯ в 2011 г., представлена в Табл. 1 (число документов[2] с соответствующими глаголами в корпусе в целом, а для 1990-х и 2000-х гг. — отдельно в художественных текстах и публицистике) и в Табл. 2 (число документов, где глаголы сочетаются с наречием *быстро*, вкл. форму *быстрее*).

Комментарии начну с употребления производящего глагола. Из 22 «документов» XIX в. 14 относятся к естественным наукам. *Документ* здесь — понятие довольно условное: все тексты, взятые из трехтомника А. М. Бутлерова (1953–1958), сведены в один документ «Теоретические и экспериментальные работы по химии», датируемый «1851–1886»[3].

Таблица 1

	реагировать	**прореагировать**	**отреагировать**	**среагировать**
всего	2003	122	744	184
1800–1899	22	1	0	0
1900–1924	68	1	0	0

[1] Ограничения на объем печатаемого текста вынуждают снять некоторые частные комментарии, а сокращение аргументация делает ее декларативной. Вчетверо больший развернутый текст доклада имеется в электронной форме; в бумажном варианте наиболее значимые сокращения обозначены астериском в квадратных скобках.

[2] Произведения, представленные в Корпусе частями, считаются за один документ; ср. Прим. 3[*]. Датировки типа «1943–1999» обычно усредняются.

[3] В других случаях единый текст может достаточно искусственно делиться, так, самостоятельными документами считаются главы «Книги воспоминаний» И. М. Дьяконова; в результате документ «И. М. Дьяконов. Книга воспоминаний. Часть вторая. Последняя глава (После войны)» по объему в 25 с лишним раз уступает «единому документу» Бутлерова.

	реагировать	**прореагировать**	**отреагировать**	**среагировать**
1925–1949	120	1	2	0
1950–1989	412	31	77	20
1990–1999	249	22	152	46
2000–	1132	66	513	118
худ., 1990–99	97	4	65	26
худ., 2000–	115	12	96	28
публ., 1990–99	137	14	84	22
публ., 2000–	746	41	366	72

Вне естественнонаучной литературы примеры на *реагировать* появляются лишь с 1890 г., при этом примеры из беллетристики достаточно физиологичны: *зрачок не реагировал на свет* ‹…› (Мамин-Сибиряк), *органическая ткань* ‹…› *должна реагировать на всякое раздражение* (Чехов), *способность кожных нервов реагировать на температурные колебания* (Вересаев). В более обобщенном значении глагол фигурирует лишь у М. Горького («Мужик», 1899: *Сурков делает свои дерзости* ‹…› *на них никто не реагирует*) и текстах, отнесенных к публицистике.

Что касается глаголов совершенного вида, то *прореагировать* появляется первым (в «документе» с датой 1851–1886), и в течение длительного времени его можно считать узкопрофессиональным. В художественной прозе он фиксируется довольно поздно (1968), а до этого однажды фигурирует в дневниковой записи 1944 г. и 11 раз в химических текстах. Глагол *отреагировать* появляется в НКРЯ также как узкоспециальный — в первой половине XX в. он встретился лишь в двух текстах по психиатрии, а с 1950 г. проникает в художественную литературу. Несколько позднее (с 1960/64) в корпусе фиксируется глагол *среагировать*.

Отвлекаясь от профессиональных текстов, делаем вывод, что в повседневный узус все три глагола вошли **одновременно**, а даты появления в беллетристике — 1950–1960/64–1968 — аккуратно объясняются различиями в частотности: разрыву в 12 лет соответствует соотношение 77/31, а разрыву в 6 лет — 31/20. Новизну глаголов подтверждают лексикографы: ни одного из трех нет ни в словаре под ред. Д. Н. Ушакова, ни 17-томном БАСе, ни в МАСе.

О частоте употребления приставочных глаголов можно говорить лишь за последние 60 лет; если принять за единицу число текстов с наиболее редким глаголом *прореагировать*, то окажется, что *отреагировать* в «доперестроечном языке» появлялся в два с половиной раза чаще, а *среагировать* — на треть реже. В последующие годы употребимость двух более новых глаголов резко возросла: *отреагировать* стал абсолютным лидером, а тексты со *среагировать* стали в два раза более частотными, чем с *прореагировать*. Но сводная картина несколько маскирует реальное положение вещей. По числу текстов с глаголом несовершенного вида довольно ясно видно резкое увеличение доли публицистики: за 1950–1989 она составляла 36%, в 2000-х годах — 66%, а доля художественных текстов сократилась с 29% до 10%[*].

Попробуем выйти за пределы НКРЯ. Насколько верно, что глагол *реагировать* начал употребляться учеными-естественниками, а в общее употребление

попал не без воздействия докторов и писателей-медиков, причем лишь в самом конце XIX в.? Обращение к сегменту «Классика» Библиотеки Мошкова заставляет в этом усомниться.

Она не знала, что в этом кротком ребенке ее женственная ласка возбудит подземный огонь эротической страсти; а раз эта страсть возбудилась во мне — она реагировала на нее, и чем я больше рос, тем больше ее материнская любовь переходила в иную любовь, а моя детская привязанность в козлоногую похоть сатира, откровенничает Н. П. Огарев о своей гувернантке (1862). До первых внеестественнонаучных фиксаций в НКРЯ глагол не так уж редко встречался и в отечественной беллетристике, и в художественных переводах, и в судебных речах, и в литературной критике[*]. Видовая параллель, которую и в первой половине XX в. мы пока знаем лишь по психиатрическим текстам, столь же «древняя», ср.: *Прежде и мои «обидчики» отреагировали бы на укол в печати, а теперь все прошло без отклика* (Лесков, 1884); *Ха, ха! — отреагировал Риенцо своим странным смехом* (Э. Бульвер-Литтон, пер. 1875 г.). Об укорененности этих глаголов в общелитературной норме говорит их использование в книгах для девочек: *Эти слова встревожили Эстер. Она прижала к себе сестру и почти не реагировала на ее лепетание ‹…› Никто из подруг не отреагировал на ее слова, и после минутного возбуждения от полученных новостей девочки замолчали* (Э. Мид-Смит, пер. 1900 г.).

Глагол *прореагировать* попадает в детскую литературу также уже в начале XX в.: *Шакал не прореагировал; ему уже минуло три года, но нельзя же сердиться на оскорбление, нанесенное особой с клювом в ярд длины и сильным, как дротик* (Киплинг, пер. 1916 г.). А вот глагола *среагировать* в сегменте «Классика» нет вообще; это уже дает основание считать, что он утвердился в языке позже начала XX в.; когда именно — сказать будет можно лишь после появления достаточно представительного массива разнотипных оцифрованных русских текстов «средней половины» XX в.; пока его нет.

Рассмотрим результаты поисков на совместимость глаголов совершенного вида с наречием *быстро* (см. верхнюю часть Табл. 2). Согласно Табл. 1 глагол *отреагировать* за 1990-е — 2000-е гг. встретился в 665 документах, а *среагировать* — в 164, то есть в 4,1 раза реже.

За тот же период наречие *быстро* в непосредственном соположении с этими глаголами встретилось в 18 и 6 текстах соответственно (в три раза реже), однако при учете всех примеров, где наречие связано с глаголом (21/6) — лишь в 3,5 раза реже; десятые доли «разов» при таких цифрах вряд ли подлежат учету, так что намек на бóльшую «привязанность» наречия к глаголу *среагировать* выглядит сугубо формальным. Уровень риска каких бы то ни было рассуждений на эту тему для предшествующего периода очевиден, поэтому о динамике использования наречия с этими глаголами судить нельзя.

Таблица 2

быстро +	**прореагировать**	**отреагировать**	**среагировать**
НКРЯ, релевантные при поиске «/1» (релевантные при «/10»)			
всего	1 (1)	20 (23)	6 (6)

быстро +	**прореагировать**	**отреагировать**	**среагировать**
1950–1989	0 (0)	2 (2)	0 (0)
1990–1999	0 (0)	8 (10)	2 (2)
с 2000	1 (1)	10 (11)	4 (4)
Военная литература, тексты, опубликованные в 1950–1989 гг., релевантные при поиске «/1»			
	0	20	17
ЖЗ, релевантные при поиске «/1»			
1995–1999	0	3	2
с 2000	0	22	12

Но есть другие массивы оцифрованных текстов, где сходные задачи решаются успешнее, хотя работа с ними заметно более трудоемка, в первую очередь из-за отсутствия необходимого инструментария.

Сегмент «Военная литература» Библиотеки Мошкова (militera.lib.ru) содержит большой массив мемуарной, художественной, исторической и иной литературы (возможен поиск по соответствующим подмассивам), в основном второй половины XX — начала XXI в., но выявление датировки требует обращения к каждому тексту.

В сегменте в целом число документов с глаголом *отреагировать* троекратно превышает число документов со *среагировать*, однако имеется явное жанровое различие: если в мемуарах разница чуть больше, чем в два раза, то в текстах по военной истории она составляет 7,5; при наличии обстоятельства *быстро* различие во всех случаях уменьшается[*].

Хороший срез современного литературного языка дает сегмент «Журнальный зал» (ЖЗ); это в основном тексты XXI в., частично также 1990-х гг., но в толстых журналах изредка публикуются и заметно более ранние произведения. Здесь основная проблема в том, что подавляющее большинство текстов продублировано дважды. Несколько лет назад поисковый алгоритм Яндекса это учитывал и первоначально предлагал каждый текст по одному разу, а полную выдачу давал лишь при реализации опции «еще с сайта» (при этом на первом этапе были некоторые потери полноты, которые ликвидировались лишь при полной выдаче). Сейчас выдается сразу всё, так что близкое приближение к точному числу текстов можно получить делением брутто-результата[4] на два. При небольшой выдаче несложно выяснить подлинный результат вручную (см. нижние строки Табл. 2), но пока нас интересует соотношение, достаточно сопоставления брутто-результатов. В ЖЗ *среагировать* встречается в 5–6 раз реже, чем *отреагировать*, а в сочетании с наречием *быстро* — всего лишь в два раза. То есть сам глагол *среагировать* используется существенно реже, но его относительная сочетаемость с наречием в два, а то и в три раза выше, чем у *отреагировать*[*].

[4] Так я называю собственно выдачу по запросу, с дублетами и цитатами; переход к нетто-результату требует ручной обработки.

Общий недостаток поисковых машин — недостоверность объявленных цифр найденного, если они превышают 1000 — может быть компенсирован в этом сегменте поиском по отдельным журналам[*].

Можно ли что-то сказать о динамике процесса? Лингвистически полезным является раздел «Самиздат» Библиотеки Мошкова (zhurnal.lib.ru), где произведения размещает каждый, обнаруживший у себя творческие наклонности. Любые наметившиеся в языке изменения выражены в этом сегменте Интернета сильнее, чем в профессиональной литературе. Относительная частота *среагировать* здесь явно выше, чем у тех, кто пишет профессионально, а в сочетании с наречием *быстро* глагол *среагировать* используется даже чаще, чем *отреагировать*.

Наиболее отчетливо языковые инновации проявляются в блогосфере. Не трудно показать, что в этом сегменте популярность глагола *среагировать* несколько выше, чем в ЖЗ: число блогов, в которых он был использован, в разных регионах за разные временные отрезки составляет четверть или несколько более от числа тех блогов, где фигурировал глагол *отреагировать*[*].

Результаты поиска *быстро* /1 [глагол] по московским блогам представлены в Табл. 3, они довольно причудливы: изначально *быстро* /1 *среагировать* преобладало незначительно, в 2006–2007 гг. его доминирование существенно возросло, а затем процесс пошел вспять, да так, что к концу 2010 г. статистика уже в пользу глагола *отреагировать*.

Таблица 3

быстро /1	**2001–2005**	**2006–2007**	**2008–2009**	**2010**
реагировать	193	692	981	711
прореагировать	2	7	6	4
отреагировать	62	189	298	259
среагировать	80	334	365	250
среагировать к отреагировать, %	129	177	122	97

Разгадка проста: возрастная структура блогосферы подвержена существенным изменениям, в 2006–2007 гг. доля подростков была максимальной, позже их число снижается в связи с массовым уходом в социальные сети. А последние пару лет наметился приток новых блоггеров старших возрастов.

В отдельных регионах блогосфера оказывается в основном подростковой, и там такого рода статистика выглядит очень рельефно. В Астрахани за все годы по 2010 включительно *быстро /1 отреагировать* встретилось в 7 блогах, а *быстро /1 среагировать* — в 12, в Оренбурге, соответственно, 2 и 8; цифры мизерные, но показательные.

Не подлежит сомнению, что любые языковые инновации сильнее выражены в младших возрастах. Имея стратифицированную по возрасту синхронную статистику словоупотреблений, мы получим сведения о динамике языковых процессов. К сожалению, распределение блоггеров по возрасту,

декларированное в расширенном поиске Яндекса, с лета 2008 г. практически не работает.

В ЖЗ, отражающем более «старый» язык, *отреагировать* встречается в 5–6 раз чаще, чем *среагировать*, у блоггеров только в три-четыре; популярность второго глагола явно растет. А резкий рост сочетания *быстро среагировать* по-настоящему впечатляет: в младших возрастах оно явно обогнало синонимичное *быстро отреагировать* не только относительно, но и в абсолютных цифрах. Разница в употреблении двух глаголов может быть связана с полом автора (в блогосфере доминируют женщины, а среди авторов ЖЗ их меньше), или же с различиями в тематике. Данные за разные периоды показывают, что блоггеры, не зависимо от пола, применяют глагол *среагировать* **к себе** в 2–3 раза реже, чем *отреагировать*, а **к власти** — в 4–6 раз реже. Соответствующая статистика по всей блогосфере за последний квартал 2010 г. есть в Табл. 4, там же для сопоставления приведены данные, которые можно получить из НКРЯ и ЖЗ. Довольно очевидно, что 30 текстов ЖЗ позволяют лишь выдвигать какие-то гипотезы, а из втрое меньшего стилистически разнородного материала НКРЯ вряд ли что можно извлечь.

Разница в частотности рассматриваемых глаголов в разных контекстах особенно заметна на фоне того, что контекстно не привязанное соотношение употребимости этих глаголов находится как раз посредине[*]. Поведение наречия остается не вполне ясным; не исключено, что молодежь считает, что на любые перемены она способна *среагировать* **на порядок** *быстрее*, чем *отреагирует* власть.

Таблица 4

	Блоги, 10–12.2010	**НКРЯ, 2001–2004**	**ЖЗ, 2001–2010, нетто**
«я отреагировал»	218	2	11
«я среагировал»	100	1	6
«я отреагировала»	260	2	2
«я среагировала»	97	0	0
власть /З отреагир., муж.	253	4	9
власть /З отреагир., жен.	43	1	2
власть /З среагир., муж.	34	0	0
власть /З среагир., жен.	8	0	0

В цитате, с которой я начал, относительно трех глаголов были сформулированы четыре вопроса, ответы на которые в применении к текстам НКРЯ, действительно, можно получить «буквально за считанные минуты». Легко узнать, что чаще всего в корпусе встречается глагол *отреагировать*, он же охотнее двух других сочетается с наречием *быстро*, он же первым — в 1950 г. — появляется в беллетристике (в текстах по химии его опережает известное с XIX в. *прореагировать*). На любом относительно большом отрезке времени примеров на этот глагол больше, чем на конкурирующие. Исключение — дореволюционный период, когда глагола *отреагировать* еще не было, а *прореагировать* дважды

было употреблено химиками. Но эти ответы — про НКРЯ. Как выясняется, ответы на те же вопросы про русский язык могут заметно отличаться, если воспользоваться статистикой, полученной по разным относительно однородным массивам текстовых документов, представленных в разных сегментах Интернета. Эту методику я называю сегментно-статистической.

НКРЯ в совокупности является хорошим «макетом» русского языка: разметка и стоящая за нею идеология — это модель устройства языка, а комплект текстов корпуса — экспериментальная база для проверки модели. НКРЯ создан лингвистами и для лингвистов, а лингвиста интересует грамматика. Часть ее «зарыта» в лексиконе, это продуктивные и уникальные модели грамматического поведения лексических единиц. Слова, ведущие себя стандартно, лингвиста интересуют очень мало: важно знать объем некоторого класса однотипно ведущих себя слов (степень продуктивности модели) и иметь под рукой несколько представителей каждого класса, чтобы строить примеры типа *Seymour sliced the salami with a knife*. Прагматическая ценность такого рода примеров лингвиста не интересует[*]. Собственно лексикографу (составителю неспециализированного словаря) интересно **каждое** слово. Еще один тип специалиста по языку — социолингвист, на него НКРЯ также не рассчитан, поскольку ни в коей мере не может считаться сбалансированным с точки зрения решаемых им задач.

НКРЯ хорош для решения любых собственно лингвистических задач, часто полезен лексикографу и социолингвисту для предварительной оценки положения вещей, хотя в отдельных случаях может служить и основанием вполне серьезных выводов. Но исследователи с лексикографическими и социолингвистическими интересами иногда склонны принимать его за полигон принятия решений в последней инстанции, что досадно. Приведу два показательных примера.

Недавно была высказана претензия к «Русскому орфографическому словарю», куда включены единицы, «обнаружить наличие которых не удается даже при обращении к Национальному корпусу русского языка» [Николенкова 2011: 181]. Примеров необнаруженного нет, приведены лишь три слова, якобы встречающиеся там в единственных контекстах: *остервенить*, *окровенить* и *форсун*; эти слова традиционно включаются в толковые и орфографические словари — но зачем засорять словари тем, что даже в НКРЯ (почти) не встречается?[*].

Для исследования НКРЯ «с точки зрения отражения социальной дифференциации языка» из слов, «отражающих молодежный сленг школьников, студентов, были отобраны следующие: *училка*, *химица*, *студак*, *туса*» [Киеня 2010: 263–264]. Для слов *студак* и *химица* во всех подкорпусах нашлось лишь по одному примеру, тем не менее С. Н. Киеня делает вывод, «что национальные корпусы отражают все многообразие существования языка, генеральную совокупность языкового материала, что свидетельствует об их высокой социокультурной значимости» [Там же: 265].

Я же из этой работы делаю прямо противоположный вывод: НКРЯ с социолингвистическими задачами не справляется. Между тем,

сегментно-статистический метод выявляет, что в современном русскоязычном узусе Белоруссии *химица* используется в три раза чаще *химички*, что белорусскому (и московскому) *студаку* в Петербурге соответствует *студень*, а в Екатеринбурге *студик*, этот метод позволяет проследить диахронию замен *тусовка — туса* и *тусоваться — тусить* в молодежном жаргоне[*].

Метод этот пока довольно убог. «Оцифрованный мир» дал социолингвисту необъятный языковой материал, который «самоорганизовался» в динамично растущие относительно однородные в стилистическом отношении сегменты. Но для работы с ними специального инструмента нет, приходится пользоваться тем, что имеется: поисковым алгоритмом Яндекса, рассчитанным на принципиально иного пользователя.

Будучи оптимистом, надеюсь, что рано или поздно полнота описания языка, объявленного в России государственным, заинтересует кого-то из тех, кто в силах принять ответственное решение, ведущее к созданию специализированного инструмента для исследования больших массивов оцифрованных текстов.

А пока буду вручную доказывать необходимость такого инструмента. Gutta cavat lapidem.

References

1. *Kienia S. N.* 2010. Corpus in a Sociolinguistic Aspect [Korpusy v sotsiololingvisticheskom Aspekte]. Nauka-2010: Sbornik Nauchnykh Statei.
2. *Nikolenkova N. V.* 2011. Spelling Dictionary and Codification of the Modern Standart: the Problem of Non-corrdination [Orfograficheskii Slovar I Kodifikatsiia Sovremennoi normy: Porblemy Nesoglasovannosti]. Voprosy Kul'tury Rechi.
3. *Plungian V. A.* 2005. Why are We Making the National Corpus of Russian? [Zachem My Delaem Natsional'nyi Korpus Russkogo Iazyka?]. Otechestvennye Zapiski, (2).
4. *Russian* Spelling Dictionary [Russkii Orfograficheskii Slovar]. 2007.

ИТАЛЬЯНСКИЕ КОНСТРУКЦИИ С ГЛАГОЛОМ ПОДДЕРЖКИ *FARE* В СОПОСТАВЛЕНИИ С РУССКИМ[1]

В. Бениньи (benigni@uniroma3.it),
Третий Римский университет, Рим, Италия

П. Котта Рамузино (Paola.cottaramusino@unimi.it),
Государственный университет Милана, Милан, Италия

В работе исследуется особый тип конструкции с глаголом поддержки, имеющей структуру [N_1[fare [(det) N_2]]]. Авторы предлагают лексико-семантическую классификацию в межъязыковой перспективе. Ставится вопрос о том, можно ли среди таких конструкций выделить регулярные и продуктивные модели для решения разных прикладных задач.

Ключевые слова: глагол поддержки, классификация, лексико-семантическая классификация, модели, решение задач.

ITALIAN CONSTRUCTIONS WITH SUPPORT VERB *FARE* IN COMPARISON WITH RUSSIAN

V. Benigni (benigni@uniroma3.it)
Universita Roma Tre, Roma, Italia

P. Cotta Ramusino (Paola.cottaramusino@unimi.it)
State University Milan, Milano, Italia

The paper deals with Support Verb Constructions (SVC) in Italian that are formed by the verb fare ‘to make’ and its nominal object (V+NOBJ) in an interlinguistic perspective with the Russian SVC with the verb delat’. The study has been carried out for Italian on ITWac (gathered by Baroni) and, for Russian, on the Russian Web Corpus (gathered by Serge Sharoff, University of Leeds), both are available as pre-loaded corpora within The

[1] Предлагаемая статья является результатом совместной работы двух авторов, имена которых указаны в алфавитном порядке. Валентина Бениньи подготовила текст для параграфов 1, 2, 3, 5, 5.1, 6, 6.1, 7, в то время как Паола Котта Рамузино несет ответственность за параграфы 4, 5.2–5.6, 6.2–6.4.

Sketch Engine corpus query system (http://the.sketchengine. co.uk). About 280 types of SVC with a token frequency ⩾200 resulted from the query in the Italian corpus. The Italian SVC have been classifi ed into lexical-semantic patterns, on the basis of Nsubj and Nobj semantic features and the Support verb lexical-semantic meaning. Subsequently, the patterns have been grouped into the well-known actional classes of accomplishments, achievements, semelfactives, activities and states (Vendler 1967, Comrie 1976). The overall classifi cation shows that most SVCs go hand in hand with the features of telicity (as regards verbs) and of concreteness and referentiality (as regards NOBJ), and in these classes (accomplishments, achievements) there is a partial parallelism with Russian, whereas fewer Russian SVCs can be found in the activity and states verb classes. Moreover, the presence of a high number of SVCs in the Russian corpus may be considered as a further evidence of the typological shift towards the analytic type that contemporary Russian is apparently undergoing (see e. g. the simplifi cation of noun declension, the expansion of invariable words and the increasing number of bi-aspectual verbs).

Key words: SVC, Support Verb Constructions, classification, lexical-semantic patterns, models.

1. Введение

В предлагаемой работе рассматривается вопрос, касающийся т. н. *Support Verb Construction* (конструкции с опорным глаголом). Выявление в межъязыковой перспективе (итальянский-русский языки) их семантической и синтаксической структур может оказаться полезным для решения таких прикладных задач, как автоматическое определение словосочетаний в тексте и создание лексикографических справочников.

2. Понятие Support Verb Construction в лингвистике: постановка вопроса

Прежде чем приступить к предмету исследования, кратко коснемся истории самого понятия. Понятие *Support Verb Construction* (далее SVC) стало сравнительно недавно предметом лингвистических исследований: первым лингвистом, столкнувшимся с этим вопросом был Есперсен [6], который называл «легкими» (*light verbs*) те глаголы, которые в рамках определенных конструкций проявляют семантику, частично или полностью «облегченную» по сравнению со своим первичным значением.

В семидесятые и восьмидесятые годы к этому вопросу возвращались Поленц [11] (для немецкого языка) и французские исследователи [4, 5] романских языков. Несмотря на то, что сегодня более распространено определение *support verbs*, данные структуры продолжают изучаться с различных точек зрения в разных лингвистических школах. В англо-американской лингвистике

бытуют определения *light verbs*, *operator verbs*, *support verbs*, тогда как в французской и вообще в лингвистических школах романских стран речь идет о *constructions a verbe support*, а в немецкой лингвистике преобладает скорее термин *Funktionsverbgefuege*. Конструкции с опорным глаголом заняли должное место в анализе естественных языков, в *Frame Semantics* [3] и в разных областях формальной семантики и компьютерной лингвистики. В советской и российской лингвистике [7] SVC *первым исследовал* Мельчук — хотя и не называя их эксплицитно — уже с '60-х гг. в рамках своей теории о лексических функциях [9]. В некоторых работах последних лет стало употребляться определение глагол поддержки/опорный глагол. Между прочим, дело — не только в терминологическом разнобое, вопросы ставятся и по поводу самой природы данной категории: в одних случаях SVC понимается в очень широком смысле [5], в других — резко критикуется даже само право на существование данной категории, так как зачастую трудно определить границы SVC и ограничить автономное поле их исследования [12].

2.1. Что понимаем под SVC

Конструкцией с глаголом поддержки в самом широком смысле будем называть ту конструкцию, в которой присутствует глагол V семантически частично или полностью облегченный, по сравнению со своим первичным значением, вместе с существительным-предикатом N, являющимся прямым дополнением V (1), или от V зависимым посредством предлога (2), и придающим значение целой конструкции.

(1) [V [N]]

(2) [V [Prep[N]]]

Самые спорные вопросы в классификации таких конструкций возникают именно по поводу уровня десемантизации и референциальности V и N.

3. Структура работы

В работе попытаемся выявить семантические и морфосинтаксические характеристики итальянских конструкций с глаголом поддержки *fare*, представляющих следующую структуру (3):

(3) [N_1[fare [(det) N_2]]]

Выбор глагола *fare* 'делать' диктуется тем, что, скорее всего, в итальянском языке он представляет собой самый продуктивный глагол поддержки и, следовательно, является особенно проблематичным в процессе перевода.

В (§4) рассмотрим прототипическое значение глагола *fare* и опишем критерии, позволяющие отличать SVC от конструкций лишь поверхностно подобных, т. е. свободных глагольных групп (§4.1). Более того, попытаемся выявить критерии, позволяющие отличать, в рамках SVC, полностью устойчивые словосочетания от конструкций более свободных, которые будем называть «глагольными коллокациями» (§4.2).

В (§ 5) сосредоточимся на глагольных коллокациях и, на основе данных корпуса предложим классификацию (*data driven*) разных видов конструкций, в которых глагол *fare* является глаголом поддержки.

В заключение, в (§6), попытаемся сравнить итальянские и русские глагольные коллокации, и показать, какие продуктивные в итальянском языке SVC с глаголом *fare* имеют соответствие в русском языке с SVC с глаголом делать/сделать, а также выявить основные стратегии, используемые в русском языке для того, чтобы найти соответствия в тех случаях, когда глагол *fare* употребляется не в своем прототипическом значении. Покажем также разные виды данных стратегий (употребление другого глагола поддержки, использование морфологических средств — префиксации, суффиксации и лексических средств) и попытаемся делать выводы типологического характера.

4. Прототипическое значение глагола *fare*

В своем первичном значении *fare* описывает предельный акт создания, который приводит к конкретному результату. Действие может быть точечным (*fare un errore* ‘делать ошибку’), или длительным (*fare una traduzione* ‘делать перевод’).

Событийную структуру глагола в данном значении можно представить следующим образом:

Лексическое значение:	*creare* ‘создать’, *compiere* ‘завершить’ N_2
Акциональный класс:	[± длител.], [+акт-длит.], [+предел.] → совершения /достижения
Аргументная структура:	N_1: *существо* [+одуш.], [+агентив.] N_2: *существо* [+конкр.]

Когда глагол *fare* употребляется в данном значении, он создает вместе со своим дополнением свободную глагольную группу, эта «свобода» проявляется как в парадигматическом плане (то же значение реализуется при замене глагола другим глаголом того же синонимического ряда: *fare una torta* ‘сделать торт’/ *preparare una torta* ‘приготовить торт’/ *cucinare una torta* ‘испечь торт’; *fare un errore* ‘сделать ошибку’/ *commettere un errore* ‘допустить ошибку’/ *compiere un errore* ‘совершить ошибку’) так и в синтагматическом плане (словосочетание поддается синтаксическим манипуляциям: *fare una torta saporita* ‘сделать *вкусный* торт’, *fare un errore irrimediabile* ‘сделать *неисправимую* ошибку’).

4.1. SVC vs свободные глагольные группы

Первый шаг — выявление критериев, позволяющих отличать случаи, когда *fare* образовывает свободную гл. группу, от тех, в которых он образует несвободную конструкцию, являясь, таким образом, глаголом поддержки.

Первый критерий — семантический: в несвободных конструкциях глагол предствляет событийную структуру, не совпадающую с тем, что описано в §4: например в словосочетании *fare una/la doccia* ‘принять душ’ (букв. ‘сделать душ’) акциональный класс (совершение) и лексическое значение N_1 и V соответствуют первичному значению *fare*, тогда как N_2 представляет собой не конкретный предмет, а имя события. В *fare festa* ‘праздновать’ (букв. ‘делать праздник’), разница еще четче: лексическое значение глагола никоим образом не совпадает со значением ‘создать’ и акциональный класс конструкции — деятельность.

Второй критерий скорее формальный: CVS не свободны на парадигматическом плане [8], т. е. составные элементы обязательны и не могут быть заменены синонимами: *fare la doccia*, (букв. ‘делать душ’), но норма итальянского языка не позволяет замену на **eseguire la doccia* (букв. ‘совершить душ’), *fare finta* ‘делать вид’, но **fare finzione* (букв. ‘делать **притворство**’), или **compiere finta* (букв. ‘совершить вид’).

В тех немногочисленных случаях, когда замена возможна, имеется лишь несколько вариантов, имеющих диалектную, индивидуальную или социалную окраску.

4.2. SVC: глагольные коллокации vs устойчивые глагольные словосочетания

Как уже было сказано, в рамках SVC нужно отличать жестко зафиксированные конструкции, которые мы назвали устойчивыми глагольными словосочетаниями, от несвободных конструкций, которые мы определили как глагольные коллокации. Первые характеризуются очень жесткой внутренней когезией, не позволяющей ни одной из нижеуказанных синтаксических манипуляций:

- **Изменение N_2 (ввод прилагательного, относительного прид. предл.)**

(4) *Ho fatto una doccia* → *Ho fatto una doccia calda*
‘Я принял душ’ → ‘Я принял горячий душ’

(5) *Fa finta di essere ricco* → **Fa finta incomprensibile di essere ricco*[2]
‘Он делает вид, что он богат’ → ‘*Он делает непонятный вид, что он богат’

[2] Исключением являются квантификаторы, которые на самом деле изменяют семантику глагола, а не имени:

- **Замена N_2 местоимением:**

(6) *Ha fatto una doccia. L'ha fatta perché era molto stanco.*
'Он принял душ. Он его принял, потому что был очень усталым'

(7) *Giovanni fa finta di essere ricco;*la fa per conquistare Maria.*[3]
'Иван делает вид, что он богат; *он его делает, чтобы Мария его полюбила'

- **Пассивизация конструкции:**

(8) *Ha fatto la spesa al supermercato → La spesa è stata fatta al supermercato*
'Он сделал покупки в супермаркете' → 'Покупки были сделаны в супермаркете'

(9) *Ha fatto finta di essere ricco per conquistare Maria → *Finta è stata fatta di essere ricco per conquistare Maria*
'Он сделал вид, что он богат, чтобы Мария его полюбила' → '*Вид был сделан, что он богат, чтобы Мария его полюбила'

В категории устойчивых словосочетаний находим и те конструкции, которые Мельчук [10] называет идиомами, т. е. непрозрачные идиоматические словосочетания, значение которых не извлекается из значений отдельно взятых составляющих идиому слов: несмотря на то, что большинство устойчивых глагольных словосочетаний имеет идиоматический характер (*fare cilecca* 'дать осечку' (букв. 'делать осечку'), *fare acqua* 'трещать по всем швам' (букв. 'делать воду'), тем не менее, среди них есть и конструкции, чье значение можно извлечь полностью или частично из значений составляющих элементов (*fare scandalo* 'строить скандал', букв. 'делать скандал'), хотя сами конструкции являются жестко зафиксированными как на парадигматическом, так и на синтагматическом уровнях.

Таким образом, различие в рамках SVC между глагольными коллокациями и устойчивыми глагольными словосочетаниями проводится не только на основе семантики, но и в структурном плане. На основе этих двух критериев можно расположить все конструкции по линии нарастания семантической непрозрачности и внутренней когезии:

свободные глагольные группы > глагольные коллокации с глаголом поддержки > устойчивые глагольные словосочетания с глаголом поддержки

---

faccio frequente riferimento aisuoiscritti = faccio frequentemente *riferimento aisuoiscritti*
'делаю частую ссылку на его работы = делаю часто ссылку на его работы'

[3] В предложении такого типа местоимением можно заменить целиком конструкцию: *Иван делает вид, что он богат; он это делает, чтобы Мария его полюбила.*

5. Семантические классы

Исследование в итальянском языке проводилось на материалах корпуса *ITWac* (сост. Baroni, 1,909,535,703 tokens), а в русском языке — на материалах корпуса *Russian Web Corpus* (сост. С. А. Шаров, 187,965,822 tokens): поиск по обоим корпусам проводился по программе *Sketch Engine* (http://the.sketchengine.co.uk).

Поиск выдал все вхождения [*fare* [(det) N]] в итальянском корпусе и [делать/сделать [N]] в русском корпусе, после этого программа позволила их отсортировать по частоте.

Что касается итальянских данных, были учтены все вхождения, частота которых составляла ≥ 200[4].

Синтаксические тесты нам позволили сначала отличить свободные глагольные группы от SVC, а затем среди SVC были выделены с одной стороны глагольные коллокации, а с другой — устойчивые глагольные словосочетания. Мы сосредоточились на глагольных коллокациях, пытаясь разделить их по семантическим признакам на определенные подклассы, в рамках каждого из которых глагол *fare* приобретает преобладающее значение и сочетается с именами N_2, имеющими общие семантические характеристики.

Мы разделили, таким образом, выделенные семантические подклассы на 5 больших групп, соответствующих акциональным классам Вендлера [14], к которым был еще добавлен класс однократных глаголов [2].

Выделение данных акциональных классов позволяет располагать их по порядку увеличивающегося удаления от прототипического значения глагола *fare* в свободных глагольных группах.

На первом месте располагаются коллокации, обладающие результативным характером (*accomplishments* ‘совершения’), хотя в них глагол *fare* не является глаголом создания (*fare spese* ‘делать покупки’); дальше находятся коллокации, описывающие изменения (*achievements* ‘достижения’: *fare* capolino ‘выглянуть’ букв. ‘делать головку’, *fare amicizia* ‘подружиться’ букв. ‘делать дружбу’, *fare richiesta*, ‘делать запрос’), в которых полностью отсутствует предельное значение, на третьем месте *semelfactives* ‘однократные’, характеризующие точечные действия *(fare un salto* ‘зайти’, букв. ‘делать прыжок’, *fare cenno* ‘намекнуть’ букв. ‘делать намек’), на четвертом *activities* ‘деятельности’ (*fare politica* ‘заниматься политикой’, букв. ‘делать политику’, *fare lo stupido* ‘валять дурака’, букв. ‘делать дурака’), и наконец, на пятом месте, расположены *states* ‘стативы’, характеризующие действия, не имеющие ни актуально-длительного, ни предельного значений (*fare il ministro* ‘быть министром’ букв. ‘делать министра’, *fare il sindaco* ‘быть мэром’ букв. ‘делать мэра’).

Данная классификация, как известно, учитывает лишь *лексическую аспектуальность*, оставляя в стороне *грамматическую аспектуальность*, которая

[4] Ниже мы приведем примеры в порядке уменьшающейся частоты.

в итальянском языке выражается и морфологическими (категорией времени) и лексическими (присутствием/отсутствием артикля) способами.

Иными словами, не учитывается то, что называется *Aspectual shift* [13], т. е. тот факт, что глаголы, однозначно принадлежащие к одному акциональному классу, могут перейти в другой класс из-за изменений в контексте:

(10)

а.	*ha fatto una passeggiata*,	*e poi*	è *tornato*	*a casa*
	сделал прогулку [+длитель.], [+акт-длит.], [+предель.],	и потом	вернулся	домой
б.	*mentre*	*faceva una passeggiata*,	*si è messo a piovere*.	
	пока	делал прогулку[+длитель.], [+акт-длит.], [-предель.],	пошел дождь	

5.1. Совершения

В значении, не слишком далеком от прототипического, *fare* может обозначать действия речевого или интеллуктуального создания:

5.1.1. *Тип* Fare il punto *‘подвести итоги’ (букв. ‘делать точку’):*

fare il punto, *fare previsioni*, *fare un discorso*, *fare progetti*, *fare ipotesi*, *fare rapporto*, *fare programmi*, *fare valutazioni*, *fare il bilancio*.

Лексическое значение:	N_1 выполняет речевой/когнитивный акт, описанный N_2
Акциональный класс:	[+длитель.], [+акт-длит.], [+предель.] → Совершение
Аргументная структура:	N_1: существо [+одуш.], [+агент.] N_2: речевой/когнитивный акт

5.1.2. *Тип* fare la spesa *‘делать покупки’:*

fare la spesa, *fare un giro*, *fare colazione*, *fare una passeggiata*, *fare il bagno*, *fare lezione* (*di*) (=*impartire lezione di*), *fare un viaggio*, *fare rifornimento*, *fare un gioco*, *fare benzina*[5], *fare la doccia*, *fare merenda*.

Лексическое значение:	N_1 занят работой/развлечением, описанной/ым N_2
Акциональный класс:	[+длитель.], [+акт-длит.], [+предель.] → Совершение
Аргументная структура:	N_1: существо [+одуш.], [+агент.] N_2: трудовая/ развлекательная деятельность

[5] Здесь наблюдается метонимический перенос, благодаря которому событие передается именем результата: *fare benzina* ‘делать заправку бензином’, букв. ‘делать бензин’. Апресян относит такие примеры к явлению регулярной многозначности [1].

5.1.3. *Тип* fare la scuola XY *'учиться в школе XY' (букв. 'делать школу XY'):*

fare la scuola XY, fare filosofia, fare l'università, fare il liceo.

Лексическое значение: N_1 учится в учебном заведении, обозначенном N_2

Акциональный класс: [+длитель.], [+акт-длит.], [+предель.] → Совершение

Аргументная структура: N_1: существо [+одуш.], [-агент.]

N_2: учебное заведение [±референц.]

5.1.4. *Тип* fare le analisi *'делать анализы':*

fare le analisi, fare il vaccino, fare le iniezioni, fare i massaggi[6].

Лексическое значение: N_1 подвергается медицинским/косметическим процедурам, обозначенным N_2

Акциональный класс: [+длитель.], [+акт-длит.], [+предель.] → Совершение

Аргументная структура: N_1: существо [+одуш.], [-агент.] = пациенс

N_2: медицинская/косметическая процедура

5.2. Достижения

5.2.1. *Тип* fare capolino *'выглянуть' (букв. 'делать головку'):*

fare capolino, fare irruzione, fare un salto (= *andare per breve tempo da qualche parte*), *fare marcia indietro, fare breccia, fare ingresso, fare rientro, fare retromarcia.*

Лексическое значение: N_1 совершает точечное движение, обладающее предельное значение

Акциональный класс: [-длитель.], [+акт-длит.], [+предель.] → Достижение

Аргументная структура: N_1: существо [+одуш.], [+агент.]

N_2: действие/результат

5.2.2. *Тип* fare amicizia *'подружиться' (букв. 'делать дружбу'):*

fare amicizia, fare pace, fare conoscenza

Лексическое значение: N_1 завершает действие, изменяющее его положение

Акциональный класс: [-длитель.], [+акт-длит.], [+предель.], [+инхоативный] → Достижение/Инхоатив

Аргументная структура: N_1: существо [+одуш.], [+агент.]

N_2: результат [-предм.]

5.2.3. *Тип* fare obbligo *'обязывать'(букв. 'делать обязанность'):*

fare obbligo, fare divieto, fare concessione.

Лексическое значение: N_1 навязывает свою волю N_2 (кому-л. N_3)

[6] Данный подкласс является очень продуктивным, поэтому мы решили включить в классификацию и элементы имеющие частоту < 200.

Акциональный класс: [-длитель.], [+акт-длит.], [+предель.], [+ каузатив] → Достижение
Аргументная структура: N_1: существо [±одушевл.] [+агент.]
N_2: принудительный акт [-предм.]
(N_3: существо [+одушевл.] [+агент.])

5.2.4. *Тип* fare un esempio *'делать пример':*

fare richiesta, *fare un esempio*, *fare ricorso* (= *presentare un ricorso*), *fare domanda*, *fare appello*, *fare menzione*, *fare una domanda*, *fare nome e cognome*, *fare causa*, *fare rinvio*, *fare distinzione*, *fare una proposta*, *fare un bilancio*, *fare un paragone*, *fare parola*, *fare testamento*, *fare ammenda*, *fare denuncia*, *fare segnalazione*.

Лексическое значение: N_1 выполняет речевой/когнитивный акт N_2
Акциональный класс: [-длитель.], [+акт-длит.], [+предель.], → Достижение
Аргументная структура: N_1: существо [+одушевл.], [+агент.]
N_2: речевой /когнитивный акт [-предм.]

5.2.5. *Тип* fare carriera *'делать карьеру':*

fare carriera, *fare strada* (= *avere successo*), *fare soldi*, *fare cassa*, *fare affari*, *fare strage*, *fare centro*, *fare fortuna*.

Лексическое значение: N_1 достигает результата, обозначенного N_2
Акциональный класс: [-длитель.], [+акт-длит.], [+предель.] → Достижение
Аргументная структура: N_1: существо [+одушевл.], [+агент.]
N_2: результат

5.2.6. *Тип* fare chiarezza *'выяснить'(букв. 'делать ясность'):*

fare chiarezza, *fare luce*, *fare giustizia*, *fare strada* (= *lasciar passare*), *fare coraggio*, *fare pulizia*, *fare ordine*, *fare posto*.

Лексическое значение: N_1 совершает действие, производящее изменение в аргументе N_3, который может быть не обозначен эксплицитно
Акциональный класс: [-длитель.], [+акт-длит.], [+предель.] → Достижение
Аргументная структура: N_1: существо [+одушевл.], [+агент.]
N_2: результат
(N_3: существо, конкретное или абстрактное, претерпевающее изменение)

5.3. Однократные

5.3.1. *Тип* fare clic *'кликнуть' (букв. 'делать клик'):*

fare clic, *fare cenno*, *fare un passo*, *fare un salto*, *fare segno*, *fare salti*, *fare un sorriso*, *fare movimenti*, *fare un gesto*, *fare un tuffo*.

Лексическое значение: N_1 совершает движение (конкретное или метафорическое), не имеющее указания на срок завершения.
Акциональный класс: [-длитель.], [+акт-длит.], [-предель.] → Однократный
Аргументная структура: N_1: существо [+одушевл.], [+агент.]
N_2: однократное действие

5.4. Деятельность

5.4.1. *Тип* fare polemica *'делать полемику':*

fare polemica, *fare propaganda*, *fare campagna, fare discorsi*, *fare confronto*, *fare paragone*, *fare indagini*, *fare scherzi*, *fare complimenti*, *fare astrazione*.

Лексическое значение: N_1 осуществляет речевую/когнитивную деятельность, обозначенную N_2
Акциональный класс: [+длитель.], [+акт-длит.], [-предель.] → Деятельность
Аргументная структура: N_1: существо [+одушевл.], [+агент.]
N_2: речевая/когнитивная деятельность

5.4.2. *Тип* fare ricerca *'исследовать'(букв. 'делать исследование'):*

fare l'amore, *fare ricerca*, *fare esperienza*, *fare compagnia*, *fare musica* (= *occuparsi di musica*), *fare sesso*, *fare sport*, *fare politica*, *fare shopping*, *fare lezione di* (= *seguire lezione di*), *fare cinema* (= *occuparsi di cinema*), *fare cultura*, *fare volontariato*, *fare ginnastica*, *fare sciopero*, *fare jogging*, *fare commercio*, *fare turismo*, *fare un corso* (= *seguire un corso di*).

Лексическое значение: N_1 занимается физической/интеллектуальной деятельностью (трудовой или развлекательной), обозначенной N_2
Акциональный класс: [+длитель.], [+акт-длит.], [-предель.] → Деятельность
Аргументная структура: N_1: существо [+одушевл.], [+агент.]
N_2: физическая/интеллектуальная деятельность

5.4.3. *Тип* fare riferimento *'делать ссылку':*

fare riferimento, *fare ricorso* (=*ricorrere*), *fare uso*, *fare affidamento*, *fare economia*.

Лексическое значение: N_1 выполняет действие, обозначенное отглагольным N_2
Акциональный класс: [+длитель.], [+акт-длит.], [-предель.] → Деятельность
Аргументная структура: N_1: существо [+одушевл.], [+агент.]
N_2: имя действия/результата

5.4.4. *Тип* fare tappa *'делать остановку':*

fare visita, fare tappa, fare scalo, fare sosta, fare rotta.

Лексическое значение: N_1 движется

Акциональный класс: [+длитель.], [+акт-длит.], [-предель.] → Деятельность

Аргументная структура: N_1: существо [+одушевл.], [+агент.]
N_2: движение [-предм.]

5.4.5. *Тип* fare fronte *'справиться' (букв. 'делать фронт'):*

fare fronte, fare attenzione, fare fatica, fare leva, fare pressione, fare forza, fare perno, fare opposizione, fare guerra.

Лексическое значение: N_1 применяет интеллектуальные способности/физическую силу, обозначенные N_2

Акциональный класс: [+длитель.], [+акт-длит.], [-предель.] → Деятельность

Аргументная структура: N_1: существо [±одушевл.], [+агент.]
N_2: деятельность

5.4.6. *Тип* fare rumore *'делать шум':*

fare rumore, fare casino, fare confusione, fare silenzio, fare ombra, fare chiasso.

Лексическое значение: N_1 занимается деятельностью, изменяющей окружающее пространство

Акциональный класс: [+длитель.], [+акт-длит.], [-предель.] → Деятельность

Аргументная структура: N_1: существо [+одушевл.], [+агент.]
N_2: существо, относящее к чувствам [+предм.]

5.4.7. *Тип* fare il furbo[7] *'хитрить' (букв. 'делать хитрого'):*

fare il furbo, fare la vittima, fare la spia, fare il figo, fare il moralista.

Лексическое значение: N_1 ведет себя, как обозначено N_2

Акциональный класс: [+длитель.], [+акт-длит.], [-предель.] → Деятельность

Аргументная структура: N_1: существо [+одушевл.], [+агент.]
N_2: существо [+одушевл.]

5.5. Стативы

5.5.1. *Тип* fare il presidente *'быть президентом' (букв. 'делать президента'):*

fare il presidente, fare il sindaco, fare il ministro, fare il medico, fare il giornalista, fare il direttore.

Лексическое значение: N_1 работает/является N_2

[7] См. сноску 4.

Акциональный класс: [+длитель.], [-акт-длит.], [-предель.] → Статив
Аргументная структура: N_1: существо [+одушевл.], [-агент.]
N_2: профессия / социальная роль

5.5.2. *Тип* fare piacere *'быть приятным' (букв. 'делать удовольствие'):*

fare piacere, *fare paura*, *fare schifo*, *fare pena*, *fare invidia*, *fare gola*, *fare impressione*, *fare rabbia*, *fare tenerezza*, *fare orrore*, *fare meraviglia*, *fare pietà*.

Лексическое значение: N_1 вызывает ощущение, обозначенное N_2
Акциональный класс: [+длитель.], [-акт-длит.], [-предель.] → Статив
Аргументная структура: N_1: существо [-агент.], [±одуш.]
N_2: чувство/ощущение
N_3: экспериент [+ одуш.][-агент.]

6. Сравнение итальянского и русского языков

В своем первичном значении итальянский глагол *fare* соответствует русскому глаголу делать:

fare un lavoro = делать/сделать работу

А в тех случаях, когда *fare* выступает как глагол поддержки, он имеет целый ряд лексических значений, которые в русском языке редко выражаются глаголом делать:

fare una doccia ≠ принимать душ; *fare sport* ≠ заниматься спортом

Ниже мы попытаемся выяснить, существует ли в русском языке доминирующая стратегия для передачи значения каждого выделенного в итальянском языке класса конструкций с глаголом поддержки *fare*.

В данной работе мы ограничились только исследованием коллокаций, и не рассматривали устойчивые глагольные группы, поскольку они, будучи часто лексикализованными и идиоматичными, не позволяют сравнения двух языков.

6.1. Совершения

Коллокации с глаголом fare, входящие в класс глаголов совершения, не всегда имеют *соответствующие формы в русском языке*.

Что же касается глаголов, обозначающих речевые и интеллектуальные действия (*fare il punto*), наблюдается *высокая* степень соответствия двух языков:

fare pronostici = делать прогнозы, *fare un progetto* = делать проект, *fare una valutazione* = делать оценку.

В то время как коллокации, обозначающие виды трудовой/развлекательной деятельности актуально-длительного типа (*fare la spesa* ‘делать покупки’), не имеют соответствия в русском языке*:* из 12 коллокаций с глаголом *fare* с показателем частоты ≥ 200 только 4 *переводятся* на русский язык коллокацией с глаголом делать:

делать покупки, делать игр-у/-ы, делать запас/-ы, делать прогулку

В тех случаях, когда используются аналитические структуры, предпочитается глагол *с семантически более узким значением*: например, для обозначения понятия *fare il bagno/la doccia* (букв. ‘делать ванну/душ’) в русском языке употребляется опорный глагол принимать (принимать ванну/душ), который подчеркиват низкую агентивность N1.

Коллокация типа *fare la scuola XY* (букв. ‘делать школу XY’), которой в итальянском языке обозначается посещение учебного заведения, передается на русский язык с помощью глаголов состояния (учиться, заниматься в N2) и деятельности (посещать N2): более того, итальянская конструкция в перфективных временах приобретает предельное значение (*ho fatto l’universit*à *in 5 anni* ‘я закончил(а) университет за пять лет’), которое отсутствует в соответствующих русских конструкциях: *я посетил(а) университет за пять лет.

Конструкции с глаголом *fare*, обозначающие получение косметических/медицинских услуг (тип *fare le analisi*), вполне соответствуют русским конструкциям с глаголом делать, которые присутствуют в русском корпусе в относительно большем количестве, чем в итальянском (частота указана в скобках):

делать операцию (182), аборт (115), массаж (113), укол/-ы (110), прививку/-и (60), анализ/-ы (51),УЗИ (41).

6.2. Достижения

Что касается класса глаголов достижения, который в итальянском языке является особенно продуктивным, нужно отметить, что для передачи на русский язык возможны два способа. Для обозначения точечных предельных движений русский язык в большинстве случаев пользуется морфо-лексическими ресурсами, т. е. синтетическими формами с суффиксами инхоативного или недуративного значений[8]: *fare capolino* = выглянуть (букв. ‘сделать головку’), fare irruzione = **ворваться**, *fare un salto* = зайти. Однако в корпусе встречаются далеко не единичные аналитические структуры, соответствующие итальянским: сделать набег, налет/-ы, наезд.

[8] Для краткости мы здесь решили представить только совершенную форму русских глаголов достижения, которая в большей степени подчеркивает предельное значение этих конструкций.

Коллокации, *обозначающие перемену состояния* N1 (*fare amicizia*, букв. ‘сделать дружбу’, *fare pace*, букв. ‘сделать мир’, *fare conoscenza*, букв. ‘сделать знакомство’ передаются в русском языке как синтетически (лексическими элементами с инхоативным значением: помириться, подружиться, познакомиться), так и аналитически (аналитическими конструкциями, подчеркивающими результат: стать друзьями, заключить мир).

Для обозначения регулирующих и принудительных актов, используемых в итальянском официально-административном стилистическом регистре, русский язык пользуется синтетическими структурами (запретить, обязать, допустить), что соответствует, между прочим, и употреблению в разговорном и литературном стилях речи итальянского языка: *vietare*, *obbligare*, *concedere*.

Что же касается обозначения точечных речевых и интеллектуальных актов, в большинстве случаев итальянские коллокации соответсвуют русским (*fare domanda* = делать запрос; *fare l’appello* = делать перекличку, *fare differenza* = делать различия, *fare una proposta* = делать предложение); то же самое происходит с коллокациями, обозначающими достижение результата: *fare carriera* = делать карьеру, *fare soldi* = делать деньги.

События, производящие модификацию состояния, необязательно выраженного аргумента N3 (*fare chiarezza*, букв. ‘делать ясность’, *fare luce* букв. ‘делать свет’, *fare giustizia* букв. ‘делать справедливость’, *fare ordine* букв. ‘делать порядок’), передаются на русский язык морфо-лексическими ресурсами (выяснить, осветить) или конструкциями с разными глаголами поддержки: отдать справедливость, привести в порядок.

6.3. Однократные

В классе однократных глаголов в русском языке наблюдаются три закономерности:

а) использование морфологической стратегии, т. е. в основном употребление суффикса *-ну-*: прыгнуть, щелкнуть, улыбнуться;
б) сосуществование синтетической формы с суффиксом *-ну-* с аналитической: делать шаг/шагнуть, делать намек/намекнуть, делать движение/двигаться;
в) только конструкция с глаголом поддержки: делать знак.

6.4. Деятельность

В этом классе, как и можно было предвидеть, наблюдается мало соответствий, преобладающими являются:

а) лексическая стратегия: используется один глагол;
б) CVS в тех случаях, когда N_2 представляет собой речевой или когнитивный акт (делать пропаганду, делать комплименты), или движение/прекращение движения (делать остановку, перерыв, пересадку), или же представляет собой занятие (делать музыку, кино, фильм).

6.5. Стативы

Для передачи этого чрезвычайно продуктивного в итальянском языке класса, русский язык регулярно прибегает к именному составному сказуемому, делая акцент на N_2, выраженном существиетельным в творительном падеже: работать журналистом.

7. Заключение

Благодаря извлечению из корпуса конструкций с глаголом *fare/делать* стало возможно оценить продуктивность данной конструкции в итальянском языке и значение соответствующего явления в русском. На основе материалов, полученных в результате работы над корпусами стало возможным разделить все конструкции с глаголом *fare* на три подгруппы: свободные глагольные группы, глагольные коллокации и устойчивые глагольные словосочетания с опорным глаголом. На данном этапе работа ограничилась классификацией глагольных коллокаций на основе преобладающего лексико-семантического значения глагола в каждом подклассе. Сравнение с русским языком показывает, как и предполагалось, что нет бинарного соответствия между двумя языками: более или менее регулярные стратегии соответствия наблюдаются только в передаче глаголов совершения и достижения, а чем больше значение глагола *fare* удаляется от своего первичного значения (в глаголах деятельности и состояния), тем больше русский язык пользуется соответствующими синтетическому типу стратегиями. Однако, данные показывают растущую продуктивность аналитических конструкций и в русском языке.

Что касается прикладных выходов данного исследования, то уже на этой стадии работы можно сказать, что семантическая разметка аргументов глагола (в частности N_2) каждого подкласса позволила бы осуществлять автоматический перевод коллокаций разных подклассов.

References

1. *Apresian Iu. D.* 1974. Lexical Semantics: Synonimical Linguistic Media [Leksicheskaia Semantika: Sinonimicheskie Sredstva Iazyka].
2. *Comrie B.* 1976. Aspect: An Introduction to the Study of Verbal Aspect and Related Problems.
3. *Fillmore C. J., Johnson C. R., Petruck M. R.* Background to Framenet. International Journal of Lexicography, 16 (23): 235–250.
4. *Gross M.* 1981. Les Bases Empiriques de la Notion de Predicat Semantique. Langages, 15, 63 : 7–49.
5. *Gross G., Pontonx S. De.* 2004. Les Verbs Supports. Nouvel Etat de Lieu. Special Issue of Lingvisticae Investigationes, 27 (2).

6. *Jespersen O.* 1965. A Modern English Grammar on Historical Principles. Part VI. Morphology, VI : 117.
7. *Langer S.* 2004. A Linguistic Test Battery for Support Verb construction. Verbes Supports. Nouvel état des lieux. Special issue of Linguisticae Investigationes, 27 (2) : 171–184.
8. *Masini F.* 2009. Combinazioni di Parole e Parole Sintagmatiche. Spazi Linguistici. Studi in Onore di Raffaele Simone : 191–209.
9. *Mel'chuk I. A.* 1982. Lexical Functions in Lexicographic Description. Proceedings of the Eighth Annual Meeting of the Berkeley Linguistics Society : 427–444.
10. *Mel'chuk I. A.* 1995. Phrasemes in Language and Phraseology in Linguistics. Idioms: Structural and Psychological Perspectives : 69–252.
11. *Polenz P.* 1963. Funktionsverben im Heutigen Deutsch. Sprache in der Rationalisierten Welt.
12. *Pottelberg J.* 2001. Verbonominale Konstruktionen, Funktionsverbgefüge.
13. *Rothstein S.* 2004. Structuring Events. A Study in the Semantics of Lexical Aspect.
14. *Vendler Z.* 1967. Linguistics in Philosophy.

ЭЛЕКТРОННАЯ ПОЧТА VS. ЧАТ: ВЛИЯНИЕ КАНАЛА КОММУНИКАЦИИ НА ЯЗЫК

А. Бердичевский (alexander.berdichevsky@if.uib.no)

University of Bergen, Norway

Влияет ли смена канала коммуникации, не сопровождающаяся изменением других ситуационных параметров, на лингвистические характеристики коммуникации? Количественный анализ двух корпусов русских текстов, отличающихся исключительно каналом (электронная почта vs. чат) показывает, что влияет.

Ключевые слова: канал коммуникации, ситуационные параметры, характеристики коммуникации, лингвистические характеристики, анализ.

E-MAIL VS. CHAT: THE INFLUENCE OF THE COMMUNICATION CHANNEL ON THE LANGUAGE[1]

A. Berdichevskii (alexander.berdichevsky@if.uib.no)

University of Bergen, Norway

Does the mere change of the communication channel, unaccompanied by any other changes in situational characteristics, affect the language? Quantitative analysis of two corpora of Russian texts that differ solely by the communication channel from which they originate (e-mail vs. chat) proves that it does.

Key words: communication cannel, situational characteristics, communication characteristics, linguistic characteristics, analysis.

[1] This work was carried out as part of the project “The Future of Russian: Language Culture in the Era of New Technology”, supported by the Norwegian Research Council and the University of Bergen.

1. Introduction

Linguists are paying ever increasing attention to computer-mediated communication (CMC) and "the language of the Internet". At the "Dialogue" conference the Internet is usually viewed as a tool, not an object of linguistic research; however, even here one can find papers that focus on the linguistic properties of electronic communication (Макаров, Школовая 2006; Зализняк, Микаэлян 2006; Бурас, Кронгауз 2007; Богданов 2008; Anni 2008; Занегина 2009; Людовик 2010). In order to pursue a study of this kind, the scholar has to assume that the linguistic properties of CMC are somewhat different from those of other media (oral speech and written speech, for example) and thus worthy of separate research.

This assumption is often based on a more general one: the physical properties of the communication channel affect the linguistic properties of communication taking place in this channel, acting either as constraints or as enablements (see Hård af Segerstad 2002: 10–11 for the history of this term). This hypothesis has been well researched in the context of the differences between written and oral speech:e. g. see the classic works of Chafe (1982) and Biber (1988). Later, interest in this field was reinvigorated by the emergence and spread of a new channel, namely CMC. The constraints there seem to be heavier than in "traditional" channels, and the enablements wider, so that one might expect that their influence on the language would be clearly visible and detectable by quantitative methods.

Since the 1980s there have been quite a few studies that have used quantitative approaches to examine differences and similarities between CMC and other channels. It is important to keep in mind that CMC is not monolithic, and that we are in fact speaking about a set of different communication channels, united by the same physical medium: these channels have been compared to each other as well. See Collot and Belmore 1996, Yates 1996, Hård af Segerstad 2002 and review therein, Jensen 2007, Ling and Baron 2007, Tagliamonte and Denis 2008 and review therein. The results showed that CMC (or rather the specific channel studied — instant messaging, e-mail, computer conferencing and so on) is indeed a new linguistic register, neither oral speech nor written speech, and often looks like a hybrid of these two. Asynchronous communication channels with unlimited buffer size (e. g. e-mail) tend to be more similar to traditional written speech, whereas synchronous channels, especially with limited buffer size (instant messaging), are more similar to oral speech. However, Ko (1996) showed that, in certain parameters, CMC is even more "spoken" than speech and more "written" than writing.

2. Aim of this study

My intention is to compare two communication channels within CMC: e-mail and a certain type of instant messaging. A principal novelty of this study is that the

registers compared differ just by one parameter, namely the communication channel, whereas all other parameters (communicators and their relation to each other, subject matter, time of the discussion etc.) are controlled for as much as possible.

The studies mentioned above are often criticized precisely because of the lack of a control for additional parameters. Critics claim that the differences ascribed to the influence of the communication channel might in reality depend on other factors, e. g. the subject of discussion. Androutsopoulos (2006) takes this criticism even further: he states that the focus of attention should be the social context of a discourse and not its channel-specific properties. He even raises doubts about the existence of any linguistic features which might be ascribed to a communication channel: "It is empirically questionable whether in fact anything like a 'language of e-mails' exists, simply because the vast diversity of settings and purposes of e-mail use outweigh any common linguistic features" (Androutsopoulos 2006: 420).

The question I am addressing is the following: *does the communication channel per se have any influence on the linguistic properties of communication?*

Another novelty of my study is that I am analyzing Russian: it seems important to take CMC studies beyond the Anglophone world.

3. Materials

As a data source, I am using the contents of my own Gmail mailbox. Gmail provides not only the usual e-mail communication, but also a chat system (called Gmail chat). Since the chat is integrated into the same window (Fig. 1) and is easy to use, it is becoming increasingly popular.

Hence, it is common for the same two people to communicate both via e-mail and via chat. I collect my chat and e-mail conversations with three persons from my contact list. In order to avoid the observer's paradox, I am only using conversations which took place after June 2007 (after I graduated) and before March 2009 (before I submitted a proposal for my current PhD position), that is, when I was neither studying nor working as a linguist and did not have an idea of the current study (or anything similar) in mind.

That allows me to control for all the parameters except the communication channel itself. Indeed, the interlocutors are always the same, the setting is always the same, the subject matter may, of course, vary, but in general it might be quite clearly seen that the same things are discussed in both chat and in e-mail messages. There is no distribution of topics (such as chat for personal matters, e-mail for business). Conversation topics include mostly personal, business, scholarly and educational matters, and none of these four classes is restricted to a particular channel.

There are four subjects in my corpora: all male, native speakers of Russian, at the moment of communication aged 18 to 32, one university student, two journalists and one researcher. The e-mail corpus consists of 12 260 words and the chat corpus of 17 671 words, giving 29 931 words in total. The communication is always one-to-one.

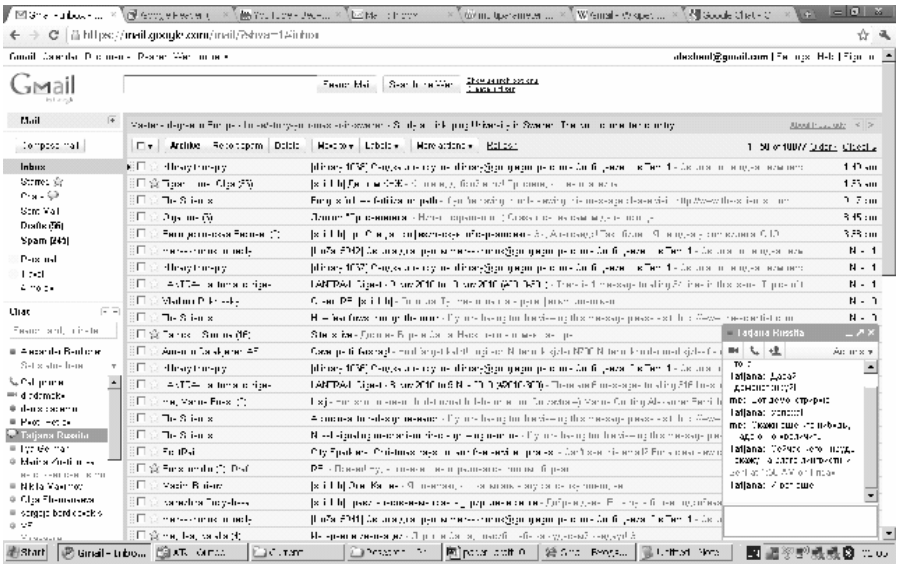


Fig. 1. Gmail chat

The Chat window is in the bottom right-hand corner. The contacts list can be seen in the bottom left-hand corner, and the name of the person who has sent a new message is highlighted.

4. Methods

Biber (1994) outlines a framework for the comparison of two registers. The framework consists of three components: analysis of the situational characteristics of the registers, analysis of the linguistic characteristics of the registers, and analysis of the functional and conventional associations between situational and linguistic characteristics. This section includes the situational analysis and lists the parameters for linguistic analysis . The “Results” section provides the results of the comparison of these parameters . The “Conclusions” section discusses the associations between situational and linguistic characteristics . This approach might be viewed as behavioural reductionism: I try to look at the influence of simple situational parameters on linguistic behaviour.

4.1. Differences between the situational characteristics of e-mail and Gmail chat

First, chat messages are delivered instantly. E-mails are also delivered quickly, but it might take a few seconds (or even minutes) for a letter to come.

Second, when you type an e-mail, your text is being auto-saved on a regular basis, so you do not have to worry about losing it should your browser crash, your

Internet connection be lost, or your computer stop working. When you type a message in a chat window, it is not saved anywhere until you send it.

Third, the chat window is narrow and small (see Fig. 1), while e-mail can occupy almost the whole screen. It is possible to open the chat in a separate window (and make it as large as one wants), or to install additional software in order to make chatting more convenient, but my subjects typically use the basic small window.

Fourth, when your interlocutor is typing a chat message to you, you can see an info message "XXX is typing..." (or "XXX has entered text") in the chat window.

Fifth, chat is more prone to technical failures: messages are more likely to get lost.

These are the real and primary differences between the two channels. They lead to the emergence of numerous secondary differences. For instance, in theory you may use chats to write long complex texts, but that would also be awkward : first, you always risk losing everything you have typed, second, it is inconvenient to read (and type, and edit) a large text in a small window. Some of these secondary differences are not, in fact, driven by physical reality, they are conventional. Strictly speaking, you do not have to answer to chat messages immediately, but that is what you are expected to do and what you usually do (and info messages contribute to users staying online and waiting for a reply to come).

Thus, chatting is usually a more synchronous, faster form of communication, implying immediate responses and rapid changes of turn. It is also somewhat less reliable and more volatile.

According to Biber, one of the principal oppositions in register comparison is informational versus involved production: "discourse with interactional, affective, involved purposes, associated with strict real-time production and comprehension constraints, versus discourse with highly informational purposes, which is carefully crafted and highly edited" (1988: 115). Oral speech is usually located closer to the "involved" pole of this dimension, while written speech — closer to the "informational" pole. It seems natural to expect that the same would be true forchat and e-mail respectively. Thus, many of the linguistic parameters discussed below are those that allow one to estimate the position of a register on this scale.

4.2. Quantitative parameters for discovering linguistic characteristics of e-mail and chat[2]

1. Mean length of an utterance (MLU)

Utterance here means 'sentence', with one exception: in chat, each turn is considered a separate utterance, i. e. a turn[3] might consist of several utterances (=sentences), but not vice versa. If a user chooses to split one sentence into nine turns (this is known to happen, although in my corpus they are rare), they are counted as nine utterances.

---

[2] Qualitative differences are not analyzed in this study.

[3] A turn is one chat message. In terms of Baron (2004: 408): 'composition (i. e., by typing) and transmission of an instant message'.

Otherwise, periods, exclamation, interrogation and ellipsis marks as well as emoticons were considered as marks to end an utterance. MLU is measured in symbols. High MLU is typical of informational speech production.

2. Mean length of a word (MLW)

High MLW is typical of informational speech production. Ko (1996) found MLW to be equal in speech and in instant messaging, but different from that in writing.

3. Lexical density (LD)

The ratio of lexical items (nouns, adjectives, verbs, adverbs, pronouns, numerals, as opposed to conjunctions, interjections, particles and prepositions) to the total number of words in a text. High LD is typical of informational speech production. Yates (1996: 35–39) showed that the LD of computer conferencing is close to that of writing, although still significantly different.

4. Type/token ratio (TTR)

The ratio of *different* words (types) in the text to the total number of words (tokens) in a text. Different word forms of the same lexeme were considered the same type, but different tokens. This measure depends on the text length, so it was calculated using two sub-corpora of equal size: 4 000 words.

High TTR implies a rich vocabulary and is typical of informational speech production. Yates (1996: 33–35) showed that the TTR of computer conferencing is close to that of writing, although still significantly different.

5. Sentence end marks

The percentage of sentences with any visible end marks: period, exclamation, interrogation or ellipsis marks. Sentences ending with an emoticon were also considered to have an end mark: sentence end is the most typical position for emoticons, and the period is usually omitted before them, so they can be viewed as an explicit signal of sentence end.

6. Capitals

The percentage of sentences beginning with a capital letter, as required by the rules of Russian punctuation/orthography.

7. Personal pronouns (first person, singular)

The ratio of the number of occurrences of the pronoun *я* ('I, me') (in all its forms) to the total number of words in a text. A high ratio is typical of "involved" speech production. Yates (1996: 40–41) found that the proportion of first-person pronouns in total pronoun use in CMC is higher than in speech, and in speech higher than in writing. Tagliamonte (2008: 16) confirmed the first part of this finding for instant messaging.

8. Brackets

The ratio of the number of brackets to the total number of words in the text. Brackets, too, serve as an indicator of informational production: a complex embedded

structure (both semantic and syntactic) is difficult to create (and perceive) when text is produced (and read) "on the fly".

9. Emoticons

The ratio of the number of emoticons to the total number of words in the text. The functions of emoticons are quite broad, but it is possible to state that, in general, speakers use them to compensate for the lack of non-verbal cues. Thus, high emoticon ratio would imply higher involvement.

10. Complex sentences

The ratio of complex sentences (i. e. sentences containing more than one clause) to the total number of sentences. Complex sentences are typical of informational production. This measure could not be calculated automatically, so it was calculated manually using the same sub-corpora that were compiled for measuring TTR.

Results

The results are summarized in Table 1. All the parameters were computed for each person and each pair separately, but only the results for the whole corpus are reported, since patterns were nearly the same in all cases.

The results of significance testing are reported, as well as effect sizes[4]. Parameters which are manifestly different for the two channels (difference is both significant and important) are highlighted in bold.

Table 1. Results

					% (1 per 100)				‰ (1 per 1000)	
	Utterance length	**Word length**	**LD**	**TTR**	**Sentence end marks**	**Capitals**	**1sg pronouns**	**Complex sentences**	**Brackets**	**Emoticons**
Chat	**33.8**	4.88	74.1	31.4	**54.7**	**78.3**	3.1	**22.9**	4.4	**22.9**
E-mail	**56.5**	5.03	75.1	30.2	**98.0**	**97.3**	3.0	**42.9**	9.3	**7.3**
Δ	**22.7**	0.15	0.9	1.2	**43.3**	**19,0**	0.1	**20.0**	4.9	**15.6**
Significant	**yes***	yes	no	no	**yes**	**yes**	no	**yes**	yes	**yes**
Effect size	**medium****	none	none	none	**large**	**medium**	none	**small**	none	**small**

[4] Significance testing shows how likely it is that the observed effect is random. It does not show how large and important it is. Since large samples can make very small effects visible, it is becoming increasingly common to report not only traditional significance, but also effect size (APA 2010: 33, Perry 2005: 224).

**yes* means p≤0.05 (in fact p is smaller than 0.001 in all the cases), *no* — p>0.05
***large* means h>0.80, *medium* — h>0.50, *small* — h>0.10, *none* — h≤0.10
Welch two-sample t-test (two-sided) applied for MLU and MLW; two-sample proportion test, for all the other cases. Effect size calculated as Cohen's d for MLU and MLW and as Cohen's h (arcsine transformation) in all the other cases.

Conclusions

Since five parameters appeared to be truly different for e-mail and chat, we can give a positive answer to the main research question: *yes, the communication channel does influence the language*.

The sentences are shorter in chat, due to the higher speed of communication: since an immediate answer is expected, people try to be quick rather than elaborate, and do not waste much time on editing and improving their texts (especially given that chat is not the best place to do that). Interestingly, that does not affect word lengths: the pressure is probably not strong enough to make that happen.

The lack of sentence end marks and capital letters occurs for two reasons. First of all, the need for speed leads to a weakening of the norm. Second, the norm actually turns out to be unnecessary: if a turn contains only one sentence (and that is usually the case), then even without capitals and periods it is clear where the sentence begins and where it ends. It would be different in a letter or in a turn containing several sentences, but in these cases the norm is usually not ignored.

It might also be supposed that chat is considered to be a less formal channel where norm violations are more appropriate, but this claim is hard to prove or disprove using my data.

Emoticons are more numerous in chat, since in a synchronous mode it is more important to show a "polite smile" to an interlocutor. They also have a phatic function: you are showing that you are interested in what your partner is saying, and you might reply to a message with a single smiling emoticon if you do not have anything else to say. As one of the subjects of this study put it, when questioned, "...I also want to be polite, so in chat I actually use a smiley instead of a period :)".

It is interesting to compare my results to those reported in Baron 2004 for English instant messaging (IM). Baron has found 49 instances of emoticons in her corpus of 11 718 words (Baron 2004: 413), the ratio being 0.004. My ratios are 0.023 (405/17671) for chat, 0.007 (90/11260) for e-mail, and 0.017 counted together — that is, much higher. This seems unusual, since the participants in my study are older and more educated than in Baron's. Besides, Baron's sample includes female subjects, and women tend to use more emoticons than men (Baron 2004: 416). It is unlikely that Russian IM is so much richer in emoticons than English IM. One explanation might be the observer's paradox: Baron's subjects knew they were being recorded while chatting, and this might easily have influenced their speech production (they might have tried to avoid "informal" traits like emoticons). Alternatively, it is possible that emoticons were less popular when Baron's study was conducted[5].

[5] This possibility was suggested to me by Alexander Piperski.

For brackets, the difference is significant, but the effect size is too small. Most likely this means that there actually is a difference, but the sample is too small to show it.

As for the other parameters, we might be quite sure that there are no differences, or that they are really tiny. This means that the influence of the communication channel should not be overestimated.

Further development of this study might include analysis of more complex parameters and data from the other social groups: less educated, less language-aware and not including myself, the researcher. It would also be useful to compare another set of channels, but it would be difficult, if not impossible, to reduce the distinction between registers to this single parameter.

References

1. *American* Psychological Association. Publication manual of the American Psychological Association, 6th edition. 2010.
2. *Androutsopoulos J.* 2006. Introduction: Sociolinguistics and Computer-mediated Communication. Journal of Sociolinguistics, 10 (4): 419–438.
3. *Anni O.* 2008. Choosing Language in Internet Conversations Between Russians and Estonians. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2008" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2008"), 7 (14): 602–605.
4. *Baron N.* 2004. See You Online: Gender Issues in College Student Use of Instant Messaging. Journal of Language and Social Psychology, 23 : 397–423.
5. *Biber D.* 1994. An Analytical Framework for Register Studies. Sociolinguistic Perspectives on Register : 44–56.
6. *Biber D.* 1988. Variation Across Speech and Writing.
7. *Bogdanova A. V.* 2008. Spelling in Internet: Analysis of one Misspelling Case [Orfografiia v Internete: Analiz odnoi orfograficheskoi oshibki]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2008" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2008"), 7 (14): 50–56.
8. *Buras M. M., Krongauz M. A.* 2007. The Language of Corporation Websites: Game, Parody, Provocation [Iazyk Korporativnyh Saitov: Igra, Parodiia, Prokatsiia]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2007" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2007") : 109–114.
9. *Chafe W.* 1982. Integration and Involvement in Speaking, Writing, and Oral Literature. Spoken and Written Language: Exploring Orality and Literacy : 35–53.
10. *Collot M., Belmore N.* 1996. Electronic language: A new variety of English. Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives : 13–28.
11. *Hård af Segerstad Y.* 2002. Use and Adaptation of Written Language to the Conditions of Computer-Mediated Communication.

12. *Jensen B. U.* Syntactic Variables in Pupils' Writing: a Comparison between Handwritten and PC-written Texts, available at http://privat.hihm.no/buj/dokumenter/2007-06-04 %20Bergen%20artikkel%20BJ.doc.
13. *Ko K.-K.* 1996. Structural Characteristics of Computer-Mediated Language: A Comparative Analysis of InterChange Discourse. Electronic Journal of Communication, 6 (3).
14. *Ling R.,Baron N.* 2007. Text Messaging and IM: Linguistic Comparison of American College Data. Journal of Language and Social Psychology, 26 (291): 291–299.
15. *Liudovyk T. V.* 2010. SMS Analysis with the Imrovement of Quality Target [Analiz tekstov SMS-soobshchenii c tsel'iu povysheniia kachestva]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2010" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2010"), 9 (16): 313–317.
16. *Makarov M. L., Shkolovaia M. S.* 2006. Linguistic and Semiotic Aspects of the Identity Construction in Electronic Communication [Lingvisticheskie I Semioticheskie Aspekty Konstruirovaniia Identichnosti v Elektronnoi Kommunikatsii]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2006" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2006") : 364–369.
17. *Perry F. L.* 2005. Research in Applied Linguistics: Becoming a Discerning Consumer.
18. *Tagliamonte S., Denis D.* 2008. Linguistic Ruin? Lol! Instant Messaging and Teen Language. American Speech, 83 (1): 3–34.
19. *Yates S. J.* 1996. Oral and Written Linguistic Aspects of Computer Conferencing: A Corpus Based Study. Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives : 29–46.
20. *Zalizniak A. A., Mikaelian I. L.* 2006. Emaling as a Linguistic Object [Perepiska po Elektronnoi Pochte kak Lingvisticheskii Ob"ekt]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2006" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2006") : 157–162.
21. *Zanegina N. N.* 2009. I Didn't Say it: on Lituratives, Strikeout and Quasi-texts [Ia etogo ne govoril: O Liturativakh, Zacherkivaniiakh ili Mnimukh Tekstakh]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2009" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2009"), 8 (15):112–115.

СОВРЕМЕННЫЙ РОССИЙСКИЙ ПУБЛИЧНЫЙ ДИСКУРС: ВЕДУТ ЛИ ТЕХНОЛОГИЧЕСКИЕ НОВШЕСТВА К НОВЫМ ДИСКУРСИВНЫМ СТРАТЕГИЯМ ИЛИ НОВОМУ МИРОВОЗЗРЕНИЮ

М. Б. Бергельсон (mirabergelson@gmail.com)

Филологический факультет МГУ, Москва, Россия

В данном исследовании современный русский публичный дискурс рассматривается в его электронном варианте, что позволяет взглянуть на используемые в блогах дискурсивные стратегии с точки зрения отражаемых ими изменений в социокультурных и лингвопрагматических паттернах коммуникативного поведения.

Ключевые слова: публичный дискурс, блог, дискурсивные стратегии, паттерны, коммуникативное поведение

MODERN RUSSIAN PUBLIC DISCOURSE: DO CHANGES IN INFORMATION TECHNOLOGY LEAD TO NEW DISCOURSE STRATEGIES, OR TO NEW WORLDVIEW?

M. B. Bergel'son (mirabergelson@gmail.com)

Lomonosov Moscow State University, Moscow, Russian Federation

This study aims at looking into various formats of modern Russian-language internet communication in order to discover changes in sociocultural patterns and models of the discourse behavior that characterize values and norms of the contemporary Russian public life. Specifi c public discourse genres — high offi cials' internet blogs — are analyzed with a special emphasis on whether the public discourse represented in the modern electronic modes is different in the language used from that of the traditional offi cial discourse. This analysis should allow to better understand ideas and

beliefs prevailing in the Russian public opinion, to trace its changes and emerging linguistic patterns.

Key words: public discourse, blog, patterns, discourse patterns, linguistic patterns, communicative behavior

Project goals

This study aims at looking into various formats of modern Russian-language communication to discover sociocultural patterns and models of the discourse behavior that characterize values and norms of the contemporary Russian public life. Specific public discourse genres are singled out and analyzed with a special emphasis on the public discourse represented in the modern electronic modes. This analysis should allow to better understand ideas and beliefs prevailing in the Russian public opinion, to trace its changes and emerging linguistic patterns.

The project is a part of a broader comparative cross-cultural study of sociocultural foundations of discourse interpretation. I see discourse interpretation as one important instrument of analysis to discover values and norms prevailing in a given sociocultural group, which, in its turn, is essential for understanding and predicting any sound and sensible public behavior, including communication. This connection between discourse analysis revealing changes in cultural patterns that underlie public discourse, and well-grounded predictions of the changes in public discourse itself makes this project relevant for multidisciplinary studies of modern Russian-language communication.

Context

All linguistic models admit that language functions both as an informational and an interactional system. At the same time an overwhelming majority of mainstream linguistic models occupy themselves with various mechanisms of information processing and consider the interactive component as some sort of additional topping on the informational substance of a message. However, if we admit that communication is a goal-driven activity, then study of the ways we reach our goal in communication, of the ability of natural language speakers to communicate more than what is explicitly stated and be successful in their communication by maintaining interpersonal harmony and complying with sociocultural norms, lies in the heart of any comprehensive linguistic model.

As knowledge-based interaction communication can only be successful when the participants share culturally determined communicative competence acquired in the processes of primary socialization. Thus, cognitive schemas of sociocultural knowledge and competences are central to any studies of the so-called 'national communication styles'. Though the very term of a 'national communication style' is quite misleading and sometimes even considered obsolete, it is possible to avoid

the kind of false generalizations it implies by introducing a concept of specific communication contexts, or *discourse genres*. If *discourse events* are units (not necessarily elementary ones), to be used to describe communication process, then discourse genres will be relevant *types* of discourse events that will allow to discriminate between the communication contexts. Then generalizations of various kinds, including that of national communication styles, can be made in terms of specific discourse genres (DG) and allow comparisons between both various DGs within one languaculture (Agar 1994) and similar DGs in different cultures. Furthermore it allows to arrive at a classification and to relate DGs to the existing genres of modern Russian public communication (cf. with oral speech genres in Russian National Corpus www.russcorpora.ru). Discourse communities (Swales 1990), (Scollon and Scollon 2001) have their specific clusters of DGs and thus can be determined by the latter.

Hypothesis

Communication patterns that may emerge as different DGs are organized along interactional and information-handling dimensions. Both dimensions are regulated by means of special type of rules, namely *pragmatic principles*, which determine choice of linguistic form not on the basis of grammar or world-knowledge, but based on the fact that language in its communicative function is a form of sensible and goal-driven activity.

Information-handling dimension embraces various strategies ranging in scope from a clause to larger discourse units level. It deals with fore- and backgrounding and cognitive accessibility of information reflected in the information structure of a clause/sentence, and content organization beyond the sentence boundaries. One way to understand which ideas are highly topical for and prevalent in the Russian society is to study information-handling public discourse strategies.

Interactional dimension deals with 'politeness phenomena' (Brown and Levinson 1987) which involve presentation of self, distribution of talk, and Face Threatening Acts with numerous politeness strategies to mitigate them.

I assume that there is an underlying principle of *Pragmatic Control* that is responsible for various aspects of interaction between participants in discourse; for both linguistic politeness and its conscious and accepted absence. Pragmatic Control (PC) is a degree of the Speaker's assessment of her/his right to certain communicative behavior towards the Addressee. This right motivates the Speaker's decision to use politeness strategies and to choose among them. Politeness is but an instance of Pragmatic Control principle. Incidentally, the politeness strategies hierarchy is based on speakers' assessment of the degree of pragmatic control they possess in a current discourse event with a given addressee. In certain cases even highly face-threatening acts are performed without any mitigation. The way pragmatic control is expressed in various public discourse events and shared between various discourse participants sheds light on distribution of power and potential changes in contemporary Russians' worldview.

Research objectives

1. To analyze public discourse in terms of its content, message and targeted areas.
2. To arrive at a typology of public discourse genres based on the message, area and participants.
3. To analyze Russian public discourse from the interactional point of view: participants, way and degree of interaction, linguistic mechanisms of interaction.
4. To analyze Russian public discourse in electronic media as one highly interactional channel of communication and elicit changes in interactional strategies.

Research data and primary results

The crucial data for this project can be found in various instances of public electronic discourse — a new (at least for Russia), dynamic and highly interactive discourse genre. It is well represented in various blogs of Russian public figures, officials (see for example http://gosblogi.ru/opml.xml) and especially that of the champion of the public internet discourse — President Medvedev (http://blog.kremlin.ru, http://community.livejournal.com/blog_medvedev). These blogs demonstrate changes in different strategies pertaining to public discourse. The information-handling side of the electronic public discourse is well represented in the personal livejournals of the high officials (primary data is taken from the livejournals of governors, mayors, and vice-mayors) and is mostly related to a funny mix of formal and informal registers, where the formal register is abundant with the typical bureaucratic expressions and constructions. The informal register is characterized by use of interactional discourse markers addressed to the audience, borrowings from the oral speech and specific internet jargon, which is an absolute innovation in the public figures' written discourse. It is worth noting that the degree of informality is never close to that of the private persons in their blogs or livejournals. One nice difference is that the officials promote spelling and punctuation rules.

1. (1)[1]

Многие, абсолютно справедливо, возмущаются состоянием тротуаров. Проведенные в декабре конкурсы в районах определили подрядные организации по содержанию тротуаров в 2011 году. ***Увы****, многие из них не имеют опыта работы в городском хозяйстве.*

…….

С учетом предложений районных администраций, Депутатов Городской Думы, ГИБДД и читателей блога, был сформирован перечень объектов площадью свыше 1-го миллиона квадратных метров. ***Всем большое спасибо!***

В декабре, ознакомившись с опубликованным Бюджетом РФ, выяснилось,

[1] Underlined are official, specifically bureaucratic style constructions and expression, while informal, oral speech pieces are in **bold**.

что сумма 1 033 млн рублей сохранилась, но была поделена на 2 части. 344 млн. рублей на ремонт дорог, 689 млн. рублей на ремонт дворовых территорий. ***Вот и приходится 2/3 списка вырезать. Искренне жаль.*** *Несомненно, ремонтировать дворы надо, но за счет дорожного ремонта,* ***на мой взгляд****, не разумно.*

http://lipovich.livejournal.com/

1.(2).

Весь прошедший месяц был неактивен в ***инете****. Думал,* ***«опахивал» столицу****, формируя программу развития экономики региона на ближайшие 2011 и 2012 годы. Нужно привлечь значительные инвестиционные ресурсы, чтобы обеспечить серьезный рост бюджетных поступлений. Достаточно напряженный, но продуктивный месяц, пока говорить о намерениях не буду, увидите позже.*

……..

С Новым Годом! Думаю, что у всех моих ***френдов*** *и читателей все хорошо и праздники прошли весело.* ***Я же просто спал, читал и молчал, за год наговорился :)))***

Традиционно в конце года выступаю с отчетом перед общественностью: что сделано, что не получилось и почему, планы в наступающем году.

http://alexandr-jilkin.livejournal.com/

Both Medvedev's blogs are moderated for obscenities and off-topic content only. The Medvedev's livejournal blog in particular allows posts and free discussion (comments and new posts within a thread, starting a new thread, etc). One should yet see what the political and social implications of this, so unusual for Russian politics, enterprise will be. One may only hope that the desired outcomes of openness, transparency of decision-making, so much expected shift to the e-government, will be visible and will give tangible results. Along with this there are certain linguistic phenomena related to politeness and pragmatic control. Analyzing the data from the Dmitry Medvedev's bog at http://blog.kremlin.ru I am looking primarily at modes of address and degrees of informality.

In Russia, with its highly hierarchical, high-distance-power culture, vertical communication in public discourse (especially when addressing high officials) is extremely deferential and formal. On the other hand, normally electronic discourse in blogs is an example of the exactly opposite interactive behavior. It is probably no surprise that the President Medvedev's blogs give evidence of something in-between. And not in a mixed way — like working out rules for some 'intermediate level' of politeness — but in a split way. The examples of different posts from the kremlin.ru blog illustrating these tendencies are shown below. Some posts are quite deferential, with traditional greeting and leaving formulae — *(Глубоко)уважаемый Дмитрий Антольевич!* , *Спасибо за внимание, Жду ответа* (approx. 30%), while others are following the rules of a typical electronic discourse leaving out greetings and goodbyes, using conversational language and even slangish expressions — see (1), (3). Of course when the author of a comment addresses not the President, but

some other participant on the blog (they mark it by putting the nick of the Addressee at the beginning of their own comment) the style may be considered even more informal — see (4).

2. (1).
непейвода евгений, Красноярский край 1 декабря 2009 20:27
здрав дмит анат*. в милиции беспредел. хотят сажают, хотят сами стреляют. А прокуратура требует с них показателей. а суды? просто умора. сделают по закону , так прокурор оспаривает и себе галочку. все повязаны. адвокатов ни во что не ставят. в лесосибирске вообще полный беспредел. все менты коммерсанты. даже начальник ГОВД лесом занимается , а служба безопас его покрывает. осетриной торгуют. а простых под суд. в крае вообще творится нечто. хлопонин бизнесмен всех под себя подмял. народ в нищете. они жируют. малый бизнес закрывается. главное что молодежь уезжает, значит перспективы нет. вмешайтесь в проблему малых городов типа нашего. все федеральные органы продажны. а прокурор молодой себе карьеру делает. стряпает не существующие дела. ну и про себя не забывает. тем более он засланный казачок, не местный. живет в служебной квартире. на выходные в центр уезжает. бросается только общими словами. главный мент весь в торговле. неугодных убирает. под суд. у них одна проблема. лишь бы посадить чиновника. а между собой разберуться или поделят . для галочки. город у нас небольшой, но перспективный. увы, все в руках непонятных людей с москвы и красноярска. развития нет. нужно ваше вмешательство.*

2. (2).
Filinova, Московская область 1 декабря 2009 18:25
Глубокоуважаемый Дмитрий Анатольевич*, только благодаря обращению к Вам лично мне удалось получить ответ о гражданстве в РФ(многие госучреждения отписываются, при том очень даже изобретательно). Теперь новый вопрос опять к Вам, как юристу, по защите прав садоводческих некоммерческих товариществ…..*

2. (3).
Влад, Республика Саха (Якутия) 4 декабря 2009 11:48
Читаю комментарии и думаю: дурак начальник-горе для подчинённых (русская народная пословица). А вывод такой: дурак подчинённый, который, пытаясь избавиться от горя, идёт к начальнику.

2. (4).
Приятель, Санкт-Петербург 3 декабря 2009 00:32
Диме Рудакову (Калужская область, 02.12.2009, 13:32):
Дима, ну о каких налоговых льготах вы говорите! Не будут они этого делать. Наоборот, как мне сказали налоговые инспектора (кстати, в суде по поводу взыскания налогов), сейчас дана установка тянуть по полной

не только с бизнеса, но даже с обычных граждан.
Потому что федералы сейчас обеспечивают соцобязательства субъектов РФ, которые формально находятся в ведении этих самых субъектов, а на практике — уже давно финансируются Москвой. В условиях кризиса нагрузка в этой части на Ф. Бюджет всё больше. Вот и собирают с мира по нитке. А вы говорите о налоговых льготах. Если надо -они последнее с бизнеса и с нас снимут, лишь бы потом эти бабки в рамках социалки частично раскидать, чтобы народ на баррикады не пошел, а частично освоить. Надо понимать, у ребят в Москве думалка работает по-советски: в одном месте взять, в другое место отдать, а если в этом месте разворуют — ну и х... с ним. Главное — создать видимость, что они нас поддерживают. Это чтобы мы не замечали, что всё разваливается.
http://blog.kremlin.ru/post/50?page=2

Even these few examples demonstrate how split is Russian society both in terms of public issues and linguistic behavior. To work out parameters and find models that will allow to describe various vectors of potential changes in the ways relevant public issues are raised and discussed are among the main goals of this study. Still, as this project has been taken up as a longitudinal study, a year later the data from the presidential blog (again, I am looking only at the comments to the Medvedev's posts) shows more serious level of discussion.

About 50% of 121 comments to the post of January 17, 2011 on corruption (http://blog.kremlin.ru/post/136?) don't address Medvedev. They are addressed to the multiple Addressee — the community. Sometimes it is done explicitly (*Добрый вечер всем!*). Another half addresses President himself. Just a few of these greetings (3) use a highly deferential form (*Глубокоуважаемый Дмитрий Анатольевич!*), others are mostly low Power, neutral Distance greetings (Уважаемый *Дмитрий Анатольевич!, Дмитрий Анатольевич*,). A few others will be low Power, wide Distance greetings (*Господин Президент!*).

Leaving formulae are skipped, which is typical of electronic discourse. Emotionality is much higher than accepted in the traditional public discourse, but evidently more in compliance with acceptable ways of expressing anger, frustration and irony than a year ago. There is significantly less direct complaints, which brings discussion to a more professional level. Specific cases of corruption are brought in with names and places, but more as examples and arguments to the case.

To sum it up, there is evidence for special markers of a specific discourse genre being developed in front of our eyes. This, in its turn, allows to single out parameters that may be used to define and describe a given discourse genre — public electronic discourse: modes of address, presence or absence of leaving formulae, presence or absence of persuasive type of discourse, level of formality, interactional markers in addressing the community, the 'normality' of discourse, and adherence to the spelling and punctuation rules.

References

1. *Abramova A. A.* 2005. Linguistic Characteristics of Electronic Communication [Lingvisticheskie Osobennosti Elektronnogo Obshcheniia].
2. *Agar M.* 1994. Language Shock: Understanding the Culture of Conversation.
3. *Bakhtin M. M.* 1953.The Problem of Speech Genres [Problema Rechevykh zhanrov]. Estetika Slovesnogo Tvorchestva : 237–280.
4. *Bergel'son M. B.* 2002. Linguistic Aspects of Virtual Communication [Iazykovye Aspekty Virtual'noi Kommunikatsii]. Vestnik MGU. Ser.19. Lingvistika I Mezhkul'turnaia Kommunikatsiia.
5. *Bergel'son M.* 2011. Russian Cultural Values and Workplace Communication Patterns. Intercultural Communication: A Reader.
6. *Palmer G. B.* 1996. Toward a Theory of Cultural Linguistics.
7. *Ratmayr R.* 1998. Hoeflichkeit als kulturspezifisches Konzept. Russisch im Vergleich :174–182.
8. *Riazantseva T. N.* 2007. Some Peculiarities of the Realization of Communicative Principles and Strategies in the face of Computer Media Communication [Nekotorye Osobennosti Realizatsii Kommunikativnykh Printsipov I Strategii v Usloviiakh Komp'iuterno-obuslovlennogo Obshcheniia]. Vestnik MGU. Ser. 19, Lingvistika I Mezhkul'turnaia Kommunikatsiia :202–211.
9. *Scollon R., Scollon S. B. K.* 2001.Intercultural Communication : a Discourse Approach.
10. *Swales J.* 1990. Genre Analysis: English in Academic and Research Settings.
11. *http://alexandr-jilkin.livejournal.com/*
12. *http://blog.kremlin.ru*
13. *http://lipovich.livejournal.com/*

ИНСТРУМЕНТЫ КОНТРОЛЯ КАЧЕСТВА ДАННЫХ В ПРОЕКТЕ ОТКРЫТЫЙ КОРПУС

В. Бочаров (bocharov@opencorpora.org)
С. Бичинева (bichineva@opencorpora.org)
Д. Грановский (grand@opencorpora.org)
Н. Остапук (nataxan90@gmail.com)
М. Степанова (mariarusia@gmail.com)
OpenCorpora

Ключевые слова: русский корпус, аннотирование, комментарии.

QUALITY ASSURANCE TOOLS IN THE OPENCORPORA PROJECT

V. Bocharov (bocharov@opencorpora.org)
S. Bichineva (bichineva@opencorpora.org)
D. Granovskii (grand@opencorpora.org)
N. Ostapuk (nataxan90@gmail.com)
M. Stepanova (mariarusia@gmail.com)

OpenCorpora is a project that aims at creating an annotated corpus of Russian texts, which will be fully accessible to researchers, the annotation being crowd-sourced. The article deals with annotation quality assurance tools.

Key words: corpus, Russian corpus, annotation, assurance tools.

Introduction

One would think that finding a corpus of Russian texts, including annotated one, could not be a problem at the moment. Moreover, some of the corpora are accessible on the Internet (see survey in [1]), which we think considerably increases their value for the linguistic community. To the best of our judgment, accessibility on the Internet usually means that there is an interface that may be used to submit parameterized search queries to the corpus. Of course, this makes possible various theoretical linguistic research, which requires analysis of word usage, frequency etc. However, this is not sufficient for the corpus to be used as a resource for machine learning or testing applications such as morphological parsers or disambiguation systems, because

in these cases we need the annotated texts themselves rather than the search results. In principle it is possible to obtain the annotation from the existing corpora, but one faces a lot of trouble of either technical, administrative or proprietary kind.

This state of affairs motivated the OpenCorpora project, intending to create an annotated corpus of Russian texts. The content of the corpus will be accessible to everyone under a free license, the annotation being crowd-sourced. This has so far proved to be a good editing model in a number of projects, the best known of them, perhaps, Wikipedia [2]. It has been recently demonstrated that crowd-sourcing is a suitable method for obtaining linguistic data and «the quality is comparable to controlled laboratory experiments, and in some cases superior» [3] (another survey on crowd-sourcing in linguistics is provided in [4]). In OpenCorpora project we are working on crowd-sourcing linguistic annotation. In the near future the primary issue is the morphological annotation, the syntactic and semantic ones are to follow. Our goal is to collect high-quality manual annotation which will then be used to train disambiguation tools and other kinds of software.

OpenCorpora will include texts obtained from sources where copying and redistribution is legally allowed, i.e. texts with CC-BY-SA compatible license or public domain texts. CC-BY-SA compatible sources are: Wikimedia projects (Wikipedia and WikiNews [5]), www.chaskor.ru news agency. Many texts of Russian classic literature are in public domain and are available from WikiSource [6].

Crowdsourcing being the main method of aggregating materials, it is essential to make the quality of the results a priority since it is impossible to know users' qualification in advance. Moreover, quality assurance should be as automated as possible. Drawing experts into the annotation verification is unacceptable (due to the high cost of human labor) although it may be necessary in certain cases. The main QA tools of the morphological annotation are a model of grammatical labels compatibility and morphological dictionary. Both instruments are used to describe acceptable combinations of grammatical labels. Whenever an unacceptable input combination occurs, the software produces an error message. In OpenCorpora we use warnings as error reports, which do not block user's activity. The aim of such warnings is to draw user's attention to the potentially erroneous input. It is the user who has the final say.

The article will further deal with the OpenCorpora project itself, its morphological dictionary structure and compatibility models for grammatical labels.

About the OpenCorpora project

OpenCorpora includes a linguistically annotated corpus of Russian texts freely distributable under CC-BY-SA [7] licence and software to annotate it.

Under linguistic annotation we understand the following:

- segmentation of a character sequence into word forms, sentences and paragraphs
- morphology: specifying the part of speech and grammatical features of word forms, morphological disambiguation
- local syntax:
  - specifying the borders of multiword entities (analytical verbal structures, compound comparatives and superlatives, compound numerals,

compound conjunctions, prepositions and particles, adverbial, predicative and parenthetical clauses, names of people and objects, dates) and their grammatical features
  - combining word forms into syntactic groups (nominal, prepositional, verbal, adjectival groups) and marking dependencies within groups
- lexical semantics: specifying a particular word sense for a word form

Division into paragraphs is taken from the source and is intended mainly to facilitate adding new texts, though it well may be useful for a researcher, too. Division of a paragraph into sentences is done automatically and checked by users.

A sentence is divided into tokens also automatically with a manual check to follow. As a token we regard a minimal meaningful symbol sequence without spaces. Any token may be either a word form (that is present in the dictionary) or not. The latter case applies to e. g. punctuation marks, web addresses, formulae (chemical, mathematical etc.) and other character combinations not kept in the dictionary, for example, due to their infinite number.

The unit of morphological annotation is a token. The annotation of a token consists of one or several (in the case of homonymy) interpretations. Each interpretation must include the token class (present in the dictionary or not). For dictionary tokens it also includes:

- lemma ID from the dictionary
- part of speech
- a set of values of the obligatory grammatical categories (e. g. number for nouns)
- a set of labels marking the features of the particular word form used in the text (e. g. "misprint", "verb used impersonally")

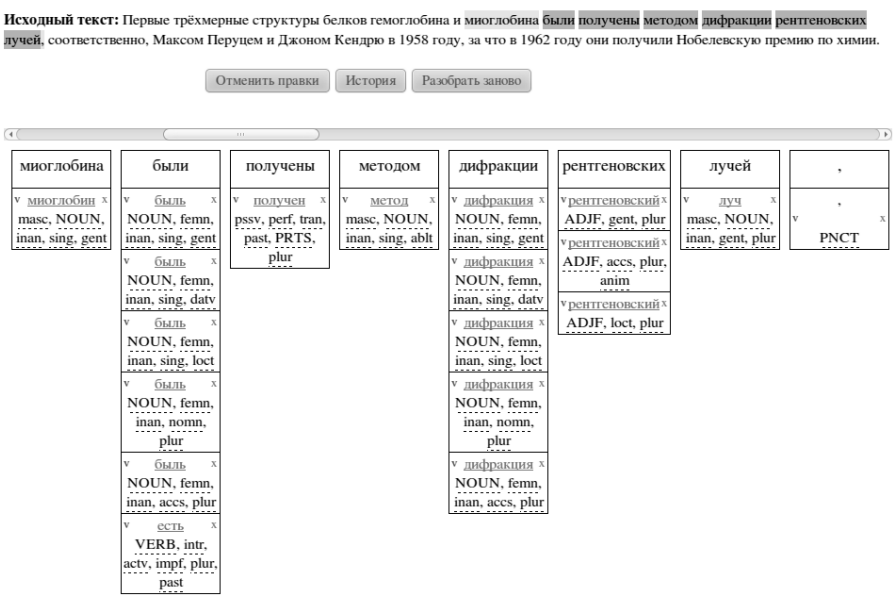


Fig. 1. Morphological markup editor web-based user interface

Figure 1 shows a sentence morphological annotation fragment in annotation editor interface. The whole sentence is in the upper part. The part annotation of which fits the width of the edit window is highlighted. The annotation itself is represented in columns. Each column includes all the versions of morphological analysis of a word form. The word underlined is a hyperlink to the morphological dictionary. The "v" button in the left upper part is used to mark an option as the correct one and all the others as incorrect. The "x" button marks the selected option as incorrect.

The same information is available in the XML format. For the word form "были" it is as follows:

```
<tfr t="были">
 <v>
  <l id="4342" t="быль">
   <g v="NOUN"/> <g v="femn"/> <g v="inan"/> <g v="sing"/> <g v="gent"/>
  </l>
 </v>
 <v>
  <l id="4342" t="быль">
   <g v="NOUN"/> <g v="femn"/> <g v="inan"/> <g v="sing"/> <g v="datv"/>
  </l>
 </v>
 <v>
  <l id="4342" t="быль">
   <g v="NOUN"/> <g v="femn"/> <g v="inan"/> <g v="sing"/> <g v="loct"/>
  </l>
 </v>
 <v>
  <l id="4342" t="быль">
   <g v="NOUN"/> <g v="femn"/> <g v="inan"/> <g v="nomn"/> <g v="plur"/>
  </l>
 </v>
 <v>
  <l id="4342" t="быль">
   <g v="NOUN"/> <g v="femn"/> <g v="inan"/> <g v="accs"/> <g v="plur"/>
  </l>
 </v>
 <v>
  <l id="52243" t="есть">
   <g v="VERB"/> <g v="intr"/> <g v="actv"/> <g v="impf"/> <g v="plur"/>
   <g v="past"/>
  </l>
 </v>
</tfr>
```

The software being developed within the OpenCorpora project is designed as a set of web applications, i.e. it is run on a web server and provides access to its functions via a web browser. The OpenCorpora web server implements the following functions:

- storing annotated texts and their edit history
- storing the dictionary and its edit history

- a web interface for annotation and dictionary editing
- quality assurance software

The project is currently at the stage of developing the software for graphematic and morphological levels of annotation. The current version is available at http://opencorpora.org. At the moment the data on the website is for demonstrating the software functionality and for debugging.

Morphological dictionary

OpenCorpora uses AOT's [8] morphological dictionary, customized for annotation quality assurance tasks. The dictionary is closely integrated with the corpus itself by numerical descriptors of lemmata — the components of word forms morphological interpretation in the annotations. This integration of text morphological annotation with the dictionary enables to solve the following problems:

- misprint checking at the stage of adding text to the corpus: trying to add a text with word forms not found in the dictionary will result in a warning
- morphological features acceptability checking: specifying in annotation a combination of grammatical labels, not stated in the description of the said lemma in the dictionary will result in a warning as well
- possibility to change lemma interpretation throughout the whole corpus: if lemma description in the dictionary is changed, all its entries in the corpus will be marked as requiring revision
- possibility to add lexical semantic information without changing annotation structure: one can divide certain lemmata into several units at the dictionary level thus stating that these units reflect different meanings of the lemma. The user who edits annotation can choose one of the lemma meanings for each its occurence in the corpus.

The unit of the morphological dictionary is a lemma. A lemma has the following properties:

- part of speech
- values of grammatical categories obligatory for the lemma of the given part of speech
- optional labels
- list of word forms
- list of links with other lemmata

The following properties are specified for word forms:

- textual representation of the word form
- values of grammatical categories obligatory for the word form of the given part of speech
- optional labels

The list of all possible grammatical labels is closed. Grammatical labels compatibility restrictions are described in the framework of compatibility model not designed to be changed frequently.

Adaptation of AOT morphological dictionary

Since the task of annotation correctness assurance is solved by means of the morphological dictionary, we needed to bring the dictionary to a form in which the correctness of dictionary data could be also verified automatically. The main changes followed two directions: paradigm grids unification and reduction of homonymic lemmata during automated morphological parsing. Besides that, we added to the dictionary a possibility to establish links between lemmata. Despite the fact that the format of dictionary representation has changed significantly, all the data presented in AOT was included to the OpenCorpora dictionary in one form or another to retain the possibility to review any decision without relaunching the procedure of dictionary transformation.

Paradigm grids unification means that lemmata of each part of speech have a set of acceptable forms and grammatical features. Acceptable grammatical features are described within the model of grammatical categories compatibility, the list of acceptable forms being generated on its basis. Some lemmata in the AOT dictionary contain descriptions of forms of several parts of speech[1] at the same time, e. g. verbs (infinitive, finite verb form, participles and gerund) and adjectives (the full and the short forms). Such lemmata were divided into several separate lemmata with links between them, the latter enabling us to restore if needed the fact that in the source dictionary it was one lemma.

Some parts of speech from the AOT dictionary were merged since they had the same sets of word forms and grammatical categories. Thus ordinal numerals and pronominal adjectives were joined to full adjectives as classes. Pronominal adverbs were added as a class of adverbs.

Reduction of homonymic lemmata during automated dictionary-based morphological text parsing was done to simplify morphological disambiguation. We tried either to eliminate the ambiguity at all whenever it was possible without information loss or to pass from lemma ambiguity to word form ambiguity within one lemma.

For instance, the word "микроб" may be either animate or inanimate. In the AOT dictionary such cases are described by means of two lemmata, which differ only in the form of accusative case. While transforming the dictionary, we merged such lemmata adding two alternative forms of accusative case.

Another example is the full adjectives, whose forms coincide with forms of full participles. We also united such lemmata, adding to the participles a label, stating that this lemma may be used as an adjective. In the course of morphological disambiguation in the text the user should retain or remove this label instead of choosing either

[1] Here we use term «part of speech» the same way it is used in AOT project (see http://www.aot.ru/docs/rusmorph.html)

full adjective or full participle. Since distinguishing between adjective and participle needs better qualification in the field of Russian grammar than distinguishing gender, case and number, such description form is more preferable, because anyone who is not enough sure about the first question can put the correct values of gender, case and number leaving the choice between adjective and participle to someone more competent. If adjectives and participles were different lemmata, the choice between them would have to be made before the choice of gender, case and number.

At the moment the possibility to establish links between lemmata is used only for joining the lemmata that have been divided into several parts, but it also can be used to describe other phenomena, e. g. different kinds of derivational relations, whenever we need them.

The changes described above are in fact technical changes of format without changing of content.

Compatibility model for grammatical labels

The compatibility model describes a set of possible grammatical labels, their applicability to the description of different objects and their co-occurrence possibility. Possible objects of description are the word forms in the annotation, lemmata and forms in the dictionary . Some grammatical labels can describe any of possible objects, e. g. the labels masc/femn/neut (masculine, feminine, neuter) are applicable to word forms, the lemmata of nouns and forms of adjectives. Uimp (impersonal use of a personal verb) is an example of a label applicable only to the word forms (i. e. only to the text annotation in the corpus rather than description of lemmata and forms in the dictionary).

Among grammatical labels in use we specify the following types:

- labels standing for the parts of speech (NOUN — noun, ADJF — full adjective, VERB — finite verb form, ...)
- labels standing for the grammatical categories (CAse — case, NMbr — number, GNdr — gender, ...)
- labels standing for the values of grammatical categories (sing — singular, nomn — nominative, 1per — first person)
- labels marking a group of lemmata within one part of speech (Qual — qualitative adjective, Sgtm — singularia tantum, Geox — label for toponyms, ...)
- labels referring to word form features but not being the values of some grammatical category (Erro — misprint, Infr — colloquial form)

The full current list of grammatical labels is available at http://opencorpora.org/dict.php?act=gram.

Relations determining the compatibility of grammatical labels and their applicability to certain objects are defined on the set of grammatical labels. It features the following types of relations:

- relation "grammatical category — attribute" (POST (part of speech) — NOUN, CAse — nomn, ...). The label which is the first element in the relation

is a grammatical category. The second element in the relation is a value of this category.
- relation of obligatory application for lemmata (NOUN — GNdr, PRTF — TEns). This relation describes a set of classifying categories for lemmata of the given part of speech. Lemmata of a noun must have a gender, lemmata of a full participle must have a tense (full participle is a separate lemma for reasons stated above).
- relation of obligatory application for forms (NOUN — CAse, ADJF — GNdr). This relation describes inflectional categories. Nouns decline and adjectives change according to genders.
- relation of optional application for lemmata (NOUN — Pltm, VERB — Impr). The relation describes possible but not obligatory labels. Some nouns were assigned the label Pltm (pluralia tantum), some verbs were assigned the label Impr (impersonal).
- relation of optional application for forms (NOUN — nomn, VERB — 1per). This relation describes a possibility to apply a label to a word form of a given part of speech. Usually the relation is not defined manually but is determined automatically.
- relation of incompatibility (plur — masc). The relation implies that both labels can not be applied to the word form at the same time. The values of the same category are incompatible, for example, plur and sing are incompatible for they are both the values of the category NMbr.

A set of relations of optional application is automatically deduced from the relations of obligatory applicability: if a grammatical category is obligatory for a form or a lemma, all its values are applicable for the form or the lemma. For example, the relations of optional applicability for a form NOUN — sing and NOUN — plur are automatically deduced from the relation of obligatory applicability for a form NOUN — NMbr.

Relations between grammatical labels are described as a table which is an integral part of the dictionary. The full current version of this table is available at http://opencorpora.org/dict.php?act=gram_restr.

The compatibility model is used to validate the dictionary and the morphological annotation in the following way:
- grammatical labels must be applicable to the objects they describe. If there is no explicit relation of applicability (either added manually or inferred on the basis of other relations) then the grammatical label is not applicable
- grammatical labels must be compatible, i.e. any labels explicitly marked as incompatible must not occur together with the same object

The dictionary and annotation are validated automatically as the source data changes. The errors are entered in the table, which is available at http://www.opencorpora.org/dict.php?act=errata. These errors may be (or may be not) a reason for changing the dictionary or the annotation. In some cases it is reasonable not to fix the error but rather mark it as an exception to the common rules if it is really an exception.

Conclusion

The article presents the description of annotation quality assurance tools based on enumerating all possible word forms in the dictionary and on describing the grammatical labels compatibility rules as a set of typified relations.

References

1. *AOT.* Automatic Text Processing [Avtomaticheskaia Obrabotka Teksta]. http://www.aot.ru
2. *Creative* Commons — Attribution Share-Alike 3.0 Unported, available at: http://creativecommons.org/licenses/by-sa/3.0/
3. *Munro R., Bethard S., Kuperman V., Lai V. T., Melnick R., Potts C., Schnoebelen T, Reznikova T. I., Kopotev M. V.* 2005. Linguistic Annotations for Corpus of Russian Texts [Lingisticheski Annotirovannye Korpusa Russkogo Iazyka]. Natsional'nyi Korpus Russkogo Iazyka.
4. *Tily H.* 2010. Crowdsourcing and Language Studies: the New Generation of Linguistic Data. Proceedings of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk.
5. *Wang A., Hoang C. D. V., Kan M.-Y.* 2010. Perspectives on Crowdsourcing Annotations for Natural Language Processing, available at: http://dl.comp.nus.edu.sg/dspace/bitstream/1900.100/3266/1/TRB7-10.pdf
6. *Wikinews* — The Free News Source, available at: http://en.wikinews.org/
7. *Wikipedia* — The Free Encyclopedia, available at: http://en.wikipedia.org
8. *Wikisource* — The Free Library, available at: http://en.wikisource.org

О НЕКОТОРЫХ ЛЕКСИЧЕСКИХ «ОТКРЫТИЯХ» НА МАТЕРИАЛЕ РУССКОЙ СПОНТАННОЙ РЕЧИ (КОРПУСНОЕ ИССЛЕДОВАНИЕ)[1]

Н. В. Богданова (nvbogdanova_2005@mail.ru)

Н. А. Осьмак (nataly.androsova@gmail.com)

Филологический факультет Санкт-Петербургского государственного университета; Санкт-Петербург, Россия

В докладе предлагаются результаты первого опыта многоаспектного лексикографического описания материалов Звукового корпуса русского языка (в первую очередь — блока «Один речевой день»): представлены новые лексические единицы (слова и словосочетания), отсутствующие пока в толковых словарях русского языка, новые значения или коннотации старых слов, отмечаются особенности функционирования в устной речи ряда единиц.

Ключевые слова: спонтанная речь, лексикография, лексикографическое описание, корпус русского языка, «Один речевой день»

SOME LEXICAL “DISCOVERIES” ON THE MATERIAL OF RUSSIAN SPONTANEOUS SPEECH, A CORPUS STUDY

N. V. Bogdanova (nvbogdanova_2005@mail.ru)

N. A. Os´mak (nataly.androsova@gmail.com)

Faculty of Philology, Saint-Petersburg State University, Saint-Petersburg, Russian Federation

The article presents results of the first attempt at lexicographical description of Russian spontaneous speech. Analysis is based on the material from the Corpus of Spoken Russian "One Speech Day". New linguistic units (words

[1] Исследование выполнено при поддержке гранта РФФИ «Изучение зависимости речевых характеристик от условий коммуникации (корпусное исследование на материале повседневной русской речи)» (проект 10-06-00300).

and phrases) not represented in dictionaries yet, new meanings and definitions or connotations of “old” words are described along with the trends of use in everyday speech. It is shown that a new area of lexicography, which could be called “speech lexicography”, is emerging. Its overall principles have not been completely determined yet, although some of the directions can already be specifi ed: 1) creation of a dictionary of common Russian colloquial speech, which should refl ect linguistic units used in everyday speech; 2) creation of a dictionary of context-dependent expressive units; 3) creation of a dictionary of discursive units, and 4) collection of a corpus of aphetic and reduced units. The paper outlines controversial problems for each direction and provides linguistic examples.

Key words: spontaneous speech, lexicography, lexicographical description, Corpus of Spoken Russian, ¨One Speech Day¨.

Известно, что «как бы и о чем бы мы ни говорили, спонтанно порождая свою речь, в ней всегда так или иначе будут сочетаться элементы кодифицированной и устно-разговорной нормы» (*Лаптева* 1992: 153). Увидеть все это многообразие современной русской речи, с «борьбой языков, отражающей непрестанное столкновение и скрещивание <…> разнородных культур» (*Ларин* 1977: 177), можно, как представляется, только при корпусном подходе к сбору и анализу речевого материала. Именно такую возможность предоставляет Звуковой корпус русского языка (ЗКРЯ), особенно та его часть (блок «Один речевой день» — ОРД), которая фиксирует повседневную бытовую речь носителей русского языка в ее наиболее естественной форме. Методика 24-часовой записи, используемая при формировании данного блока ЗКРЯ, обеспечивает почти полную свободу говорящего от диктофона или от тех или иных коммуникативных заданий (чтение, пересказ, описание, рассказ на заданную тему), которые неизбежно усложняют его задачу и обеспечивают «на выходе» хоть и вполне спонтанную, но все же экспериментальную речь (см. подробнее о Звуковом корпусе: *Богданова и др.* 2008а,б).

Материал Звукового корпуса может быть использован в самых разных теоретических и прикладных аспектах изучения живой речи: от общего описания специфики устной спонтанной речи на всех уровнях и анализа разных типов внутриязыковой интерференции до пересмотра нормативных требований к построению устного монологического текста и изучения грамматики речи в русской филологической аудитории. Не последнее место среди этих возможностей занимает *создание лексикографического описания бытовой спонтанной звучащей речи*. Думается, что такое описание на материале ЗКРЯ может вестись в нескольких различных направлениях. Рассмотрим эти направления и первые результаты соответствующих исследований.

1. ***Словарь русской бытовой разговорной речи***. Возможности для создания такого словаря дает *электронная картотека E-Kar*, одно из программных средств, использованных при построении Звукового корпуса. Эта программа автоматически создает конкорданс по выбранным текстам и позволяет решать многообразные задачи классификации и описания языковых единиц,

обеспечивая, в частности, возможность новых содержательных форм интерпретации лингвистического материала[2]. Словник такого словаря, построенный по дифференциальному принципу, во многом повторяет словники обычных толковых словарей, кроме того, значения или особенности функционирования некоторых лексических единиц часто оказываются отличными от традиционных словарных дефиниций. Это мы и назвали первой группой небольших лексических «открытий», сделанных на звучащем речевом материале. Приведем ряд примеров.

Так, слово *хлеб*, согласно МАС (*Словарь русского языка* 1984: 601–602), имеет 4 значения:

1) *только ед. ч.* Пищевой продукт, выпекаемый из муки. *Ржаной* ***хлеб***.
2) *только ед. ч.* Зерно, из которого приготовляется мука для выпечки такого продукта. *Везти* ***хлеб*** *на элеваторы*.
3) *(мн.* хлеба́*), перен., разг.* Пища, пропитание. *Противен мне* ***хлеб*** *ваш.*
4) *перен., только ед. ч., разг. Средства к существованию. Добывать* ***хлеб*** *трудом.*

В материалах ОРД формы этого слова встретились 16 раз[3], причем пять из этих словоупотреблений (31,3 %) фактически не укладываются в предложенные словарем значения:

- *а он меня(:) не то пощекотал / не то что-то говорит / а я* ***хлеб*** *маслом на(:)… мазывала говорит / я говорит и дернулась*;
- *говорит / *П я хлеба … э я* ***хлеб*** *намазывала маслом / а он меня не то пощекотал / не то отвлек / говорит / и получилось / что говорит / я его пырнула ножом // *П в руку.*

Конечно, можно интерпретировать эти словоупотребления как явления *метонимии* (перенос названия с одного класса объектов или единичного объекта на другой класс или отдельный предмет, ассоциируемый с данным, в частности, по партитивности, т. е. перенос с целого на часть), но, с учетом большой распространенности в нашей речи именно такого значения слова *хлеб*, логичным представляется внести в список его значений в будущем словаре *русской бытовой разговорной речи* и семантему ‘кусок хлеба’ (на формулировке не настаиваем). Надо отметить, что и многие другие употребления в ОРД слова *хлеб*

---

[2] Реально сравнение значений слов из корпуса ОРД проводилось только с их описанием в Малом академическом словаре (*Словарь русского языка* 1984) (МАС). Выбор данного словаря объясняется тем, что в дальнейшем в исследовании использовалась компьютерная программа eLex, предназначенная для составления словарных статей на основе значений слов именно по МАС.

[3] По данным на январь 2011 г., из всего массива расшифровок объемом около 280 тыс. словоупотреблений. Наиболее частотная форма этого слова (Им. п.) встретилась 9 раз и занимает в частотном словнике ОРД место от 1643 до 1836, наряду с другими единицами, имеющими такую же частотность.

вынуждают задуматься, к какому из предлагаемых словарем значений их отнести, ср.:

- ***хлеб*** */ это тоже живое существо / # да // # а вы его ножом*;
- *ну что / думаю(?) / тебе булки не предложили / булки нет в доме / есть только* ***хлеб***;
- *не знаю /* ***хлеб*** *булка есть ?*
- *Коль / у меня есть пакет / я (...) пакеты как* ***хлеб*** *покупаю на работу // *П у меня в одном (...) пакет кармане / и второй*;
- *буду как в детстве сейчас / есть сосиски с* ***хлебом;***
- ***хлеба*** *надо тебе* ?
- *но он такой // *П назыв... он назывался / бабушкин* ***хлеб***;
- *не знаю // так называется // бабушкин* ***хлеб*** *почему-то.*

Последние два примера иллюстрируют еще и появление идиомы *бабушкин хлеб*, значение которой из контекста ОРД вывести довольно затруднительно[4], ср.:

- *— хлебушек есть // но он такой // *П назыв... он назывался /* ***бабушкин хлеб*** *//*
  — а / почему бабушкин ?
  — не знаю // так называется // ***бабушкин хлеб*** *почему-то //*
  *— *Н так называется? *П ёшкин /кот!*
  — ну вот // ну давай же шить !

Интересными оказались и употребления форм *девочки* и *девчонки* (только во мн. ч.) в значении ‘лицо одного возраста с говорящим’, например:

- *я жалею / что я не пошла с* ***девочками*** */* ***девочки*** *как раз ходили на концерт / а я чем-то была развлечена другим*;
- *я вот этой* (...) *по совету* ***девчонок*** */ они же опытные / всё знают // купила вчера капли анти... антибиотические // специальный антибиотик // ага* !
- *у меня вот приличные* ***девочки*** */ таких вещей не рассказывают*;
- *так нет / так у нас было то / давно ещё / в Испании / *В (э-э) там* ***девчонки*** *были / одна / *П такая () русичка бл... блондинистая / а другая татарочка чёрненькая // *В а они перепутали паспорта // и *Н говорит / вы говорит девочки / вы так это / нас так примитивно проверяете ?*
- *пф... где-то в час ночи наверное // ну что я пришла(:) (э-э-э)* ***девчонки****(:) ждали *Н / чай попить // *П вчера мы ведь безь день говели.*

МАС ограничивает возрастные рамки для данных лексем, определяя их как ‘ребенок или подросток женского пола’ (*Словарь русского языка* 1981: 375). Из приведенных контекстов отнюдь не следует, что *девчонка* или *девочка* — именно ребенок. С равной степенью вероятности речь может идти как о семилетнем ребенке, так и о пятидесятилетней женщине.

---

[4] В настоящей статье мы сознательно ограничиваемся постановкой вопросов и проблем лексикографического описания русской устной спонтанной речи, понимая, что поиск ответов на эти вопросы требует дополнительного, весьма серьезного и, может быть, пословного исследования. Такая работа по отдельным словам на материале ЗКРЯ уже начата, см.: *Богданова* 2010а, 2011а, в, г, д; *Осьмак* 2011а, *Хан* 2011.

Любопытно, что лексема *мальчик* в подобном значении (и только в ед. ч.) употребляется значительно реже, чем *девочки/девчонки*. Ее значение можно определить, скорее, как ‘лицо одного возраста с говорящим или младше’. Кроме того, если *девочки/девчонки* можно услышать из уст женщин-девушек-девочек, т. е. в их разговорах о себе, то *мальчиком* мужчину одного с собой возраста или младше назовут только женщины:

- *так вот у нас был* ***мальчик*** */ который работал / у нас все работали () в равных;*
- *там сели поехали / собралась / там как бы / компания там / четыре девчонки из Питера там /* ***мальчик*** *потом подсоединился из Москвы // ну вот как бы / тусили // а / прикольно // нормальные ребята такие.*

Мужчины в этой ситуации употребляют другие слова, ср.:

- *и мы же* ***мужики*** */ друг друга понимаем // # бежит к тебе / знает // *П так значит там лучше*;
- *— а есть полтаха* ? по полтахам разменять ?*
  — вон спроси / может ***парни*** *разменяют ?*
- ***пацаны-****то рубят же в этом деле.*

Думается, что в словаре *русской бытовой разговорной речи* все эти нюансы употребления лексических единиц должны найти свое отражение, даже если для их уточнения понадобится провести серию специальных психолингвистических экспериментов.

Интересным оказалось и функционирование в устной речи лексем *тётя* и *дядя*. При сравнении с однокоренными *тётка/тётенька* и *дядька/дяденька* выявляется отчетливая тенденция употребления *тёти* и *дяди* перед именем собственным:

- *там *П мама если бы сидела* ***тётя*** *Лена и моя бабушка Соня вот тогда бы @ *С @ всё исчезло через час;*
- **Н колготки тебе /* ***тётя*** *Таня подарила / я помню;*
- *и вот пришёл к маме / и вот просто понимаешь / вот банальщина // она говорит / С…* ***дядя*** *Сергей / типа выхлопайте / ну вот у нас / мужа-то как бы у неё нету / выхлопайте типа матрас // он говорит / *П хорошо // так я уж раз пойду / так давайте вы мне сразу / несколько вам выхлопаю.*

При этом совершенно необязательно, что речь идет о родственнике — скорее это форма именования близкого человека, возможно, друга семьи, старше говорящего.

Кроме этого, в ОРД встретился особый случай употребления лексемы *тетя* в составе устойчивого выражения *тетя Мотя*, также не зафиксированного в толковых словарях[5]:

- *и они вот задрали эту цену // а* ***тётя Мотя*** *тут / они сейчас критикуют / вон тут в подъезде / вон встал утром / вон президент сказал / чтоб на музеи не повышали*;

[5] См. о подобных сочетаниях слов *дядя* и *тетя*: *Богданова* 2010а, 2011а.

Среди контекстов лексемы *девка*, кроме указанных в МАС значений возраста и семейного положения (‘незамужняя’) (*Словарь русского языка* 1981: 375), возникают новые различно окрашенные (негативная и положительная) коннотации, практически не отраженные в словаре[6]:

- *приводит / просто какую-то* (...) ***девку*** *// какую-то такую гужбанскую какую-то пэтэушницу*;
- *ну вот / и* ***девка*** *такая а-а-а / типа дерьмо*;
- *ну ладно //* ***девка*** *такая / о ! ништяк !*

В лексико-семантической группе слов — наиболее общих наименований лиц мужского или женского пола — отдельного рассмотрения в аспекте их коннотаций и функционирования, не отраженных в существующих толковых словарях, заслуживают еще многие, отнесенные нами к лексическим «открытиям» спонтанной речи[7], ср.:

- *я говорю конечно / если она с ребёнком не умеет разговаривать / что это /* ***бабулька****-то* ?
- *там сидят* ***бабуськи*** *такие / на остановке // у них глаза такие по полтиннику // ну представляешь / они сидят / ждут автобуса;*
- *есть тут жилые дома вообще ? (э-э) дак (э) никто не знает // а потом () какой-то вот* ***мужичок*** *мне указал / что вот вам туда надо;*
- *ну как всегда у неё // достаёт классный телефон / тут ни с того ни с сего / начинает мне говорить / вот этот* ***мужчинка*** */ я уже не раз от неё такое слышала / она вот () рассказывала // да?*
- *ну там такие* ***бабцы*** */ конечно призы... / я в шоке // откуда у них грудь* ? *неужели у них всех силикон* ? *я не понимаю*;
- *ну вот / ну тем не менее // а* ***дедун*** *этот всё время / этот хохол-то (э-э) значит Майкл / Михаил по-моему его звали.*
- *пусть* ***молодой человек*** *от... отдохнет от нас / да ? Настя / а куда вот эти ? что вот этими штуками делают / не знаешь ?*

Нашлись такого же рода небольшие «открытия» и в других лексико-грамматических классах слов. Так, в словарной статье МАС на глагол *учесть/учитывать* (*Словарь русского языка* 1984: 543) не нашлось места конструкции *учитывая что*, достаточно широко распространенной и функционирующей в устной (не деловой, а повседневной, бытовой!) речи в значении союза причины, синонимичного *потому что*, *поскольку*, *так как*:

- *Алексей / вы извините ради бога / но / мне Максим сказал что вы еще пока здесь / в Питере // *П у меня один вопросик / буквально минутку / я думаю займет // *П у нас тут вот компьютер / который внизу у аналитиков // *П у нас сейчас просто оперативная очень работа / нам через пять дней надо группу выслать на заказ // *П вот / мы обрабатываем информацию /*

---

[6] См. подробнее о данной лексеме в ОРД: *Осьмак* 2011а.

[7] См. о них подробнее: *Осьмак*, *Чен* 2010; *Осьмак* 2011б.

*и он у нас конечно *Н зверски(?) тормозит // *П вот* ***учитывая (…) что (…)*** *Марина (…) аналитик (…) загружается(?) // что / может / ничего *Н // *П по крайней мере *Ш*;

- *на самом деле / я бы очень хотела себя послушать // *П я думаю / что я на работе скину (…) на компьютер // *П а потом послушаю как я говорю // # а скинь у нас тоже / *П интересно знаешь посмеяться // ну ()* ***учитывая / что*** *там будут записаны / почти сутки // *П там / (…) длинно всё очень // # знаешь это к кому надо ? Вике в комнату // там очень интересно для исследования // @ *С*;
- *но / туда ты идешь спокойно еще / вроде не замечая дороги / но назад // это ужасно //* ***учитывая что*** *приходится обычно тащить мне две корзины / так как муж тяжести носить не может после операции* (САТ)[8]*;*
- *хотя цирк я не очень-то и люблю // но /* ***учитывая*** */* ***что*** *у меня маленький ребенок шести лет естественно мы туда пошли* (САТ).

В следующих контекстах из ОРД можно видеть не учтенный словарями повторяющийся разделительный (не уступительный, присоединяющий придаточное предложение! — см. *Словарь русского языка* 1984: 622) союз *хоть.., хоть*:

- *и вот блин я как вспомню / у меня постоянно всё в душе сжимается // понимаешь как мы едем блин / *П с Ириной / всё вроде нормально // я нажимаю на тормоза / а машина не останавливается // *П вот как *Н вспоминаю я вообще так (…) просто у меня сердце сжимается каждый раз // *П так беспонтово / по тупому // причём аккуратно / главное мог бы куда-нибудь там /* ***хоть*** *чего-то /* ***хоть*** *ручник дёрни /* ***хоть*** *скорость мог включить просто / вторую / третью / а машина бы стала тормозить // *С но я / так нахально просто в общем / *П впилился без скрипа тормозов*;
- *значит мы одну штуку покупаем / *П и обкладываем // её можно* ***хоть*** */ это самое /* ***хоть*** *под металл / *П* ***хоть*** *вот такого цвета // @ *Н;*
- *а я вот / например / с Барчуковой разговариваю / мне* ***хоть*** *в лоб /* ***хоть*** *по лбу // (@*Н …) я понимаю / мне эти шины там / эти капоты / ну зам… замки на капоты / мне как вот (…) ну вообще.*

В последнем примере можно видеть вариацию известного просторечного выражения *что в лоб, что по лбу* (‘все равно, одинаково’ — см. *Словарь русского языка* 1982: 194): разговорная речь, как и следует ожидать, увеличивает количество вариантов выражения того или иного значения, свойственных уже кодифицированному языку.

Было выявлено в ОРД и не зафиксированное в словарях значение ‘мимо’ для предлога *через*[9]:

---

[8] Примеры, маркированные таким образом — из второй части ЗКРЯ — сбалансированной аннотированной текстотеки (САТ); см. о ней подробнее: *Богданова и др.* 2008; *Богданова* 2010б.

[9] См. об этом: *Сковпень* 2011.

- *я мимо этих уродов хожу вот так вот // я их тоже очень не люблю // но все время приходится* ***через них*** *ходить*;
- *а где это ? вам такая блин / там-то там-то там-то / у него пи…дец был // за пустырь там блин / (...) через кладбище /* ***через собак*** */ которые тебя покусают*.

Само же слово *мимо* выявляет в ОРД возможность употребления в функции предиката (не предлога и не наречия, что определяет для него МАС, — см. *Словарь русского языка* 1982: 271):

- *ну так это* ***мимо*** *// *П это совершенно () а ! на третьем этаже ? ну там я не знаю // *П там я не знаю.*

Подобное употребление близко к зафиксированному в «Большом словаре русской разговорной экспрессивной речи» В. В. Химика выражению *мимо денег (кассы)*: «*шутл., жарг. молод.* Опрометчиво, неудачно, невпопад, некстати» (*Химик* 2004: 139), однако его можно интерпретировать лишь как вариант этого последнего, не нашедший пока словарной фиксации.

2. ***Словарь контекстных экспрессем русской разговорной речи***. Это направление возможного лексикографического описания спонтанной речи во многом пересекается с предыдущим, о чем свидетельствуют уже вышеприведенные примеры на *бабец/бабца*, *девка*, *мимо* и некоторые другие, однако его потенциальный словник практически полностью выходит за рамки кодифицированного языка и требует не только тщательной пословной фиксации, но и большой работы (вплоть до серии психолингвистических экспериментов) по определению значений лексических единиц, их стилистической окраски и степени идиоматичности. Ограничимся здесь лишь некоторыми фрагментами ОРД, размеченными в интересах этого будущего словаря:

- *здесь раньше / () ацтекский календарь висел // *П тоже хорошо // *П но потом возобладал более научный подход / *Н карты звёздного неба // @ а потом кто-то () кто-то вернулся из Мексики / *П сказал что это всё* ***фуфло*** *// *С*;
- **Н # вот / и я всё время прихожу к нему в гости / он начинает мне рассказывать про () там (...) теории / *П которые во времена Третьего рейха / *П бытовали // то есть теория полой Земли там // или чего-нибудь там / или какие-то специальные отделы там // *П то исследование / *П ведических культур // ну в общем что-нибудь такое // такой* ***загруз*** *// *С*;
- *слушай / этот контакт /* ***глючит*** */* ***глючит*** *просто*;
- *а(:) / это (...) москвичи ко мне приезжали / я почему спрашивала что тебя до... дома бу... будешь ты или нет // *П тут вот / ночью я их* ***вписывала*** *// и они накупили / *П какого-то конопляного пива*;
- *ну то есть имеет смысл / да / как бы не... @ просто я ... @ не ... не* ***гонево*** */ тебе(:) / больше не кажется / что это всё обман*;
- **Н # это* ***ориентировано на западный (:) манер*** */ понимаешь это большинство* ***ориентировано на западный манер*** *// *В когда (...) люди /*

сотрудники в коллективе / перестают быть хорошими приятелями / а становятся / ***нездоровыми конкурентами***;

- *понимаешь / я () могу привести тебе массу примеров // у нас например работали на складе / вот (э) ещё до того когда у нас вот стала да такая компания //* ***откуда*** *ещё / вот знаешь* ***ноги растут***;
- *(э-э) у нас работали в компании там / например грузчики на складе // ну это основная как бы да / всегда* ***гуляющая штатная единица*** */ и всегда требующая какого-то денег*;
- *не / ну то что вот Ленка *Н ушла ушла / *П хотя на самом деле она сделала очень много // человек-то на самом деле / по первости я просто… / я* ***шизею*** *как она работала *П столько сделать /* ***перелопатить*** */* ***мама не горюй*** *грубо говоря*;
- *они (э) дела… стараются делать для всех / то есть / невзирая на личности / да ? @ угу @ просто / что видит человека / вот* ***из него*** *можно* ***доить денежку*** *// вот например торговый представитель / определённый //* ***с него*** *можно* ***тянуть денежку*** */ определённую // вот он сейчас… / его запустить @ угу @ /* ***выжать из него все соки*** */* ***получить с него максимум*** */ (а…) прибыли / и его можно* ***выкинуть***.

Из примеров видно, что чаще в устной речи встречаются не просто экспрессивные лексические единицы, а экспрессивные выражения разной степени идиоматичности. Последний контекст иллюстрирует даже целый синонимический ряд таких идиом, что создает прекрасную базу для лексикографического описания.

3. Великолепный материал предоставляет Звуковой корпус и для возможного ***словаря дискурсивных единиц***. Наблюдения показывают, что на текстовом, дискурсивном, уровне наша устная речь, вопреки весьма распространенным представлениям, крайне неэкономна и обнаруживает нерегулируемое расширение, приращение звуковой формы, без всякого приращения смысла. Это происходит в результате того, что говорящий, вынужденный в условиях временно́го дефицита решать сразу две задачи — обдумывать речь и собственно говорить, — вербализует (попросту озвучивает) весь процесс порождения текста, включая поиск слова, разного рода хезитации (колебания), самокоррекцию и прочие проявления своей «борьбы» с текстом, а зачастую и с самим собой[10]. В Звуковом корпусе примеров такого рода множество:

- ***а ну*** *тут* ***значит тут*** */* ***это самое*** *//* ***значит*** */ дед // с бабушкой сидят / беседуют / слушают радио // и обсуждают /* ***ох*** *конечные новости что по радио говорят* ***значит*** *на следующей картиночке /* ***э-э это самое*** */* ***значит*** *старичок / смотрит / на картинку и /* ***э-э*** *хочет тоже так же прокатиться //*

[10] Ср.: «Устная речь необратима — такова ее судьба. Однажды сказанное уже не взять назад, *не приращивая к нему нового* (курсив автора. — *Н. Б.*); «поправить» странным образом значит здесь «прибавить»» (*Барт* 1989: 541). Подробнее о принципе экономии в устной речи см.: *Богданова* 2011б.

*с / **с** горы // **так значит** / собрал / **значит** взял лыжи / шапку / **ну вот** / и **значит это самое** / собрался / поехал // приехал **значит** / в горы* (CAT);

- *у меня () подружка / считает / **(э)** что (...) Масяня / **это очень** / **(и-и)** как интеллектуальный питерский диалект // так вот именно тот язык / на котором в этих мультиках разговаривают / то он именно **вот** / какой-то питерский / **(и-и)** (...) **ну такой очень** / **ну** / вот именно так и говорят*;
- ***ой** / **ха-ха** я чего-то / **я чего-то** / **я чего-то** запомнила только конец // как они кота накормили / **это самое** // **он начал** / **это самое** / **э-э ну это** / **как его // э-э ну з…** / **ж…** / **ну** жареной свининой // **значит** / окунями // и он начал кататься валять по полу // **кататься и валяться по полу** // **ну вот ну** он **значит** был / грязный / рыжий / как они его **в общем это самое** нашли / подобрали / **так вроде бы*** (CAT);
- ***а вот (э-э) на**... **н**... **вот наше вот это вот (э-э) вот это вот** / **вот тут** / тут сложнее гораздо / **да** // потому что / **значит** / я **вот вот (э-э) вот эти** / **ну в принципе** / **значит** / **ну** / **п**... по моим / понятиям **значит** / **я** же не отличу **так скажем** / таджика от узбека **что называется** / **да да да** // да* ?

По большей части такое приращение формы происходит за счет единиц, которые на этом основании можно назвать дискурсивными (в известном смысле они же могут быть интерпретированы и как «слова-паразиты»). Явление это не новое, о нем много писали и пишут, но корпусный подход к сбору и систематизации речевого материала ставит задачу и открывает возможность пословного исследования таких единиц и даже лексикографического их описания. Так, анализ употребления дискурсивной (а также хезитационной и отчасти метакоммуникативной) конструкции *это самое* (в разных ее грамматических формах) выявил целый ряд ее функций, реализующихся в спонтанной речи:

1) *маркер поиска* (наиболее распространенная функция, допускающая дальнейшую систематизацию):

- *нам кстати / *П нам вот **эту самую** (...) стенку / где дверь висит / надо новый(?) рулон там где-то // *П красивые / но дорогие;*
- *да / @ Тимка смотрит крокодила Гену // @ сам пошёл ! @ Настя / за меня / за меня мы пьём **этот самый** / бокал // @ да*;

2) *маркер смены темы* в речевом фрагменте:

- *разрежь половину ягодкой @сейчас @ а то ему ... **это самое** / а половину можно бабушке отнести // @ оно полезное !*
- *вчера с матерью тоже когда разговаривали / *П **это самое** // *П она ну приедете / что делать // *П я говорю мать / но это же не маленькие дети*;

3) *маркер самокоррекции*:

- *яркая солнечная погода // говорить можно? так был ярк… ∫ **это самое** ∫ был ∫ июльский день / вот / небо было чистым / безоблачным / солнце ∫ светило ();*

4) *дискурсивный маркер* (может быть стартовым или финальным):

- *знаешь / **это(:) самое** / *Н сидят на месте / значит надо требовать / да / жалобы писать;*
- *как ты переводишь ? # нет / я не могу короче (...) **это самое***;

5) *маркер хезитации* без явной мотивации:

- *главное / *П (э-э) десятого были / *П и... и **это самое** / следующее слушание будет (...) пятнадцатого / это два дня вот выпадает / по **это самое** / *Н // *П ну.*

Своеобразной отрицательной формой данной конструкции является *не это самое*, имеющее, как показали наблюдения, свои особенности употребления[11], ср.:

- *нормально ? *П ничего там тебе / **не это самое** (...)*;
- *без фанатизма наверное ? здесь не не не / так // я тоже **не это самое**.*

Наконец, можно говорить о двух идиомах, включающих *это самое*, вообще не требующих заполнения «пустой позиции» и свойственных почти исключительно устной разговорной речи: *а не пошли бы вы в это самое?* (своеобразный эвфемизм) и *как это самое*:

- ***А не пошли бы вы** все **в это самое**?* [Ю. О. Домбровский. Факультет ненужных вещей, часть 1 (1978)][12];
- *между прочим у французов тоже простая кухня / просто () это нам теперь здесь её () подают **как это самое** / как будто пре-парте какое-то.*

Такому же детальному анализу в целях создания грамматики и словаря русской бытовой спонтанной речи могут быть подвергнуты и все другие дискурсивные единицы. Такая работа на материале ЗКРЯ уже начата, и первые результаты получены еще по двум конструкциям:

➢ ***(я) думаю (что)***[13]:

- *ну / я не знаю / в шесть или в семь / **я думаю что** / я одну оставлю*;
- *я Сашу / Са… / с… с… с… / Сашу знаю / вот / а **думаю что** там / ты говорит учти / что ты думай / что если он там узнает / что у него объёмный процесс / типа того / думай что он будет / лететь он будет / да / вот он говорю / да*;

➢ ***(я) не знаю***[14]:

---

[11] См.: *Богданова* 2011д.

[12] Данный пример — из Национального корпуса русского языка, в ОРД такой конструкции пока не встретилось. Подробнее о функциях выражения *это самое* в спонтанной речи см. *Богданова* 2011в.

[13] Подробнее см.: *Богданова* 2011г.

[14] Подробнее см.: *Хан* 2011.

- *в общем Кирилл / **я не знаю** / это просто / притча во языцех // я / вообще не знаю как с ним общаться уже*;
- *тут приехали пожарные // на самом деле наверное [нрзб.] вызвать какую-нибудь скорую / **не знаю** из психбольницы кого-нибудь* (CAT).

Специальные исследования ведутся в настоящее время на материале слов *вот* и *там* в спонтанной речи[15], на очереди и другие единицы этого класса.

4. Наконец, еще одно направление лексикографического описания материалов ЗКРЯ может быть связано с ***корпусом редуцированных (аллегровых) форм*** русской речи (АФ), создание и описание которого представляется самостоятельной и весьма перспективной задачей[16]. В известном смысле это направление пересекается со словарем русской бытовой разговорной речи (см. выше п. 1), поскольку электронная версия такого словаря предполагает возможность прослушать любое слово в любом контексте:

- *просто не горит // а чек ты в сумку не **хошь** положить?*
- ***здрасть** / отдел кадров уже закрыт? ага / **сёння** же пятница*;
- *а-а вот мы / живем / на Смолячкова / **ничо** не знаем;*
- *мы как раз здесь побудем / за час уже выболтается всё // а он же там **буит**… это*;
- *м-м хорошо / давай // ты во **скоко** будешь дома?*

Наличие словаря АФ, словник которого построен по алфавиту исходных словарных форм, позволит целенаправленно искать ту или иную единицу в материале ЗКРЯ и получать, например, статистику не только употребительности того или иного слова или словосочетания, но и представленности в нашей повседневной речи всех реальных форм языковых единиц.

Мы полностью отдаем себе отчет, что наши лексические «открытия» на самом деле всем говорящим по-русски хорошо знакомы, находятся у всех, что называется, и на слуху, и на языке, однако только корпусный подход к анализу живой устной речи позволит не только получить отдельные (порой весьма и весьма интересные) наблюдения над функционированием в речи тех или иных лексических единиц, но и дать систематическое описание речевого лексикона, близкое к тому, что мы имеем на материале русского литературного языка.

---

[15] Речь идет о студенческих работах: *Кислощук* 2011 и *Гайворонская* 2011.

[16] См. об этом подробнее, например: *Богданова* 2009; *Богданова*, *Пальшина* 2010.

References

1. *Bart R.* 1989. The Noise of tha Language [Gul Iazyka] . Proceedings. Semiotics. Poetics. : 541–544.
2. *Bogdanova N. V., Asinovskii A. S., Rusakova M. V., Stepanova S. B., Sherstinova T. Iu.* 2008. The Corpus of Spoken Russian "One Speech Day": Concept and State of Formation [Zvukovoi Korpus Russkogo Iazyka Povsednevnogo Obshcheniia "Odin Rechevoi Den'": Kontseptsiia I Sostoianie Formirovaniia]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2008" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2008") , 7 (14) : 488–494.
3. *Bogdanova N. V., Brodt I. S., Kukanova V. V., Pavlova O. V., Sapunova E. M., Filippova N. S.* 2008. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2008" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2008"), 7 (14): 57–61.
4. *Bogdanova N. V.* 2009. Database of Reduced Froms of Russian Speech as a Natural Language Fixation Method (Linguistic and Methodological Aspects) [Baza Dannyh Redutsirovannykh Form Russkoi Rechi kak Sposob Fiksatsii Iazyka v ego Estestvennoi Forme (Lingvisticheskii I Metodologicheskii Aspekty)]. Informatsinnoe I Obrazovatel'noe Prostranstvo: Mezhdunarodnaia Planeta "Russkii Iazyk". II Elektronnaia Nauchno-Prakticheskaia Konferentsiia, (Information and Education Space: International Planet "Russian Language". II Electronic Scientific-Practical Conference), 10-13.10.2009 : 28–30.
5. *Bogdanova N. V.* 2010. Who are you, Tetia Motia? [Kto Vy, Tetia Motia?] Mir Russkogo Slova, 4 : 53–62.
6. *Bogdanova N. V.* 2010. On Spoken Texts Corpus: New Materials and the 1st Research Results [O Korpuse Tekstov Zhivoi Rechi: Novye Postupleniia I Pervye Rezul'taty Issledovaniia]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2010" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2010"), 9 (16) : 35–40.
7. *Bogdanova N. V.* 2011. Tetias Motias, Diadias Vadias and Problems of Translation and Teaching of the Russian [Teti Moti, Diadi Vadi I Problemy Perevoda I Prepodavaniia Russkogo Iazyka]. Trudy po Russkoi I Slavianskoi Filologii. Lingvistika XII. HUMANIORA: LINGUA RUSSICA. Razvitie I Variativnost' Iazyka v Sovremennom Mire.
8. *Bogdanova N. V.* 2011. Is our Spoken Language Really Economical in its Resources? [Deistvitel'no li Nasha Ustnaia Rech Ekonomna v Sredstvakh?]. Iazyk I Rechevaia Deiatel'nost'.
9. *Bogdanova N. V.* 2011. ETO SAMOE: Grammatic Froms and Functioning in Russian Spontaneos Speech [ETO SAMOE: Grammaticheskie Formy I Funktsionirovanie v Russkoi Spontannoi Rechi].
10. *Bogdanova N. V.* 2011. The Construction "(Ia) DUMAIU (CHTO)"in Russian Spontateos Speech: Correlation of Different Functional Types [Konstruktsiia (Ia) Dumaiu (Chto) v Russkoi Spontannoi Rechi: Sootnochenie Raznykh

Funktsional’nykh Tipov]. Materialy Nauchnoi Konferentsii “Permskaia Sotsiopsikholingvisticheskaia Shkola: Idei Trekh Pokolenii”. K 70-letiiu so Dna Rozhdeniia Ally Solomonovny Shtern (Proc. of the Scientific Conference "Perm´s Social-Psycho-Linguistic School: Ideas of Three Generations").

11. *Bogdanova N. V.* 2011. On NE ETO SAMOE [Pro Ne Eto Samoe]. Materialy XL Mezhdunarodnoi Filologicheskoi Konferentsii. Vypusk 24. Polevaia Lingvistika. Integral’noe Modelirovanie Zvukovoi Formy Estestvennykh Iazykov, (Proc. of the XL International Philology Conference, Serie 24, Field Linguistics. Integral Modelling of Natural Languages Phonic Form), 23-25.03.2011.
12. *Bogdanova N. V., Brodt I. S., Kukanova V. V., Pavlova O. V., Sapunova E. M., Filippova N. S.* 2008. On Spoken Texts Corpus: Principles of Formation and Possibilities of Description. Komp’iuternaia Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii “Dialog 2008” (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference “Dialog 2008”), 7 (14) : 57–61.
13. *Bogdanova N. V., Pal’shina D. A.* 2010. Reduced Forms of Russian Speech (Experiment of Lexicographic Descroption) [Redutsirovannye Formy Russkoi Rechi (Opyt Lekxikograficheskogo Opisaniia)]. Slovo. Slovar’. Slovesnost’. Tekst Slovaria I Kontekst Leksikografii. Materialy Vserossiiskoi Nauchnoi Konferentsii, (Word. Vocabulary. Philology. Vocabulary Content and the Context of Lexicography. Proc. of Russian Scientific Conference): 491–497.
14. *Gaivoronskaia D. S.* 2011. Analysis of TAM: Functional, Semantic and Grammatic Characteristics of the Word “TAM”in Russian Spontaneos Speech.
15. *Khan N. A.* 2011. Construction “(IA) NE ZNAIU” in Russian Spontaneos Speech: Correlation of Different Function Types [Konstruktsiia (IA) NE ZNAIU v Russkoi Spontannoi Rechi: Soontnoshenie Funktsional’nykh Tipov]. Mir Russkogo Slova.
16. *Khimik V. V.* Large Dictionary of Russian Expressive Speech.
17. *Kisloshchuk A. I.* 2011. Experiment of Sistematization of the Functions of the Word “VOT” in a Spontaneos Monologue [Opyt Sistematizatsii Funktsii Slova “VOT” v Spontannom Monologe]. XIII Nevskie Chteniia “Kommunikatsiia v Usloviiakh Global’noi Informatizatsii”.
18. *Lapteva O. A.* 1992. Debatable Questions of Researches of Spoken Literature Language under the Principles of the Theory of Norm [Voprosy Izucheniia Ustnoi Literaturnoi Rechi v svete Teorii Normy]. Status Stilistiki v Sovremennom Iazykoznanii :150–203.
19. *Larin B. A.* 1977. On Linguistic Research of a City [O Lingvisticheskom Izuchenii Goroda]. Istoriia Russkogo Iazyka I Obshchee Iazykoznanie : 175–189.
20. *Os’mak N. A.* 2011. New Connotations of Old Words (On Lexicographical Description of Russian Spontaneos Speech) [Novye Konnotatsii Starykh Slov (O Leksikograficheskom Opisanii Russkoi Spontannoi Rechi)]. Obrazovatel’nye Tekhnologii v Virtual’nom Lingvokommunikativnom Prostranstve. IV Mezhdunarodnaia Virtual’naia Konferentsiia po Rusistike, Literature I Kul’ture, (Educational Technologies in the Virtual Lingvocommunicational Space. Proc. of the IV International Virtual Conference on Russian Philology, Literature and Culture), 2-4.03.2011.

21. *Os'mak N. A.* 2011. On the Problems of Lexicographic Description of Russian SPontaneos Speech (On the Material of Person Definition in the Russian Spoken Texts Corpus) [O Problemakh Leksikograficheskogo Opisaniia Russkoi Spontannoi Rechi (Na Primere Obboznachenii Litsa v Zvukovom Korpuse Russkogo Iazyka)]. Mir Russkogo Slova.
22. *Os'mak (Androsova) N. A., Chen Ch. V.* 2010.Experiment of Lexicographic Description of Russian Spontaneos Speech (Words Defining Female Persons, in the Russian Spoken Texts Corpus) [Opyt Leksigograficheskogo Opisaniia Russkoi Spontannoi Rechi (Slova, Oboznachaiushchie Litso Zhenskogo Pola, v Zvukovom Korpuse Russkogo Iazyka)]. Materialy XXXIX Mezhdunarodnoi Filologicheskoi Konferentsii. Vypusk 23. Polevaia Lingvistika. Integral'noe Modelirovanie Zvukovoi Formy Estestvennykh Iazykov (Proc. of the XXXIX International Philology Conference, Serie 23, Field Linguistics. Integral Modelling of Natural Languages Phonic Form), 15-19.03.2010 : 54–70.
23. *Russian* Language Dictionary, in IV vv. I v. A-I.1981.
24. *Russian* Language Dictionary, in IV vv. II v. A-I.1982.
25. *Russian* Language Dictionary, in IV vv. IV v. A-I.1984.
26. *Skovpen' O. P.* 2011. The Functioning of the Preposition "CHEREZ" in Modern Russian (On the Material of Spoken Texts Corpus "One Speech Day") [Funktsionirovaniie Predloga CHEREZ v Sovremennom Russkom Iazyke (Na Materiale Zvukovogo Korpusa "Odin Rechevoi Den'")]. Vestnik Sankt-Peterburgskogo Universiteta. Filologiia. Vostokovedenie. Zhurnalistika, 9.

БОЛЬШОЙ ЭЛЕКТРОННЫЙ СЛОВАРЬ КАК ПОЛИТЕМАТИЧЕСКИЙ СПРАВОЧНИК И ФОРМИРОВАТЕЛЬ ЗАПРОСОВ К ИНТЕРНЕТУ

И. А. Большаков (bolshakov34@mail.ru)

Независимый исследователь, Москва, Россия

А. Ф. Гельбух (gelbukh@gelbukh.com)

Национальный политехнический институт, Мехико, Мексика
и Университет Васеда, Токио, Япония

Описывается большой электронный словарь, содержащий как фундаментальные сведения о русском языке (грамматические свойства слов, их комбинаторику, семантические и паронимические связи слов), так и обширные энциклопедические сведения о географических объектах, известных персоналиях, организациях и артефактах. В словаре содержатся технические термины, базовые понятия точных и гуманитарных наук, бизнеса и экономики. Словарь позволяет быстро формировать и прямо направлять в Интернет запросы медицинского, коммерческого, туристического и др. характера.

Ключевые слова: большой электронный словарь, электронный словарь, справочник, интернет-запросы, запросы к Интернету

A LARGE ELECTRONIC DICTIONARY AS A POLYTHEMATIC GUIDE AND SHAPER OF QUERIES TO THE WEB

I. A. Bol'shakov (bolshakov34@mail.ru)

Independent researcher, Moscow, Russian Federation

A. F. Gel'bukh (gelbukh@gelbukh.com)

National Polytechnic Institute, México

A large Russian electronic dictionary is presented. It contains both fundamental information on the Russian language (grammatical and combinatory properties of words, semantic and paronymic relations between words) and ample encyclopedic information on geographical objects, famous people,

organizations, and artifacts. The dictionary includes technical terms and basic concepts of science, humanities, business, and economy. Among its applications is the possibility to form queries for Internet search engines on medicine, commerce, tourism, and other topics.

Key words: large electronic dictionary, electronic dictionary, guide, Internet search, queries

1. Введение

Лингвистика оперирует знаниями двух типов: чисто лингвистическими (о грамматике слов и их комбинаторике) и энциклопедическими (о внешнем мире). Общеобразовательные словари в основном содержат знания первого типа, хотя часто включают перечни географических имен и персоналий, т. е. энциклопедических объектов. Все более необходимые в обыденной жизни энциклопедические знания помещают в специализированные словари и энциклопедии.

Для экономии поисковых усилий возникает желание иметь электронный словарь, содержащий как можно больший объем лингвистических знаний вместе с необходимым минимумом энциклопедических сведений.

Современному человеку всегда требуется и информация о текущем состоянии внешнего мира: *Где купить X? Каков срок действия документа Y? Как лечить болезнь Z? Каковы достопримечательности страны W?* Традиционно все это отражается в прессе, радио- и телевизионных передачах, а теперь еще и в Интернете.

Казалось бы, Интернет покрывает сейчас информационные потребности любого пользователя. Но поиск в сети порой требует определенных усилий: нужно уметь разумно составить запрос и быстро выбрать нужный сайт среди появившихся на экране компьютера сниппетов. Поэтому хочется иметь лингвистическое средство, которое сочетало бы свойства универсального словаря-справочника указанного выше типа с возможностью формирования релевантного запроса к Интернету.

Данная работа содержит описание КроссЛексики (КЛ) — большого электронного словаря русского языка, адекватно отвечающего поставленным выше требованиям: выдавать любые лингвистические и многие энциклопедические справки, а также формировать типовые запросы к Интернету. Предыдущие публикации о КЛ [2] и [4] представляли этот словарь как чисто лингвистический ресурс, и содержащиеся в нем энциклопедические знания не упоминались. Новая функция КЛ как формирователя запросов к Интернету оставалась неосознанной и практически не реализованной. В данной статье мы восполняем этот пробел.

Наше изложение имеет целью описать:

- Структуру КЛ в целом — для облегчения понимания дальнейшего;
- Части структуры КЛ, содержащие особо богатые лингвистические сведения;
- КЛ как словарь-справочник энциклопедического характера;
- Одноступенчатое формирование запроса к Интернету: через словник КЛ;
- Двухступенчатое формирование запроса к Интернету: через элементы выдачи КЛ;

- Преимущества и недостатки КЛ как формирователя запросов в сравнении с непосредственным обращением в Google.
  Указываемые ниже статистические параметры соответствуют марту 2011 г.

2. Краткий обзор КроссЛексики

В основе КЛ лежит квадратная матрица {словник × словник}, где словник — это вектор $\{t_1, t_2, ..., t_n\}$ из титулов, являющихся отдельными словами или коллокациями (рис. 1). Размер словника n по состоянию на март 2011 г. превысил 250 тыс. Элемент D_{ij} этой матрицы является дескриптором односторонней связи между t_i и t_j. Связь бывает синтагматической (титулы образуют коллокацию), семантической (титулы связаны смысловым сходством типа синонимии, антонимии, гипонимии / гиперонимии и др.) или паронимической (титулы связаны буквенным или морфемным сходством). Например, титулы t_i = *борьба* и t_j = *суеверия*, образующие коллокацию *борьба против суеверий,* имеют $D_{ij} = \{t_i$ УПРАВЛЯЕТ '*против*' $t_i\}$, а титулы t_i = *диплом* и t_j = *документ* имеют $D_{ij} = \{t_i$ ГИПОНИМ $t_i\}$.

Запрос к КЛ в виде титула t_i (на рис. 1 возможные запросы представлены словником в крайнем левом столбце) ведет к выдаче всех тех титулов, связь которых с t_i зафиксирована в матрице. Совокупность потенциально выдаваемых титулов дается словником в верхней строке. Если связь является синтагматической, то в выдаче автоматически формируется коллокация в ее грамматически правильной форме, учитывающей род, число, падеж и одушевленность русских существительных и прилагательных.

Рассматриваемая гигантская матрица очень разрежена: непустым оказывается примерно один дескриптор на 7600. Однако уже выявлено 6,93 миллионов непустых ее элементов. Вот важные особенности матрицы:

- Титул любой из четырех частей речи может быть как одиночным словом, так и многословным оборотом. При этом знаменательные составляющие оборота входят в словник и автономно, а декомпозиция может быть многоступенчатой: (((*судебные*) (*прения*)) и (((*последнее*) (*слово*)) (*подсудимого*))).
- Если непуст дескриптор D_{ij}, то непуст и обратный ему D_{ji}. Например, для t_i = *диплом* и t_j = *документ* имеем $D_{ji} = \{t_i$ ГИПЕРОНИМ $t_i\}$.

В КЛ встроены входы в Google, Яндекс и Национальный корпус русского языка. Два поисковика являются самыми мощными в Рунете и обслуживают миллионы пользователей. НКРЯ, крупнейший корпус частично размеченных русских текстов, может служить источником примеров словоупотреблений в контексте.

3. КроссЛексика как лингвистический справочник

Хотя словарь КЛ является единым целым, именованные секции его выдачи, рассматриваемые совместно, можно называть подсловарями. Ряд подсловарей КЛ содержит особо богатую лингвистическую информацию.

- Подсловарь **Коллокаций** покрывает все типы русских коллокаций и включает их в количестве 1,96 млн. (3,92 млн. односторонних связей), см. [2, 4].
- Подсловарь **Модели управления** характеризует более 20 тыс. управляющих глаголов и 15 тыс. существительных. Модели управления имеют также многие прилагательные и наречия.
- Подсловарь **Синонимов** уникален по объему: 115 тыс. синонимов разбиты на 21 тыс. групп, и общее число связей превышает 1,2 млн. (ср. с 600 тыс. связей в [1]).
- Подсловарь **Ассоциаций** не имеет прецедентов. Он создан на основе сочиненных пар, образующих частые запросы к Интернету или многократно представленных на интернет-сайтах [3]. На данный момент найдено 58,1 тыс. ассоциаций для 15,3 тыс. одно- и многословных титулов, в среднем 3,8 ассоциаций на титул.
- Подсловарь **Морфопарадигм** дает все возможные формы для каждого изменяемого титула. Падежные формы многословных именных титулов могут включать до шести частей, из них до пяти изменяемых, например, *десять заповедей и семь смертных грехов*.
- Подсловарь **Семантических деривативов** составлен из подсекций титулов, выражающий одно и то же понятие четырьмя главными частями речи: существительными, глаголами, прилагательными или наречиями. Наиболее просты здесь случаи, когда семантическая деривация в основном осуществляется морфологическими средствами:

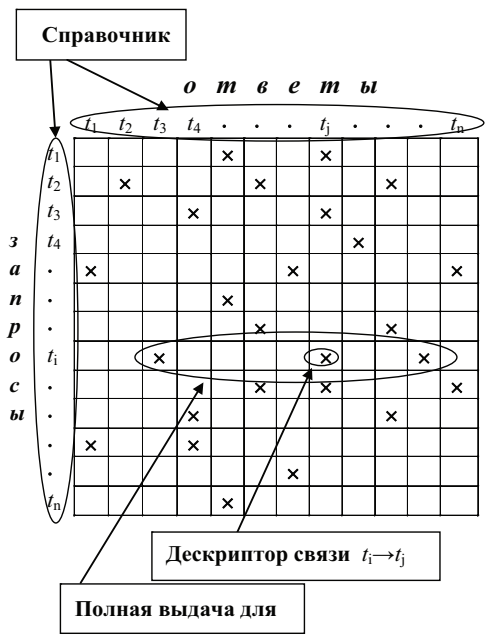


Рис. 1. КроссЛексика как матрица связей

Сущ *владение, владения, владелец, владельцы, собственник, собственники, собственность*

Глаг *владеть, завладеть, иметь в собственности, находиться во владении, являться собственностью*

Прил *владеющий, владевший, завладевший, находящийся во владении, собственный, являющийся собственностью*

Прич *будучи собственностью, владея, в качестве собственности, в собственности, в собственность, во владении, завладев, как собственник, овладев*

В подсекции существительных встречаются два числа одного существительного, а в подсекции глаголов — разные виды одного русского глагола.

- Подсловарь **Гипонимов и гиперонимов** является обширной полииерархией понятий. Например, *Франция* имеет два гиперонима — *европейская страна* и *средиземноморская страна*, а *цветы* имеет гиперонимами *предметы быта* и *растения*. Наличие нескольких уровней иерархии позволяет транзитивно переносить свойства гиперонимов на гипонимы несколькими уровнями ниже, что в КЛ активно используется.
- Поскольку однозначно принятого соотношения «часть Vs. целое» не существует, понятия **меронима** и **холонима** у нас многозначны. Так, существительное множественного числа считается холонимом единственного числа. Меронимом считается часть физического объекта, например, *нога* по отношению к *тело* или *рулевое колесо* по отношению к *автомобиль*. Для неисчисляемых объектов используется понятие **квантификатора**. Так, для *вода* квантификаторами являются *капля / стакан / бочка / цистерна воды*, а для *гнев* — *приступ гнева*. Перенос свойств от холонимов к меронимам не предусмотрен.

4. КЛ как справочник энциклопедического характера

КЛ покрывает следующие тематические области:

- Экономика и бизнес;
- Общественно-политическая тематика (политика, социология);
- Техника и технологии (радиоэлектроника, компьютеры, программирование, Интернет, бытовая техника, автомобили, строительство);
- Точные и естественные науки (математика, физика, химия, биология, геология, география);
- Медицина (не только бытовая);
- Гуманитарные науки (лингвистика, психология, философия, история), искусство, религия;
- Бытовой язык (включая бранную лексику, но без нецензурной).

Термины и имена из указанных областей и представляют собой энциклопедическую информацию. Остановимся на этом подробнее.

В подсловаре **Семантических деривативов** (далее ПССД) имеется несколько десятков наборов, характеризующих наиболее известные страны. В подразделе

существительных здесь даются наименования государства и наций, ее представителей мужского и женского пола, официального языка, столицы, титула главы государства, единицы административного деления, денежной единицы, официальной или ведущей религии (если таковая существует). Например, для титула *Франция* даются *Париж, француз, французы, француженка, француженки, французский язык, президент Франции, провинция Франции, евро, франк*. Данные здесь же коллокации могут служить отсылками к дополнительной информации внутри КЛ или вне ее: *город Франции, провинция Франции, население Франции, площадь Франции, туры во Францию, отдых во Франции, достопримечательности Франции.*

В ПССД имеется также несколько десятков наборов, характеризующих российские города и включающих наименования их жителей. Не все говорящие по-русски сразу вспомнят, как называются жители Архангельска, Смоленска, Тулы, Курска, Нижнего Новгорода или Пскова, а в отношении жительниц этих городов ситуация еще сложнее, поскольку для ряда городов их называют только описательно.

В подсловаре **Гипонимов и гиперонимов** (далее ПСГГ) содержатся важные географические понятия (моря, океаны, континенты, горы и др.) и их имена. Даются также имена городов России и множества зарубежных стран. Приведены также имена единиц административного деления, например, области или края России, провинции Франции, штата США, Канады, Мексики и Индии, графства Великобритании, воеводства Польши.

В ПСГГ содержатся также наиболее известные персоналии, в основном, из прошлого. Это ученые, композиторы, поэты, политические и общественные деятели, миллиардеры. Здесь же содержатся имена наиболее известных культурных артефактов, как то романов, опер, оперетт, мюзиклов, фильмов, мультфильмов.

В ПСГГ приведены 48 видов цветов, 152 оттенка цвета, 183 специализации заводов, 124 специализации промышленности. Широко представлены предметы быта, включая мебель, электрооборудование и гаджеты.

В подсловаре **Определений** содержится более 200 тыс. научно-технических терминов со структурой «прилагательное ← существительное», где существительное-гипероним служит терминообразующим ядром. Наиболее продуктивны *режим*$_2$ (514 терминов-гипонимов), *покрытие*$_1$ (438), *анализ* (423), *препараты* (416), *конструкция*$_1$ (386), *вещества* (377), *контроль* (364), *детали*$_2$ (364).

Продемонстрируем извлечение из КЛ энциклопедической информации о Франции (рис. 2). Войдя в КЛ, мы сразу оказываемся в словнике, где набором пяти начальных букв попадаем в строку *Франция*. Нажав *Enter*, получаем на экране соответствующую Франции выдачу. В секции **Синонимы** приведены известные фигуральные названия страны — *галльский петух* и *страна гурманов*. **Гиперонимы** свидетельствует, что Франция и европейская и средиземноморская страна. Ее **Когипонимы** — это многочисленные страны, образующие эти две группы — среди них, например, *Австрия* как страна Европы и *Алжир* как страна Средиземноморья. В секции **Семантические деривативы** особо интересна подсекция существительных. Выбрав *город Франции*, получаем новую выдачу, где **Гипонимы** — длинный список французских городов: *Абвиль, Авиньон, Авориаз, Адье,...* Выбрав *президент Франции*, получим **Синонимы** в виде перифраз *Николя Саркози, президент Николя Саркози, президент Саркози*.

Выбрав *провинция Франции*, получим список *Аквитания, Бретань, Бургундия, Лангедок, Лотарингия* и т. д.

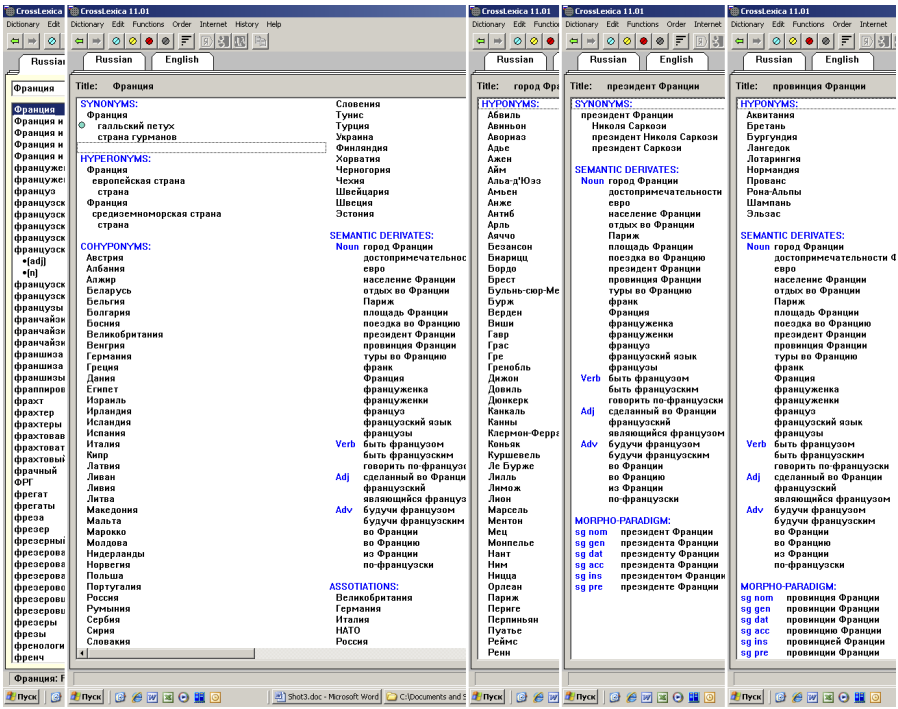


Рис. 2. Поиск в КЛ информации о Франции

5. Одноступенчатое формирование запроса к Интернету

Размер словника достигает в настоящее время 250 тыс. единиц. Каждый его титул можно сразу направить в Интернет в качестве запроса. Это одноступенчатое его формирование. Скорее всего, это будут именные титулы, которых в словнике более 90 тыс.

Одним шагом могут быть сформированы запросы типа *боль в горле / желудке / ногах / шее*, *туры в Египет / Испанию / Италию / Израиль / Турцию / Финляндию.* Особенно интересны для одноступенчатых запросов сочиненные пары типа *простуда и антибиотики, радиация и иммунитет, лечебное питание и климатолечение, герметизация и теплоизоляция, боль и тяжесть в желудке, боль и хруст в коленях, светильники и абажуры / бра / люстры / лампы, курага и диабет / запор / сердце / чернослив.* Таких пар в словнике 34 тыс.

Продемонстрируем одноступенчатое формирование запроса на примере титула *простуда и антибиотики* (рис. 3). Достаточно набрать в словнике шесть первых букв, как на экране появляется искомый титул. Подведя к нему курсор и нажав иконку Google, получим совокупность релевантных сниппетов.

Для диверсификации конечного вида запроса в КЛ предусмотрены четыре опции: 1) отсылка запроса *Q* как есть; 2) отсылка «*Q*» в кавычках; 3) отсылка *Q это*, т. е. с постфиксом «это»; 4) отсылка «*Q это*» в кавычках. Первая опция очень мало ограничивает поиск. Вторая выдает только буквальные вхождения запроса в тексты сайтов. Третья опция ищет определения запрашиваемого понятия (например, *валоризация это*) и, как правило, выдает определение из Википедии. Четвертая опция примерно эквивалентна третьей.

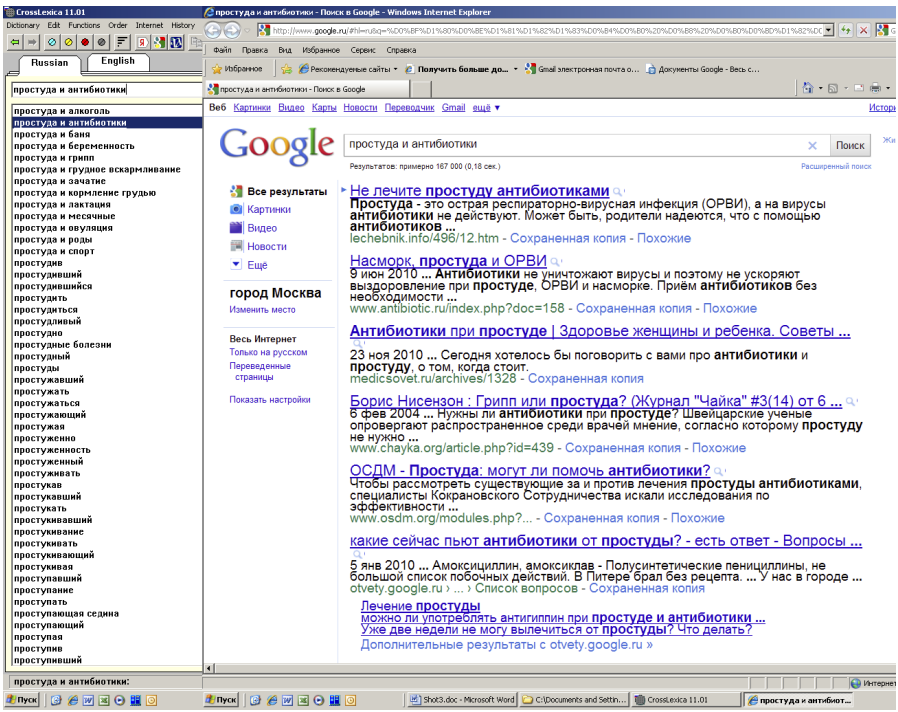


Рис. 3. Одноступенчатый поиск для *простуда* и *антибиотики*

6. Двухступенчатое формирование запроса к Интернету

Можно брать в качестве запросов к Интернету и любой элемент выдачи для произвольного титула. Число таких элементов равно суммарному числу связей в КЛ, т. е. 6,93 млн. Это двухступенчатое формирование запроса. Берется титул в словнике, вызывается на экран его выдача, и какой-либо ее элемент отсылается в Интернет. В первую очередь интересно использовать в качестве запроса любую из 1,96 млн. коллокаций.

На рис. 4 отражены этапы формирования запроса *купить бинокль*. В окошке Словника вводим начальные буквы слова *купить*. В секции **Модели управления** его выдачи выбираем подсекцию **купить что / кого?**, в ней подводим курсор к строке *бинокль* и нажимаем иконку поисковика. Точно так же

ищется *лечить артроз / варикоз / гастрит / диатез / холецистит*, *найти врача / няню / тренера*, *воспаление лимфоузлов / печени / почек / слизистой / сосудов*.

Для запросов в виде коллокации всегда существуют два эквивалентных двухступенчатых пути в Интернет. Так, запрос *купить бинокль* можно сформировать и через секцию **Управляется_глаголами** в выдаче для *бинокль*.

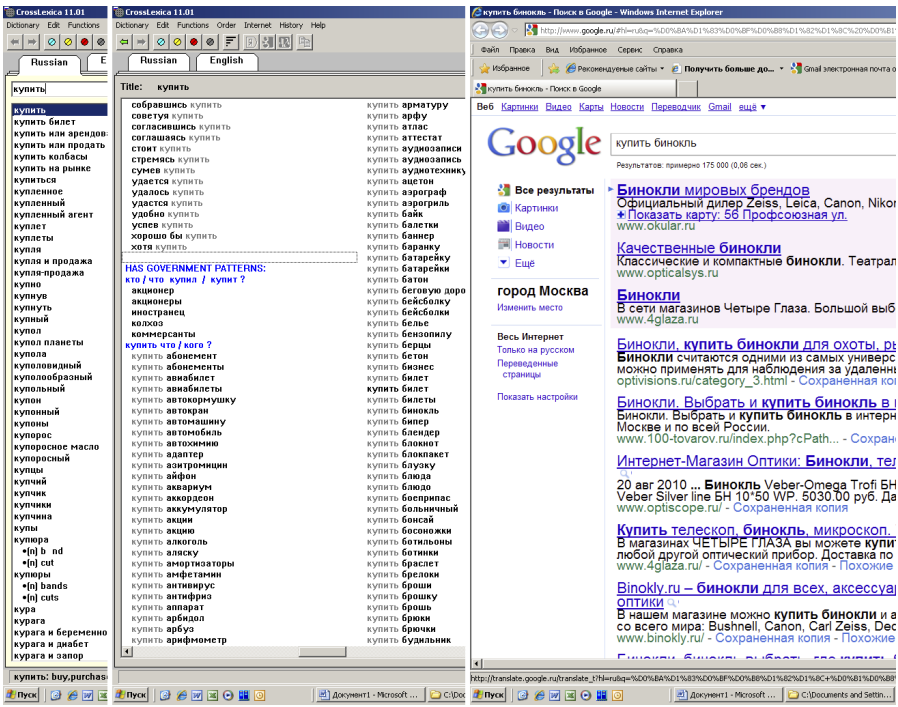


Рис. 4. Двухступенчатый поиск для *купить бинокль*

7. Преимущества и недостатки КЛ как формирователя запросов

Поясним преимущества и недостатки КЛ как формирователя запросов к Интернету по сравнению с непосредственным вводом в Google.

Как только пользователь входит в Google, он соединяется с накопленной базой прежних запросов, из которых Google оперативно формирует меню подсказок. Своей начальной частью элементы меню повторяют уже введенную строку и наиболее частые из прежних запросов. Подсказок всегда немного, обычно 4–6. Одна подсказка иногда вкладывается в другую, так что реально элементов в меню оказывается еще меньше.

Ввод существенно зависит еще от того, включен ли режим так называемого «живого поиска». При широкополосном Интернете живой поиск включен

всегда, при узкополосном — когда поисковик недогружен. Если живой поиск отключен, поисковик не ищет нужные сайты, пока пользователь не нажмет *Enter*. При включенном живом поиске Google постоянно ищет сайты и мгновенно выдает их сниппеты на экран даже при самой короткой или абсурдной строке, введенной пользователем.

И в базе запросов, и в основном массиве Google за годы накопилось много неверных начальных цепочек. Поэтому при вводе первых букв запроса возможно появление неграмотных или абсурдных подсказок и сниппетов, соответствующих накопленным в прошлом ошибкам. Это цена, которую приходится платить за статистические методы работы поисковика.

КроссЛексика не пользуется ни базой запросов, ни живым поиском. Фактически, она заменяет базу запросов. Достоинствами ввода запроса через КЛ является то, что:

- В окошке ввода КЛ появляется продолжение введенной цепочки букв в виде существующего титула, ближайшего в словнике по алфавиту. Если пользователь начал вводить абракадабру, поиск в Словнике останавливается там, где пользователь свернул с правильной дороги.
- Меню в КЛ всегда много обширнее, чем у Google. Ввод первых 3–6 букв обычно позволяет видеть в пределах экрана строку, нужную пользователю. Остается только подвести к ней курсор и нажать иконку Google. Это значит, что КЛ позволяет быстрее перейти от буквенного ввода к перемещению курсора вдоль выведенных на экран упорядоченных по алфавиту альтернатив.
- Содержимое экрана КЛ, в отличие от случая живого поиска, не меняется суетливо по мере ввода очередной буквы и не отвлекает пользователя от процесса ввода, ввод проходит более плавно.
- Независимо от того, включен ли живой поиск или нет, при вводе запроса в Google в любой момент может возникнуть период ожидания до нескольких секунд, когда процесс ввода блокируется. Это может раздражать пользователя.

Если говорить о недостатках ввода запросов через КЛ, то ими является следующее:

- КЛ, как и положено словарям, является ресурсом инерционным и избирательным. Если идет речь о событии, продукте или лице, внезапно возникшем в информационном поле, то сведения о них не будут найдены в словаре, и составлять запрос относительно них через КЛ нецелесообразно.
- В КЛ мало конкретных названий артефактов, даже устоявшихся за годы (например, марок технических изделий). Поэтому запрос, включающий торговую марку, нужно направлять поисковику напрямую.
- В КЛ отсутствует подсказка в виде более вероятного варианта, очень похожего на введенный пользователем запрос. Иногда подсказка помогает, но для опытного пользователя, не допускающего тривиальных ошибок, может явиться «медвежьей услугой»: он ищет одно, а ему подсовывают другое, забивая начальную часть экрана ненужным материалом.

8. Заключение

Современный пользователь электронных словарей скорее предпочтет единый словарь, содержащий как всесторонние знания о языке, так и спектр энциклопедических познаний. Предлагаемый словарь в существенной мере является таковым. В нем хранятся имена, неизменные за годы и десятилетия. В целом структура, содержание и объем хранящейся в КроссЛексике информации не имеют аналога в электронном мире в виде единого продукта.

Однако для получения актуальной информации, изменяющейся в течение недель, дней или часов, необходимо обращаться в Интернет. При желании пользователя иметь в своем компьютере единый информационный интерфейс можно потребовать от электронного словаря еще и умения помочь быстро и удобно составить запрос к Интернету по темам, которые сам словарь отражать не может из-за их изменчивости. КроссЛексика предоставляет средства формирования и отсылки запросов по самой широкой тематике, в первую очередь — медицинской, коммерческой или туристической.

Пополнение и совершенствование КроссЛексики продолжается.

Работа выполнена при частичной поддержке правительства Мексики (CONACYT 50206-H, SIP-IPN 20113295, SNI) и Индии (DST India).

References

1. *Aleksandrova Z. E.* 2005. Russian Synonims Dictionary [SLovar' Sinonimov Russkogo Iazyka].
2. *Bol'shakov I. A.* 2009. KrossLeksika: Large Electronic Dictionary of Russian Words Combinations and Connections [KrossLeksika: Bol'shoi Elektronnyi Slovar' Sochetanii I Smyslovykh Sviazei Russkikh Slov] . Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2009" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2009"), 8 (15) : 45–50.
3. *Bol'shakov I. A., Bol'shakova E. I., Gel'bukh A. F.* 2010. Associative Net of Concepts forming Queries to the Web [Assotsiativnaia Set' Poniatii, Obrazuiushchikh Zaprosy k Internetu]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2010" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2010"), 9 (16): 55–61.
4. *Bol'shakov, I. A.* 2004. Getting One's First Million Collocations. Computational Linguistics and Intelligent Text Processing. Proc. 5th Intern. Conf. on Computational Linguistics CICLing-2004 : 229–245.

КОРПУСНОЕ ИССЛЕДОВАНИЕ КРИПТОКЛАССОВ АНГЛИЙСКОГО ЯЗЫКА

О. О. Борискина (olboriskina66@mail.ru)

Воронежский Государственный Университет,
Воронеж, Россия

Реконструкция скрытых лексико-грамматических именных классов (криптоклассов) английского языка опирается на критерии выявления явных именных классов языков мира. Изучение криптоклассов позволит увеличить объем знаний для более глубокого понимания общих закономерностей в организации лексики и поиска путей формализации языковой семантики.

Ключевые слова: криптоклассы, именные классы, скрытые именные классы, английский язык, языковая семантика.

A CORPUS-BASED STUDY OF NOUN CRYPTOTYPES IN ENGLISH

O. O. Boriskina (olboriskina66@mail.ru)

Voronezh State University, Voronezh, Russia

We develop a method of identifying noun cryptotypes in English, relying primarily on the Corpus of Contemporary American (COCA) and the results of typological studies. The study uses data-oriented and theory-oriented approaches to linguistic description. A cryptotype is referred to the principle of distribution of nouns among classes in accordance with a certain semantic feature and with reference to the typological principle of contrastive grammar. The class membership of a noun is evidentially revealed in syntax, particularly in collostructions which bear the classifying function of the noun class. The semantic, morphologic and syntactic criteria for identifi cation of a noun class are discussed. The study of cryptotypes concerns the issues of grounding, recognition, and reasoning. An adequately formalized description of cryptotypes can be used in computational modeling and text processing.

Key words: cryptotypes, noun classes, English, language semantics, computational modeling, text processing.

1. Introduction

The main point of this paper is to exemplify a method of exploration of covert noun classes, and to justify its applicability to applied linguistic technologies and its significance for the study of linguistic categorization, metaphor, and lexical systems in typological perspective. The method exploits the term *cryptotype* introduced by B. Lee Whorf (1956), and develops his idea of *phenotype* (overt word class) vs. *cryptotype* (covert word class) distinction.

> *"I call a cryptotype a linguistic classification like English gender, which has no overt mark actualized along with the words of the class but which operates through an invisible "central exchange" of linkage bonds in such a way as to determine certain other words which mark the class, in contrast to the phenotype, such as gender in Latin" (Whorf, 1956, p. 72)*

Throughout this study the Whorfian understanding of a noun cryptotype (largely confined to "English gender") has been "reincarnated" and advanced. Cryptotype is widely understood as an implicit rule of language use which can be extended to noun categorization. As V. A. Vinogradov (1991) and A. A. Kretov (2010) have argued, noun classes in world languages vary in degree of grammatical representation. One can think of grammar structure as a kind of scale. At the one end of the scale is the most grammaticalized type of a noun class that of gender in European languages. At the other end is the least grammaticalized class of so called numeratives (e.g., lexical count words in Asian languages). A variety of noun classifications in world languages is in between. This is why, a noun class (NC) is approached in this research as a class of Lexical Grammar. Such approach to noun classes evoked in different types (cryptotype vs. phenotype) allows for a more holistic view of language. The reconstruction of cryptotypes of English nouns is meant to offer the paleo-semantic interpretation of noun classes' formation.

The main challenge of pursuing the common grounds for the classification of nouns in typological perspective is that, presumably, there should be a universal set of noun properties reflected in grammatical meanings of some languages (morphologically tagged) vs. lexical meanings of others (marked by lexical selection, e.g., in a verb stem in syntactical constructions)[1].

A cryptotype is regarded as a covert noun class of Lexical Grammar, where the class membership is marked by lexical selection in the construction rather than a morphemic tag. A cryptotype can be identified owing to the "typological principle of contrastive grammar". The principle is as follows: if a NC is grammatically mapped (represented) in a language (L1), and is NOT represented in the Grammar of L2, it is viewed as a grammatical lacuna of L2 and studied in latent Grammar or Lexical Grammar of L2.

In this research of cryptotypes we have adopted the practice of describing noun classes in the world languages (Dixon 1968, Rosch 1973, Givon 1996, Vinogradov et al, 1996, 2000, Gillon 2005, just to name a few); the main achievements in the investigation

[1] About the universal menu of grammatical signs (cf., e. g. Chvany 1995, Itkonen 1994). The problem of discovering lexical universals is under consideration in, e. g., "Aquamotion" (2007).

of metaphoric representation of concepts (Arutjunova 1976, Uspenskij 1979, Johnson 1980, Lakoff 1986, Lakoff & Johnson 1987, Kövecses 2002, 2005) and the primary metaphors theory (Grady 1997); the implications of Covert Grammar (Katsnelson 1972); the basic assumptions of construction grammar (Fillmore 1985, Kay & Fillmore 1999), specifically, (Lakoff 1987, Goldberg 1996), Pattern Grammar (Hunston & Francis 2000) and the concept of collostruction (Gries & Stefanowitsch, 2003, 2004).

The paper is structured as follows: Section 2 deals with the main criteria for identifying a noun class proposed by the Russian Typological School. In Section 3 we will focus on the semantic and syntactic criteria for recognizing the English cryptotype 'Res Longa'. Section 4 will conclude the paper.

2. The main criteria for identifying a noun class

According to the Russian Typological School, there are three aspects of noun classes' exploration: semantic, morphologic and syntactic (Toporova 1996, 25). Since a cryptotype lacks morphological marking, it is meant to be identified via the syntactic and the semantic criteria.

2.1. Semantic criterion for identifying a noun cryptotype

Like many overt noun classes of African languages (see Vinogradov 1996, Toporova 1996, Koval' 1996 & others) a cryptotype is semantically heterogeneous. It means that the nouns a class incorporates bear diverse lexical meanings but they are united owing to the "prototypical categorical attribute or feature" (PCA) which underlies the noun class. While the PCA is indeed part of the core meaning of the nouns that are called class prototypes, there are nouns that are only associated with the class because the PCA is the element of their metaphorical meaning, i.e. they are conventionally attributed to this class of nouns.

To illustrate, let us take the PCA 'being liquid'. It underlies the overt noun class DAM (so called class of liquids) in the Pular-fulfulde language (Koval' 1996) and languages of Bantu family (Toporova 1996). With reference to the "typological principle of contrastive grammar" we can identify the cryptotype "Liquidus" in English which incorporates class prototypes (*water, blood, milk*) as well as class metaphor-types, i.e. metaphor-driven members of the class (*life, color, truth, information*, etc.). The latter are attributed to the class in accordance with their metaphorical meaning that bears the traces of the nouns' former occurrences, which in their turn reflect the cognitive associations of these abstract entities with the properties of liquids.

This semantic diversity of the class can be accounted for by the analogy principle of human cognition as well as the "law of metaphor", both of which would regulate the linguistic categorization of human experience at the early stages of mind and language development. Since analogical mapping is a key cognitive operation (Fauconnier 1997), it can be assumed that such concepts as 'life', 'time', 'truth', and 'color' were cognitively associated with the properties of liquids and categorized as liquid, so that the lexemes

for these concepts were attributed to the class of nouns which share such properties. The consensus view emerging from a large body of synchronic and diachronic research acknowledges abstract concepts being largely metaphorical which means that most of our nonphysical reality is conceptualized via physical reality, that is, in terms of physical domains of experience. This approach to categorization and embodiment of immaterial, incorporeal, intangible, insubstantial, impalpable concepts of abstract dimension was independently introduced in the works of Arutunova 1976, Uspenskiy 1979, Lakoff & Johnson 1980, 1987, and further refined in the framework of cognitive linguistics.

Similarly, the restoration of another four cryptotypes for English has been attempted so far:

- Round objects (prototypes: sun, egg)
- Solid, long, pointed objects (prototypes: stick, rod, spear, arrow)
- Containers (prototypes: container)
- Hand-fit objects (prototypes: stone, fruit, seed).

In return to the semantic issues of the problem, most scholars state that the semantic foundations of noun classes in African languages have been fading; so, the semantic principle of class formation is losing its validity for the study of language systems (Vinogradov et al 1996). Hence the tradition in Africanistics to code noun classes in numerals to lessen the value of vague and obscure lexical semantics of the class. Curiously enough, infants start out with semantic categorization criteria, but at some point they abandon them (Gillon 2005: 447).

By contrast, the morphological principle of class formation remains strict enough: all nouns of the class are marked in themselves by a morphemic tag. For example, the class DAM and class 6 in Bantu languages include names of liquids and names of abstract entities, the latter (e.g. *life* or *truth*) are marked by a morphemic tag. However, the loanwords that enter African languages do not necessarily receive the morphemic tag of the class but are colligated in syntax. (e.g. Lutskov, 1996). Thus, the semantic and morphemic criteria for NC identification are fairly relative while the syntactic criterion appears to be the most relevant one to NC identification.

2.2. Syntactic criterion for identifying a noun cryptotype

V. A. Vinogradov (1996, 9) shows that when the semantic criterion is ambiguous, and the morphemic tag of the class has been obliterated, the syntactic agreement, even though implicit, becomes the only reliable warrant of a noun classification. Schematically, the cryptotype recognition procedure is as follows: class prototypes $\Leftrightarrow$ classifiers $\rightarrow$ class metaphor-types. Identification of noun cryptotypes in English starts from the contextual analysis of the class prototypes in corpora. The detailed investigation of the prototype occurrences in corpora allows recognizing the constructions, or rather collostructions, which can be further used for the purposes of classification recognition. Collostruction is understood as the pattern of co-occurrence between the collocating items (collexemes) which occur preferably in certain constructions (cf. Gries and Stefanowitsch 2003, 2004). The prototypical nouns of a cryptotype

and the words which assign the meaning to the construction, are called collexemes. It should be noted, that this research is not concerned with measuring collostructional strengths between collexemes; collostructions are treated as classifiers of noun classes, or a means of lexico-syntactic recognition of the class. The main focus of the study is collostructional patterning observable within the argument structures associated with predicating words. Additionally, N of N constructions are taken into account. Then we undertake a corpus analysis of abstract nouns' aptitude for filling the syntactical slots of classifiers. The class membership of a noun is diagnosed when it occurs in the collostruction(s) associated with a particular cryptotype.

All things considered, a cryptotype can be defined as the principle of distribution of nouns among classes in accordance with a certain semantic feature (PCA) and with reference to the typological principle of contrastive grammar, evidentially revealed in syntax, particularly in classifying collostructions of the noun class.

Let us consider an example of cryptotype identification.

3. Case study

A corpus-based study of the English cryptotype "Res Longa" relied primarily on Corpus of Contemporary American (COCA[2]). The work started with corpora data retrieval and manual sorting out relevant results from the accidental ones, received from *List* display of collexemes (e.g. <stick> + <prick>) as well as *KWIC* concordances for "Res Longa" prototypes. A series of informal experiments (in the shape of questionnaires for native speakers of American English) have been designed in order to gather data unavailable in the corpus that might have tested the hypotheses of the research.

3.1. Classifiers of the cryptotype RES LONGA

The cryptotype "Res Longa" is a lacuna in the Grammar of English, which we attempted to approach with the help of the principles described in Section 2.

Most languages of Bantu family, Fula family and languages of South-eastern Asia (Vietnamese, Japanese, dialects of Chinese, Burmese) have an overt grammatical NC with the PCA 'inanimate, hand-fit, long object of fixed pointed shape'. Furthermore, Burmese differentiates 'long objects' and 'pointed instruments' (cf. Hla Pe 1965). Clearly, the shape and size of inanimate objects as features, deeply entrenched in the human bodily experience and cognition, tend to emerge in language systems in either way: phenotype or cryptotype. According to the contrastive principle this fact can be regarded as a basis for the identification of the cryptotype "Res Longa" in English.

Among the class prototypes are such nouns as *stick, rod, prickle, arrow,* etc. All of them are polysemantic with all their meanings related to the solid long pointed shape (SLPSh) that fits the hand. Noun *stick* (OE **sticca*) is derived form the verb

[2] Davies, Mark. (2008-) The Corpus of Contemporary American English (COCA): 410+ million words, 1990-present. Available online at http://www.americancorpus.org.

(OE *stician* ← Teut. Root **stik* — 'pierce, be sharp'). Apparently, the entities that the prototypes named were strongly associated with the main functions of the pointed shape: its penetrating capacity. Hence the cultural value of (SLPSh) that of weapon, tools, symbols of power including religious ones.

Let us consider the classifiers of the cryptotype "Res Longa".

- Subject transitive classifiers [a pointed object pierces smth.], [a pointed object penetrates smth.], [a pointed object punctures smth.], [a pointed object pricks smth.], [a pointed object punches smth.], [a pointed object spears smth./smb.], [a pointed object sticks smb.].

The collostruction [smth/smb stings smb.] requires special attention. The question arises if it is apt to diagnose the class membership of a noun? The argumentation for its status as a classifier of the cryptotype Res Longa is as follows. The OED defines *to sting* as 'to stick, stab, pierce with a sharp-pointed weapon, to prick with a small point'; in O.E. *stingan* "to prick with a small point" (of weapons, insects, plants, etc.). This verb is used in different senses in Contemporary American, but the primary (or the core) meaning, which contributed to metophorization of abstract entities, remains and regulates the abstract nouns usage. Most native speakers would agree that this verb is associated not only with insects and jellyfish, but also with the long-pointed form as in Example 1.

(1) A sharp object stung me as I passed through the meadow.
A pointed object stung me as I backed into a cactus.
A tip from the broken glass shards stung me.
The needle stung me. A prickle stung me.

- Subject intransitive classifiers [a pointed object stabs (in/into/through smth.)], [a pointed object thrust through /toward smth.], [a pointed object jabs at smth./smb.], [a pointed object penetrates into smth.], [a pointed object sticks (in/into/from smth.)], [a pointed object points at smth.].
- Object transitive classifiers [smb. pricks a pointed object into smth.], [smb. punches a pointed object into smth.], [smb. stabs a pointed object in/into/somewhere], [smb. thrusts a pointed object in/inside/into/at smb./smth.]. [smb. jabs a pointed object at/into smth./smb.]. [a pointed object is thrust in/inside/into/at smb./smth.].
- Instrumental classifiers [to pierce smth / to be pierced with/by a pointed object], [to puncture smth. with a pointed object], [to prick/be pricked with/by a pointed object], [to punch smth. with a pointed object], [to stab smb. with a pointed object], [to thrust with a pointed object], [to stick smth./smb. with a pointed object], [spear/be speared with/by a pointed object], [to jab/be jabbed with a pointed object].
- Predicative classifiers [a pointed object is sharp], [a pointed object is poignant].
- Attributive classifiers [a sharp object], [a poignant object], [a pungent object].

The attributive collostruction [a pungent object] merits its own discussion, as neither the corpus data nor the native speakers associate its meaning with SLPSh

of an entity. According to the OED *pungent* in the meaning 'pricking, piercing, sharp-pointed' substituted *poignant* in many of its senses in the 17th century (Example 2). This meaning of the adjective has been specified in the course of time as 'the effect on the organs of smell or taste resembling that produced by pricking' or 'of the penetrating nature of smell, taste and sensation'. The direction of the broadening of its core meaning[3] can be presented as follows: from SLPSh entities (Example 2) to substances with 'smell or taste resembling that produced by pricking' (Example 3) and then further to abstract entities such as speech acts (word, advice, language), emotions (shock, grief) and memories (Example 4). Apparently, both senses of the word are metaphorical extensions of the core meaning. This argument could be in favor of the classifying status of the collostruction [pungent smth.].

(2) Between three thornes pungent (1601) (OED);
was whilome used for a pungent spear (1606) (OED);
< ...> cutting or pungent instruments (1750) (OED).

(3) A strong, sharp taste of pepper greets my tongue, along with something creamy. Risotto, I guess.
The mild flavor of the beans plays off the pesto's pungent taste and adds a delicious creaminess.
it had <…> sharp bitter flavour;
You can also coat wiring with a spray that deters biters with its pungent flavor, such as Grannick's Bitter Apple;
The flower gave off a sharp scent at night;
<..> a small creeping plant of a pungent scent;
The sharp smell of her perfume made me dizzy;
Beau recognized the sweet, pungent smell of something stronger than tobacco.

(4) <…> perform oral sex. But U. S. District Judge Susan Webber Wright, in unusually sharp language, ruled Jones, " has filed to demonstrate that she has a case;
<…> little they know or have guessed about sex. Ma doesn't use such pungent language with her daughters. They might remember her by those words alone;
Removed from the transcript, these sources say, were Ambassador Glaspie's sharp advice to Saddam not to use force in his dispute with Kuwait <…>;
Congressman James Traficant of Ohio offered some of the most pungent advice imaginable, and believe me you don't want to imagine it;
I lost two very close friends of mine, and that was a real sharp shock because I realized if I carried on, I'd be next;
and we followed, holding our noses. The bathroom was a pungent shock of Lysol;
That was four years ago, but the memory was as sharp as ever;
treated and painted their curved inner walls until the gas stink was just a pungent memory. In an instant, memory pierced him.

[3] I thank an anonymous review for pointing out this issue.

- Substantive classifiers [the prick of a pointed object], [the punch of a pointed object], [a stab of/from/by a pointed object], [the thrust of a pointed object].

The collocation [a point of smth.] doesn't look reliable for diagnosing purposes as it has dubious acceptability. The noun *point* refers to the end of a pointed object as well as to a trace (a wound or a sign or a series of signs) a pointed object leaves on/in another object. Such signs tend to be used for measuring something that is viewed as a continuum (emotions, conditions, progress in studies, efficiency, etc.). However, it can be a feasible classifier in a wider context. In example 5, *point* being the collexeme of the noun *pain*, is described as thin, capable of sliding, and penetrating. Hence, pain is categorized as a pointed object. Conversely, in Example 6, the N slot of the construction [be frustrated to N] is associated with the limit or a level of a continuum. Thus, depression is categorized as a (relatively high) level of an unpleasant state of being, rather than a pointed entity. Example 7 demonstrates constructions in which the noun *point* is assigned the meaning 'the sense of smth.'. Neither (6) nor (7) can be treated as classifiers of the cryptotype.

(5) the thin point of pain slid around her throat. It <pain> penetrated.
A sudden sharp point of pain at the back of her head <…>.

(6) Last year, frustrated to the point of depression, he secretly recorded a pop album.

(7) <…> this was the point of the dream;
What was the point of the reputation economy?

3.2. Core-periphery structure of RES LONGA

In languages with scarce morphology and lack of overt noun classes, nouns of abstract semantics in order to be communicated are likely to be attributed to one of the covert noun classes. Thus, cryptotype categorization is of non-taxonomic nature, which means that a noun tends to belong to more than one cryptotype but its proximity to the core of different classes is different. The cryptotype is structured on core-periphery basis in accordance with the *CRI* of nouns[4].

The cryptotype "Res Longa" incorporates nouns of abstract semantics, whose class membership is evidentially revealed in communication when these nouns "choose" the classifying collostruction(s) of the class.

---

[4] *Cryptotype Radius Index (CRI)* indicates the proximity of a noun to the Core of Cryptotype in terms of core–periphery proximity. In a model a cryptotype on core-periphery basis, the formula to calculate the CRI is the following: $CRI = \frac{V_{cor}}{\sum cl-s}$. V^{Cor} is the number of cryptotype classifiers the noun has co-occurred with in the corpus, and $\sum cl\text{-}s$ stands for all classifiers of a certain cryptotype. The algorism calculating the proximity of a noun to the core of cryptotype is described in (Boriskina 2010).

The restoration of "Res Longa" in English shows that out of 500 nouns of abstract dimension 124 nouns are attributed to this cryptotype. Below is the wordlist of some class members in decreasing sequence[5]:

- *light*
- *memory pain*
- *word question fear*
- *guilt image music reality*
- *anger attack force hope image reality relief sense sight song advice answer case change criticism disappointment evidence experience influence loss opposition problem voice work*
- *act awareness belief challenge comment condition contrast defense description desire dream equipment idea industry issue joy language love method mind pleasure pride quality recession style terror thought truth view etc.*

These nouns demonstrate different degree of linkage with the classifying collostructions. According to the COCA data there are nouns with strong ties (Example 8), such as *pain* (433 occurrences), *evidence* (413), contrast (622), *voice* (182), *light* (277), criticism (122), relief (116), *mind* (107). The nouns which occur in the cryptotype collostructions less than three times are those of with weak ties (Example 9).

(8) A piercing stab of bright yellow pain;
after the murder of his sister, pain punched him in the gut;
The figure's deep voice pierces the darkness;
He wants you, said Tony, his voice piercing Pat's thoughts like a spear;
"Jerry" she said, her voice sticking in her throat <…>;
Whose voice and judgment penetrates to the depths of a person's soul;
"Step away from the dead" said Maggot, his voice as sharp as his weapon;
As light pierced the dust he found himself inside the fort;
the ferocious light stabbed her eyes;
this exaggeration was deliberate, but <…> evidence points in that direction;
Hard evidence that sticks out like rock, evidence that can break bones;
Frequently good arguments and good evidence will puncture them, <…>;
as if those extra few inches of elevation could help his sight pierce the fog;
more he believed it, until it seemed to him like a sharp point of truth;
And I was left alone with truth too poignant to deny;
In Alice's case the truth had penetrated farther and revealed to her that;
A sharp thought scratched his cheek;
I experienced a stab of doubt and thought;
Since pure thought can penetrate the universe's mysteries;

[5] The decreasing sequence shows the proximity of a noun to the core of cryptotype. The whole wordlist is not included due to the limits on the paper format.

(9) She remembered the stab of terror she had felt when Chase came <…>;
I only remember the sudden and sharp terror of that moment;
With its mixture of idealism and limited but sharp violence, this latest uprising was more like the Ukrainian Orange Revolution than the Castro-style putsch.

4. Conclusion

The courpus-based study of cryptotypes of English nouns is aimed at systematization and formalization of separate ancient relicts of the naïve mapping of the world in language. RES LONGA is one of five cryptotypes reconstructed in English so far. The method demonstrated in the paper has several advantages. First, it can be applied to the study of covert noun classes in other languages and thus contribute to pursuing common grounds for the classification of nouns in typological perspective. Second, the results of the exploration of noun cryptotypes of English can have some implications for the theory of metaphor and Lexical Grammar of English. The data base of English cryptotypes (http://ling.dentry.ru) can be used in lexicography, for educational purposes, and in formulating and testing the hypotheses, related to a noun's potential collocability. Finally, studying cryptotypes can advance our understanding of the criteria necessary to computational representation of lexicon. We believe, that the adequately formalized description of cryptotypes can find application in computational modeling and text processing issues.

References

1. *Arutiunova N. D.* 1976. Sentence and its Sense.
2. *Boriskina O., Marchenko T.* 2010. An Algorithm for Analysis of Distribution of Abstract Nouns in Cryptotypes. Proceedings of the 2010 International Conference on Artificial Intelligence (ICAI 2010). July 12–15, 2010, II : 907–913.
3. *Dixon R. M. W.* 1968. Noun Classes. Lingua, 21 : 64–95.
4. *Fauconnier G.* 1997. Mappings in Thought and Language.
5. *Fillmore C. J.* 1985. Syntactic Intrusions and the Notion of Grammatical Construction. Proceedings of the Eleventh Annual Meeting of the Berkeley Linguistic Society : 35–55.
6. *Foundations* of African Linguistics: Noun classes. 1996.
7. *Gillon B.* 2005. Semantic Categorization. Handbook of Categorization in Cognitive Science.
8. *Givon T.* 1986. Prototypes: between Plato and Wittgenstein. Noun classes and categorization.
9. *Goldberg E.* 1996. Construction Grammar. Concise Encyclopedia of Syntactic Theories : 68–71.
10. *Grady J.* 1997. Foundations of Meaning: Primary Metaphors and Primary Scenes.
11. *Gries St., Stefanowitsch A.* 2004. Extending Collostructional Analysis: a Corpus-based Perspectives on Alternations. International Journal of Corpus Linguistics, 9(1) : 97–129.

12. *Hla Pe.* 1965. A Re-examination of Burmese Classifiers. Lingua,15.
13. *Hunston S., Francis G.* 2000. Pattern Grammar: a Corpus-driven Approach to the Lexical Grammar of English.
14. *Johnson M.* 1987. The Body in the Mind: The Bodily Basis of Meaning Imagination and Reason.
15. *Katsnelson S. D.* 1972. Language Topology and Thinking.
16. *Kay P., Fillmore C. J.* 1999. Grammatical Constructions and Linguistic Generalizations: the 'What's X doing Y' Construction. Language, 75 (1) : 1–33.
17. *Koval' A. I.* 1996. Noun Classes in Pular-fulfulde. Foundations of African Linguistics: Noun classes : 92–220.
18. *Kövecses, Z.* 2002. Metaphor. A Practical Introduction.
19. *Kövecses, Z.* 2005. Metaphor in Culture. Universality and Variation.
20. *Kretov A. A., Titov V. T.* 2010. The Role of Covert Categories in Typological Study of Grammar of Roman Languages. Proceedings of VSU. Series: Linguistics and Intercultural Communication, 1 : 7–12.
21. *Lakoff G.* 1986. Classifiers as a Reflection of Mind. Noun classes and categorization.
22. *Lakoff G., Johnson, M.* 1980. Metaphors We Live By.
23. *Lakoff G.* 1987. Women, Fire and Dangerous Things: What Categories Reveal about the Mind.
24. *Lutskov A. D.* 1996. Noun Classes in Bantu and Loanwords. Foundations of African Linguistics: Noun Classes : 75–91.
25. *OED:* Oxford English Dictionary on CD. Version 3.1. 2009.
26. *Stefanovich A., Gries St.* 2003. Collostructions: Investigating the Interaction of Words and Constructions. International Journal of Corpus Linguistics, 8(2): 209–243.
27. *Toporova I. N.* 1996. Noun Classes in Bantu Family. Foundations of African Linguistics: Noun Classes : 24–74.
28. *Uspenskii V. A.* 1997. Material Connotations of Abstract Nouns. Semiotics and Informatics, 35 : 146–152.
29. *Whorf B. L.* 1956. Language, Thought and Reality. Selected Writings of Benjamin Lee.

ПАРАМЕТР БЛИЗОСТИ В МЕТАФОРИЧЕСКОМ ПРОСТРАНСТВЕ

Е. Г. Борисова (egbor@mail.ru)
МГПУ, Москва, Россия

Т. Е. Овчинникова (teomax@ya.ru)
МГЛУ, Москва, Россия

В работе разрабатывается представление об использовании дейктических средств как показателей пространственных отношений для функции модальных (усилительных) частиц. Связь с дейктической функцией позволяет говорить о метафоризации пространственных отношений, переносе их параметров на отношения в области дискурса, общих знаний говорящего и слушающего и более тонких смысловых отношений, которые связаны со степенью важности для говорящего и слушающего и могут метафоризоваться через понятия «пространство говорящего», «общее пространство говорящего и слушающего».

Ключевые слова: дейктические средства, дейктическая функция, модальные частицы, усилительные частицы, метафора.

PARAMETER OF NEARNESS IN THE METAPHORICAL SPACE

E. G. Borisova (egbor@mail.ru)
The Moscow City Teachers' Training University, Moscow, Russian Federation

T. E. Ovchinnikova (teomax@ya.ru)
Moscow State Linguistic University, Moscow, Russian Federation

In this work a conception of using deictic means as indicators of spatial relations for function of modal (intensifying) particles is developed. The relation with the deictic function implies metaphorisation of spatial relations and transfer of their parameters on relations in the fi eld of a discourse, the speaker's and listener's general knowledge and more delicate semantic relations connected with the degree of importance for Speaker and Listener. They are able to be metaphorised through the concepts "speaker's space", "speaker's and listener's common space". Obviously, the modal particles VOT and VON are connected with index (spatial) particles. The modal

meanings, different from index ones, are the approximation meanings, the intensifying meanings and a number of other ones. We are guided by the opposition of the spatial index particles VOT and VON connecting “indication near subject — indication distant subject” with opposition. As identifi cation is an action quite widespread in metaphorical space, we often meet with the use of VOT for indicating an object or phenomena so that we can speak about metaphorical “sense space” (the thesaurus of conversation participants) or “speech space” (i. e. the semantic network of discussed events).

Kew words: deictic means, deictic function, modal particles, spatial index particles, metaphora.

Пространственная метафора — одна из самых базовых в когнитивной науке (что, возможно, связано с тем, что отражение реальности происходит на пространство, каковое образует кора головного мозга). Ее изучение может идти с двух сторон: со стороны понятий (изучение представлений о пространствах на основании общих моделей когнитивной деятельности) и при анализе конкретных языковых данных, интерпретация которых допускает пространственное представление. В нашем случае мы идем от языковых данных, для которых метафорические связи достаточно прозрачны.

Очевидно, связаны с указательными (пространственными) частицами модальные частицы ВОТ и ВОН. (История их изучения была рассмотрена в наших работах Овчинникова 2007 и Борисова 1999 и здесь не описывается). Модальные значения, отличные от указательных, это значения примерности *(1) Вон любовь сколько страданий несет*, значение усиления *(2) Вон ты какая сдобная*, а также значения, близкие к синтаксической обусловленности *(3) Вот не захочу — и не сделаю, (4) Мне пообещали, вот я и пошел* и т. п. и ряд других. Последние свойственны только частице ВОТ.

Мы опираемся на противопоставление пространственных указательных частиц ВОТ и ВОН, которые связывают с оппозицией «указание на ближний предмет — указание на дальний предмет». Реально в русском языке это противопоставление теряет четкость, в некоторых текстах, цитируемых по памяти, одну частицу заменяют на другую: в Интернете представлены варианты *(5) Вон моя деревня* («Вот моя деревня, вот мой дом родной» из стихотворения «Деревня» И. З. Сурикова). Видимо, противопоставление между этими частицами осуществляется по другому признаку, который коррелирует с удаленностью, но не совпадает с ней полностью. Отметим, что несовпадение оппозиции ВОТ — ВОН с противопоставлением «близкое — далекое» было сделано еще Ю. Д. Апресяном в работе «Дейксис в лексике и грамматике» [Апресян 1986]. Академик Апресян связывал это противопоставление с важностью/ неважностью указываемого объекта для говорящего.

Схожие соображения мы можем найти у представителя иной научной дисциплины — психологии. Н. А. Алмаев, рассматривая различение между указательными частицами ВОТ и ВОН, говорит следующее: «Об объектах, о которых говорится ВОТ, хочется сказать, что они находятся вблизи, тогда как об объектах, находящихся «вдали», хочется сказать ВОН. Легко убедиться, что «близь»

и «даль» в данном случае мало зависят от реального расстояния. При повышенной значимости объектов вполне можно сказать: *(6) Вот за десять километров отсюда скачет всадник*, но — *(7) Вон у стены, в полутора метрах отсюда, лежит тапок* — при незначительной заинтересованности. Поэтому мы должны определить «близь» и «даль» с чисто формально-интенциональной точки зрения: «близь» как область относительно большего, «даль» — малого прироста импрессий» [Алмаев, 2007]. (Под импрессиями психолог понимает возможность восприятия новых порций сведений о реальности). И хотя примеры, приведенные психологом, представляются не вполне естественными, в принципе с возможностью использования частицы ВОТ для обозначения далекого объекта, а ВОН — близкого, можно согласиться.

Более точным было бы считать, что ВОТ и ВОН подразумевают два разных вида указаний. ВОТ отождествляет предмет (лицо, место и даже действие или качество) с наименованием: *(8) Вот это стул. На нем сидят*, а ВОН задает способ нахождения нужного объекта.

Поскольку процесс отождествления может рассматриваться с разных позиций, уточним, что имеется в виду момент непосредственного соотнесения (дейксиса), при котором наименование предполагается уже известным *(9) Вот сканнер, а вот это принтер* или, по крайней мере, относящимся к известному множеству: *(10) А теперь познакомимся с типами рубанков. Вот шерхебель, цензубель, а вот это — фуганок*. Как замечено в [Падучева 1985, с.158], указательная частица ВОТ способна обеспечить дейктическое употребление и тех слов (в частности, личных местоимений), которые без нее едва ли могут использоваться в дейктическом (а не референциальном либо ином) смысле: *(11) Вот он пусть подойдет*.

Как отмечено и в словарях, указание возможно не на реально видимый, а представляемый предмет: *(12) Вот злонравия достойные плоды*. Однако во всех случаях отождествление не предполагает особых усилий по поиску нужного объекта. Чаще всего такое имеет место при достаточной близости предмета.

Если теперь попытаться дать толкование частице ВОТ, то лучше всего прибегнуть к так называемому процедурному типу толкования [Паршин 1988], когда описание смысла подается в виде инструкции для адресата. Именно такой тип толкований достаточно часто используется последние двадцать-тридцать лет для описания семантики служебных (дискурсивных) слов. Тогда толкование могло бы принять приблизительно такой вид: «ВОТ X = ‘Говорящий дает возможность адресату установить, что в реальности (или в представлении, разделяемом участниками коммуникации) соответствует имени (или высказыванию) X’».

Психологическая интерпретация Н. А. Алмаева отмечает только одну, не обязательную характеристику этой частицы: в случае ВОТ «объект сперва опознается, затем воспринимается в несколько большей подробности. Формально говоря, эта частица содержит момент идентификации 1) до и 2) после увеличения числа удерживаемых импрессий объекта 3) — соответствующее ожидание увеличения импрессий. В увеличении числа удерживаемых

импрессий — различие между ВОТ и ВОН. В последнем случае такого ожидания нет» [Алмаев, 2007]. Заметим, что такое описание вполне подходит для примеров, когда фраза типа *Вот X* имеет продолжение. Однако не исключено, что продолжение не последует: *(13) Вот эта проходная, кто спрашивал*, и тогда об ожидании или о большей подробности говорить уже трудно.

Иначе обстоит дело с указанием ВОН, которое служит признаком (или инструкцией) поиска нужного объекта. Естественно, вопрос о поиске встает обычно тогда, когда этот объект расположен удаленно. Однако возможны и ситуации поиска в относительно близком пространстве: *(14) Вон, видишь, мой дом. Вон светится моё окно* (анекдот из сети). Нередко одна и та же задача — указать на требуемый объект — может решаться обоими способами*: (15) Вот мое окно (вон мое окно).* Это объясняет и отмеченную выше замену *(16) Вон мой дом родной*. Очевидно, что автор стихотворения имел в виду распространенный в художественной литературе способ образного представления объекта: одно за другим всплывают перед глазами сначала деревня в целом, потом дом и т. п. (Тот же способ использован, к примеру, А. А. Блоком: *(17) Ночь. Улица. Фонарь. Аптека*). А не очень искушенный читатель мог воспринять подачу этой информации более тривиальным способом, как если бы ему указали на «дом родной» из окна автомобиля, что действительно предполагало бы поиск, выбор из ряда нескольких похожих объектов.

Если давать толкование частице ВОН, то его тоже следует выполнить в процедурной традиции: «ВОН X= ‘Говорящий предлагает адресату каким-то образом (скорее всего, последовав взглядом за движением руки или головы) найти X’».

А теперь вернемся к использованию указательных частиц в метафорическом смысле, что описано в работах [Овчинникова 2007, Овчинникова 2009], где дейктический характер частиц ВОТ и ВОН в исходном смысле позволил рассматривать их переносное употребление как расширение пространственного значения на семантические объекты (тезаурус, содержание речевого акта и т. п.).

Поскольку отождествление — действие вполне распространенное в метафорическом пространстве, мы часто встречаем употребление ВОТ для указания на какие-либо объекты или явления, что позволяет говорить о метафорическом «пространстве смысла» (тезаурус участников беседы), ср. *(18)* ***Вот*** *Вася какой хороший* или «пространства речи» (т. е. семантическая сеть обсуждаемых событий), ср. *(19)* ***Вот*** *об этом я и хочу поговорить*.

Поэтому описание разницы в употреблении частиц ВОТ и ВОН в модальных (усилительных) функциях можно сопоставлять с понятиями близости и отдаления, но с учетом описанных различий в их семантике.

Начнем с так называемого «примерного» употребления. Оно четко выделено для ВОТ.

(20) — *Как может в клинике не быть телефона? Ну, а если сейчас что-нибудь случится?* ***Вот*** *со мной, например.*

(21) — *Ефрем! Хватит скулить. Возьми-ка* ***вот*** *книжку почитай.*

(22) — *Сколько свободного времени! И на дежурствах — можно книжечку почитать, можно* ***вот*** *поболтать. (Солженицын. Раковый корпус.)*

Однако имеется некоторое количество примеров, показывающих, что и частица ВОН здесь возможна. Причем, употребления можно признать синонимичными, но полного смыслового совпадения нет:

Ср. *(23) — И зачем ты ввязался?* ***Вон*** *Николай не стал ни о чем сообщать — его и не трогают*. Пример с ВОТ явно требует большей актуализованности для собеседников приводимого в пример Николая.

Интересно, что даже указание на близость объекта говорящему или слушающему не может помешать употреблению ВОН:

(24) — *Да я-то что?* ***Вон*** *Соня у нас… А Соня всё чаще оставалась дома или, посортировав почту полдня, уходила с работы. (Виктор Астафьев. Обертон (1995–1996) НКРЯ)*

(25) — *Да уж задавай-не задавай, а домой не вернешься. Очки* ***вон****, можешь вернуть. Пижаму новую. (Солженицын. Раковый корпус.)*

Однако важно, что обращение к этому объекту неожиданно для собеседников, что означает удаленность, необходимость поиска не в реальном, а метафорическом «пространстве речи» (Отметим и различие в интонировании фраз с ВОН и с ВОТ). Поэтому в примерном значении ВОН с названиями участников общения обычно не употребляется, ср. нежелательное *(26)* ***Вон*** *я, например*. Здесь явным образом актуализуется тот фрагмент значения ВОН, который связан с удаленностью, необходимостью поиска, отмеченный нами для прямого указательного значения этой частицы.

Но вполне возможно употребление ВОН с местоимениями первого и второго лица в значении подчеркивания качества (в сочетании с *какой, где* и т. п.):

(27) ***Вон*** *я какой — в шляпе, при галстуке… [Василий Шукшин. Калина красная (1973) НКРЯ]*

(28) ***Вон*** *ты какой у меня вырос, чисто отец. [А. И. Мусатов. Стожары (1948)] НКРЯ]*

Примеры с частицей ВОН содержат оттенки удивления, нечто вроде приглашения взглянуть со стороны [Овчинникова 2007], что соответствует значению удаленности. Употребление же ВОТ, предполагающее отождествление с уже представляемым свойством, вносит иные оттенки. Для первого лица *(29) Вот я какой*! это скорее самодовольство, хвастовство, бахвальство. Например:

(30) *А у тени ушки были длинные-длинные, как большие рога. —* ***Вот*** *я какой, — крикнул зайка. — Давай бодаться! [Валентин Постников. Колыбельная сказка (1997) НКРЯ].*

Для второго лица *(31)* ***Вот*** *ты какая*! — более разнообразная гамма эмоций, которая появляется вследствие импликатур из факта сообщения об отождествлении говорящим слушателя с соответствующим свойством.

(32) *Но, присмотревшись, Катерина Федосеевна заметила открытую форточку и следы грязных лап на стекле изнутри и снаружи окна. —* ***Вот*** *ты какая у меня лазунья! — сказала она. [Александр Яшин. Подруженька (1965) НКРЯ]*

Таким образом, получается, что особенности метафорического употребления ВОТ и ВОН достаточно прочно связаны с их значениями в указательном смысле, причем не просто с противопоставлением по степени дальности, а именно со способами указания. Поэтому это противопоставление тоже метафоризуется. И мы можем говорить, что частицы указывают на разные способы поиска объектов в метафорических пространствах. Перед нами были пространства смысла (тезаурус говорящего и слушающего), пространство общения — смысла сообщений, появившихся в общении. Во всех случаях понятие близости и дальности (как производные операций отождествления и поиска) связано не с участниками общения, а с расстоянием между актуальными представлениями, реализуемыми в данный момент общения, и новыми привлекаемыми для семантических, эмоциональных и других задач понятиями, объектами, свойствами.

References

1. *Almaev N. A.* 2007. Meaning Understanding and Expression in Human Speech [Ponimanie I Vyrazhenie Znachenii v Rechi Cheloveka].
2. *Apresian Iu. D.* 1986. Proceedings. V.2 : 636–637.
3. *Borisova E. G.* 1999. Principles of Reserved Words Description [Printsipy Opisaniia Sluzhebnykh Slov]. Vestnik Moskovskogo Universiteta, 9 (2) : 71–85.
4. *Borisova E. G., Ovchinnikova T. E.* 2005. Intensification Spaces (Spatial Metaphora and the Origin of Intensification Particles) [Prostranstva Usileniia (Prostranstvennaia Metafora I Vozniknovenie Usilitel'nykh Chastits)]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2005" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2005").
5. *Ovchinnikova T. E.* 2009. Spatial Metaphora in the Semantics of the Particles of Deictic Origin [Prostranstvennaia Metafora v Semantike Chastits Deikticheskogo Proiskhozhdeniia].
6. *Ovchinnikova T. E.* 2007. Meaning Space Segmantation (On the Material of Intensification Particles) [Chlenenie Prostranstva Smysla (Po Dannym Usilitel'nykh Chastits)]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2007" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2007").

7. *Paducheva E. V.* A 1985. Statement and its Relation with the Reality (Referential Aspects of Pronouns Semantics) [Vyskazyvanie I ego Sootnesennost' s Deistvitel'nost'iu (Referentsial'nye Aspekty Semantiki Mestoimenii)].
8. *Parshin P. B.* 1988. Comparative Separation as a Communicative Category (Experiment of Procedural-Semantic Description) [Sopostavitel'noe Vydelenie kak Kommunikativnaia Kategoriia (Opyt Protsedurno-Semanticheskogo Opisaniia)].

УЗНАТЬ ИЛИ КУПИТЬ? КЛАССИФИКАТОР СТРАНИЦ ОБЗОРОВ И ИНТЕРНЕТ-МАГАЗИНОВ

П. И. Браславский (pb@yandex-team.ru)
Яндекс

Ю. А. Киселев (yurikiselev@yandex-team.ru)
Уральский Государственный Университет,
Екатеринбург, Россия

Ключевые слова: обзоры, интернет-магазины, запросы, классификация

TO FIND OUT OR TO BUY? PRODUCT REVIEW VS. WEB SHOP CLASSIFIER

P. I. Braslavskii (pb@yandex-team.ru)
Yandex

Iu. A. Kiselev (yurikiselev@yandex-team.ru)
Ural Federal University, Ekaterinburg, Russian Federation

In this paper we examine two categories of search results retrieved in response to product queries. This classification reflects the two main kinds of user intents — product reviews and online shops. We describe the training and test samples, classification features, and the classifier's structure. Our findings demonstrate that this method has good quality and performance suitable for real-world applications.

Key words: review, web shop, queries, classifier

1. Introduction

Recently the diversity of search results has gained the attention of information retrieval researchers and practitioners. When considering the diversity of search results we shift the emphasis from the relevance of a single query-document pair to the relationships between documents in the result list and search engine results

page (SERP) as a whole. The diversity of search results has many aspects [1] that are associated with the incompleteness of available information. Ambiguous queries are a classic example, when a diverse SERP could compensate for the lack of knowledge about the actual needs of the user. For instance, the Russian query [*алые паруса*] (*scarlet sails*) may refer to the novel by Alexander Grin, its screen version of 1961, a retail chain, a residential complex in Moscow, or a school graduation day celebration in St. Petersburg. When it is impossible to disambiguate the query, we can try to organize the result list the way it reflects the different intents of the query (how to identify these intents and what method to choose to structure SERP are other problems). Another aspect is the ambiguity associated with the actual users. For example, the query [*fixed assets amortization*] might be issued by an experienced accountant or a college student doing her/his homework. Accordingly, if we cannot obtain additional information, the search results may contain documents addressing topics on different levels. Another consideration is the *genre* variety of documents in the results: documents on the same topic, but of different types. For example, the results for [*large hadron collider*] may contain both news and popular science articles.

In our work, we consider the problem of diversity for queries about the products traditionally offered in online shops. The spectrum of these products is well presented on the shopping comparison service Yandex.Market (http://market.yandex.ru) and includes electronics, photo and home appliances, mobile phones, computers, etc. The typical examples of online shopping queries are queries like [*samsung g400*], [*home air conditioner*], and [*netbooks review*]. We estimate the share of such queries at about 4% of the whole query stream on Yandex. The range of users' needs behind such queries can be quite broad. However, the majority of users either want to: 1) know what is being offered, make the choice, examine the product's characteristics, compare it with similar products — these are the steps usually leading to the purchase of a product presented on Yandex.Market — or 2) make the actual order or purchase. These user intents correspond to two types of documents: 1) online product surveys and reviews and 2) webshop pages where users can make an order. Of course, these intents do not cover the full range of users' needs — people may use the same queries to search for technical documentation, spare parts, service and repair shops, accessories, software for devices, classified ads, etc. However, the two aforementioned needs are prevalent.

In our work, we are not offering methods to achieve search result diversity, but showing, instead, how to create preconditions for it by addressing the problem of classifying web documents into reviews and online shops (see for example [2] on diversity-based search results optimization). We define as "reviews" detailed and thorough professional or editorial reviews, while excluding short user opinions. Digital Photography Review (http://dpreview.com/) is a good example of such content.

In the rest of the paper we give an overview of related work on web document classification (Section), describe the requirements for and the resulting structure of classifiers (Section), specify our data (Section), define classification features (Section 0), and present evaluation results (Section 6). Section 7 is the conclusion.

2. Related work

Various web page classifications are widely used in web applications, including web search. Web document categorization is used to improve search quality, build vertical searches, filter spam, to categorize user queries, etc. In contrast to traditional methods of text document classification, web page classification can be based on a wider range of features including those based on document structure, HTML tags, metadata, hyperlinks, URLs and user behavior. The problem of classifying web documents can be complicated by clutter such as advertisements, navigation bars, etc. Since the pioneering work by Joachims [4] SVM (Support Vector Machine) is a method widely used to classify text documents.

Page classification into reviews and online shops is an example of genre classification. A detailed survey of the approaches to and methods of genre classification is presented in [5]. At least two noteworthy papers dealing with the analysis of web documents appeared after the survey had been published. Meyer zu Eissen and Stein [6] conducted a user study spawning a set of eight web genres useful for web search, and built a corpus containing these genres. Along with the surface and linguistic features traditionally used in genre analysis, their study employed HTML-based features. The method was implemented as a plug-in for the Firefox browser that enriches Google snippets with genre labels [7]. Lim et al. [8] expanded this approach even further and made use of a wider range of features (326 in total), including various surface, lexical, syntactic, HTML, and URL features.

Mindset, a Yahoo! research project [9], allowed users to rank search results based on their commercial or informational value. In addition to the standard query box, Mindset had a slider that the user could move between «shopping» and «researching», changing the appearing results from less to more commercial. Unfortunately, the project is now closed, and the implementation details have not been published.

Dai et al. [10] solved the problem of detecting user's online commercial intention. In order to accomplish this task they constructed a classifier of commercial and noncommercial web pages. Classification was performed using SVM in the space of terms, term occurrences in the document's body and HTML-tags were counted separately (thus, n terms generated $2n$ features). The training sample contained 5,375 pages, 2,820 of them were labeled as commercial. The authors obtained good results with precision 93.0 % and recall 92.5 % for the commercial pages class. The demo classifier is available online [11].

Paper [12] describes a simple client-side tool that classifies commercial (i. e. online shops' product pages) vs. noncommercial pages. Classification is performed based on different features: presence of images and product descriptions, indication of price, "buy" button, URL, etc. Classification is followed by product name and price extraction.

The problem of filtering product reviews from search results is addressed in [13]. The task was solved based on result snippets: experimental dataset contained 1,200 Google snippets for queries in the form [*product_name* + *"review"*]. The features used

for classification were terms in the title, URL, and snippet itself. The final classifier combined the result of SVM classifier and heuristic rules.

Product review classification, based on label propagation over click graphs was considered, among other classification problems, in [14]. A sample of 10,000 positive and negative examples was used for learning with gradient boosting of decision trees. Different features were used: text (unigrams and bigrams in various structural parts of the document, the number of words in the document, the number of capitalized words), link (properties of incoming and outgoing links), URL (length and presence of certain tokens), and HTML features (presence of specific tags). The best results for review class reported by the authors: precision — 63.96%, recall — 73.97%.

3. Classifier

Our goal was to build a classifier suitable for a large-scale web search engine capable to process billions of web pages in reasonable time. So, performance was as crucial as the quality of classification. Consequently we were restricted to employ only light-weight features that could be extracted by one-pass page scan. We opted for embedding the classifier into the search engine's indexing pipeline. Even though it led to even harder efficiency restrictions, we could easily employ tokenization, lemmatization, language detection and other results available at indexing time.

For learning we used LIBSVM [15], an implementation of SVM. To compose a three-class classifier out of binary classifiers (`shop − other`, `review − other`) we had two options:

- ***Parallel classifier.*** The page is processed by both classifiers independently. As a result, some pages can be assigned to both classes (`shop` and `review`).
- ***Sequential classifier.*** Negative (`other`) output of the `shop` classifier is fed to `review` classifier.

In fact, these options differ insignificantly. In both cases we had to extract all features at once (see Section 5). Since web shop pages account for about 4% (see Section 6) of the web (the reviews share is much less), sequential scheme does not save much computations.

4. Data

To classify a significant portion of the indexed documents (excluding only documents in a language other than Russian and very short documents), we constructed problem-driven training and test sets consisting of the documents returned to product queries on Yandex. This approach supposes that we can automatically detect queries of the target class. The problem of classifying queries is beyond the scope of this paper (for example, [3] describes a method for detecting product queries with high precision and recall).

In order to build the training sample, we randomly sampled 100 queries from the list of manually tagged product queries. For each query we downloaded top10

documents from Yandex SERP. The total number of downloaded pages was 979 (some pages were inaccessible and other were filtered out as non-Russian). The set was labeled by a Yandex assessor. Each web page had to be assigned to exactly one class: `shop`, `review` or `misc`. If a page had properties of both the `shop` and the `review` class (e. g. a shop page with a detailed product description), then it had to be labeled as `shop` (i. e. `shop` label overrides `review` label). Table 1 shows the break-down of the sample.

Table 1. Learning sample for shop classifier

Class	# of pages
Shop	301
Review	87
Misc	591
Total	979

Initial experiments with this sample showed that its size does not allow for a learning review classifier of a satisfactory quality. So, we used this sample for the learning `shop` classifier only. To train the review classifier, we composed a synthetic learning sample. It contained 150 `review` pages, 150 miscellaneous pages from the initial training sample labeled as `misc`. Also, we added 50 long documents collected manually (biographies, encyclopedia entries, etc.). Table 2 shows this breakdown.

Table 2. Learning sample for review classifier

Class	# of pages
Review	150
Misc	150
Long docs	50
Total	350

The test sample was obtained the same way as the `shop` training sample: we downloaded and labeled top10 from the Yandex results for 100 product queries. Table 3 shows the structure of the test sample.

Table 3. Test sample

Class	# of pages
Shop	431
Review	101
Misc	557
Total	1089

5. Classification features

5.1. Shop classifier

We used different feature groups for classification: term, textual, lexical, HTML, and URL features.

Term features. We identified the most informative term-features based on *mutual information*. For performance reasons, we did not consider the semantic or the visual structure of a document (document's main content, navigation, headers, footers etc.). As expected, the most contrasting terms were *магазин, рубль, каталог, цена, прайс,* and *корзина* (*shop*, *ruble*, *catalog*, *price*, and *basket*). The full list of terms used for classification consisted of about one hundred terms.

HTML features. The main high-level feature of a shop page is a possibility to make an order. We used two features aimed at detecting the "buy" button:

- number of specific keywords (*купить — buy, заказать — order*, etc.) in links and buttons;
- number of HTML-tags (*img*, *button*, etc.) with words "*cart*", "*basket*", "*order"* etc. in attributes.

Lexical features. We used the list of trademarks and brands on the Yandex. Market comparison shopping service (excluding commonly used words and the names consisting of two and more words). This list generated two features: the number of words from the list on the page and the number of unique words from the list.

URL feature. Many tokens in URLs are good cues for classification of a page as a web shop. This feature reflected the number of specific terms, such as *product*, *shop*, *itemID*, etc. in the URL.

5.2. Review classifier

Term features. By analogy with the shop classifier, we selected the most informative terms for the review classification. Since lexical variety of reviews is much higher than that of shop pages, the list of contrasting words was much longer and exceeded 7,000 words. The most informative terms for review class were *рынок, взгляд, автор, обзор, комментарий, маленький,* and *китайский* (*market*, *view*, *author*, *review*, *comment*, *small*, and *Chinese)*.

Textual features. Textual features were document's length in words and characters and sentence length distribution.

Lexical features. The list of 165 manually collected appraisal adjectives — *хороший, прекрасный, великолепный, плохой, отвратительный, ужасный,* etc. (*good*, *excellent*, *magnificent*, *bad*, *disgusting*, *awful*, etc.) — produced two features: the total number of words from the list and the number of unique words.

6. Results

Classification results with various feature groups for the test sample are presented in tables 4 and 5.

Table 4. Online shop classification results

Set of features	Precision	Recall
Terms only	0.918	0.809
HTML features only	0.894	0.491
Term + HTML features	0.934	0.800
Term + lexical features	0.910	0.807
Term + URL features	0.876	**0.856**
All features	**0.937**	0.837

Table 4 shows that classification based on terms only produced good results. Adding HTML markup features, i. e. detecting "buy" button, increased precision of classification. These findings support the results shown by Dai et al. [10]: term features and HTML tags work well even with a learning sample of a modest size. The observation that lexical features generated from the list of vendors and brands impair the quality of classification can be explained by the fact that almost all pages returned in response to a commercial query already contain brand names. The features would probably have increased the quality, if we evaluated classification results on a sample of random web pages. Adding URL features reduced precision, but increased recall. A set of all presented features provided the best precision for `shop` class (0.937).

Table 5. Review classification results

Set of features	Precision	Recall
Terms only	0.644	0.861
Term + URL features	0.643	0.841
Term + lexical features	0.625	0.861
Term + textual features	**0.681**	**0.891**

As expected, the quality of `review` classifier was much lower, considering the diversity of the class members and the shallow features we used. The lexical and URL features did not contribute to classification quality. The term and textual features provided the best precision for `review` class (0.681).

Tables 4 and 5 show the results of parallel classification (i. e. the entire test sample was processed by both classifiers, see Section). Superimposition of classifiers' results showed that only 16 pages were assigned both to `shop` and `review` (all these 16 pages were labeled as `shop` by a human assessor). The results of the three-class

classifier (`shop` label overrides `review` label) are shown in Table 6 (true classes in rows, classification output in columns).

Table 6. Confusion matrix of the three-class classifier

	Shop	**Review**	**Misc**	**Recall**
Shop	361	3	67	**0.84**
Review	1	90	10	**0.89**
Misc	23	23	511	
Precision	**0.94**	**0.78**		

To check the hypothesis that the `shop` classifier will perform well even on arbitrary documents (not only on documents returned to specific queries), we sampled randomly 56,768 Russian pages from Yandex's index. 2,071 pages were automatically labeled as `shop`, 1,908 of the labels (3.6% of the initial sample) were approved by a human, which resulted in precision 0.92.

7. Conclusion and future work

In this paper we presented a genre classifier classifying search results retrieved by product queries into two classes reflecting the two main intents of the user — product reviews and online shops. The aim of this classification is to compensate for the lack of knowledge about the actual needs of the user by providing a diversity of search results.

In the future our work will center around:

- information extraction from web shop and product review pages: product name, its category, price, etc.;
- improving quality of the product review classification. To bootstrap the results, we plan to calculate linguistically richer features in off-line mode;
- investigating the possibilities on taking page segmentation into account (i.e. page main content, navigation, etc.) to improve classification accuracy, as some studies on web page classification suggest.

References

1. *Agrawal Rakesh, Gollapudi Sreenivas, Halverson Alan, Leong Samuel.* 2009. Diversifying Search Results. WSDM '09 : 5–14.
2. *Cattelan Renan, Kirovski Darko, Vijaywargi Deepak.* 2009. Serving Comparative Shopping Links Non-invasively. Proceedings of the Web Intelligence and Intelligent Agent Technologies : 498–507.
3. *Dai Honghua (Kathy), Zhao Lingzhi, Nie Zaiqing, Wen Ji-Rong, Wang Lee, Li Ying.* 2006. Detecting Online Commercial Intention (OCI). WWW'06 : 829–837.

Detecting Online Commercial Intention, available at: http://adlab.msn.com/Online-Commercial-Intention/Default.aspx
4. *Eissen Sven Meyer zu, Stein Benno.* 2004. Genre Classification of Web Pages. KI 2004 : 256–269.
5. *Joachims Thorsten.* 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. ECML-98 : 137–142.
6. *Kim Soo-Min, Pantel Patrick, Duan Lei, Gaffney Scott.* 2009. Improving Web Page Classification by Label-Propagation Over Click Graphs. CIKM '09 : 1077–1086.
7. *Li Xiao, Wang Ye-Yi, Acero Alex.* 2008. Learning Query Intent from Regularized Click Graphs. SIGIR '08 : 339–346.
8. *LIBSVM,* available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm/
9. *Lim C. S., Lee K. J., Kim G. C.* 2005. Multiple sets of Features for Automatic Genre Classification of Web Documents. Information Processing & Management, 41 : 1263–1276.
10. *MindSet,* available at: http://research.yahoo.com/node/1912
11. *Radlinski Filip, Bennett Paul N., Carterette Ben, Joachims Thorsten.* 2009. Redundancy, Diversity and Interdependent Document Relevance. SIGIR Forum 43, 2 (December 2009) : 46–52.
12. *Santini M.* 2004. State-of-the-art on Automatic Genre Identification. Technical Report
13. *ITRI-04–03,* 2004, ITRI. available at: ftp://ftp.itri.bton.ac.uk/reports/ITRI-04-03.pdf
14. *Thet Tun Thura, Na Jin-Cheon, Christopher S. G.* 2007. Khoo Automatic Classification of Web Search Results: Product Review vs. Non-review Documents. ICADL'2007 : 65–74.
15. *WEGA* (Web Genre Analysis) project, available at: http://www.webis.de/research/projects/wega

«ФУНКЦИОНАЛЬНЫЙ» СТАНДАРТ В РУССКИХ И АНГЛИЙСКИХ СТЕПЕННЫХ КОНСТРУКЦИЯХ

Е. Г. Былинина (e.g.bylinina@uu.nl)

Институт лингвистики OTS, Утрехтский Университет, Нидерланды

В работе рассматривается ранее подробно не исследованный компонент степенных конструкций, который мы называем «функциональный стандарт» («Этот зал маловат для игры в баскетбол»). Предлагается анализ, включающий целевую пропозицию в число аргументов градуального прилагательного.

Ключевые слова: степенные конструкции, «функциональный стандарт», градуальное прилагательное, целевая пропозиция.

"FUNCTIONAL" STANDARD IN RUSSIAN AND ENGLISH DEGREE CONSTRUCTIONS

E. G. Bylinina (e.g.bylinina@uu.nl)

Institute for Linguistics OTS, Utrecht University, Netherlands

We develop a notion of functional standard, which refers to the 'functional standard degree construction'(John is a little bit too tall for this job). The construction involves a 'purpose' proposition parameter that determines the set of degrees compatible with the purpose. The maximal degree belonging to this set serves as a standard in the construction. We argue against contextual and comparative analyses either explicitly or implicitly assumed in the literature. Instead, we propose that the purpose is an argument of (certain) gradable adjectives, and the whole construction is a positive construction. We try to pinpoint the difference between Russian and English functional standards.

Key words: degree constructions, 'functional standard', gradable adjective, purpose.

1. Introduction

The starting point of this work is an observation made in (Kagan and Alexejenko 2011) that Russian adjectival suffix *-ovat* means something like 'slightly too' (1a) in certain environments and simply 'slightly' (1b) in other cases:

(1) a. Takije kabluki dlja menja vysok-ovat-y. (=1b Kagan and Alexejenko 2011)
such heels for me high-ovat-PL.NOM
'Such heels are somewhat too high for me.'
b. Lena protjorla mebel' vlažn-ovat-oj trjapkoj. (=2b Kagan and Alexejenko 2011)
Lena wiped furniture wet-ovat-INSTR duster
'Lena wiped the furniture with a wettish duster.'

(1a) says that the degree that the heels reach on the scale of height is slightly greater than the highest degree that would be good for me to wear. Crucially, there is a 'purpose' proposition involved, defining a degree interval; its maximum is used as a standard of comparison. We call this "functional standard". (1b) doesn't have anything like that — it just states that the duster possesses a low degree of wetness.

We discuss briefly the technical details of degree semantics we will be using.

2. Degree semantics background

I follow (Bartsch and Vennemann 1972, 1973) and (Kennedy 1999, 2005) and analyze gradable adjectives as measure functions: functions of type <e, d> from the domain of individuals to degrees on a certain scale:

(2) a. [[tall]] = λx.**tall**(x)
b. [[expensive]] = λx.**expensive**(x),

where **adj**(x) is 'the degree on the appropriate scale that represents x's measure of adjective-ness'

Measure functions are converted into properties of individuals by degree morphology, which includes comparative morphemes, intensifiers and so forth. For (morphologically) unmarked positive form (*John is tall*) null POS morpheme is introduced, with a denotation along the lines of (3), where $\mathbf{d}_s$ is 'contextually appropriate standard of comparison, whatever that is':

(3) $[[[_{\text{Deg}}\ \textbf{pos}]] = \lambda g \lambda x.g(x) \geq \mathbf{d}_s$ (=9 Kennedy 2005)

To be more precise, various evidence (which we omit here) shows that there are several homonymous POS morphemes, at least this is one of the straightforward ways to capture the distinct behavior of gradable adjectives with different scalar structure. Relative adjectives (gradable adjectives encoding a scale with neither minimum nor maximum — *tall*, *wide*) combine with POS_{rel}, which is analogous to (3). The difference

is the '**significantly**' component which is necessary for the positive form of the relative adjective to be true of an individual:

(4) $[[_{\text{Deg}}\ \text{POS}_{rel}]] = \lambda g\, \lambda c{\in}D_{\langle e,t\rangle}\, \lambda x.g(x)\ !{\succ}\ \mathbf{norm}(c)(g)$ (in lines of Kennedy 1999, 2005)
c = comparison class, g = gradable property, !≻ = significantly exceed

Absolute adjectives (encoding scales with minimum — *sick*, *wet* — or maximum — *healthy*, *dry*) combine with POS_{min} or POS_{max}. For an adjective like *wet* to hold of an object, it suffices for the object to possess any small degree of wetness, while for *dry* to hold it should be completely dry:

(5) a. $\lambda g \lambda x.g(x) \succ \min(\text{SCALE}(g))$ (=76bc Kennedy 2005)
b. $\lambda g \lambda x.g(x) = \max(\text{SCALE}(g))$

Comparative clauses involve comparative elements (*more*) that are often treated as expressions that establish an ordering relation between two degrees: one derived by applying the adjectival head to its subject, the other by applying it to the 'standard' constituent, marked by *than* (Hankamer 1973; Hoeksema 1984; Heim 1985; Kennedy 1999). So that the predicate 'larger than Rome' would have semantics like λx.large(x) ≻ large(Rome).

3. Functional standards: possible analyses

Functional standards are not limited to *–ovat* (1a). More examples from Russian and English:

(6) a. Vasja nemnogo vysokij. (≈22 K&A2011)
Vasja slightly tall
'Vasja is slightly too tall.'
b. Etot zal malenkij / mal dlja igry v basketbol.
This gym small for play in basketball
'This gym is too small for a basketball game'

(7) a. These pants are *({a little bit / slightly / somewhat}) long for me.
b. This soup is hot for me.

There are two straightforward ways to approach functional standard semantics. We discuss them in turn.

3.1. Contextual view

The straightforward view on the functional standard composition would not make substantial difference between distributional and functional standards: for

a gradable predicate to hold of an individual (= for a positive form of an adjective to be true), the individual should exceed a standard degree on a relevant scale; whether the standard degree is fixed distributionally or 'functionally', can be a matter of contextual salience and prominence.

This is in fact the view adopted in (Kagan and Alexejenko 2011). They develop a unified semantics for *-ovat* that covers both the cases when it means 'slightly too' and just 'slightly', depending on which kind of standard of comparison (d') is used — a **distributional** ('slightly') or a **functional** one ('slightly too'):

(8) $\lambda P_{<d,et>} \lambda d' d \lambda x_e$. max{d: P(d)(x)} > **d'** $\wedge$ (max{d: P(d)(x)} – d' < d_c) (=9 K&A2011)

In prose, *-ovat* states that a degree an entity reaches on the scale provided by a gradable predicate exceeds a certain standard d' and that the interval between these two degrees is small. The distributional standard is calculated on the basis of distribution of the relevant property (height, price etc.) within a comparison class (as in *expensive for a studio*), while the functional standard is the max degree on the interval of degrees that are compatible with the requirements of the situation (as in *expensive for me*): max{d: ∃w› ∈ Acc(w): P(w›)(d) = 1}; see similar treatment of 'too' construction in (Heim 2000). The choice between two options for d' is (implicitly) treated as pragmatic rather than semantic. We argue against this view.

No matter how strong the context is, it is not always the case that one can freely choose which standard to use. (1a), (6) and (7a) strongly disallow distributional standard interpretation. The low-level generalization here is that it's the modification of low degree (*slightly*, *somewhat* etc.) that bans the distributional standard reading with relative adjectives (= gradable adjectives with open scales). Thus, availability of functional vs. distributional standards depends on the scale structure of the gradable adjective. We think that low degree modifiers are not acceptable in positive constructions with relative adjectives precisely because they are not compatible with a POS morpheme for relative adjectives.

The contextual view is not capable of treating scale structure sensitivity of functional standards — one is forced to conclude that construction with functional standards is not just the same positive construction, and we need to find another analysis.

3.2. Comparative analysis

The other obvious analysis would relate functional standards to *too*, either positing a silent 'too'-like element in the structure or equivalently postulating a similar type shift.

A state-of-the art analysis of *too* is along the lines of (9). It incorporates an observation that a possibility element under comparative yields a maximal degree reading (and necessity yields min) (Heim 2001):

(9) Bertha is too old to be allowed to drive a car
{d | ∃w: w ∈ H_{w^*} & Bertha is allowed to drive a car in w & AGE_w(Bertha) ≥ d} ⊂
{d | AGE_{w^*}(Bertha) ≥ d} (von Stechow 2003, von Stechow et al. 2004)

The semantics of *too*-constructions is essentially comparative, stating the relation of inclusion between two sets of degrees: one of Bertha's age range that includes her ages in at least one of the possible worlds where Bertha is allowed to drive — and the other set is her 'actual age'. Existential quantification over worlds reflects the possibility modality. Doing it in terms of degree sets rather than points is a matter of convention. General acceptability of measure phrases (MPs) in *too*-constructions (10) follows from a comparative analysis as in (9).

(10) These pants are (10 cm) too long for me.

As Russian doesn't have MPs, at least as closely attached to the adjective phrase as in English, we will look at English to check the idea of functional standard constructions as implicit *too* constructions. Alarmingly, MPs are ok with *too* and ungrammatical with functional standards:

(11) *These pants are 10 cm long for me.

Since the acceptability of MPs in *too*-constructions follows from a comparative analysis, we can conclude that MPs are not compatible with functional standards since it is not a comparative construction, and adjectives themselves do not combine with MPs unless they undergo further shifts (Schwarzschild 2005). Thus we believe the comparative view to be false as well.

3.3. Alternative view

We propose to attempt a straightforward account of the above data. The first step is building a 'purpose' parameter into the expression. The 'purpose' phrase is quite often introduced by the for-phrase.

For-phrases are found across degree constructions and come in various sorts. The most well-known are 'comparison class' (CC) (12a) (Fults 2006, Bale 2010, Solt t.a.) and 'judge' (12b) (Lasersohn 2005, Stephenson 2007) *for*-phrases:

(12) a. Fred is tall for a 8-year-old.
b. The movie was fun for me.

If, as it has been argued, CC *for*-phrases are arguments of POS morpheme, POS has semantics in (13a) and the *for*-phrase provides a degree for contextual standard; 'judge' *for*-phrase is taken to be an argument of an adjective itself and 'Skolemize' the degree expression in a sense (13b):

(13) a. $[[\text{POS}]] = \lambda C_{\langle et \rangle} \lambda P_{\langle d,et \rangle} \lambda x. \exists d\, [P(x,d) \wedge d > R_{Std:CC}]$ (=32 Solt ta)
b. $[[\text{fun}]]^{c;\, w,t,j} = [\lambda x.\, [\lambda y.\, y \text{ is fun for } x \text{ in } w \text{ at } t]\,]$

We believe that functional standard for-phrases are not arguments of POS, rather they are arguments of the adjective directly, similar to 'judge' PPs, since they

appear in non-POS constructions (for comparatives, one needs to find an adjective that doesn't preserve order under different purposes):

(14) a. This book is more suitable for a 3-year-old than that one
b. Your room is best for our meetings.
c. .. discusses how very expensive insurance is for circus performers.. (from web)

We believe 'expensive' and the like to be 'inherently purpose-relative' and have type <st <ed>> as they freely combine with purpose for-phrases without low degree modifiers:

(15) This book is expensive for a 3-year-old.

Once combined with a 'purpose' set of worlds, we get (15):

(16) $[[\text{expensive for a 3-year-old}]] = \lambda x.\ \text{PRICE}(x) - \mathbf{max}\{d \mid \exists w'.\ \text{PRICE}_{w'}(x)=d \wedge \mathbf{R}_{w'}(x)(\mathbf{3yo})\}$

In prose, we measure the extent to which the price of *x* differs from the max price that would still make *x* fit the purpose (in a world *w*, a conventionally prominent relation **R** holds between *x* and a 3-year-old). The resulting predicate has a scale with a derived minimum, which can combine with POS that is tailor-made for absolute adjectives with a minimum (see Kennedy 2005 on ambiguity of POS):

(17) $[[\text{POS}_{min}$ expensive for a 3-year-old]] $=\ldots=$ PRICE(x) $\succ$ $\mathbf{max}\{d \mid \exists w'.\ \text{PRICE}_{w'}(x)=d \wedge \mathbf{R}_{w'}(x)(\mathbf{3yo})\}$

Derivation would work in almost the same way for adjectives that are not inherently purpose-related (like *long*), though to get a purpose parameter they will need to undergo a purpose-shift:

(18) $\lambda x.\ \text{LENGTH}(x) \rightarrow \lambda P_{\langle st\rangle}\ \lambda x.\ \text{LENGTH}(x) - \mathbf{max}\{d \mid \exists w'.\ \text{LENGTH}_{w'}(x)=d \wedge P_{w'}\}$

The relation **R** used in (16–17) is determined by the purpose *for*-phrase, but in an indirect way. The complement of *for* can be either an individual (*me*) or an indefinite NP (*a 3-year-old*), and the purpose proposition is recovered on the basis of what relation is typical between the kind introduced by a subject NP and the kind introduced by a complement of *for*, restricted by the adjective. Say, in (15) **R** is a relation between books and 3-year-old children s.t. the price is relevant to it; it is very likely to be an OWN or BOUGHT-FOR relation. One might want to introduce generic semantics into **R**, and have an add-on for individual *for*-complement, but we will not attempt that now.

4. English vs. Russian

In English, a sentence like (6b) is ungrammatical without low degree modification. When the low degree modifiers are present, their incompatibility with POS_{rel}

forces the functional standard reading. To get a purpose parameter *long* will need to undergo a purpose-shift.

But it's not true for all relative adjectives in English — adjectives like *expensive* and *hot* do not need to appear in an environment that excludes normal POS_{rel} to be interpreted as purpose-relative (7b). Thus there is a lexical distinction between 'inherently purpose-relative' adjectives in English and the rest of the adjectives.

Russian doesn't seem to exhibit this lexical contrast, compare (6b) and (7a) — all relative adjectives enter functional standard construction easily without modifiers. One would argue for view A for Russian, but we believe it not to be the case because scale structure sensitivity is still present (1a and 6a do not allow for distributional standard interpretation). There are several possible ways to account for this. First, one could say that the difference is lexical — Russian has more inherently purpose-relative adjectives than English. Second, it can actually be the case that English possesses of two similar constructions rather than one: one that is sensitive to a lexical class of the adjective but does not require a type-shift, and the other one requires a type-shift and is insensitive to the lexical difference the former is sensitive to; Russian exhibits only one of the two constructions. We will argue for the second view.

5. Summary

We introduced functional standards as propositional arguments of gradable adjectives (some have them from the start, some have to type-shift), which use functional standard to derive a *min* with a possibility modality inside. Combination with POS_{min} explains that: a) low-degree modifiers are fine with functional standards but not with a contextual standard that uses POS_{rel}, b) MPs are not compatible with functional standards since it is not a comparative construction, and adjectives themselves do not combine with MPs unless they undergo further shifts (Schwarzschild 2005). **A potential problem** is that *expensive* is quite often used as a relative adjective. We believe that the 'purpose' parameter can in these cases be suppressed. Crucially, however, *expensive* comes with a purpose parameter from the lexicon, while *long* doesn't.

As a working hypothesis, we propose that the difference between English and Russian is not lexical, but rather English has two similar constructions only one of which is represented in Russian.

References

1. *Bale A.* 2010. Scales and Comparison Classes. Natural Language Semantics.
2. *Bartsch R., Vennemann T.* 1972. The Grammar of Relative Adjectives and Comparison. Linguistische Berichte 20: 19–32.
3. *Bartsch R., Vennemann T.* 1973. Semantic Structures: A Study in the Relation between Syntax and Semantics.
4. *Fults S. W.* 2006.The Structure of Comparison: an Investigation of Gradable Adjectives.

5. *Hankamer J.* 1973. Why there are two 'Than's in English. Proceedings of the 9th Annual Meeting of the Chicago Linguistics Society.
6. *Heim I.* 1985. Notes on Comparatives and Related Matters.
7. *Heim I.* 2000. Degree Operators and Scope. Semantics and Linguistic Theory, 10 : 40–64.
8. *Hoeksema J.* 1984. Negative Polarity and the Comparative. Natural Language & Linguistic Theory, 1 : 403–434.
9. *Kagan O., Alekseenko A.* Degree Modification in Russian Morphology: The Case of the Suffix –ovat. Proceedings of IATL.
10. *Kennedy C.* 1999. Projecting the Adjective: The Syntax and Semantics of Gradability and Comparison.
11. *Kennedy C.* 2005. Vagueness and Grammar.
12. *Lasersohn P.* 2005. Context Dependence, Disagreement, and Predicates of Personal Taste. Linguistics and Philosophy, 28 (6) : 643–686.
13. *Schwarzschild R.* 2005. Measure Phrases as Modifiers of Adjective. Recherches Linguistiques de Vincennes, 35, L'adjectif : 207–228.
14. *Solt S. Notes on the Comparison Class.*
15. *Stechow A. von.* 2003. Different Approaches to Semantics of Comparison. Lublin lecture notes.
16. *Stechow A. von, Krasikova S., Penka D.* 2004. The Meaning of German um zu: Necessary Condition and enough/too. Tübingen workshop on modal verbs and modality hand-out.
17. *Stephenson T.* 2007. Judge Dependence, Epistemic Modals, and Predicates of Personal Taste. Linguistics and Philosophy, 30 : 487–525.

ИССЛЕДОВАНИЕ ЗАДАЧИ КЛАССИФИКАЦИИ ОТЗЫВОВ НА ТРИ КЛАССА

И. И. Четверкин (ilia2010@yandex.ru)
МГУ, Москва, Россия

Н. В. Лукашевич (louk_nat@mail.ru)
МГУ, Москва, Россия

Ключевые слова: отзывы, классы, классификация, обзор.

THREE-WAY MOVIE REVIEW CLASSIFICATION

I. I. Chetverkin (ilia2010@yandex.ru)
Faculty of Computational Mathematics and Cybernetics
Lomonosov Moscow State University, Moscow, Russian Federation

N. V. Loukachevitch (louk_nat@mail.ru)
Research Computing Center, Lomonosov Moscow State University, Moscow, Russian Federation

In this paper, we consider a three-way classification approach for Russian movie reviews. All reviews are divided into groups: “thumbs up”, “so-so” and “thumbs down”. To solve this problem we use various sets of words together with such features as word weights, punctuation marks and polarity influencers that can affect the polarity of the following words. Besides, we estimate the maximum upper limit of automatic classification quality in this task.

Key words: classification, review, review classification, movie, movie review.

1. Introduction

Actually, users can find any type of information in the Internet. Tentatively, it can be divided into two classes: factual information and user opinions. Most of current information processing techniques (e. g., search engines) work with facts and have satisfactory quality. Processing of user's opinions is a more complicated problem. Ranking of the reviews according to their sentiment is a very difficult and urgent task.

The easiest subtask is to classify reviews into two classes: *positive* and *negative*. Quality of two-way classification using topic-based categorization approach for reviews exceeds 80 % [9]. In [12] the quality of review classification, based on the so-called appraisal taxonomy, was described as 90.2 %.

However, when we turn to the problem of review division into three classes («thumbs up», «thumbs down», «so-so»), the quality of automatic classification decreases significantly [7]. This is partly due to the subjectivity of human evaluation. In [8] the authors conducted a study on the possibility of a human to distinguish reviews rated on a ten-point scale. They describe that if the difference between review scores is more than three points, the accuracy is 100 %, two — 83 %, one point — 69 % and zero points, correspondingly, 55 %. Thus, if to classify reviews into a large number of classes, even a human will show low classification accuracy.

In addition, in that paper the difference between evaluation styles of various people was indicated: a review estimated in 5 points (on a ten-point scale) by one person, may express the same opinion and be estimated as 7 points by the other [8]. It was shown that after adjustment to an individual author's style, the quality of the classification increased significantly and reached 75 %. But in the classification of 5394 reviews from a large number of authors (494), the achieved accuracy was 66.3 %.

In this paper, we analyze various features to improve three-way classification of movie reviews in Russian. For Russian language, studies of this task practically *do not exist*.

We used the following classification features:

- word weights based on different sources,
- single word polarity,
- use of polarity influencers: they may reverse or enhance (*not, very*) polarity of other words,
- length and structure of reviews,
- usage of punctuation marks — as for example in [11] authors used punctuation to reveal sarcastic sentences.

2. Features for review classification

For our experiments, we chose movies' domain. We collected 28 773 film reviews of various genres from online recommendation service *www.imhonet.ru*. For each review, we extracted user's score on a ten-point scale.

Example of the review:

Nice and light comedy. There is something to laugh — exactly over the humor, rather than over the stupidity... Allows you to relax and gives rest to your head.

2.1. Word weights

As the main elements of a feature set we used lemmas (words in the normal form) mentioned in the reviews. Word weights can be binary and reflect only word presence in a review or TFIDF formula can be used.

TFIDF is the most popular method of word weighting in information retrieval [6]. For each term in a text, its TFIDF weight can be represented by multiplication of two factors: TF that defines the frequency of this term in the text and IDF specifying occurrence of the term in documents of a text collection. The more frequently such occurrences are, the smaller resulting IDF will be [6]. TF and IDF factors can be defined by various formulas. We used two variants of TFIDF for calculation.

First, we used the simplest form of TFIDF [6]:

$$\mathrm{TF} = \frac{n_i}{\sum_k n_k} \quad \mathrm{IDF} = \log \frac{|D|}{|(d_i \supset t_i)|} \tag{1}$$

- n_i is the number of occurrences of a term in a document, and the denominator is the sum of occurrence number of all terms in the document,
- $|D|$ — total number of documents in a collection,
- $|(d_i \supset t_i)|$ — number of documents where term t_i appears (that is $n_i \neq 0$).

In addition, we used TFIDF variant described in [1] (based on BM25 function [6]):

$$\textit{TFIDF (l)} = \beta + (1 - \beta) \cdot \textit{tf(l)} \cdot \textit{idf(l)}$$

$$tf_D(l) = \frac{\mathrm{freq}_D(l)}{\mathrm{freq}_D(l) + 0.5 + 1.5 \cdot \frac{dl_D}{\mathrm{avg\_dl}}} \qquad idf(l) = \frac{\log\left(\frac{|c| + 0.5}{df(l)}\right)}{\log(|c| + 1)} \tag{2}$$

- *freq(l)* — number of occurrences of *l* in a document,
- *dl(l)* — length measure of a document, in our case, it is number of terms in a review,
- *avg_dl* — average length of a document,
- *df(l)* — number of documents in a collection (e. g. movie descriptions, news collection) where term *l* appears,
- $\beta = 0.4$ by default, in our case $\beta = 0$,
- $|c|$ — total number of documents in a collection.

2.2. Opinion words

We considered opinion words as an important type of features for review classification.

We use the automatically extracted list of opinion words [3]. To generate this list, we exploited four text collections: a movie review collection (review corpus), a collection of film descriptions (description corpus), a special small corpus and a collection of general news. On the basis of these collections we calculated a set of statistical features for words mentioned in reviews. All features were calculated separately for adjectives and not adjectives (verbs, adverbs, nouns). At the next step, we used machine learning to classify terms' feature vectors. As a result we obtained term lists (adjectives and not adjectives), ordered by predicted probability of their opinion orientation.

Let us look at some examples of opinion words with high probability value:

- adjectives: *dobryj (kind), zamechatel'nyj (wonderful), velikolepnyj (gorgeous), potrjasajushij (stunning), krasivyj (beautiful), smeshnoj (funny), ljubimyj (love)* etc.,
- not adjectives*: fuflo (trash), naigranno (unnaturally), fignja (junk), fil'm-shedevr (masterpiece film), tufta (rubbish)* etc.

In our study of three-way review classification, we used the most probable opinion words and automatically obtained opinion probability weights. In addition, we manually labeled a set of opinion words [3].

2.3. Polarity influencers

Intuitive is the fact that there are some words, which can affect polarity of other words — polarity influencers. To find them the manually compiled set of opinion words (3200 units) was used [3]. From the review corpus (see section 2.2), we automatically extracted words directly preceding the manually labeled opinion words and ordered them by decreasing frequency of their occurrence.

Then from the first thousand of words from this list, potential polarity influencers were manually chosen (74 words). To assess how significant the effect of these polarity influencers can be, the following procedure was made: we calculated the average score of opinion words in two cases, when they follow the potential polarity influencers and when they occur without them. The average score of a word is the average value of numerical scores of reviews where this word occurs.

After comparison of these average scores, two significant groups of polarity influencers were discriminated. If an opinion word had the high average score (>8) and changed it to the lower when used after a given polarity influencer, and an opinion word with the low average score (<6.7) changed it to the higher one, it means that this polarity influencer *reverses* word polarity (operator –).

If after a polarity influencer, an opinion word with the high score increased its average score, and an opinion word with the low average score decreased its score, it means that this polarity influencer *magnifies* polarity of other words (operator +).

In our review corpus, we found the following polarity influencers:

- operator (–): *net (no), ne (not);*
- operator (+): *polnyj (full), ochen' (very), sil'no (strongly), takoj (such), prosto (simply), absoljutno (absolutely), nastol'ko (so), samyj (the most).*

On the basis of this list of polarity influencers we substituted sequences *"polarity influencer_word"* using special operator symbols («+» or «–») depending on an influencer, for example:

NE HOROSHIJ (NOT GOOD) → –HOROSHIJ (— GOOD)
SAMYJ KRASIVYJ (THE MOST BEAUTIFUL) → + KRASIVYJ (+ BEAUTIFUL)
NASTOL'KO KRASIVYJ (SO BEAUTIFUL) → + KRASIVYJ (+ BEAUTIFUL)

Modified lemmas were added to the feature set. Now if in a text a word with a polarity influencer occurs, then only the corresponding modified lemma would be added to the review's vector representation, but not both words. This allows us to take into account the impact of polarity influencers.

2.4. Review length and structural features

Movie reviews can be long or short. We chose a threshold on the review length to be 50 words. If a review is long, it often contains overall assessment for a movie at the beginning or at the end. This was the basis for separate consideration of short and long reviews and dividing long reviews into three parts: the beginning (first sentences of a review with total length less than 25 words), the end (last sentences of a review with total length less than 25 words) and the middle (all that is left). We classified each part separately and then aggregated obtained scores in various ways (voting, average).

2.5. Punctuation marks

In addition we included punctuation marks «**!**», «**?**», «**…**» as elements of the feature set.

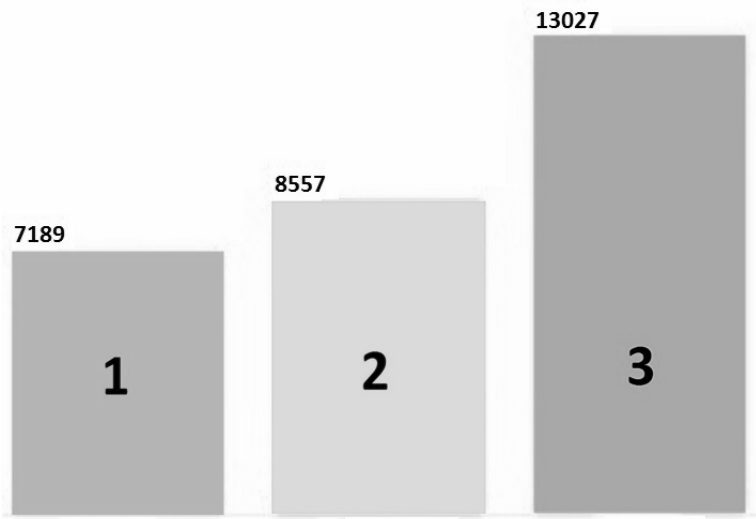
3. Experiments

Reviews in the working dataset are provided with authors' scores from 1 to 10 points. To map from the ten-point scale to the three-point scale we used the following function: {1–6} → «**1**» (thumbs down), {7–8} → «**2**» (so-so), {9–10} → «**3**» (thumbs up). The resulting distribution of reviews by grade is shown on Picture 1. Thus, the number of reviews belongs to class «3» is approximately 45 % of the total.

All reviews from the collection were preprocessed by a morphological analyzer and lemmas with part of speech tagging were extracted.

Authors of previous studies almost unanimously agreed that Support Vector Machine (SVM) algorithm works better for text classification tasks (and review classification task in particular). We also decided to use this algorithm. In view of the fact that we had a large amount of data and features, library LIBLINEAR was chosen [10]. This

library had sufficient performance for our experiments. To obtain statistically significant results five fold cross-validation was used. All other parameters of the algorithm were left in accordance with their default values.



Pic. 1. The distribution of reviews into three groups by sentiment: "thumbs down"(1),"so-so" (2),"thumbs up"(3)

We used the following word sets in our classification experiments:

- Finding an optimal set of opinion words produced by the method described in Section 2.2. From the list of adjectives and not adjectives (ordered by the probability of their opinion orientation — ***opinweight***) we selected the optimal opinion word combination. We iterated over words in these lists and compared quality of classification. We denote this experiment set ***OpinCycle***,
- set of words, which was used in [4] to achieve the best results (***OpinContrast***). This set contains near 500 the most frequent words with high opinion probability weight [3] and 400 words with the highest TFIDF score calculated using review and news collections (see Section 2.2),
- set of opinion words (3200 units), obtained by manual labeling by two experts (see Section 2.2) (***OpinIdeal***),
- set of all words occurring in the review corpus four or more times (***BoW***). The set includes prepositions, conjunctions and particles as well.

From all these word sets, we chose one set, which yields the best classification accuracy, and analyzed the effect of other features: word weights (***tfidf***), opinion weights (***opinweight***), punctuation marks (***punctuation***), polarity influencers (***operators***), review length (***long*** and ***short***).

TFIDF word weights were calculated relying on two formulas: the most well known formula (1) (***tfidf simple***) and formula (2) (***tfidf)*** (see Section 2.1). IDF factor was calculated on the basis not only the review corpus, but also two other collections: the news corpus (***tfidf news)*** and the description corpus (***tfidf descr)***.

To assess the quality of classification we used *Accuracy measure*. It is calculated as the ratio of correct decisions taken by the system to the total number of decisions [2].

The results of algorithms using different sets of words and features are listed in Table 1. It is worth mentioning that different sets have different coverage area. All reviews without any features from the set were considered as strongly positive ("thumbs up") in accordance with review distribution between classes. The basic weight of each word is its presence in a review.

The results obtained by using ***BoW + tfidf simple*** were taken as a *basic line*. The best results were obtained using bag of words (***BoW***) with TFIDF, opinion weights and polarity influencers. This is clear improvement over *62.52* where ***BoW + tfidf simple*** is applied; indeed the difference is highly statistical significant ($p < 0.001$, $\alpha = 0.05$, Wilcoxon signed-rank test/Two-tailed test). Punctuation marks did not give any quality improvement, although their usage gave slightly better coverage. Formula (2) usage gives slightly better quality than the first one (1). The choice of the news corpus for IDF calculation in (2) draws better results than using the description corpus (***BoW + tfidf descr***) and the review corpus (***BoW + tfidf***).

Table 1. The classification results using various features

Feature set	Feature number	Accuracy %
OpinCycle	1000 *adj* + 1000 *not adj*	58.00
OpinContrast	884	60.33
OpinIdeal	3 200	57.62
BoW	19 214	57.37
OpinCycle + tfidf simple	1000 *adj* + 1000 *not adj*	59.13
OpinContrast + tfidf simple	884	59.43
OpinIdeal + tfidf simple	3200	59.72
BoW + tfidf simple	19 214	*62.52*
BoW + tfidf	19 214	61.71
BoW + tfidf descr	19 214	61.74
BoW + tfidf news	19 214	62.90
BoW + tfidf news + operators	22 218	63.46
BoW + tfidf news + punctuation + operators	22 221	63.17
BoW + tfidf news + opinweight + operators	22 218	**64.48**

Feature set	Feature number	Accuracy %
BoW + tfidf news+ opinweight + operators + short	22 218	63.56
BoW + tfidf news + opinweight + operators + long	22 218	62.37
BoW + tfidf news + opinweight + operators + avg	22 218	63.14

To increase weights of opinion word in contrast with the other words we used the list of opinion words with probability weights from 0 to 1 (see Section 2.2). We took 800 the most probable adjectives and 200 not adjectives (we have tried another combinations also) as opinion words. All other words from the feature set were considered with ***opinweight*** 0. We modified the weight of each word in the feature vectors in the following manner:

$$wordweight(x) = TFIDF(x) \cdot e^{(opinweight(x) - 0.5)}$$

Thus, we want to increase weights of the words with high ***opinweight***, and decrese for the other words.

The classification accuracy for short reviews (***BoW + tfidf news + opinweight + operators + short***) is better than for long one (***BoW + tfidf news + opinweight + operators + long***). Although, in average (in accordance with review number in each part) the results were not improved (***BoW+ tfidf news + opinweight + operators + avg***).

For the method with the best results of classification ***BoW + tfidf news + opinweight + operators***, we made additional evaluation with so-called *soft borders*, that is if in the basic scale the author of a review puts a boundary score («8» or «6»), then classification of this review as either class «3» or «2» in case of basic «8», and class «2» or «1» in case of basic «6», was not considered as an error. Such weakening of conditions was made on the assumption that even a human distinguishes boundary classes unsatisfactory. The classification accuracy with *soft borders* reaches **76.48 %.**

4. Evaluation of reviews by assessors

We also studied the human's ability in three-way review classification. We wanted to know what the maximal quality of classification we could expect from automatic classification algorithms. Significance of such quality upper bound evaluation is declared, for example, in [5]. For a benchmark, we selected one hundred short reviews (with length less than 50 words) and one hundred long reviews (with length more than 50 words) from the review corpus. Assessors did not know the initial score of a review set by its author. Reviews were extracted in such a manner, as to retain original class distribution. All explicit references to the initial score were removed.

Two assessors evaluated the selected reviews. The results of their evaluation are given in Table 2. The last row of the table indicates the agreement in scores between two assessors.

Table 2. The results of humans' estimating

Assessor	Assessors accuracy relative to the author of the review	Accuracy with soft borders %	Accuracy of the best classification algorithm relative to the assessor
1	72.5	86.5	69.5
2	72.5	78.5	63.5
1 AND 2	71.5	—	—

Thus, we see that human assessors can reproduce the original scores or be consistent with each other only at the level of 71–72 %, which is the absolute upper limit to improve the quality of automatic algorithms. Note that quality of the automatic classification with soft borders, taking into account the possible ambiguity of the border scores, is 76.48 %, which is very close to the classification quality of the second assessor (78.5 %).

The percentage of coincident scores between the best algorithm and assessor's scores confirms the results obtained by cross-validation.

5. Conclusion

In this paper, we investigated influence of various factors on the quality of three-way classification of movie reviews in Russian. The most significant impact on the quality of classification had the choice of TFIDF formula, polarity influencers accounting and opinion words information usage. We estimated the upper limit of classification quality, which is very close to the results of the best automatic algorithm. This fact makes it difficult to reach further quality improvement of automatic three-way review classification.

Acknowledgements

This work is partially supported by RFBR grant N11-07-00588-a.

References

1. *Ageev M., Dobrov B., Loukachevitch N., Sidorov A.* 2004. Experimental Algorithms vs. Basic Line for Web Ad Hoc, Legal Ad Hoc, and Legal Categorization in RIRES2004. RIRES.
2. *Ageev M., Kuralenok I. Nekrest'ianov I.* 2010. Official RIRES Metrics. Kazan: Russian Information Retrieval Evaluation Seminar (RIRES 2010).
3. *Chetverkin I., Loukachevitch N.* 2010. Automatic Extraction of Domain-specific Opinion Words. Dialogue.
4. *Chetverkin I., Loukachevitch N.* 2019. Automatic Review Classification Based on Opinion Words. Tver: Conference on Artificial Intelligence.
5. *Kilgarriff A., Rosenzweig J.* 2000. Framework and Results for English Senseval. Computers and Humanities, Special Issue on SENSEVAL : 15–48
6. *Manning C., Raghavan P., Schütze H.* 2008. Introduction to Information Retrieval.
7. *Pang B., Lee L.* 2008. Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval.
8. *Pang B., Lee L.* 2005. Seeing stars: Exploiting Class Relationships for Sentiment Categorization with respect of Rating Scales. Proceedings of the ACL.
9. *Pang B., Lee L.* 2002. Thumbs Up? Sentiment Classification using Machine Learning Techniques. EMNLP.
10. *R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin.* 2008. LIBLINEAR: A Library for Large Linear Classification. Journal of Machine Learning Research, 9, 2008 : 1871–1874. Software available at http://www.csie.ntu.edu.tw/~cjlin/liblinear
11. *Tsur O., Davidov D., Rappoport A.* 2010. ICWCM — a Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews. International AAAI Conference on Weblogs and Social Media.
12. *Whitelaw C., Garg N., Argamon S.* 2005. Using Appraisal Taxonomies for Sentiment Analysis. CIKM.

КЛАССИФИКАЦИЯ ЭМОЦИОНАЛЬНОГО СОСТОЯНИЯ ДИКТОРА ПО ГОЛОСУ: ПРОБЛЕМЫ И РЕШЕНИЯ

А. Г. Давыдов (davydov-a@speetech.by)

В. В. Киселёв (kiselev-v@speetech.by)

Д. С. Кочетков (kochetkov-d@speetech.by)

ООО «Речевые технологии», Минск, Беларусь

Описан алгоритм, позволяющий автоматически определять эмоциональное состояние диктора по голосу. Обучение и тестирование проводилось на модельном корпусе эмоциональной речи, собранном в техническом университете Берлина. Точность распознавания состояний «*anger*» и «*neutral*» составила порядка 96 %.

Ключевые слова: голос, диктор, эмоции, эмоциональное состояние, эмоциональная речь.

VOICE EMOTION CLASSIFICATION: PROBLEMS AND SOLUTIONS

A. G. Davydov (davydov-a@speetech.by)

V. V. Kiselev (kiselev-v@speetech.by)

D. S. Kochetkov (kochetkov-d@speetech.by)

Speech Technologies LLC, Minsk, Belarus

An algorithm for automatic emotion recognition from the speaker's voice has been developed. A number of tests were performed using the widely known corpus of Emotional Speech — Berlin Database (Emo-DB). The classifi cation effi ciency for different acoustic features was estimated and a very small set of the most reliable characteristics was extracted in order to obtain a robust and quick emotion state classifi cation. Using the SVM classifi er with quadratic kernel and this feature set provides the recognition accuracy of approximately 96 %. between "anger" and "neutral" emotional states. GMM classifi er was less effective and demonstrates a classifi cation error of up to 6 %. A brief comparison of this feature set and SVM kernel effectiveness was performed using the Munich openEAR toolkit. A recommended set of 384 features and linear-kernel SVM was used to solve the same problem.

The classifi cation effi ciency of such algorithm reached 98%. This value is only ~2% higher than the respective value for the designed feature set and classifi er. Under the several conditions, such as in the case of obtaining a decision support factor in the systems of real-time speech analytics the simplifi ed classifi cation scheme would be more preferable than a complex one.

Key words: voice, emotions, emotional state, emotional speech.

1. Введение

Первые попытки автоматического определения эмоциональных состояний по голосу, были предприняты еще в середине 80-х годов. С тех пор эволюция компьютеров с одной стороны, и требования рынка с другой, неуклонно стимулируют дальнейшее развитие систем распознавания эмоций, а так же иных систем голосового анализа, детектирующих уровень стресса, депрессии, усталости, алкогольного опьянения и т. п. Несмотря на это, проблема взаимосвязи эмоциональных состояний диктора с параметрами его голоса до сих пор полностью не решена.

Трудности, встающие перед исследователем при решении этой задачи весьма многообразны, однако можно особо выделить две из них [1]. Прежде всего, чёткого определения эмоции не существует. Это приводит к различным формам классификации эмоциональных состояний и различной расстановке акцентов у разных исследователей [2]. Помимо этого, отсутствует однозначный ответ на вопрос о соотнесении акустических особенностей речи диктора и его эмоционального состояния. Порой различные авторы приводят полностью противоположные результаты. В настоящий момент многие исследователи приняли на вооружение классификацию эмоций, основанную на непрерывной их модели, согласно которой каждое эмоциональное состояние может быть описано точкой в эмоциональном пространстве. Чаще всего при этом координатная сетка задается двумя шкалами, определяющими уровень активации психики и валентность (устойчивость предпочтений человека относительно конкретного результата). Другой распространенной теоретической моделью является так называемая дискретная, «палитровая» теория, согласно которой любое эмоциональное состояние можно описать как совокупность действия ряда архетипических эмоций. Как правило, к ним относят гнев, раздражение, страх, радость, печаль и удивление. Выбор конкретной модели описания при решении задачи распознавания эмоционального состояния определяется в основном соображениями удобства. Вследствие отсутствия универсальной теоретической модели, и возникающей отсюда необходимости оперировать статистическими закономерностями, при практическом построении системы распознавания эмоций целесообразно проводить классификацию только наиболее существенных для решения конкретных задач эмоциональных состояний, что снижает ошибку классификации и повышает точность работы алгоритма.

Важное влияние на форму проявления эмоционального состояния оказывают культурные, языковые особенности и окружение диктора. Попытки многоязычной классификации эмоций демонстрируют значительное снижение эффективности их распознавания [3]. Помимо этого, эмоциональный корпус может

содержать специально подготовленные с участием актеров записи, либо спонтанную эмоциональную речь, полученную в реальных условиях [4]. Переход от модельных эмоциональных баз к распознаванию эмоций в спонтанной речи так же неминуемо ведет к заметному снижению эффективности работы алгоритмов. Тем не менее, существует определенная общность в выражении эмоций у различных людей, которой уделяется особое внимание в рамках эволюционной биологии [2], и которая делает возможной создание систем голосового анализа эмоционального состояния. При этом модельные эмоциональные базы, записанные при помощи профессиональных актеров, служат неплохим плацдармом для первоначальной оценки работоспособности разрабатываемых алгоритмов, позволяя на время избежать сложностей работы со спонтанной речью, хотя их репрезентативность существенно ниже, чем в случае реальных записей. Тем не менее, использование известных модельных корпусов эмоциональной речи, с которыми ранее работали другие группы исследователей, позволяет выявить относительную эффективность функционирования разрабатываемых алгоритмов.

Диктор-независимые системы голосового детектирования эмоционального состояния, работающие со спонтанной речью, могут комбинировать акустические и лингвистические информативные признаки, использовать алгоритмы сегментации речевого потока на отдельные эмоциональные фрагменты, иерархическую классификацию и т. п. Однако в любом случае основой алгоритма голосового анализа является модуль выделения информативных признаков речевого сигнала и классификатор, относящий звуковой фрагмент, согласно этим признакам, к тому либо иному эмоциональному классу. Соответственно, выделение новых, по возможности родственных человеческому восприятию, информативных признаков, а так же поиск новых высокоэффективных техник классификации на текущий момент времени являются важнейшими задачами голосового распознавания эмоционального состояния.

Исследования в области психологии и психолингвистики предоставили сведения о множестве акустических, просодических и лингвистических характеристик речи, способных служить информативными признаками при распознавании эмоционального состояния, и проявляющихся на уровне голосовых сегментов, слогов и целых слов. Чаще всего в целях дальнейшего анализа из аудио сигнала выделяют [5]:

- различные параметры частоты основного тона и формант;
- кратковременную оценку мощности;
- темп речи (количество слов произносимых в единицу времени);
- контур основного тона.

На основе выделяемого набора информативных признаков строится классификатор, который обучается на предварительно подготовленном наборе звуковых фрагментов. Классификация эмоциональных состояний производится в соответствии либо с задачами построения анализатора (оценки удовлетворенности, уровня стресса, усталости и т. п.), либо с выбранной моделью описания (набор базовых эмоций, непрерывная классификация и т. п.). Как правило, с ростом числа возможных вариантов классификации, точность распознавания эмоциональных состояний

значительно снижается. Соответственно, количество классов, используемых для обучения выбирается небольшим. Наиболее популярными техниками классификации являются следующие [5]: поиск ближайших соседей, метод опорных векторов, скрытые марковские модели, модель смеси нормальных распределений, модели на основе нечеткой логики, байесовские классификаторы максимума вероятности.

Далее в работе рассматривается набор информативных признаков и алгоритм, позволяющий с высокой степенью надежности и малой временной задержкой определять эмоциональное состояние диктора по голосу. Для оценки эффективности работы алгоритма аналогичная задача классификации решалась с использованием разработанного в мюнхенском университете инструментария распознавания эмоций openEAR [6].

2. Методы

Для автоматического распознавания были выбраны два эмоциональных состояния — нейтральное и агрессивное, отмеченных как «*neutral*» и «*anger*». Такой выбор обусловлен интересами дальнейшего применения разрабатываемой технологии. Обучение и тестирование алгоритма проводилось на записях, взятых из берлинской базы данных эмоциональной речи (Emo-DB) [7]. Данный корпус был собран в техническом университете Берлина, и неоднократно использовался исследователями при разработке систем распознавания эмоционального состояния. Он содержит записи эмоциональной речи на немецком языке, полученные с привлечением профессиональных актеров. База включает 535 записей речи 10 дикторов (5 мужчин, 5 женщин), воспроизводящих набор дискретных эмоциональных состояний, называемых иногда «архетипическими» (гнев, раздражение, страх, радость, печаль, удивление и нейтральное состояние). Авторское исследование Берлинской базы показало [7], что эмоции в ней распознаются слушателями в 80 % случаев, и в 60 % признаются естественными.

Из записей эмоциональной речи Берлинской базы данных выделялся ряд информативных признаков. Для каждого из них производилась оценка эффективности классификации. По полученным данным выбирался ряд наиболее эффективных показателей, по значениям которых производилось обучение классификатора, либо процедура классификации.

В блоке выделения признаков каждая запись предварительно умножалась на случайный коэффициент усиления от -20 до +20 дБ, чтобы исключить привязку к абсолютному уровню сигнала. В качестве возможных информативных признаков был выделен ряд прямых акустических характеристик сигнала, а так же набор метапризнаков, определяемых косвенно на основе данных прямых измерений. Прямыми характеристиками звукового сигнала являлись: оценка мощности, частота основного тона (ЧОТ), асимметрия от медианы ЧОТ, линейные спектральные частоты, кепстральные коэффициенты, вычисленные по коэффициентам предсказания, статистики высшего порядка, энергетический оператор Тигера [8]. Затем по известным данным прямых наблюдений рассчитывались 1я и 2я производные, энергетический оператор Тигера и ряд других параметров.

Оценка эффективности классификации наблюдения, необходимая для выявления наиболее существенных информативных признаков, осуществлялась следующим образом. В начале для каждого класса вычислялась медианная функция распределения на основе всех функций распределения (CDF) данного класса (Рисунок 1). Затем все функции распределения классифицировались по минимуму функции расстояния до медианных функций распределения. В качестве функции расстояния использовалась сумма модулей разностей функции распределения и медианной функции распределения. После этого вычислялась матрица неточностей. Эффективность классификации определялась как среднее значение диагональных элементов этой матрицы.

Классификация производилась с использованием двух моделей: опорных векторов и смеси нормальных распределений. В методе опорных векторов использовалось квадратичное ядро.

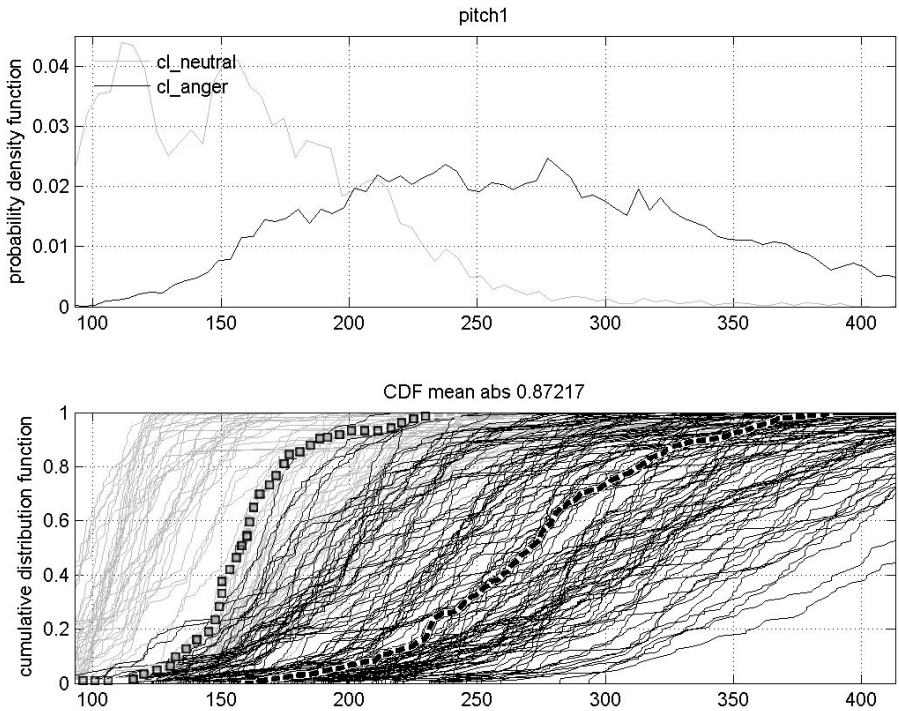


Рис. 1. Функции плотности вероятности (вверху) и распределения (внизу) частоты основного тона для состояния «neutral» (серые кривые) и «anger» (черные). Жирными прерывистыми линиями на нижнем рисунке отмечены медианные функции распределения для этих состояний. Видна область перекрытия функций распределения для двух эмоциональных классов. Оценка эффективности классификации для данного информативного признака ~0,87

Для оценки эффективности работы алгоритма, эта же задача распознавания эмоционального состояния решалась с использованием инструментария openEAR (open-source emotion and affect recognition toolkit) [6]. Данный набор программ был разработан в мюнхенском университете для нужд исследователей, работающих в сфере голосового анализа эмоционального состояния. Пакет доступен на условиях GNU General Public License, и включает в себя средства чтения и записи аудио файлов, выделения из речевого сигнала паралингвистических информативных признаков, и инструментарий для построения классификатора на основе библиотеки libSVM (широко известная библиотека, реализующая метод опорных векторов). Программа способна выделить свыше 6500 характеристик звукового сигнала, перечень которых задается при помощи файлов конфигурации. Файлы эмоциональной базы анализировались с использованием набора из 384 информативных признаков, сформированного авторами пакета openEAR по итогам конференции Interspeech'09. Количество записей в обучающей выборке при этом составляло 80 штук — 40 записей для нейтрального состояния, и 40 для состояния гнева. Документация к пакету libSVM [9], рекомендует в таких случаях, когда количество информативных признаков сравнимо либо превышает количество образцов в обучающей выборке, применять для классификации линейное ядро.

3. Результаты

Тестирование алгоритма позволило выявить ряд информативных признаков, эффективность классификации эмоциональной речи по которым оказалась максимальной (Таблица 1). Наиболее значимыми для принятия решения о принадлежности записи к классу «neutral» либо «anger» информативными признаками оказались частота основного тона, 2ой коэффициент линейных спектральных частот, вторая производная оценки мощности и эксцесс ошибки линейного предсказания (Таблица 1). В литературе существуют разногласия касательно взаимосвязи эмоциональных состояний с просодическими и акустическими характеристиками речи [1]. Однако влияние состояния гнева, помимо прочего, на частоту основного тона и форму огибающей энергии сигнала хорошо известно и не подлежит сомнению [10]. Таким образом, выделенные информативные признаки находятся в хорошем согласии с литературными данными.

Таблица 1. Эффективность классификации выделенного набора информативных признаков

Информативный признак	Эффективность классификации
Частота основного тона	0.87
Второй коэффициент линейных спектральных частот	0.90
Вторая производная оценки мощности	0.74
Эксцесс ошибки линейного предсказания	0.78

Для выбранного набора параметров, оценки точности работы алгоритма классификации оказались следующими. Ошибка классификации в случае использования метода опорных векторов составила порядка 4%. При применении модели смеси нормальных распределений соответствующее значение оказалось равным примерно 6%. Таким образом, использование выделенных информативных признаков позволило распознавать в базе Emo-DB состояния «neutral» и «anger» с точностью порядка 94–96% в зависимости от используемого алгоритма классификации.

В свою очередь, модель, построенная и обученная с использованием инструментария openEAR, задействующая набор из 384 информативных признаков и классификатор, работающий по методу опорных векторов с линейным ядром, имела для этой же задачи точность классификации порядка 98%. Иными словами, точность классификации, достигаемая с использованием этой методики лишь на 2% превосходит точность, достигнутую при помощи ограниченного набора информативных признаков, отобранных по принципу наибольшей эффективности классификации. Данный факт свидетельствует о том, что при отнесении голоса диктора к одному из классов «anger» либо «neutral» целесообразно пользоваться небольшим набором информативных признаков, важную роль в котором играют характеристики основного тона и мощности сигнала.

4. Анализ и выводы

Автоматическое распознавание эмоционального состояния окажется полезным в любой сфере человеческой деятельности, где требуется его оперативная оценка — в маркетинге, медицине, психологии, обеспечении безопасности и т. п. Разработка такой технологии позволит качественно изменить форму коммуникации между человеком и машиной. Более того, разрабатываемые здесь подходы находят свое применение не только в сфере анализа эмоционального состояния, но и при распознавании других состояний, например — алкогольной интоксикации, усталости, подавленности и т. п. Тем не менее, общие теории взаимосвязи эмоций с характеристиками голоса на данный момент отсутствуют, что вынуждает исследователей каждый раз заниматься разработкой новых и тонкой подстройкой существующих алгоритмов под условия конкретной задачи.

Из литературы известно, какие из параметров голоса могут служить индикаторами состояния гнева [10]. Прежде всего, исследователи отмечают его влияние на характеристики частоты основного тона. Возрастает его медианное значение, скорость изменений, расширяется его диапазон. В состоянии гнева диктор издает звуки с более открытым речевым трактом, что приводит к возрастанию средней частоты первой форманты. Так же, по отношению к ней, возрастают амплитуды второй и третьей формант, повышается неоднородность формантных контуров. Кроме частотных параметров голоса важную роль играют характеристики огибающей его энергии. В состоянии гнева энергия

речевого сигнала увеличивается. К вышеназванным признакам можно добавить еще увеличение скорости речи, а так же ряд других показателей.

Анализ акустических характеристик записей берлинской базы эмоциональной речи позволил выделить ряд параметров с наибольшей эффективностью классификации состояний «neutral» и «anger». Включение выделенных информативных признаков в алгоритм, работающий на основе метода опорных векторов с квадратичной разделяющей функцией, позволило добиться точности классификации порядка 96%. Данный показатель лишь незначительно (~2%) уступает точности, полученной при помощи набора из 384 информативных признаков, выделенных и обработанных при помощи стандартного инструментария пакета openEAR.

Столь высокий процент распознавания можно объяснить, прежде всего, идеальностью использованного корпуса, и тем фактом, что классифицировались только лишь два эмоциональных состояния. Как известно из литературы [1, 5], увеличение числа распознаваемых эмоций, равно как и переход от модельных эмоциональных баз данных к реальным, необходимо ведет к возрастанию ошибки классификации. Так, средствами openEAR, можно получить точность распознавания одного из семи эмоциональных состояний на берлинской базе эмоциональной речи порядка 89% [6]. Кроме того, даже в модельных условиях эффективность работы алгоритмов распознавания эмоций может существенно варьироваться для различных языков [3]. Тем не менее, предполагается, что описанный набор информативных признаков и алгоритм классификации, работоспособность которых была проверена на корпусе Emo-DB, после соответствующей адаптации, будет, в качестве составного элемента, включен в систему, работающую с русскоязычными голосовыми базами данных в реальных условиях.

References

1. *Ayadi M. El, Kamel M. S., Karray F.* 2011. Survey on Speech Emotion Recognition:
2. *Features,* Classification Schemes, and Databases. Pattern Recognition, 44(3) : 572–587.
3. *Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W., Weiss B.* 2005. A Database of German Emotional Speech. Proc. Interspeech.
4. *Chih-Chung C., Chih-Jen L.* 2001. LIBSVM: a Library for Support Vector Machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
5. *Cornelius R. R.* 1996. The Science of Emotion: Research and Tradition in the Psychology of Emotions.
6. *Eyben F., Wöllmer M. and Schuller B.* 2009. OpenEAR — Introducing the Munich OpenSource Emotion and Affect Recognition Toolkit. Proc. ACII : 576–581.
7. *Hozjan V., Kacic Z.* Context-independent Multilingual Emotion Recognition from Speech Signal. Int. J. Speech Technol, 6 : 311–320.
8. *Morrison D., Wang R., De Silva L. C.* 2007. Ensemble Methods for Spoken Emotion Recognition in Call-Centres. Speech Communication, 49 : 98–112.

9. *Ververidis D., Kotropoulos C.* 2003. A Review of Emotional Speech Databases. Proc. Panhellenic Conference on Informatics (PCI) : 560–574.
10. *Pantic M., Rothkrantz L. J. M.* 2003. Toward an Affect-Sensitive Multimodal Human–Computer Interaction. Proc. of the IEEE, 91(9) : 1370–1390.
11. *Zhuikov V. Ia., Kuznetsov N. N., Kharchenko A. N.* 2010. Signals Change Evaluation with Differential Energetic Operators [Otsenka Izmeneniia Signalov s pomoshch'iu Differentsial'nykh Energeticheskikh Operatorov]. Elektronika I Sviaz'. 3' Tematicheskii Vypusk "Elektronika I Nanotekhnologii" : 63–67.

СИНТАКСИЧЕСКИЙ АНАЛИЗ ТЕКСТА С ОРФОГРАФИЧЕСКИМИ ОШИБКАМИ В СИСТЕМЕ DICTASCOPE SYNTAX

Т. Н. Ерехинская (te@dictum.ru)

А. С. Титова (titova@dictum.ru)

В. В. Окатьев (oka@dictum.ru)

ООО «Диктум», Нижний Новгород, Россия

В статье рассматривается синтаксический анализ ЕЯ-текстов с опечатками. Предлагается способ интеграции модуля проверки орфографии и синтаксического анализатора, позволяющий с одной стороны исправлять опечатки с учетом контекста и повысить устойчивость синтаксического анализатора с другой.

Ключевые слова: синтаксический анализ, синтаксический анализатор, опечатки, ошибки, орфографические ошибки

SYNTAX PARSING FOR TEXTS WITH MISSPELLINGS IN DICTASCOPE SYNTAX

T. N. Erekhinskaia (te@dictum.ru)

A. S. Titova (titova@dictum.ru)

V. V. Okat'ev (oka@dictum.ru)

Dictum Ltd., Nizhny Novgorod, Russia

The paper deals with syntax parsing of natural language texts with misspellings and misprints in DictaScope Syntax. We propose a method for integration of a spellchecker and parser, which allows us on he one hand to correct typographical errors considering the context and on the other hand to increase the robustness of the parser. We start by outlining various types of misprints and ways to correct them, taking account of the specifi c character of keyboard typing and typical mistakes. To correct the misspellings and misprints we propose to use a modifi ed Levenshtein algorithm, in which each pair of characters involved in calculation of the Levenshtein distance is assigned a specifi c weight from the interval. This accounts for keyboard typing, phonetically similar characters,similarity between Russian and Latin alphabet symbols, numbers and other symbols.The paper states the need to take into account the

lexical context of the words to be corrected in order to achieve the maximum accuracy of correction, which helps correct words used in an unusual context. As a result we get a number of correction options for the words. The final choice is made by the DictaScope parser Basing on the modified Eisner algorithm, the parser builds a dependency tree for the sentence. The modification includes punctuation cheking and some additional linguistic limitations. In our model several vertices of interpretations correspond to one word, and variants of spell correction could be processed in the same way as morphological interpretations. The integration of misprint correction and syntactic analysis is illustrated by a simple case (correcting a single word) and a more complex case — splitting a word in two or merging two words into one. The proposed method of integration of the parser and the spellchecker modules was implemented in the DictaScope Syntax system. This made it possible to considerably increase the stability of the parser and provided an opportunity to use it as a component of the opinion mining system for monitoring of blogs and forums.

Key words: syntax parsing, syntax parser, parsing, misspellings, misprints.

Введение

Орфографические ошибки, опечатки, намеренное искажение слов — одна из характерных особенностей современного текстового контента. Борьба с опечатками давно вышла за рамки текстовых редакторов, где исправление проводилось в интересах человека — читателя. Такие небольшие недоразумения как пропущенная буква или вставленный по ошибке пробел мешают автоматической обработке текста, поэтому многие системы АОТ включают в себя исправление опечаток.

Поисковые системы вынуждены исправлять — «переколдовывать» — опечатки в запросах. Системам мониторинга блогов и форумов также требуется исправление опечаток или умение обходить их.

Многие прикладные приложения обработки текста включают синтаксический анализатор в качестве компонента. С одной стороны это позволяет добиться хорошего качества работы, с другой — делает приложение зависимым от результатов синтаксического анализа. Большинство синтаксических анализаторов не справляется с обработкой предложений, содержащих опечатки. Таким образом, само приложение также становится уязвимым.

Задача исправления орфографических ошибок имеет длинную историю [5–7,14]. Традиционные методы фокусировались на исправлении ошибок вставки, удаления, замены и перестановки символов в неизвестных словах (то есть словах, которые не содержатся в используемом словаре). Heidorn [8] и Garside [9] разработали систему, которая полагалась на синтаксические шаблоны в распознавании ошибки замены, когда как Mays [10] использовали данные совместной встречаемости слов из большого корпуса, чтобы обнаружить и исправить такие ошибки [11]. Совместная встречаемость слов используется и в исправлении запросов в поисковой системе Yandex [12].

Некоторые системы проверки орфографии (например, ОРФО) рассматривают только одноместные замены внутри неизвестного слова. Это значит, что слова, в написании которых допущено более одной ошибки, останутся без вариантов замены.

Наиболее близкий подход можно найти в статье Vosse [15], описывающей интеграцию модуля проверки орфографии и синтаксического анализатора для немецкого языка. Особое внимание автор уделил ошибкам согласования, в результате которых отдельные слова в предложении остаются словарными, но само предложение перестает быть грамматически верным. В работе использован анализатор на основе алгоритма Томиты. Для обработки структурных ошибок в грамматику были введены дополнительные правило.

По сравнению с работой Vosse, целью данной работы является повышение устойчивости анализатора к опечаткам за счет минимальных изменений.

1. Постановка задачи

Задачу синтаксического анализ текста с орфографическими ошибками можно рассматривать с двух сторон. Во-первых, в синтаксическом анализаторе возможно использование модуля, предлагающего варианты исправлений для слов с ошибками. Это позволит повысить устойчивость синтаксического анализатора. Во-вторых, учет синтаксического контекста может быть очень полезен для корректного исправления опечаток. Например, в предложении «*мне нравиться телефон*» все слова написаны правильно, однако само предложение некорректно.

Цель данной работы — расширение возможностей синтаксического анализатора за счет внедрения исправления опечаток и совершенствование исправления опечаток за счет использования синтаксического анализа. Схема взаимодействия модулей проиллюстрирована на рис. 1.

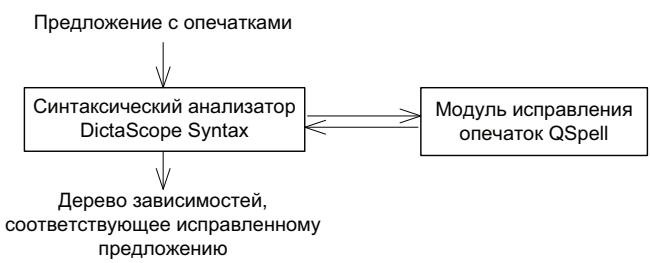


Рис. 1. Взаимодействие модулей синтаксического анализа и исправления опечаток

2. Виды опечаток

При исправлении орфографии должны учитываться следующие типы ошибок:

1. вставка лишнего символа «*будующее — будущее*»;
2. пропуск символа «*грусно — грустно*»;

3. замена одного символа на другой *«кинтакт — контакт»*;
4. транспозиция рядом стоящих символов *«рбаво — браво»*.

Также надо учитывать особенности печати слов на клавиатуре и типичные ошибки. К ним можно причислить:

1. клавиатурная близость клавиш: *«анеудот — анекдот»*;
2. ошибки в безударных гласных: *«аностасия — анастасия»*;
3. фонетическая похожесть букв: *«брюпетка — брюнетка»*;
4. парные буквы: *«автограв — автограф»*;
5. вставка лишнего пробела: *«сло во — слово»*;
6. отсутствие пробела или дефиса: *«футбольныйклуб — футбольный клуб»*;
7. идентичное написание букв в разных раскладках: «хромосом — хромосом»;
8. схожесть написания цифр и букв (ч-4, о-0, з-3): *«4естно — честно»*;
9. буквы и символы в разных раскладках: *«<лизнец — близнец»*;
10. ошибки после шипящих и ц: *«жолтый — желтый»*;
11. перепутывание и смещение рук при слепой печати: *«инвнжае — телефон»*;
12. перевод транслитерации на русское написание: *«kartinki — картинки»*;
13. исправление неправильной раскладки клавиатуры как для целого, так и для части слова: *«jlyjrkfccybrb — одноклассники»*.

3. Исправление опечаток

3.1. Базовый алгоритм

Постановка задачи исправления ошибок и опечаток в словах может быть сформулирована следующим образом:

Пусть $\sum$ будет алфавитом нашего языка и $L \subset \sum^*$ — словарь грамматически правильных слов. Для слова $w' \in \sum^* \setminus L$ требуется найти слова $w' \in L$ такие, что $dist(w, w') \le \delta$ и $F(w') = \max_{v \in L; dist(w,v) \le \delta} F(v)$. [11]

Для некоторого несловарного слова из текста требуется найти одну или несколько ближайших словоформ из словаря, и выбрать из них наиболее частотные, где:

$dist(w, w')$ — расстояние между двумя словами – мера, показывающая насколько одно слово похоже на другое. Будем использовать модифицированный подсчет расстояния Левенштейна.

δ — пороговое значение расстояния. Если $dist(w, w') \le \delta$, то считаем, что слова w и w' близки, то есть, можно сказать, что слово w' является исправлением несловарного слова w. $F(w)$ — частота употребления слова w.

В классическом варианте алгоритма Левенштейна вес операций устанавливают равным единице. Модификация заключается в том, что каждой паре символов, участвующей в определении расстояния Левенштейна, назначен

свой вес из интервала . Если нужный вес не находится среди данных матриц, то он автоматически считается равным единице. Для операций определены следующие матрицы весов:

1. замена одного символа на другой;
2. замена символов после шипящих и ц;
3. транспозиция рядом стоящих символов;
4. вставка символов;
5. удаление символов.

Подобная модификация объясняется тем, что буквы определенного алфавита используются неравномерно: некоторые символы употребляются чаще других, поэтому участвуют в большем количестве ошибок. Помимо этого, матрицы позволяют учесть особенности печати на клавиатуре (описка *«а-п»* более вероятная, чем *«а-ч»* в силу расположения клавиш на клавиатуре), назначить свой собственный вес для фонетически похожих символов (ошибка *«в-ф»* будет более частотной, следовательно, и более вероятной, чем *«г-к»*), распределить соответствие символов русского алфавита, латиницы, цифр и знаков (*«т-м»* и *«т-т»*).

3.2. Лексический контекст

У отдельного слова с опечаткой может быть несколько вариантов исправления, причем исправления могут иметь разные грамматические характеристики.

Пример. *гоод*

Варианты исправления: «*год*», «*город*», «*голод*», «*горд*», «*голд*».

Как видно из примера, при исправлении опечаток недостаточно близости символьного представления слов.

Пример. *аренда катра*

В данном случае «*катра*» может быть «*катером*» и «*картой*», но по окружению слов легко можно понять, что имеется в виду «*катер*».

При исправлении такого рода ошибок следует учитывать контекст, поскольку при неправильном выборе варианта исправления может измениться или потеряться смысл (*белый грип — белый гриб*, а *птичий грип — птичий грипп*). В этих целях используется словарь биграмм или словарь сочетаемости слов.

Данный словарь используется также и при исправлении пропуска пробела, когда два слова ошибочно написаны слитно.

Пример. *футбольныйклуб — футбольный клуб*

Считается, что от 80 % до 90 % всех опечаток содержат одну ошибку в слове и орфографический корректор, пытающийся достичь 90 % точности, обязательно должен использовать контекст [13].

3.3. Словарное слово + опечатка = словарное слово

До этого момента мы рассматривали исправление ошибок только среди несловарных слов. Однако то, что рассматриваемое слово содержится в словаре, еще не означает, что оно не содержит ошибок. Довольно распространенный тип ошибок — это ошибки в окончаниях глаголов на -тся/-ться.

Пример. *очень нравиться телефон*

Слово «*нравиться*» является словарным, но оно также содержит ошибку.

Если такая ошибка не будет исправлена, то синтаксический анализатор не сможет корректно построить дерево синтаксического разбора, поскольку будет отсутствовать связь между словами «*нравиться*» и «*телефон*». Само слово является словарным, поэтому на этапе исправления нельзя отбрасывать исходное слово. Модуль исправления опечаток предложит варианты исправления, а синтаксический анализатор сделает окончательный выбор.

3.4. Ошибки в окончаниях

Наверное, каждый из нас сталкивался с тем, что в ходе написания фразы мысленно ее переформулировал, после чего появляются ошибки в согласовании и управлении слов. Предположим, что хотели написать предложение «*Он был веселым мальчиком*», в процессе набора решили перефразировать его до «*Он был веселый мальчик*», а в результате получили «*Он был веселым мальчик*». Чтобы исправить ошибку согласования, требуется прилагательное поставить в нужный падеж, для чего необходимо подобрать правильное окончание. Для этого на этапе исправления искажений получаем возможные формы слова «*веселый*», затем синтаксический анализатор выберет нужную форму.

4. Синтаксис

4.1. Принятая модель

Синтаксический анализатор DictaScope строит дерево зависимостей для предложения [2–4]. Реализованный подход основан на следующей модели.

Пусть на вход анализатору приходит цепочка слов S: $S = w_1, w_2, ..., w_N$, где w_i — лексема при $i = \overline{1, N-1}$, w_N — слово, соответствующее фиктивному корню дерева зависимостей. По входной цепочке слов строится ориентированный взвешенный граф $G = \langle V, E \rangle$, вершинам графа назначены номера: $V = \{1, ..., n\}$. Задано разбиение V на отрезки-классы $I_k = [a_k, b_k]$, $k = \overline{1, N}$, $a_1 = 1$, $a_{k+1} = b_k + 1$, $k = \overline{1, N-1}$, $w_N = w_N = n$. Каждое множество I_k соответствует множеству омоформ слова w_k в предложении. Под омоформой понимается пара: грамматическое значение и начальная форма.

Для построения дерева зависимостей используется модифицированный алгоритм Эйснера [1]. Модификация включает проверку пунктуации и некоторые дополнительные лингвистические ограничения. Однако для целей данной работы важным является лишь то, что одному слову соответствует несколько вершин-интерпретаций.

Далее приведен модифицированный алгоритм Эйснера. Обозначим номер множества, содержащего вершину $j \in V$: $ind(j) \stackrel{def}{=} k \Leftrightarrow j \in I_k$. Введем функцию расстояния между двумя вершинами $L(r, l)$, $r, j \in V$.

$L(r, l) = |ind(r) - ind(l)| - c$,
где c — количество сочинительных союзов между словами $ind(r)$ и $ind(l)$.

Входом алгоритма является описанный выше граф $G = \langle V, E \rangle$ выходом - минимальное проективное дерево $\dot{O} = \langle V^*, E^* \rangle$: $V^* \subseteq V$, $E^* \subseteq E$, причем из каждого класса в дерево входит строго одна омоформа: $\forall k \in \{1, ..., N\}$ $|I_k \cap V^*| = 1$.

Алгоритм Эйснера представляет из себя заполнение таблицы $C[x][y][d][q]$, где x, y — левая и правая вершины, d — направление поддерева, q — показатель завершенности.

Начало алгоритма
C[s][s][d][c] = 0 ∀s; d; c
для k: 1..n
    для s: 1..n
        t = s + k
        если t > n прервать цикл
        если $ind(s) = ind(t)$
            C[s][t][d][c] = 0 ∀d; c
            следующая итерация

$$C[s][t][\leftarrow][0] = \min_{\substack{s \le r, l \le t \\ L(r,l)=1}} \left(C[s][r][\rightarrow][1] + C[l][t][\leftarrow][1] + s(t; s) \right)$$

$$C[s][t][\rightarrow][0] = \min_{\substack{s \le r, l \le t \\ L(r,l)=1}} \left(C[s][r][\rightarrow][1] + C[l][t][\leftarrow][1] + s(s; t) \right)$$

$$C[s][t][\leftarrow][1] = \min_{\substack{s \le r < t \\ \mathrm{ind}(t) \ne \mathrm{ind}(r)}} \left(C[s][r][\leftarrow][1] + C[r][t][\leftarrow][0] \right)$$

$$C[s][t][\rightarrow][1] = \min_{\substack{s < r \le t \\ \mathrm{ind}(t) \ne \mathrm{ind}(r)}} \left(C[s][r][\rightarrow][0] + C[r][t][\rightarrow][1] \right)$$

конец цикла
конец цикла
конец алгоритма

Решению соответствует клетка матрицы
$C[0][k^*][x^*][y^*] : k^*, x^*, y^* = \arg\min_{k,x,y} C[0][k][x][y]$.

Для получения результирующего дерева выполняется проход сверху вниз. Для этого нужно поддерживать обратные указатели на поддеревья, которые составляют каждый элемент таблицы.

4.2. Адаптация модели для исправления опечаток

Обсуждение интеграции исправления опечаток и синтаксического анализа начнем с наиболее простого случая — исправление одного слова на другое.

Пусть слову w_k соответствуют варианты исправления $w_1^k, w_2^k, ..., w_Q^k$. Тогда расширим множество омоформ I_k, добавив к нему омоформы, соответствующие вариантам исправления: $I_k = [a_k, b_k] \cup [\tilde{a}_1^k, \tilde{b}_1^k] \cup [\tilde{a}_2^k, \tilde{b}_2^k] \cup ... \cup [\tilde{a}_Q^k, \tilde{b}_Q^k]$, сохранив принцип нумерации: $\tilde{a}_1^k = b_k + 1$, $a_{k+1} = \tilde{b}_Q^k + 1$, $\tilde{a}_{i+1}^k = \tilde{b}_i^k + 1$, $k = \overline{1, N-1}$. Кроме того, расширим представление омоформы до тройки: вариант исправления (может совпадать с исходным словом), грамматическое значение, начальная форма.

Таким образом, множество омоформ слова w_k расширено за счет добавления омоформ вариантов его исправлений – это соответствует добавлению новых вершин в граф *G*. Потенциальные связи для новых вершин устанавливаются на общих основаниях с «обычными» вершинами – по базе синтаксических правил.

Пример. *Мне нравиться телфон.*

Слово *телфон* может быть проинтерпретировано как несловарное слово, в этом случае ему будут соответствовать 6 грамматических значений имени собственного (во всех падежах). Модуль исправления опечаток предложит два варианта: *телефон* и *тефлон*. Каждому соответствует по две омоформы — в именительном и винительном падеже. Таким образом, слову *телфон* будет соответствовать 10 вершин в графе *G*. Аналогично, для слова *нравиться* будет предложено исправление *нравится*.

В качестве результирующего может быть выбрано дерево для следующих вариантов: *«мне нравится телефон»* и *«мне нравится тефлон»*. Кроме того, возможное решение — вариант *«мне нравится телфон»*, где для слова *телфон* возможны грамматические значения имени собственного в именительном, дательном или творительном падежах. Однако вес деревьев со словом *телфон* содержит высокий штраф за несловарное слово, остаются только деревья с исправлениями. Учет частоты встречаемости слов помогает выбрать между *телефоном* и *тефлоном*.

Таким образом, в рамках принятой модели задача исправления опечаток в отдельном слове решается за счет дополнительных омоформ. Однако исправление опечаток не ограничивается коррекцией отдельного слова. Более сложный случай — разделение одного слова на два или соединение двух слов в одно. Проведем дальнейшую модификацию алгоритма.

По входной цепочке слов $S = w_1, w_2, ..., w_N$ модуль исправления опечаток может предложить варианты склейки/разделения слов: w_i, $w_{i+1} \to \overline{w}_i$ или $w_i \to w_{i1}, w_{i2}$. Как включить использование этой информации в алгоритм Эйснера?

Суть алгоритма Эйснера — построить наилучшие деревья зависимостей на отрезках, а затем собрать из них результирующее дерево.

При построении дерева на отрезке $[w_i, w_i]$можно рассматривать все слова с их омоформами, а можно исключить из рассмотрения одно или несколько слов, как это показано на рис. 2.

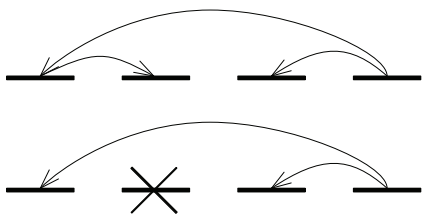


Рис. 2. Построение дерева на отрезке с исключением слова

Будем говорить, что слово, которое при некоторых условиях может не попасть в результирующее дерево, является опциональным. В контексте текущей задачи важно понять, что же мы считаем словом. На самом деле слово можно рассматривать как некоторую абстракцию, содержащую набор омоформ. Слова задают классы в графе *G*, способ нумерации вершин и ограничение на связи — вершины из одного класса не могут быть связаны.

Возьмем начальное разбиение $S = w_1, w_2, ..., w_N$, построим множество вершин *V*. Затем для каждого варианта исправления выполним следующие действия. Для исправлений вида $w_i \to w_{i1}$, w_{i2} добавим к омоформам слова w_i омоформы слова w_{i1}, добавим опциональное слово w_{i2} в цепочку *S* после слова w_i и омоформы слова w_{i2} в множество вершин *V*. Если для слова w_i есть несколько вариантов исправления вида $w_i \to w_{i1}$, w_{i2}, все омофрмы для различных вариантов w_{i2} приписываются новому опциональному слову.

Для исправлений вида w_i, $w_{i+1} \to \overline{w}_i$ к омоформам слова w_i добавляются омоформы слова $\overline{w}_i$, слово w_{i+1} помечается как опциональное.

После выполнения этих операций для всех вариантов исправления необходимо перенумеровать слова в *S*, вершины в *G* и слова в вариантах исправлений. Кроме того, при добавлении вершин необходимо сохранять информацию о несовместимости вершин из соседних слов. Если две вершины несовместимы, в дерево может попасть не более одной из них. Проиллюстрируем сказанное примером.

Пример. *Лучшепотом.*

$S = w_1$, $w_1 \to$ *«Лучшепотом»*. Варианты исправления: $w_1 \to$ *«Лучше», «потом»*, $w_1 \to$ *«Луч», «шепотом»*. С учетом исправлений $S = w_1, w_2$, w_2 — опциональное слово. Множество вершин V показано на рис. 3. Символ $\varnothing$ соответствует исключению опционального слова из рассмотрения. Пунктиром показаны несовместимые вершины.

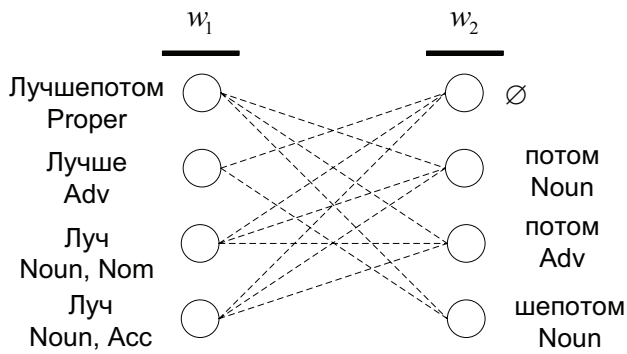


Рис. 3. Множество вершин для учета исправлений

Дополнительно в каждой ячейке таблицы $C[x][y][d][q]$ будем хранить список вершин, несовместимых с деревом, соответствующим данной ячейке. Для учета совместимости вершин при заполнении ячейки поиск минимального дерева ведется только среди тех деревьев, в которых все вершины попарно совместимы.

Заключение

Предложенный способ интеграции синтаксического анализатора и модуля исправления опечаток был реализован в системе DictaScope Syntax. Это позволило существенно повысить устойчивость анализатора и дало возможность использовать его в качестве компонента в системе извлечения мнений для мониторинга блогов и форумов.

Была проведена серия экспериментов с целью получения численной оценки качества анализа на текстах с опечатками. Полнота и точность модуля исправления ошибок и опечаток в словах составили 85.0% и 80.8% соответственно, скорость — 0.15 Кб/с. Качество синтаксического анализа (процент правильно построенных деревьев синтаксических связей) на текстах с опечатками выросло с 23% до 65%. Данные были получены на машине Athlon 3,1 GHz.

References

1. *Baitin A.* 2008. Queries Correction in Yandex. Probabilistic Linguistic Models. [Ispravlenie Poiskovykh Zaprosov v Iandekse].
2. *Cucerzan S., Brill E.* Spelling Correction as an Iterative Process that Exploits the Collective Knowledge of Web Users. Microsoft Research. One Microsoft Way.
3. *Damerau F. J.* 1964. A Technique for Computer Detection and Correction of Spelling Errors. Communications of ACM, 7(3):171–176.
4. *Eisner J.* 1996. Three New Probabilistic Models for Dependency Parsing: An Exploration. Proceedings of the 16th International Conference on Computational Linguistics (COLING) : 340–345.
5. *Heidorn G. E., Jensen K., Miller K., Byrd R. J., Chodorow M. S.* 1982. The EPISTLE Text-Critiquing System. IBM Systems Journal, 21(3):305–326.
6. *Garside R., Leech G., Sampson G.* 1987. Computational Analysis of English: A Corpus-based Approach.
7. *Mays E., Damerau F. J., Mercer R. L.* 1991. Contextbased Spelling Correction. Information Processing and Management, 27(5) : 517–522.
8. *McIlroy M. D.* 1982. Development of a Spelling List. JIEEE-TRANS-COMM, 30(1): 91–99.
9. *Norvig Peter.* How to Write a Spelling Corrector, available at: http://norvig.com/spell-correct.html
10. *Okat'ev V. V., Erekhinskaia T. N., Skatov D. S.* 2009. The Models and Methods of Punctuation Accountability for Russian Sentence Syntactic Analysis [Modeli I Metody Ucheta Punktuatsii pri Sintaksicheskom Analize Predlozheniia Russkogo Iazyka]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2009" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2009").
11. *Okat'ev V. V., Erekhinskaia T. N., Ratanova T. E.* 2010. Secret Punctuation Signs [Tainye Znaki Punktuatsii]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2010" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2010").
12. *Okat'ev V. V., Gergel' V. P., Alekseev V. E., Talanov V. A., Barkalov K. A., Skatov D. S., Erekhinskaia T. N., Kotov A. E., Titova A. S.* 2008. R&D Report on the theme: "Designing of the Russian Syntactic Analysis System Pilot Version" (# 02200803750) [Otchet o VYpolnenii NIOKR po teme: "Razrabotka Pilotnoi Versii Sistemy Sintaksicheskogo Analiza Russkogo Iazyka" (inventarnyi nomer VNTITs 02200803750)].
13. *Rieseman E. M., Hanson A. R.* 1974. A Contextual Postprocessing System for Error Correction Using Binary N-grams. IEEE Transactions on Computers, 23(5) : 480–493.
14. *Ristad Eric Sven, Yianilos Peter N.* 1998. Learning String-Edit Distance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20 (5).
15. *Vosse Theo.* 1992. Detecting and Correcting Morpho-Syntactic Errors in Real Texts. Proceedings of the Third Conference on Applied Natural Language Processing.

ЭКСПЕРИМЕНТАЛЬНЫЙ АНАЛИЗ ДИСКУРСА: РЕФЕРЕНЦИАЛЬНЫЙ ВЫБОР В СИТУАЦИИ ПОТЕНЦИАЛЬНОГО РЕФЕРЕНЦИАЛЬНОГО КОНФЛИКТА (ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ НА МАТЕРИАЛЕ РУССКОГО ЯЗЫКА)[1]

О. В. Федорова (olga.fedorova@msu.ru)

А. М. Успенская (ania.quies@gmail.com)

Московский государственный университет
имени М. В. Ломоносова, Москва, Россия

Ключевые слова: дискурс, референциальный выбор, референциальный конфликт, анализ, экспериментальный анализ.

EXPERIMENTAL ANALYSIS OF DISCOURSE: THE IMPACT OF A POTENTIAL REFERENTIAL CONFLICT ON THE CHOICE OF THE REFERRING EXPRESSION (ON THE MATERIAL OF RUSSIAN)

O. V. Fedorova (olga.fedorova@msu.ru)

A. M. Uspenskaia (ania.quies@gmail.com)

Lomonosov Moscow State University, Moscow, Russia

The paper describes an experiment carried out in order to study the referential choice in situation of potential referential conflict. The results showed that in the situation participants choose full NP. The results confirmed that referential choice depends on the participants' working memory and made some additions to the model of referential choice.

Key words: discourse, referential choice, referential conflict, analysis, experimental analysis.

[1] Работа выполнена при частичной финансовой поддержке гранта РГНФ «Когнитивные основы языковой структуры и дискурсивных явлений: теория и практика». Авторы выражают благодарность А. А. Кибрику и Е. В. Печенковой за критические замечания и советы, высказанные при подготовке работы, а также за помощь в создании некоторых примеров.

1. Введение. Экспериментальный анализ дискурса

Экспериментальный анализ дискурса (ЭАД) — это совсем молодое направление исследований, которое оформилось в 90-ых годах XX века с выходом книги Г. Кларка “Arenas of language use” (Clark 1992). В этой работе Кларк описывает две психолингвистические традиции, которые напоминают генеративный и функциональный подходы в лингвистике, — «язык как продукт» и «язык как действие». Первая традиция восходит к работам Дж. Миллера и Н. Хомского; ее сторонники занимаются отдельными языковыми репрезентациями, т. е. «продуктами» процесса понимания высказывания. Вторая традиция берет начало с работ лингвистов-философов Дж. Остина, П. Грайса и Дж. Серля; психолингвисты, работающие в рамках этой традиции, занимаются изучением речевого взаимодействия собеседников в процессе реальной коммуникации, т. е. по сути ЭАД.

Удельный вес работ по ЭАД по сравнению с другими уровневыми составляющими пока еще совсем невелик. Такое состояние дел может быть объяснено как субъективным фактором — сильной психолингвистической традицией, восходящей к идеям генеративной грамматики, ограничивающей любое исследование уровнем отдельного предложения, так и более объективными причинами, связанными с «излишней» многоаспектностью дискурса. Тем не менее, существует целый ряд работ, выполненных в данной парадигме. В числе прочих стоит отметить несколько исследований, в которых были использованы традиционные оффлайновые методы — в частности, целое направление работ Г. Кларка, объединенных общей идеей создания совместной модели взаимодействия собеседников в процессе диалога (Clark and Wilkes-Gibbs 1986), а также многочисленные работы последних лет М. Пикеринга и С. Гэррода, описывающие механизмы уподобления ситуационных моделей собеседников (Pickering and Garrod 2007). С другой стороны, Й. ван Беркум с коллегами проводит исследования дискурса с помощью новейших онлайновых методов вызванных потенциалов мозга и функциональной магнитно-резонансной томографии (van Berkum in press).

В настоящей работе мы на дискурсивном материале пройдем один полный цикл, обычный для работы экспериментального лингвиста: начнем с теоретического обзора (разделы 2 и 3), затем опишем собственно эксперимент (раздел 4), обсудим его результаты (раздел 5), а в разделе 6 вернемся к теоретической модели и предложим некоторые ее уточнения.

2. Исследование референции. Модель референциального выбора

Данная работа продолжает серию экспериментальных исследований дискурсивной референции, представленную на конференции Диалог’2010 (Федорова и др. 2010). В отличие от прошлогоднего исследования, в котором были рассмотрены механизмы понимания референциальных средств (РС), эта работа посвящена описанию процесса их порождения, т. е. референциальному выбору (РВ) говорящего.

В качестве значимых факторов, влияющих на РВ, предлагались как различным образом измеренные расстояния до антецедента — линейное, риторическое или расстояние в абзацах, так и его семантико-синтаксический статус и внутренние свойства. В работах А. А. Кибрика (напр., Kibrik 2011) процесс РВ описывается при помощи многофакторного количественного подхода, согласно которому РВ зависит от степени активации референта в рабочей памяти (РП) говорящего. Таким образом, факторы, оказывающие влияние на РВ, рассматриваются как факторы активации, дающие в сумме коэффициент активации (КА). Чем выше КА, тем больше вероятность употребления редуцированного РС (РедуцРС) — местоимения или нуля.

Кроме определения КА референта в РП говорящего модель РВ (см. рис. 1) включает фильтр референциального конфликта[2] (РК), действие которого блокирует использование РедуцРС в случае высокой активации более чем одного референта.

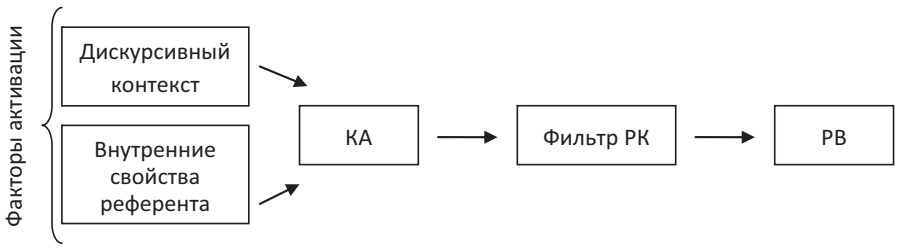


Рис. 1. Модель РВ из работы Kibrik 2011 в переводе на русский язык

3. Референциальный конфликт глазами Адресата

В этом разделе мы предложим типологию РК с точки зрения адресата. Будем называть РК такую ситуацию, при которой в пределах текущего дискурсивного фрагмента адресат может отнести использованное говорящим РедуцРС[3] к нескольким референтам, активированным в его РП. Мы будем различать временный РК, который снимается к концу дискурсивного фрагмента при помощи референциальных деконфликторов (РД), и постоянный РК (примеры 1–3). Временные РК различаются сферой действия: снятие РК может происходить в пределах клаузы (4)[4], предложения (2'), абзаца (1') и целого дискурса (2'').

[2] Этот термин был впервые использован в статье Кибрик 1987.

[3] Мы осознанно несколько упрощаем ситуацию, так как в принципе РК может возникнуть и в ситуации использования полной ИГ.

[4] С некоторой долей условности мы рассматриваем этот тип как случай реализованного, а не предотвращенного (подробнее см. ниже) РК. Однако мы считаем вполне осмысленным предположение о том, что РК, снятый в пределах одной клаузы, является предотвращенным РК. Окончательный ответ на этот вопрос, как мы надеемся, дадут будущие исследования.

(1) Кошка$_i$ почуяла собаку$_j$, только когда она$_{i,j}$ уже выбежала на дорогу.[5]

(2) Моей сестре$_i$ очень нравилась новая учительница$_j$. Она$_{i,j}$ всегда приходила в класс за десять минут до звонка, так как ø$_{i,j}$ хотела успеть получше подготовиться к уроку.

(3) В четверть девятого запыхавшийся лаборант Петров$_i$ влетел в кабинет, но тут выяснилось, что заведующий лабораторией$_j$ задерживается еще сильнее. Конечно, должность обязывала его$_{i,j}$ приходить вовремя. К тому же на половину девятого была запланирована встреча с первым посетителем, который уже сидел в коридоре.

(4) Профессор$_i$ пообещал студенту$_j$, что он$_i$/он$_j$ сможет **принять/сдать** экзамен через неделю.

(2') Моей сестре$_i$ очень нравилась новая учительница$_j$. Она$_i$/она$_j$ всегда приходила в класс за десять минут до звонка, так как ø$_i$/ø$_j$ хотела успеть **сесть на первую парту/пообщаться с учениками.**

(1') Кошка$_i$ почуяла собаку$_j$, только когда она$_i$/она$_j$ уже выбежала на дорогу. Добежав до середины дороги, она$_i$/она$_j$ вдруг громко **замяукала/залаяла**.

(2'') Моей сестре$_i$ очень нравилась новая учительница$_j$. Она$_i$/она$_j$ всегда приходила в класс за десять минут до звонка, так как ø$_i$/ø$_j$ хотела успеть получше подготовиться к уроку.
Но сегодня **сестра/учительница** была вынуждена задержаться.

В качестве РД в подобных примерах особенно часто используется контекст плюс совокупность энциклопедических знаний о мире (примеры 4, 2', 1'), а также полная ИГ[6] (2'').

Предотвращенным РК мы будем называть такую ситуацию, при которой несмотря на наличие в РП адресата нескольких активированных референтов, РК не возникает, так как в подобных случаях РД находится слева от РедуцРС или непосредственно на нем. Мы выделяем четыре способа предотвращения РК. Во-первых, он может быть предотвращен при помощи различных лексико-грамматических средств, примеры 1'', 1''':

(1'') Кошка$_i$ заметила щенка$_j$, только когда **она$_i$/он$_j$** уже выбежала/выбежал на дорогу.

(1''') Кошка$_i$ заметила собаку$_j$, только когда **ø$_i$/та$_j$** уже выбежала на дорогу.

---

[5] Все упомянутые примеры являются сконструированными.

[6] Полная ИГ, несомненно, является самым кардинальным способом предотвращения и снятия РК.

Во-вторых, РК может предотвращаться семантикой связи между пропозициями, (2'''):

(2''') Моей сестре$_i$ очень нравилась новая учительница$_j$, **поэтому/потому что** она$_i$/она$_j$ всегда приходила в класс за десять минут до звонка.

В-третьих, он может быть предотвращен пропозициональным контекстом плюс совокупностью энциклопедических знаний о мире, (3'):

(3') В четверть девятого запыхавшийся лаборант Петров$_i$ влетел в кабинет, но тут выяснилось, что заведующий лабораторией$_j$ задерживается еще сильнее. Конечно, **руководящая/лаборантская** должность обязывала его$_j$/его$_i$ приходить вовремя. К тому же на половину девятого была запланирована встреча с первым посетителем, который уже сидел в коридоре.

Четвертым типом предотвращения РК является «занятость референта-конкурента»[7]:

(5) Уже больше года Катю сильно беспокоила ее племянница. Кроме **племянницы/Кати** у нее не было никаких других родственников.

4. Референциальный конфликт глазами Говорящего. Эксперимент[8]

В начале этого раздела мы кратко опишем типологию РК с точки зрения говорящего. Рассмотрим три ситуации. В первой ситуации в РП говорящего активирован только один референт (и РК практически невозможен); во втором случае в РП активированы два референта разного пола и в случае использования РедуцРС возникает потенциальный РК; наконец, в третьем случае в РП активированы два референта одного пола, так что вероятность РК максимально высока. В этой работе мы более подробно изучим ситуацию потенциального РК и ответим на вопрос, как часто в этом случае говорящий выбирает РедуцРС и на кого в первую очередь он ориентируется — на себя самого или на адресата своего сообщения.

---

[7] Термин из работы Кибрик 1987.

[8] Наше исследование основано на англоязычном исследовании Arnold and Griffin 2007, в котором было описано три эксперимента; наше исследование также состояло из двух экспериментов — пилотного и основного, однако ввиду ограниченного объема в настоящей статье мы подробно опишем только наше собственное основное исследование, а на все остальные будем ссылаться по мере необходимости.

Итак, в зависимости от КА говорящий выбирает[9] то или иное РС[10]; использование РедуцРС может вызвать РК. Цель описываемого эксперимента состояла в проверке гипотезы о том, что наличие одного или нескольких референтов, активированных в РП испытуемого, влияет на его РВ. Арнольд и Гриффин (2007) показали, что, как и подсказывает наша интуиция, в ситуации двух активированных референтов одного пола количество РедуцРС оказалось значимо меньше, чем в двух остальных случаях. В нашем эксперименте мы исключили этот случай из рассмотрения.

Эксперимент был проведен по методике рассказывания коротких историй по картинкам. В качестве стимулов были использованы картинки с легко узнаваемыми героями мультфильмов. Каждый из 15 сюжетов состоял из двух картинок, составляющих историю, которая начиналась на первой картинке и заканчивалась на второй. Сначала испытуемый в течение 2 секунд видел на экране две картинки, расположенные одна под другой (см. рис. 1), потом нижняя картинка исчезала и он слышал предложение, которое описывало то, что происходило на первой картинке (напр., *Винни-Пух с Совой подошли к большому дереву*). Сразу после этого испытуемый должен был повторить только что услышанное предложение, а затем, вновь посмотрев на вторую картинку, закончить историю (например, *потом Винни-Пух подумал, что он хочет съесть мед*). Как и в англоязычном эксперименте, мы просили испытуемых говорить просто и ясно, чтобы было понятно и пятилетнему ребенку. По количеству персонажей каждый сюжет имел три варианта — на обеих картинках изображен: (1) один персонаж; (2) два персонажа разного пола; (3) на первой картинке изображены два персонажа разного пола, на второй — только один из них[11]. Истории составлялись таким образом, чтобы независимо от количества персонажей и на первой, и на второй картинке главным действующим лицом был один и тот же персонаж. Говоря более научным языком, у нас было три описанных выше уровня независимой переменной, в качестве зависимой переменной мы рассматривали тип РС, выбранного испытуемым для упоминания главного персонажа во втором предложении; варианты распределялись по экспериментальным листам по правилу латинского квадрата[12].

Эксперимент был проведен с 24 испытуемыми. Как и в предшествующих работах, мы учитывали только те ответы испытуемых, в которых главный персонаж был упомянут первым и являлся подлежащим[13].

---

[9] Здесь и далее, описывая процесс РВ, мы имеем в виду, что говорящий осуществляет его по всей вероятности автоматизированно и неосознанно.

[10] В некоторых случаях выбор между двумя соседними типами РС не строго детерминирован.

[11] Третий вариант был использован для того, чтобы развести две гипотезы относительно наблюдаемого эффекта во втором варианте, подробнее см. далее.

[12] Более подробно о процедуре моделирования эксперимента см. Федорова 2008.

[13] такие ответы мы вслед за Арнольд и Гриффин называем каноническими.

Табл. 1. Результаты эксперимента

кол-во персонажей на первой/второй картинке	канонические предложения, в %	тип РС, в %		
		нуль	местоимение	полная ИГ
1/1	90	34	37	29
2/2	72	1	2	97
2/1	78	0	1	99

Рис. 2. Скриншот одного из сюжетов

Результаты эксперимента (см. таблицу 1) оказались очень красноречивыми — если для первого случая мы наблюдаем примерно равное количество ответов каждого типа, то в двух других испытуемые практически всегда выбирали полную ИГ.

Результаты были проверены на статистическую достоверность с помощью R-language, языка программирования для статистической обработки данных. P-value — уровень достоверности нулевой гипотезы, а именно та вероятность, с которой при условии истинности нулевой гипотезы могла бы реализоваться наблюдаемая выборка. Нулевая гипотеза принимается при p-value > 0.05. Для

сравнения распределения ответов по разным типам использовались prop.test и binom.test. Действительно, в первом случае ответы распределены одинаково между нулями, местоимениями и ИГ, p-value = 0.4169, а в двух других — нет (p-value < 0.01).

Отсутствие различий между вторым и третьим типами (пропорции распределения типов одинаковы, p-value от 0.2364 до 0.7281) говорит о том, что при описании картинки испытуемый опирается прежде всего не на текущий визуальный ряд, где может быть как два персонажа, так и один, а на свою дискурсивную РП, в которой содержится информация о двух активированных персонажах.

5. Обсуждение результатов

Итак, в ситуации потенциального РК, который в случае выбора РедуцРС мог быть предотвращен использованием родового показателя, испытуемые тем не менее употребляли полную ИГ. Как можно объяснить подобный эффект?

Арнольд и Гриффин (2007) объясняют свои аналогичные результаты тем, что, во-первых, испытуемые ориентируются прежде всего не на адресата, а на себя, то есть используют эгоцентрическую стратегию[14]. Во-вторых, авторы считают информацию о потенциальном РК одним из факторов активации референта в РП, то есть, по их мнению, факт наличия в РП говорящего более одного референта снижает активацию этого референта до уровня, подходящего только для использования полной ИГ.

В работе Kibrik 2011 приводится обратная аргументация. Во-первых, из трех возможных стратегий — эгоцентрической, оптимальной и опекающей — А. А. Кибрик выбирает последнюю, подтверждая свои слова текстом инструкции «говорить просто и ясно, чтобы было понятно и пятилетнему ребенку». Во-вторых, одним из принципиальных положений данной модели является автономность фильтра РК, действие которого происходит уже после выбора РС; другими словами, потенциальный РК не влияет на КА.

Результаты эксперимента не могут дать нам однозначный ответ на вопрос, чья аргументация является более правильной. Для того чтобы иметь возможность развести эти гипотезы, необходимо будет провести новые эксперименты. Самый простой и очевидный путь — это повторить тот же эксперимент, убрав из инструкции слова про пятилетнего ребенка. Более серьезный, но также возможный — провести эксперимент, направленный непосредственно на определение места фильтра РК в общей модели РВ.

---

[14] Вопрос о том, какая стратегия — эгоцентрическая или ориентированная на адресата — используется говорящим в процессе РВ, вызывает в современной психолингвистике горячие споры. Некоторые авторы полагают, что любой говорящий сначала ориентируется на самого себя, но потом часто модифицирует свою стратегию в соответствии с нуждами адресата (Keysar 2007 или Fukumura and Gompel 2009); другие считают, что в процессе РВ говорящий с самого начала учитывает фактор адресата (Brown-Schmidt 2009).

Однако когда в естественно-языковой ситуации мы слышим историю, начинающуюся со слов *Мальчик поцеловал девочку*, то продолжение ее в форме *...и мальчик убежал* кажется нам гораздо менее естественным, чем *...и убежал*. Почему же в аналогичной экспериментальной ситуации испытуемые всегда выбирают полные ИГ? Возможно, сам жанр пересказывания знакомых с детства мультфильмов вызывает желание повторять имена собственные. Поэтому в следующем эксперименте имеет смысл в любом случае использовать другой стимульный материал: не имена собственные типа *Вася* и *Катя*, поскольку это может дать обратный эффект — вынудить испытуемых использовать РедуцРС[15], а нейтральные комитативные ИГ *бабушка с внуком* или *девочка с папой*.

6. Блок-схема референциального выбора

В настоящей работе мы исследовали небольшой фрагмент общей теории РВ, а именно, показали, что в ситуации потенциального РК с двумя активированными в РП референтами разного пола испытуемые предпочитают использовать полные ИГ. Для того, чтобы иметь в дальнейшем возможность тестировать тот или иной фрагмент модели РВ, а также вносить уточнения, возникающие в результате такого тестирования, нам необходимо сформировать максимально подробное представление о всех этапах процесса РВ. Ниже мы предлагаем первую попытку описания блок-схемы РВ, построенной на основе модели РВ из Kibrik 2011, а также уточнений, вытекающих из результатов проведенного эксперимента. Несомненно проигрывая в наглядности и простоте восприятия, данная схема является, на наш взгляд, необходимым подспорьем для проведения экспериментального тестирования.

Полная блок-схема приведена в приложении, в этом разделе мы кратко опишем ее основные части. Модель РВ говорящего состоит из двух блоков — блока Говорящего и блока Адресата. В первом блоке говорящий, не думая об адресате, сначала определяет текущие значения факторов активации, потом на основе факторов активации определяет КА, а затем на основе КА выбирает тип будущего РС. Если вследствие низкого КА говорящий выбрал полную ИГ, то блок Адресата не используется. Если же говорящий выбрал РедуцРС, но второй блок оказался незадействованным (развилка ‘игнорирование блока Адресата’ на схеме), мы говорим об эгоцентрической стратегии — так почти всегда ведут себя маленькие дети, а иногда и взрослые.

В начале работы второго блока говорящий впервые использует информацию об адресате — если он не задумывается специально о какой-то особой стратегии (развилка ‘выбор особой стратегии’), то стратегия получается нейтральной (оптимальной по А. А. Кибрику), если же он хочет предпринять какие-то дополнительные кооперативные меры, то он выбирает опекающую[16]

---

[15] Так как к середине эксперимента в их голове уже перемешаются все имена.

[16] Возможно, в ходе дальнейшей работы репертуар стратегий пополнится и другими возможностями.

стратегию и меняет тип выбранного РедуцРС на полную ИГ. Именно такую стратегию, как мы предполагаем, использовали испытуемые нашего эксперимента. Однако на этом этапе работает фильтр референциальной избыточности (РИ) — если в РП говорящего активирован только один референт, то выбор полной ИГ блокируется.

Наконец, когда на вход фильтра РК поступает РедуцРС, говорящий оценивает возможность РК — проверяет количество активированных в РП референтов, а при их количестве большем одного еще и возможность компенсации потенциального РК. Если РК не может быть компенсирован, то говорящий меняет тип на полную ИГ.

Подведем итоги этого раздела. Во-первых, мы эксплицитно разделили процесс РВ на два блока — блок Говорящего и блок Адресата. В-вторых, мы нашли подходящее, на наш взгляд, место для компонента, отвечающего за выбор стратегии, сохранив имеющийся в Kibrik 2011 репертуар таких стратегий. Наконец, мы включили в модель фильтр РИ, тем самым еще усилив тезис об автономности блока Адресата от блока Говорящего.

В заключении еще раз подчеркнем, что все подобные идеи остаются спекулятивными до тех пор, пока не подтверждаются в ходе дальнейшей работы. Однако именно подобная «цикличность» экспериментального процесса является одновременно наиболее сложной и наиболее интересной его особенностью.

Список сокращений

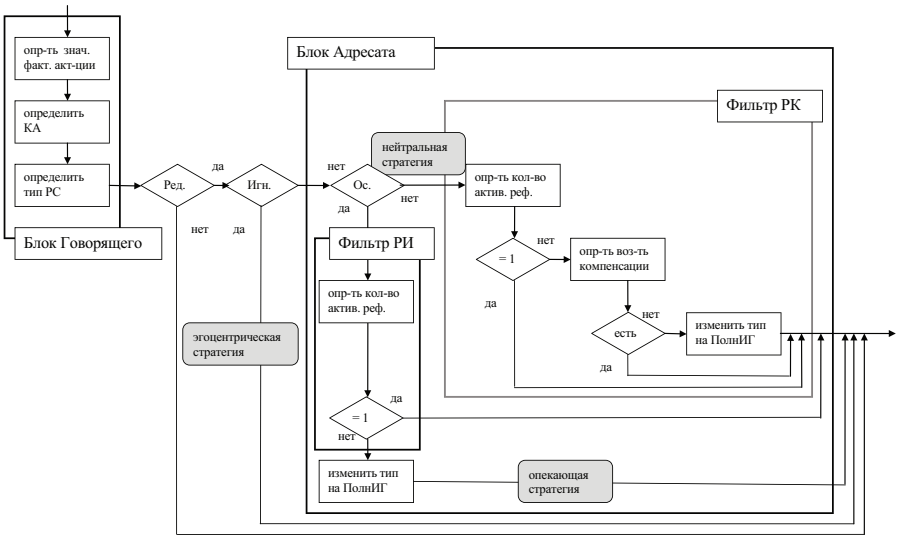
ИГ — именная группа
КА — коэффициент активации
РВ — референциальный выбор
РедуцРС — редуцированное референциальное средство
РД — референциальный деконфликтор
РИ — референциальная избыточность
РК — референциальный конфликт
РП — рабочая память
РС — референциальное средство
ЭАД — экспериментальный анализ дискурса

References

1. *Arnold Jennifer E., Zenzi M. Griffin.* 2007. The Effect of Additional Characters on Choice of Referring Expression: Everyone counts. Journal of Memory and Language, 56(4) : 521–36.
2. *Brown-Schmidt Sarah.* 2009. Partner-specific Interpretation of Maintained Referential Precedents During Interactive Dialog. Journal of Memory and Language, 61(2) : 171–190.

3. *Clark Herbert H.* 1992. Arenas of Language Use.
4. *Clark Herbert H., Wilkes-Gibbs Deanna.* 1986. Referring as a Collaborative Process. Cognition, 22(1) : 1–39.
5. *Fedorova O. V.* 2008. The Fundamentals of Experimental Psycholinguistics: The Principles of Organization of an Experiment [Osnovy Eksperimental'noi Lingvistiki: Printsipy Organizatsii Eksperimenta].
6. *Fedorova O. V., Delikishkina E. A., Maliutina S. A., Uspenskaia A. M., Fein A. A.* 2010. Experimental Approach to Discourse Reference Research: Interpretation of Anaphoric Pronoun depending on the Rethorical Distance of its Antecedent [Eksperimental'nyi Podhod k Issledovaniiu Referentsii v Diskurse: Interpretatsiia Anaforicheskogo Mestoimeniia v Zavisimosti ot Ritoricheskogo Rasstoianiia do ego Antetsedenta]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2010" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2010"): 525–30.
7. *Fukumura Kumiko, Roger P. G. van Gompel.* 2009. Speakers Use their Own, Privileged Discourse Model to Determine Referents' Accessibility during the Production of Referring Expressions. Bridging the Gap between Computational and Empirical Approaches to Reference.
8. *Keysar Boaz.* 2007. Communication and Miscommunication: The Role of Egocentric Processes. Intercultural Pragmatics, 4 : 71–84.
9. *Kibrik A. A.* 1987. Mechanisms of Referential Conflict Elimination [Mekhanizmy Ustraneniia Referentsial'nogo Konflikta]. Modelirovanie Iazykovoi Deiatel'nosti v Intellektual'nykh Sistemakh : 128–45.
10. *Kibrik A. A.* 2011. Reference in Discourse.
11. *Pickering Martin J., Garrod Simon.* 2007. Do People Use Language Production to Make Predictions during Comprehension? Trends in Cognitive Science, 11(3) : 105–10.
12. *van Berkum, Jos J. A.* (In press). The Electrophysiology of Discourse and Conversation. The Cambridge Handbook of Psycholinguistics.

Блок-схема референциального выбора



Ред. — редуцированное РС

Игн. — игнорирование блока Адресата

Ос. — выбор особой стратегии

ЛЕКСИКО-ФУНКЦИОНАЛЬНАЯ РАЗМЕТКА РУССКИХ ТЕКСТОВ

Т. И. Фролова (tfrolova@iitp.ru)

О. Ю. Подлесская (olga@iitp.ru)

Институт проблем передачи информации им. А.А. Харкевича, РАН, Москва, Россия

Ключевые слова: разметка текстов, лексико-функциональная разметка текстов, СинТагРус, лексические функции.

TAGGING LEXICAL FUNCTIONS IN RUSSIAN TEXTS OF SYNTAGRUS[1]

T. I. Frolova (tfrolova@iitp.ru)

O. Iu. Podlesskaia (olga@iitp.ru)

Laboratory of Computational Linguistics, Kharkevich Institute For Information Transmission Problems, RAS, Moscow, Russian Federation

The present paper describes the process and the results of tagging with Lexical Functions the texts of SynTagRus (Syntactic Russian corpus available at www.ruscorpora.ru). The present work started in 2009 and it is still in progress in the Laboratory of Computational Linguistics Kharkevich Institute For Information Transmission Problems, RAS. The lexical items which are identified as values and arguments of collocate Lexical Functions (LFs) are tagged in syntactically annotated Russian sentences. At the moment about 4300 sentences (about 5500 LF-phrases) have been processed and all of them were supplied with LF-annotation. At the end of the paper some examples of possible linguistic and educational uses for the corpus with LF tagging are offered.

Key words: tagging, lexical functions, SynTagRus, tags.

[1] This work has been supported in part with the program «Текст во взаимодействии с социокультурной средой: уровни историко-литературной и лингвистической интерпретации» of the Division of History and Philology of RAS.

1. Introduction

The paper describes tagging Russian texts with collocate Lexical Functions, especially the process of work, some results and possible uses in linguistics and education. The work is being done for the last two and a half years in the Laboratory of Computational Linguistics, Kharkevich Institute For Information Transmission Problems, RAS.

1.1. Concept of collocate lexical functions

Collocate Lexical Functions (LFs) are certain meanings which can be expressed by different lexemes of the given language, the choice among them being determined not only by the meaning itself, but also by the keyword (argument) with regard to which this general meaning is expressed. For example the value of LF MAGN ('a high degree of what is denoted by X') is HEAVY for noun FOG (*heavy fog*) and GRAVE for noun DISEASE (*grave disease*).

Besides this idiomaticity within the given language (different values for different arguments) the values of LFs are often idiomatic between two different languages, thus the English adjective HEAVY in *heavy fog* is the value of LF MAGN for the noun FOG, but it is not a translation for Russian ГУСТОЙ (literally 'dense, thick') which is the value of LF MAGN for the Russian noun ТУМАН in *густой туман* ('heavy fog').

The apparatus of LFs was proposed in [1,2] alongside with the list of presumably universal LFs meanings which are expressed in non-trivial way in combination with the keywords, such as 'high degree' (MAGN), 'good' (BON), 'right' (VER), 'opposite' (ANTI), 'existence' (FUNC), 'beginning' (INCEP), 'end' (FIN), 'causation' (CAUS), 'liquidation' (LIQU), 'normal use' (REAL), 'normal functioning' (FACT), 'manifestation' (MANIF) etc.

1.2. Processing of LFs in ETAP

In multipurpose linguistic processor ETAP-3 developed in the Laboratory of Computational Linguistics of Kharkevich Institute For Information Transmission Problems, RAS, LF-information is recorded in special zone of combinatorial dictionary. This information is used for lexical and syntactic ambiguity resolution in analysis and for improvement of quality of translation, as well as for automatic paraphrasing. There is a special block of rules for identification LFs, which process syntactic tree structure of each sentence. For more information on ETAP-3 system and the use of LFs in this system see [3–5] and [6–7] respectively.

2. Appearance and techniques of LF-tagging

Analysis of LF-collocations is carried out for phrases already tagged morphologically and syntactically. Syntactic tagging of Russian texts is being done in the Laboratory of Computational Libguistics already for several years. These syntactically tagged texts

form an integral part of Russian National Corpus. The lexico-grammatical search in these texts is available at http://www.ruscorpora.ru/search-syntax.html. Syntactically annotated corpus contains about 45 000 sentences from fiction and journalistic texts.

Each sentence of the corpus is represented with its syntactic tree structure, where each word forming its node is provided with full set of morphological features, and the links are marked with the names of syntactic relations. Such a view of syntactic structure of the sentence goes back to "Meaning ⇔ Text" theory by Melchuk and Zholkovsky [see 1 and 2]. The procedure of syntactic tagging is semi-automatic: at first morphological analyzer and syntactic parser of ETAP-3 make up the syntactic tree structure of the sentence [more on tagging scheme and tools for corpus creation see 8 and 9]. Then a trained linguist checks and if necessary manually corrects the result of machine analysis. See below the morphologically and syntactically annotated sentence — the result of semi-automatic procedure described above:

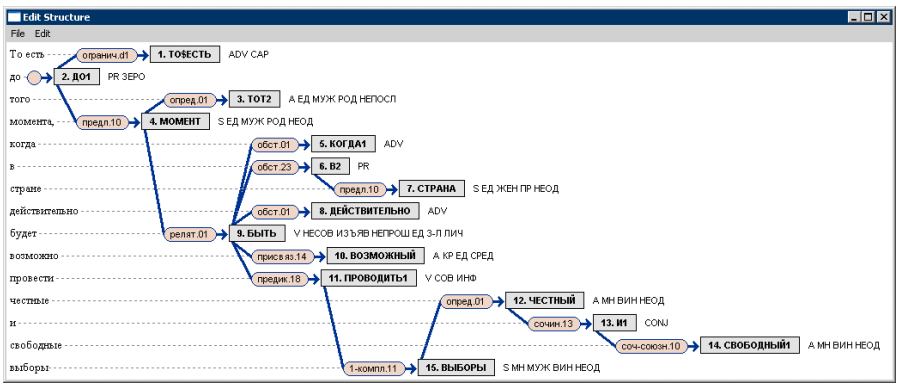


Fig. 1. Syntactic structure of Russian sentence: «То есть до того момента, когда в стране действительно будет возможно провести честные и свободные выборы» ('That is till the moment when it'll be really possible to hold honest and free elections in the country')

The LF-analysis of sentences is also done semi-automatically: syntactically and morphologically tagged sentence (already checked by a linguist) is put into LF-analyzer. This analyzer uses the information assigned to lexemes in LF-zones of Russian combinatorial dictionary in ETAP, and also the rules of recognition of LF-phrases (i.e. phrase which consists of an argument of certain LF and its value) in the text.

The rules of LF-recognition were created for machine translation and automatic paraphrasing. These rules define syntactic conditions for establishing LF-link between the argument and the value of certain LF.

For example, despite the presence of the record for LF MAGN:КРЕПКИЙ in LF-zone of article ЗДОРОВЬЕ in Russian combinatorial dictionary, the LF-link is not set between adjective КРЕПКИЙ ('firm') and noun ЗДОРОВЬЕ ('health') in coordinate phrase *крепкий сон и отличное здоровье* ('sound sleep and perfect health'). Adjective КРЕПКИЙ is used here as value of LF MAGN but of another noun, namely

СОН (‘sleep’). Correct automatic recognition is possible here due to the absence of syntactic links, specific for adjectival LFs, between adjective КРЕПКИЙ and noun ЗДОРОВЬЕ (see examples of such syntactic contexts below in part 3).

For the purpose of LF-tagging, these rules underwent some technical changes. In addition, the format of presenting LF-information in tagged texts was elaborated. This format in the form present on Fig. 2 was elaborated by members of the Laboratory V. G. Sizov and V. V. Petrochenkov.

The result of automatic LF-analysis of each sentence was checked and if necessary corrected and supplemented by linguists. See below on Fig. 2 one example of LF-structure marked for one Russian sentence. One can see LF-collocations: LOC — *в стране* (‘in the country’), OPER1 — *провести выборы* (‘hold elections’) и VER — *честные выборы* (‘honest elections’):

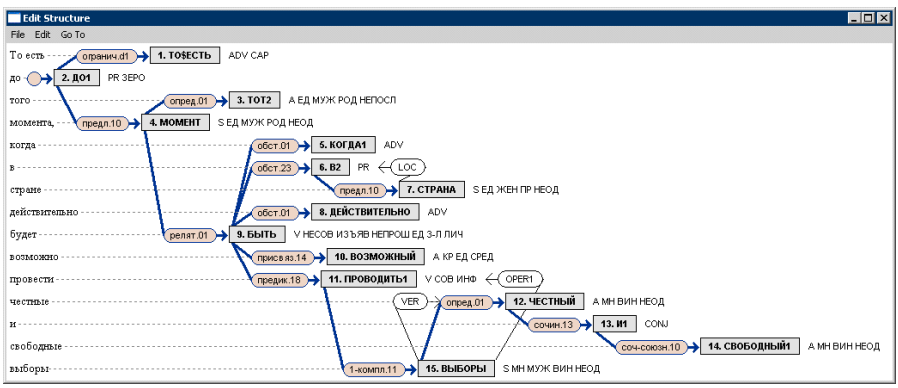


Fig. 2. Syntactic and LF-structure of Russian sentence: «То есть до того момента, когда в стране действительно будет возможно провести честные и свободные выборы» (‘That is till the moment when it’ll be really possible to hold honest and free elections in the country’)

Results of LF-tagging are now available in the Laboratory, but they are also expected to be available on-line.

3. Principles of LF-tagging

All phrases which can be described in terms of standard LFs are marked in the process of tagging. For the list of standard LFs see for example work [10].

As already mentioned above, the information about LF-collocations is recorded in LF-zone of the combinatorial dictionary in ETAP-3 system. Syntactic conditions for identifying LF-phrases are described in ETAP-3 in the rules of LF-recognition. These rules (and these conditions) are composed in terms of syntactic links. For example, for LF MAGN (and other adjectival LFs) for nominal arguments possible contexts are the following:

- attributive phrase: *крепкое здоровье* (‘sound health’);
- nominal clause: *здоровье крепко* (‘the health is sound’);
- phrases with copulative verbs: *здоровье было (казалось, могло быть) крепким* (‘the health was (seemed, could be) sound’);
- and so forth.

All phrases found in these contexts are marked in the tagged sentences.

The cases are not marked and not recognized automatically when it is necessary for establishing correct LF-link to set correct anaphoric relations between sentences or within one sentence. For example, in sentence ***здоровье*** *не доставит проблем, потому что оно будет* ***крепким*** (‘The health won’t cause any problems because it will be sound’) LF-link between noun ЗДОРОВЬЕ (‘health’) and adjective КРЕПКИЙ (‘sound’) won’t be marked.

Cases are not marked when incorrect or nonstandard values of LFs are used in text. For example, the phrase *оплачивают налоги* (‘(they) pay for taxes’) is incorrect. The verb ОПЛАЧИВАТЬ (‘pay for’) is used in the sense of LF REAL1-M (‘To do with regard to X that which is normally expected of P1’) instead of the correct value of LF REAL1-M for НАЛОГ (‘tax’) — ПЛАТИТЬ (‘pay’). The correct phrase is *платят налоги* (‘(they) pay taxes’). In this case the phrase *оплачивают налоги* is not tagged as LF-phrase. Also the phrase *производить влияние* (‘produce influence’) is not tagged as LF-phrase, because in the sense of LF OPER1 (‘To do X, to have X or to be in the state X’) for the noun ВЛИЯНИЕ (‘influence’) the verb ПРОИЗВОДИТЬ (‘produce’) is used incorrectly instead of the correct value ОКАЗЫВАТЬ (‘exert’). The phrase *плотно подсели*, nonstandard phrase used in the text in the sense of ‘become strongly dependent on’ is also not marked as LF-phrase, though adverb ПЛОТНО here could be interpreted here as value of LF MAGN (‘high degree’). Decision not to include this phrase into the results of tagging was made due to the fact that the verb ПОДСАЖИВАТЬСЯ in the sense ‘become dependent’ is not the lexeme of the literary language.

Another cases excluded from LF-tagging are LF-phrases with arguments that are phrases and not single lexemes. For example the phrases *придерживаться точки зрения* or *иметь точку зрения* (‘hold to or have the point of view’) are not tagged as LF-phrases, though the verbs ПРИДЕРЖИВАТЬСЯ (‘hold to’) and ИМЕТЬ (‘have’) are values of LF OPER1 (‘To do X, to have X or to be in the state X’) for an argument ТОЧКА ЗРЕНИЯ (‘point of view’).

The question of including values of LFs determined by pragmatic component of arguments’ meaning in LF-tagging is under discussion. For example, in the phrase *затяжная война* (‘protracted war’) the adjective ЗАТЯЖНОЙ (‘protracted’) could be interpreted as the value of LF MAGN (‘high degree’) for the component DURATION which is not a part of the meaning of the noun ВОЙНА (‘war’).

4. Results of LF-tagging

The main result of the work is the corpus of 4300. The corpus which was examined contains journalistic texts having from twenty to more than two hundred

sentences each. Sentences, containing LF-collocations, make up approximately one third of all analyzed sentences (about 4300 sentences). These sentences contain about 5500 LF-collocations. Some additional information about quantities of LFs in corpus is given below. It should be noticed, however, that due to relatively small amount of current LF-corpus, it is now unreasonable to jump to almost any substantial linguistic conclusions on basis of this information: for example, the fact that the phrase *в понедельник* has 24 occurrences in corpus, whereas *в пятницу* has 11 occurrences, doesn't mean that the first phrase is twice more common in Russian. The data below is given only to provide more detailed view of the corpus.

The most frequent LF in the corpus is LOC ('A preposition denoting the normal spatial or temporal localization of something with regard to X', more than 1800 occurrences). Here are the most frequent collocations with this LF in corpus (see in brackets the number of occurrences for each of these collocations in corpus): *в году* (250), *в России* (172), *в стране* (121), *во время* (109), *в Москве* (48), *в мире* (47), *в школе* (41), *на рынке* (31), *в городе* (28), *в сфере* (28), *на территории* (27).

The most frequent verbal LF is LF OPER1 ('To do X, to have X or to be in the state X', about 950 occurrences). The most frequent arguments of this LF are the following (see in brackets values of this LF for each argument and the number of occurrences): РЕШЕНИЕ (ПРИНИМАТЬ/ВЫНОСИТЬ, 34), ВЫВОД (ДЕЛАТЬ/ПРИХОДИТЬ К, 29), РОЛЬ (ИГРАТЬ, 27), ИССЛЕДОВАНИЕ (ПРОВОДИТЬ/ВЕСТИ, 27), ВНИМАНИЕ (ОБРАЩАТЬ, 25).

There are some other verbal LFs present in corpus relatively frequently: FUNC0 ('X exists or is taking place', 191 occurence), CAUSFUNC0 ('To cause X to happen or to exist', 188 occurrences), INCEPOPER1 ('To start to do X, to have X or to be in the state X', 154 occurrences). See below the most frequent arguments of these LSs in corpus:

FUNC0: РЕЧЬ (ИДТИ, 36), ПРОЦЕСС (ИДТИ/ПРОХОДИТЬ, 15), ВОЗМОЖНОСТЬ (БЫТЬ/СУЩЕСТВОВАТЬ, 13), ПРОБЛЕМА (ИМЕТЬСЯ/СУЩЕСТВОВАТЬ, 8).

CAUSFUNC0: ЗАДАЧА (СТАВИТЬ, 13), РЕЗУЛЬТАТ (ДОСТИГАТЬ/ПОЛУЧАТЬ, 13), ИТОГ (ПОДВОДИТЬ, 7), ПАМЯТНИК (УСТАНАВЛИВАТЬ/ВОЗДВИГАТЬ, 7), ФИЛЬМ (СНИМАТЬ, 6).

INCEPOPER1: ЗНАНИЕ (ПРИОБРЕТАТЬ/ПОЛУЧАТЬ, 8), ИНФОРМАЦИЯ (СОБИРАТЬ/ПОЛУЧАТЬ, 7), ВЛАСТЬ (ПОЛУЧАТЬ/ПРИХОДИТЬ К, 6), ДЕНЬГИ (ЗАРАБАТЫВАТЬ, 6), ДИПЛОМ (ПОЛУЧАТЬ, 6), РАБОТА (НАЧИНАТЬ, 6).

The most frequent adjectival LF is MAGN ('a large degree or a high intensity of X', 522 occurrences). The most frequent arguments with this LF are: УРОВЕНЬ (ВЫСОКИЙ, 14), РОСТ (БЫСТРЫЙ/СТРЕМИТЕЛЬНЫЙ, 12), РАСТИ (БЫСТРО/СТРЕМИТЕЛЬНО, 11), БИЗНЕС (БОЛЬШОЙ/КРУПНЫЙ, 10), ЗАРПЛАТА (ВЫСОКИЙ/БОЛЬШОЙ, 9), ЗНАТЬ (ХОРОШО/ТВЕРДО, 8), ИЗВЕСТНЫЙ (ХОРОШО, 6), КОЛИЧЕСТВО (БОЛЬШОЙ, 6), ПОНИМАТЬ (ХОРОШО/ЯСНО, 6).

For the purpose of additional illustration some sentences are given below from one of the texts in corpus, namely «Расслабьтесь и наслаждайтесь (Relax and Enjoy)», by Yevgeny Grigoryevich Yasin, from the newspaper "Trud" of the 22nd of September 2008. This text contains 48 sentences, of which only 19 with LF-phrases. For each of the sentences the picture of syntactic and LF-structure (or its fragment) is given.

Sentence 1 (see Fig. 3)

The sentence contains three LF-collocations:
НЕДЕЛЯ, LOC: НА,
КРИЗИС, INCEPFUNC0 ('X starts to exist or to be taking place'): РАЗРАЖАТЬСЯ,
ПРОГНОЗ, OPER1: ДАВАТЬ.

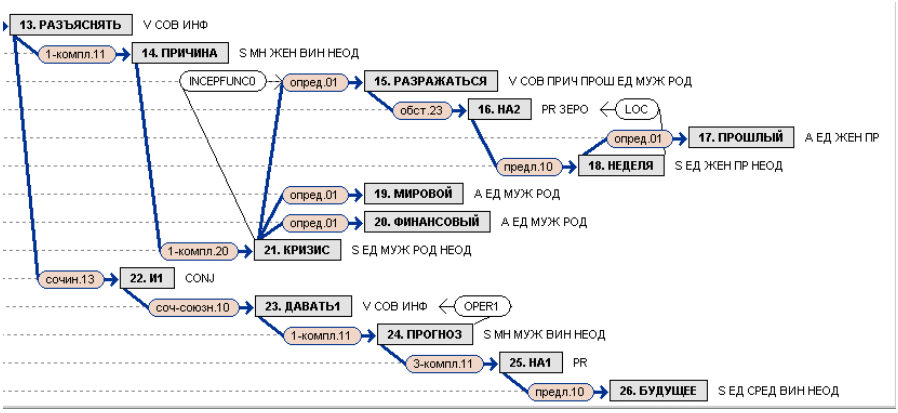


Fig. 3. Syntactic and LF-structure of the sentence: «С сегодняшнего дня "Труд" начинает публикацию статей известных экспертов, которых мы попросили разъяснить причины **разразившегося на** прошлой **неделе** мирового финансового **кризиса** и **дать прогнозы** на будущее» (fragment)

Sentence 4 (see Fig.4)

The sentence contains two LF-collocations:
РЫНОК, LOC: НА,
СДВИГ, FUNC0: ПРОИСХОДИТЬ.

Sentence 5 (see Fig. 5 and 6)

The sentence contains three LF-collocations:
ПОДЪЕМ, FUNC0: НАЧИНАТЬСЯ,
РЫНОК, LOC: НА,
НАПРЯЖЕНИЕ, CAUSFUNC0: ВЫЗЫВАТЬ.

Sentence 43 (see Fig. 7)

The sentence contains one LF-collocation:
ДОХОДНОСТЬ, ANTIMAGN ('a small degree of X'): НИЗКИЙ.

Sentence 44 (see Fig. 8)

The sentence contains one LF-collocation:
ДОЛГ, INCEPOPER1: ВЛЕЗАТЬ В.

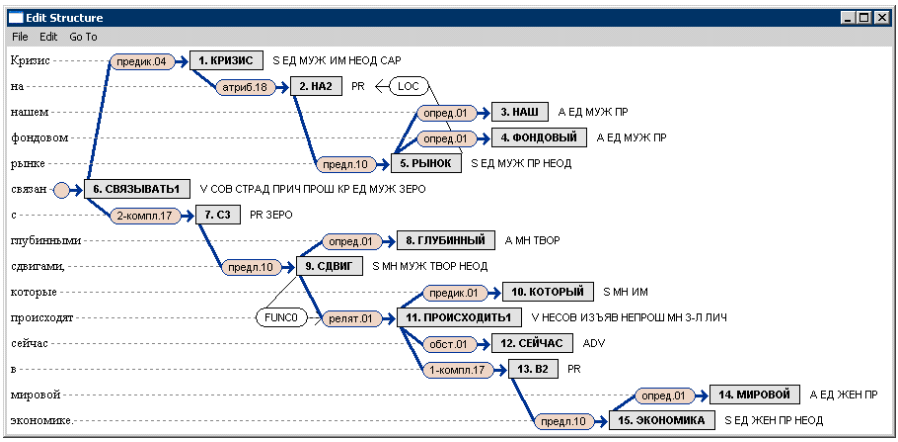


Fig. 4. Syntactic and LF-structure of the sentence: «Кризис **на** нашем фондовом **рынке** связан с глубинными **сдвигами**, которые **происходят** сейчас в мировой экономике»

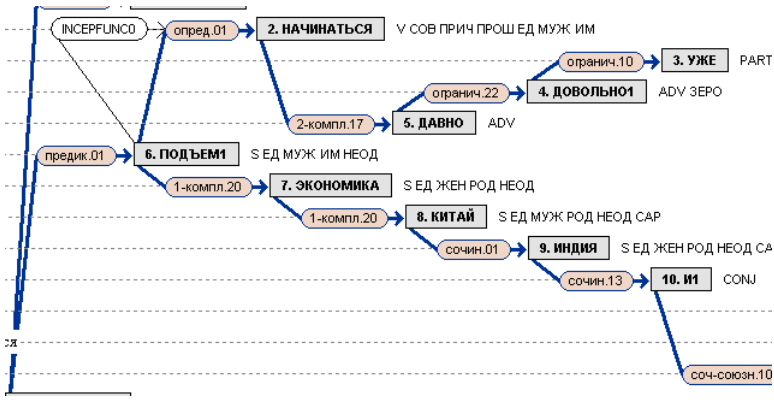


Fig. 5. Syntactic and LF-structure of the sentence: «Во-первых, **начавшийся** уже довольно давно **подъем** экономики Китая, Индии и других развивающихся стран усилил конкуренцию **на рынках** развитых стран и **вызвал** там определенное **напряжение** и с работой, и с доходами» (fragment 1)

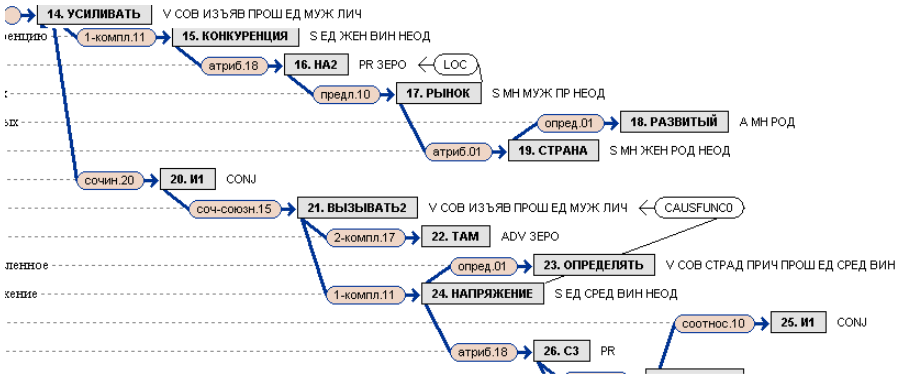


Fig. 6. Syntactic and LF-structure of the sentence: «Во-первых, **начавшийся** уже довольно давно **подъем** экономики Китая, Индии и других развивающихся стран усилил конкуренцию **на рынках** развитых стран и **вызвал** там определенное **напряжение** и с работой, и с доходами» (fragment 2)

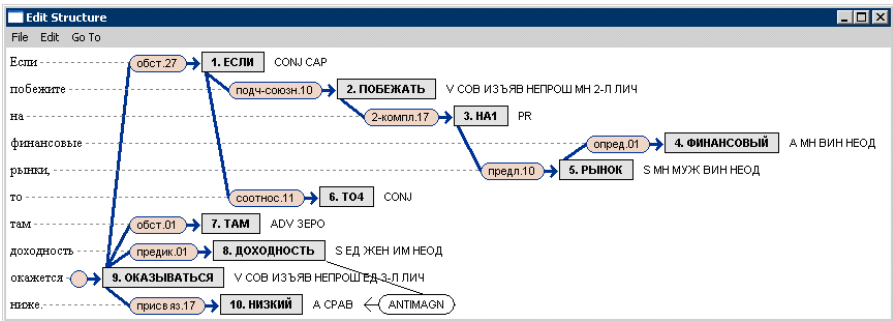


Fig. 7. Syntactic and LF-structure of the sentence: «Если побежите на финансовые рынки, то там **доходность** окажется **ниже**»

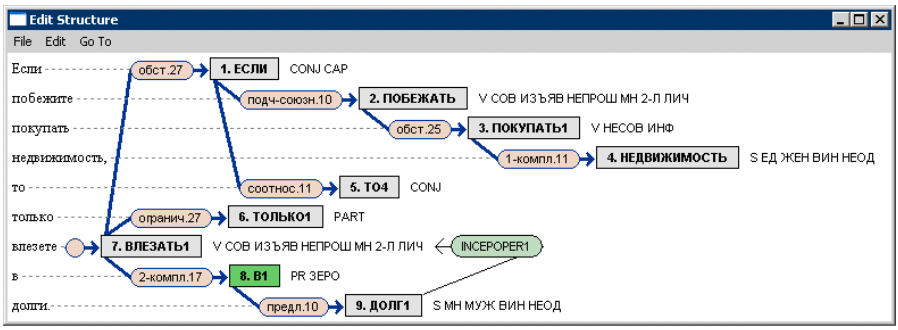


Fig. 8. Syntactic and LF-structure of the sentence: «Если побежите покупать недвижимость, то только **влезете в долги**»

4.1. Additional results. Improvements in ETAP-3 system

In the course of work some inaccuracies in description of not widespread syntactic constructions in rules of LF-recognition in ETAP-3 were revealed and corrected. Due to mass work with the texts, the LF-zones of both Russian and English combinatorial dictionaries of ETAP-3 were widened considerably. For more detailed description of changes in ETAP-3 see [11].

5. Perspectives and possible uses

The LF-tagging is still in progress. At the same time a new program for LF-search in texts is being elaborated. There is a program in Laboratory enabling LF-search in texts, but operating this program requires the knowledge of ways to record linguistic data in ETAP.

The new program is expected to let every user concerned to set and to solve different linguistic problems related to LFs. It is worth mentioning that search in LF-corpus will differ from search of collocations in corpora without syntactic and LF-tagging (see, for example, two dictionaries [12] and [13] published in the Internet in 2008). Search of Lf-collocations gives an opportunity to search not only adjacent words in certain types of phrases (such as NV, i.e. noun + verb, or VN, i.e. verb + noun, or VAN, i.e. verb + adjective + noun), but also phrases with free word order and long distances between words.

Collected corpus can be also used in education. See below several examples of possible uses of corpus in each of the two ways.

5.1. Questions about use of LF-phrases in texts

1) It is obvious that the same LF often has several values for the same argument. For example, the noun МНЕНИЕ (‘opinion’) has the values of LF OPER1: ИМЕТЬ (‘have’) and ПРИДЕРЖИВАТЬСЯ (‘hold to’). It is now possible to investigate the influence of lexical and syntactic environment on the choice of one of the two aforementioned variants. Here are some other examples for phrase with different values for the same LF and the same argument: *проводить исследование* vs. *вести исследование* (‘conduct research’); *достигать результата* vs. *получать результат* (‘obtain result’); *проводить эксперимент* vs. *ставить эксперимент* vs. *производить эксперимент* vs. *вести эксперимент* vs. *проделывать эксперимент* (‘to carry out *or* conduct *or* perform *or* run experiment’).

1.1) The same task can be set also for a class of arguments. For example, the same question seems to be sensible for values of LF FUNC2: ДОСТИГАТЬ (‘reach’), РАВНЯТЬСЯ (‘equal’), СОСТАВЛЯТЬ (‘make up’) for nouns denoting parameters: СКОРОСТЬ, ВЫСОТА, РАЗМЕР and so on.

2) It is well known that LFs can be used in paraphrasing. Special block of paraphrasing rules works in ETAP-3 system [see 6–7]. If the corpus with tagged LF-phrases is available the use of paraphrases can be investigated. For example, what are the differences that determine the choice between phrases *начал испытывать воздействие* ('started to be influenced') vs. *попал под воздействие* ('became influenced'), that is *начал* ('began') + *OPER1 X* vs. *INCEPOPER1 X*. Another example of such paraphrases is *испытывать давление* ('experience pressure') vs. *быть под давлением* ('be under pressure'), i. e. *OPER2 X* vs. *copula + ADV2 X*. One more example: не+LF ('not') vs. ANTILF (where LF is the value of any adjectival or some verbal LFs), cf. *медленное движение* ('slow motion') vs. *небыстрое движение* ('not quick motion'), that is *ANTIMAGN X* vs. *неMAGN X* ('not MAGN X').

3) There appears an opportunity to observe stylistic (or some other) peculiarities of the texts with more or less LF-phrases.

5.2. Uses of texts with LF-tagging for educational purpose

1) For students of Russian language the corpus may be used as a source of exercises for studying LF-collocations. For example, it is quite easy to pick out of corpus some sentences for the tasks like the following: — "fill the gap, marked with asterisks in the following sentence

*Евросоюз на минувшей неделе прописал евро радикальное лечение: надо *** ошибку, допущенную в 1999 году при введении единой валюты*.

with one of the verbs: УЛУЧШИТЬ, ПОЧИНИТЬ, ИСПРАВИТЬ".

(The correct answer is *исправить*).

The translation of the sentence is: 'Last weak the European Union prescribed euro curative treatment: it is necessary to correct the mistake made in 1999 in the time of creating the single currency'.

To make such exercises it is enough to pick out of the corpus sentences with LFs and put asterisks instead of values of LF.

2) For students of linguistics, studying the theory of LFs the same sentences may be a source of another two types of exercises:

a) "identify the name of LF in boldface phrase

Евросоюз на минувшей неделе прописал евро радикальное лечение: надо ***исправить ошибку****, допущенную в 1999 году при введении единой валюты*".

(The correct answer is REAL1-M).

b) "fill the gap, marked with asterisks in the following sentence with value of LF REAL1-M for the boldface argument

*Евросоюз на минувшей неделе прописал евро радикальное лечение: надо **** ***ошибку****, допущенную в 1999 году при введении единой валюты*".

(The correct answer is *исправить*).

References

1. *Apresian Iu. D., Boguslavskii I. M., Iomdin L. L., Lazurskii A. V., Pertsov N. V., Sannikov V. Z., Tsinman L. L.* 1989. Linguistic Supply of the System ETAP-2 [Lingvisticheskoe Obespechenie Sistemy EETAP-2].
2. *Apresian Iu. D., Boguslavskii I. M., Iomdin L. L., Tsinman L. L.* 2007. Lexical Functions in Actual NLP-Applications. Selected Lexical and Grammatical Issues in the Meaning-Text Theory. In honour of Igor Mel'cuk, 84 : 199–230.
3. *Apresian Iu. D., Boguslavskii I. M., Iomdin L. L., Tsinman L. L.* 2002. Lexical Functions in NLP: Possible Uses. Computational Linguistics for the New Millennium: Divergence or Synergy? Festschrift in Honour of Peter Hellwig on the occasion of his 60th Birthday : 55–72.
4. *Apresian Iu. D., Diachenko P. V., Lazurskii A. V., Tsinman L. L.* 2007. On Electronic Textbook for Russian Lexic Studies [O Komp'iuternom Uchebnike Leksiki Russkogo Iazyka]. Russkii Iazyk v Nauchnom Osveshchenii : 48–112.
5. *Biriuk O. L., Gusev V. Iu., Kalinina E. Iu.* Dictionary of Russian Non-Object Nouns' Verbal Compatibility [Slovar' Glagol'noi Sochetaemosti Nepredmetnykh Imen Russkogo Iazyka]. Slovar' na osnove Natsional'nogo Korpusa Russkogo Iazyka dict.ruslang.ru.
6. *Boguslavskii I. M., Chardin I. S., Grigor'eva S. A., Grigor'ev N. V., Iomdin L. L., Frolova T. I., Shemanaeva O. Iu.* 2010. Lexical-Functional Texts Markup in SinTagRus [Leksiko-funktsional\naia Razmetka Teksta v SinTagRus].Informatsionnye Tekhnologii I Sistemy (it is'10): Sbornik Trudov Konferentsii, (Proc. of the Conference "Information Technologies and Systems") : 320–324.
7. *Boguslavskii I. M., Grigor'ev N. V., Grigor'eva S. A., Iomdin L. L., Kreidlin L. G., Frid N. E.* 2002. Development of Syntactically Marked Out Russian Corpus [Razrabotka Sintaksicheski Razmechennogo Korpusa Russkogo Iazyka].Doklady Nauchnoi Konferentsii "Korpusnaia Lingvistika I Lingvisticheskie Bazy Dannykh", (Proc. of Scientific Conference "Corpus Linguistics and Linguistic Databases"): 40–50.
8. *Kreydlin L. G., Frid N. E.* 2002. Development of a Dependency Treebank for Russian and its Possible Applications in NLP. Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002), III : 852–856.
9. *Kustova G. I.* Russian Idiomatic Expressions Dictionary. Words Combinations with the Meaning of a High-Scale [Sochetaniia Slov so Znacheniem Vysokoi Stepeni]. Slovar' na osnove Natsional'nogo Korpusa Russkogo Iazyka dict.ruslang.ru.
10. *Mel'chuk I. A.* 1974. Experiment of the Theory of Linguistic Models "Meaning - Text" [Opyt Teorii Lingvisticheskikh Modelei "Smysl - Tekst"].
11. *Sizov V. G., Tsinman L. L.* 2003. ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT. MTT 2003, First International Conference on Meaning — Text Theory : 279–288.
12. *Zholkovskii A. K., Mel'chuk I. A.* On Semantic Synthesis [O Semanticheskom Sinteze]. Problemy Kibernetiki, 19 : 177–238.

КУРСАЧ В АТТАЧЕ: ОСОБЕННОСТИ ЭЛЕКТРОННОЙ КОММУНИКАЦИИ МЕЖДУ ПРЕПОДАВАТЕЛЕМ И СТУДЕНТОМ[1]

К. А. Гилярова (hilaris@yandex.ru)

Институт лингвистики РГГУ, Москва, Россия

В работе переписка по электронной почте между преподавателями и студентами российских вузов рассматривается с точки зрения поля (field), модуса (mode) и тональности (tenor). По содержанию письма подразделяются на «письма-контейнеры», «организационные письма» и «письма-по-существу». По модусу организационные письма оказываются ближе к устной речи, чем к письменной. Противопоставление писем студентов и преподавателей по тональности лежит в области социальных отношений между корреспондентами.

Ключевые слова: преподаватель, студент, электронная коммуникация, социальные отношения, тональность.

CHARACTERISTICS OF STUDENT-PROFESSOR E-MAIL COMMUNICATION

K. A. Giliarova (hilaris@yandex.ru)

Russian State University for the Humanities, Moscow, Russian Federation

We analyse student-professor e-mail interaction in Russian universities in terms of Field, Tenor and Mode [Halliday 1978]. According to their content, we classify all e-mail messages into three types: “container e-mails”, “organizational e-mails” and “essential e-mails”. Even though the e-mail correspondence is a variety of the written communication mode, in organizational e-mails many speech-like features are present. They contain temporal and spatial deixis, anaphora and references to common ground. The word order is typical for colloquial speech, which makes organizational e-mails

[1] Автор признателен всем предоставившим ему свою личную переписку, а также благодарит за помощь в подготовке доклада Л. Д. Беклемишева, В. И. Киммельмана, А. Ч. Пиперского и А. А. Сомина.

closer to phone calls. E-mails series resemble oral dialogues. Both students and professors use different discourse styles: formal, informal, slang, etc. The mode of writing depends more on the authors' age and computer skills rather than on their social status. However, the differences in tenor between the e-mails of students and professors do exist. They are explained by the different perceptions of the norms of social communication and politeness. The analysis of opening and closing formulae also shows that there is no signifi cant difference between the mode of writing e-mails by students and professors. Nevertheless, some specifi c traits can be found.

Key words: professor, student, e-mail, e-mail communication, social relations, communication mode.

1. Введение

В последнее двадцатилетие электронная почта привлекает все более и более пристальное внимание лингвистов. Так, в [Herring 1996] на базе рассылок *Linguist List*[2] и *WMST*[3] исследуется женский и мужской электронный дискурс. Анализ показал, что женщины чаще выражают в письмах согласие и поддержку, в то время как мужчины более категоричны и склонны к критике. Дэвид Бэнкс, анализируя небольшой корпус писем из рассылки *Sysfling*[4], находит, что для академического электронного дискурса характерен книжный, а не разговорный стиль [Banks 2001]. К сходным выводам приходит и *Shin Ja J. Hwang* в своей работе с рассылками *LinguistList u Funknet*[5]. Согласно ее результатам, электронный академический дискурс можно отнести к объяснительному (экспозиторному) типу [Hwang 1998]. В [Hård af Segerstad 2000] сравниваются бумажные и электронные письма горожан к мэру Гётеборга (Швеция). По мнению автора, решающими факторами, влияющими на стиль электронных писем, является их бóльшая анонимность, легкость и быстрота в отправке и доступность данного средства коммуникации. Одной из немногих работ по электронной коммуникации на материале русского языка является исследование [Зализняк, Микаэлян 2006]. В нем выявляются особенности электронной почты как коммуникативного жанра, стоящего особняком как от традиционного эпистолярного жанра, так и от личного и телефонного общения.

Общение по электронной почте между преподавателями и студентами стало объектом пристального внимания социологов [Taylor & al. 2011], [Wrench, Punianunt-Carter 2005], [Wrench, Punianunt 2004], [Frey & al. 2003], [Duran & al. 2005]. Так, в [Taylor & al. 2011] показано, что, несмотря на активное использование студентами электронной почты для контакта с преподавателями, они все же предпочитают «живое» общение для достижения и карьерных, и личных целей.

---

2 http://linguistlist.org/

3 http://userpages.umbc.edu/~korenman/wmst/wmst-l_index.html

4 https://mailman.cf.ac.uk/mailman/listinfo/sysfling

5 https://mailman.rice.edu/mailman/listinfo/funknet

Однако в потоке исследований по электронной коммуникации в учебной среде не хватает лингвистических работ, и тем более на материале русского языка.

В настоящей работе предполагается исследовать русскоязычную переписку по электронной почте между преподавателями и студентами, опираясь на семиотическую модель М. Халлидея [Halliday 1978]. В этой модели лингвистические ситуации рассматриваются в трех измерениях: 1) ***поле*** (field) — предметная область общения; 2) ***тональность*** (tenor) — стиль дискурса, определяемый отношениями между участниками общения; 3) ***модус*** (mode) — канал общения и связанные с ним особенности коммуникации [Карасик 1992].

2. Материал исследования

Материалом исследования послужил собранный нами корпус из 630 электронных писем. 332 из них написаны преподавателями и обращены к студентам, а 290, наоборот, написаны студентами преподавателям. В переписке задействовано 30 преподавателей и 77 студентов (или групп студентов, ибо значительная часть писем была отправлена на общие адреса курса или группы). Количество писем, написанных одним и тем же автором, колеблется в пределах 3–15 для преподавателей и 1–8 для студентов. 45% писем написаны филологами (литературоведами и специалистами по германским и романским языкам), 30% — лингвистами, 15% — математиками, 5% — биологами, 5% — всеми остальными (философы, историки, специалисты по менеджменту и рекламе). Все письма относятся к интервалу 2006–2010 гг[6].

Кроме того, в качестве пилотного исследования нами был произведен небольшой опрос, в котором приняли участие 38 преподавателей и 32 студента. Информантам были предложены вопросы о предпочитаемых ими формах приветствия и прощания, об употреблении эмотиконов (или *смайликов*), а также об их коммуникативных ожиданиях в процессе электронного общения с корреспондентом другого статуса (преподавателем или студентом соответственно).

3. Поле (field)

Согласно [Taylor & al. 2011], американские студенты пишут преподавателям как для решения организационных вопросов, связанных с учебой, так и для налаживания личных отношений. В российских вузах электронная коммуникация между учениками и преподавательским составом пока еще не столь развита. Чтобы очертить предметную область этой переписки, следует задаться вопросом о целях, которые преследуют студенты и преподаватели, обращаясь до или после занятия к своему почтовому ящику.

Согласно целям, все письма нашего корпуса можно условно подразделить на три группы. Первый тип (190 писем) — это письма с вложениями, или

[6] Автор не предполагает проводить строгий количественный анализ корпуса.

письма-контейнеры. Студенты пересылают свои эссе, рефераты, курсовые и дипломные работы, преподаватели — тексты, аудио и видеоматериалы к урокам, списки вопросов к экзамену, проверенные или отредактированные работы, таблицы с оценками и т. д. Второй тип (229 примеров) — это электронные письма, в которых преподаватель договаривается с учащимися о времени и месте занятий, переносах пар, датах экзамена и пересдач, встречах во внеучебные часы для работы над курсовыми и дипломными работами, домашних заданиях, книгах, которые надо взять в библиотеке и т. д. Одним словом — в этих письмах решаются организационные вопросы, так что мы назовем их ***организационными***. Наконец, примерно равную долю (211 писем) составляют ***письма-по-существу***. В них, например, преподаватель отвечает на вопросы студента по теме, объясняет, как решается задача, или же студент представляет план своего доклада или рассказывает о мероприятии, на котором побывал. В организационных письмах обсуждаются не научные, а мета-вопросы, что делает их контекстно-зависимыми и сильно привязанными к адресату, адресанту, месту и времени. Напротив, в письмах-по-существу решаются проблемы мироустройства (будь то даже всего лишь вопрос о транслитерации имени малоизвестного художника), так что содержание письма представляет ценность не только для адресата, но и для стороннего читателя, вне контекста. Письма-по-существу не устаревают сразу после прочтения и могут быть интересны даже следующим поколениям. Они, в среднем, более развернуты по сравнению с письмами первых двух видов, и в них возможна комбинация дискурсов разных типов: нарративного, дескриптивного, экспозиторного, инструктивного, аргументативного.

Ясно, что некоторые письма можно отнести сразу к двум, а то и всем трем группам: в одном письме могут содержаться сразу вложение, назначение встречи и конкретные вопросы по учебе. Небольшой процент писем более личного характера не попал ни в какую из трех категорий и был вынесен из рассмотрения по этическим соображениям.

4. Модус (mode)

Строго говоря, канал передачи информации во время электронной коммуникации определяется однозначно как письменный. Так что в узком смысле правильно говорить о письменном модусе или, учитывая отличия электронной почты от обычной, электронном субмодусе [Кибрик 2009]. Однако, согласно нашему исследованию, в одних случаях электронное общение более приближено к разговорной речи, чем в других. Что позволяет нам проводить противопоставление по модусу внутри электронного дискурса.

4.1. Так, письма-контейнеры обладают свойством, невозможным для писем других типов: они могут вообще не содержать текста (в нашем корпусе «пустые» письма составляют 20% от всех писем-контейнеров). Основная ценность такого письма во вложении, а тексту отводится периферийная роль. Сколь бы информативным ни был прикрепленный файл, для корреспондентов это не столько обмен

информацией, сколько обмен товаром. Тому, что вложение воспринимается как материальный объект, мы находим подтверждение и в лексике соответствующего семантического поля: *прикрепить / приложить* (файл), *загрузить* (аудиозапись), *держите, ловите, слишком тяжелый* (файл) и т.д. Аналог письма-контейнера — посылка, отправленная обычной почтой. Сопроводительное письмо, как правило, очень краткое, призвано обратить внимание адресата на отсылаемый файл:

(1) *<Имя>, вот Ваш текст наконец-то. Читайте исправления, дополняйте комментарий.*

В сопроводительных письмах часто встречаются глаголы *отправлять*, *прилагать*, *слать* и др. в форме первого лица единственного числа настоящего времени в употреблении, близком к перформативному. Конечно, эти глаголы не заменяют действие, а лишь сопровождают его. Однако употребленное в письме *Отправляю Вам мою курсовую работу* обладает большей иллокутивной силой, чем констативное *Каждый вечер я отправляю ему очередную главу*. В первом случае высказывание не может быть оценено в терминах истинности или ложности, а может — в терминах успешности и неуспешности речевого акта. Ср. слово *шах*, произнесенное во время шахматного хода. Не являясь перформативом в чистом виде, оно, тем не менее, неотделимо от определенного действия и не имеет смысла без него [Остин 1986, с. 75]. Так и слова *отправляю, прилагаю, шлю* предполагают подкрепление слова делом[7]. Если же файл, тем не менее, в приложении отсутствует, остается нереализованным отношение иллокутивного самовынуждения [Баранов, Крейдлин 1992], и наступает коммуникативная неудача. Ср.:

(2) *Добрый вечер, <Имя Отчество>,*
Вы, кажется, забыли прикрепить сам файл с литературой.:)

Кроме того, оправляя файл, адресант нередко ждет от адресата подтверждения о получении и даже иногда специально запрашивает его, переводя этим «посылку» в ранг заказных писем. Отсутствие подтверждения о получении также ведет к коммуникативной неудаче и заставляет адресанта повторно искать контакта:

(3) *<…>Я бы хотела уточнить, получили ли Вы мое письмо с переводом (курсовая за 1 курс). <…>*

Таким образом, для писем-контейнеров вообще не всегда правомочно говорить о модусе: сама по себе передача объекта не является ни письменной, ни устной коммуникацией, а может только сопровождаться ею.

---

[7] Этот факт уже закреплен в некоторых почтовых серверах. Так, при попытке отправить письмо без приложения с сервера www.gmail.com всплыло окно с подсказкой: *Вы собирались прикрепить файлы? В вашем сообщении есть фраза «прилагаю», однако прикрепленные файлы отсутствуют. Все равно отправить?*

4.2. Если письма-контейнеры сравнивать с посылкой, то аналог организационных писем — телефонный разговор.

Во-первых, им, как и разговорной речи, свойственны дейксис, анафора и отсылки к фоновым знаниям корреспондентов:

а) временной дейксис [Зализняк, Микаэлян 2006] и анафора (в письмах не уточняется, какие именно дата и время имеются в виду под «завтра», «сейчас», «праздниками»):

(4) *сейчас посмотрю то, что вы прислали сейчас*

(5) *через полчасика-часик готовый вариант диплома пришлю*

(6) *встретимся после праздников*

(7) *да я приеду я сегодня на кафедру подходил*

б) пространственный дейксис и анафора:

(8) *В понедельник встречаемся там же, где обычно*

(9) *сможете сюда подъехать?*

в) другие отсылки к имплицитной информации (фоновым знаниям корреспондентов, прагматическим предусловиям текста, предыдущим ситуациям речевого общения и т. д.):

(10) *вот песня, о которой мы говорили*

(11) *в продолжение нашего с Вами разговора об авторах*

г) межписемная анафора [Зализняк, Микаэлян 2006]:

(12) *<Имя>, насчет первого вопроса — конечно, могло. Насчет порядка слов — все правильно <…> Насчет второго — filth и foul в этимологическом плане однокоренные слова <…>.*

Организационные письма часто не понятны вне экстралингвистического контекста и скоро теряют актуальность, по мере того как уменьшается активация референтов из фоновых знаний участников диалога.

Во-вторых, в организационных письмах чаще, чем в других, встречается порядок слов, характерный для устной речи:

(13) *Дело в том, что я пишу по лексикографии у проф. Крейдлина работу об особенностях в названиях картин импрессионистов.*

(14) *Не было Интернета у меня, только сейчас появился <…> Но мне исправно одногруппники давали задания — у Вас там есть взрослый такой… Миша вроде бы.*

(15) *Вы мне напишите, пожалуйста, что наша группа по грамматике проходит, может быть, упражнения какие-то.*

В-третьих, междометия и восклицания, также принадлежащие к разговорной речи:

(16) *Мда, жалко;*

(17) *…фуф…*

(18) *ну Вы меня и озадачили!*

(19) *Ой, а и верно. Наверное, Стефани и видела, да.*

В-четвертых, как и телефонный разговор, адресант иногда начинает электронное письмо с представления, особенно если обращается к адресату по почте впервые:

(20) *Здравствуйте, <Имя Отчество>! Это <имя фамилия>, я подходила к вам насчет курсовой.*

(21) *Здравствуйте, это студентка второго курса <фамилия имя>, вы ведете у нас французский язык.))*

В подобных письмах нередко опускается подпись, остается только прощальная формула (например, *заранее большое спасибо*; *жду ответа; хороших выходных* и т. д.), что также приближает электронные письма к телефонному разговору.

4.3. Серийность

Организационные письма чаще, чем другие, предполагают скорый ответ, ведь чтобы договориться о чем-либо, обычно требуется как минимум по одной реплике с каждой стороны. Так возникают серии писем и ответов на них, которые образуют единый коммуникативный акт [Зализняк, Микаэлян 2006]. К особым правилам, обеспечивающим связность серии, авторы относят уже упоминавшиеся здесь межписемную анафору и временной дейксис, а также возможное отсутствие обращения, приветствия и прощальной формулы в неначальных письмах цепочки. «Организационные» серии, кроме того, зачастую характеризуются скорым уменьшением длины писем с последующей их редукцией до отдельных однострочных реплик, что сближает разговор по электронной почте с общением в режиме чата. В таком режиме переписка максимально

приближена к устному диалогу. Приоритетом становится уже не форма письма, а скорость, с которой оно отправлено, что приводит к опечаткам, неправильным грамматическим формам, пренебрежением прописными буквами и тем более эпистолярными формальностями. Кроме того, начинают активнее использоваться эмотиконы (или *смайлики*) и авторская пунктуация, призванные заменить невербальный компонент устного общения. Рассмотрим серию писем между студенткой и ее научным руководителем.

(22) С: *<Е.В.>, добрый вечер!*
Я Вам завтра утром пришлю оставшуюся часть курсовой. Скажите, пожалуйста, а Вы первую часть еще не проверили?
<Имя Фамилия>[8].
П:. *<Л.>, проверяю. Шлите вторую часть, ночью/утром все вышлю.*
С: *<Л.>, ловите. Обратите внимание на изменения на титуле.*
С: *<Е.В.>! Вот вторая часть рассказа.*
P. S. Не понимаю, о чем я думала, когда написала "Вашингтон Черчилль")))))))))) Порадовала сама себя
<Имя Фамилия>.
П: *Ловите.*
С: *Спасибо!*
А скажите, пожалуйста, в комментариях, когда я буду приводить примеры переводческих трансформаций, мне надо подборно описывать, как и зачем я воспользовалась тем или иным способом. Или просто привести примеры предложений и указать, какой прием использовался?
<Имя Фамилия>.
П. *Не надо подробно. Прием — пример, прием — пример.*

Диалог (22) соответствует минимальной диалогической единице (МДЕ) [Баранов, Крейдлин 1992]. В пределах этой серии писем все отношения иллокутивного вынуждения и самовынуждения выполнены, она начинается с абсолютно независимого и кончается абсолютно зависимым речевым актом, и все письма в ней связаны одной темой. Однако для применения понятия МДЕ к диалогам по электронной почте следует добавить еще одно условие: ограничение временно́го расстояния между репликами. Негласная этика электронного общения предписывает отвечать на письма в течение 1–2 суток, а в случае коротких сообщений, содержащих вопрос, — в течение нескольких часов. Если перерыв между репликами слишком большой, происходит коммуникативная неудача, и не получивший ответа начинает писать повторно или звонить.

Таким образом, ситуации электронной коммуникации между студентом и преподавателем можно если не противопоставить, то по крайней мере различить по модусу. Организационный тип более других типов тяготеет

[8] <Имя Фамилия> в конце каждого письма студентки остаются, так как подпись автоматически ставится в конце ее писем, согласно настройкам почтового ящика.

к устной речи. И, возможно, можно было бы говорить о вытеснении перепиской по электронной почте телефонных разговоров, если бы она сама не была вытесняема системами мгновенного обмена сообщениями (*Instant Messager*) и общением в социальных сетях. Письма-по-существу чаще других остаются в рамках письменного модуса. Для писем-контейнеров вообще не всегда правомочно говорить о модусе, ибо они не являются коммуникацией как таковой.

5. Тональность (tenor)

«Тональность относится к отношениям между участниками общения и характеризует степень формализованности этих отношений, наличие старшинства и иерархии, степень знакомства, сходство или различие по личностным характеристикам» [Карасик 1992]. Сначала рассмотрим подробнее особенности стиля электронного общения между преподавателями и студентами, а потом посмотрим на принятые при этом общении приветственные и прощальные формулы.

5.1. Особенности стиля

Как отмечалось выше, электронный дискурс находится на перекрестке письменной и устной речи. Весь континуум электронных писем из нашего корпуса можно распределить по шкале «официальности» или «литературности». На одном конце шкалы будут электронные письма, написанные в лучших традициях эпистолярного жанра, на другом — корявые строки будто из чата, лишенные орфографии и грамматики. По нашей гипотезе *a priori* письма преподавателей тяготеют к первому полюсу шкалы, а студенты в ситуации общения с преподавателем либо также придерживаются формального стиля, либо остаются поблизости от «разговорного» края поля. Однако анализ корпуса опроверг наши предположения. Оказалось, что в ситуации электронной коммуникации между преподавателями и студентами стиль их писем почти не различается. И в группе преподавателей, и в группе студентов встречаются как приверженцы официального языка, так и любители сниженного регистра и сетевого жаргона. Язык писем скорее коррелирует с возрастом пишущего и его «продвинутостью» как пользователя Интернета, нежели с его статусом. Молодые преподаватели, не так давно вышедшие из студенческого возраста, сохраняют свой молодежный стиль, а электронные письма преподавателей среднего и пожилого возраста приближены по своим характеристикам к бумажным.

Не останавливаясь на примерах высокого стиля, рассмотрим подробнее следы разговорной речи и языка Интернета, которые встретились во множестве в нашем корпусе, несмотря на кажущуюся официальность ситуации общения преподавателей со студентами.

1. Сниженная разговорная лексика: *пошерстить (по нашим изданиям); понавтыкать (примеров); почеркать (текст); прорвемся; не заморачивайтесь; отмазка; несовпадуха; ни фига себе; торможение (с ответом); иначе повешусь; задевать (письмо, файл); иссесьно (=естественно)* и т. д.
2. Сетевой жаргон: *ловить (приложенный файл), в аттаче (= в приложении), скачать, залить (загрузить), ссыль (=ссылка), скинуть (=переслать)* и т. д.
3. Студенческий жаргон: *курсач, препод, появиться на занятии, завалить (экзамен), валить (студента на экзамене), хвост (=задолженность)* и т. д.
4. Сокращения: *инф-ция, проф., Инет, пт (пн, вт), неуд, ДЗ.*
5. Уменьшительные: *вопросик, текстик, полчасика-часик, (выпейте) валерьяночки.*
6. Игнорирование прописных букв, многочисленные ошибки и опечатки:

(23) *добрый вечер а скажите когда и где будет перездача по риторике у третьего курса а то не висит у деканата*

(24) *Здравстуте!я еек союжалению болею отому не была на занятих=(<Имя Фамилия>*

7. Эмотиконы и дополнительная авторская пунктуация:

(25) *С: Вот так?))) Простите за мою дотошность)))*
П: угу, почти :) ловите :)

(26) *И последний вопрос (простите, пожалуйста, что так много :(Можно оппоненту послать работу без заключения?.. Она же еще не совсем готова, выводы делать трудновато….*

Наряду с избыточной пунктуацией, использованием цветных и выделяющих шрифтов, написанием отдельных слов прописными буквами с целью его эмфатического выделения, растягиванием слов (*урааааа!)* и другими стратегиями электронного модуса общения, эмотиконы используются для передачи на письме невербального компонента устного дискурса [Derks & al. 2007], [Walther, D'Addario 2001], [Макаров, Школовая 2006]. Не зря они официально называются «эмотиконы» (англ. *emoticons* от *emotions*), а в быту — «смайлы» или «смайлики» (от англ. *smile*). В [Макаров, Школовая 2006] справедливо отмечается, что эмотиконы могут дублировать вербальный компонент высказывания, замещать его, расширять или углублять, разъяснять, а могут и ему противоречить. А порой употребляются без каких-либо очевидных синтаксико-семантических связей с вербальным текстом, являясь аналогом слов-паразитов в речи [там же, стр.366]. Анализ нашего материала показывает, что именно последняя функция эмотиконов — самая частотная в письмах студентов. Смайлики выступают в роли своеобразных

дискурсивных маркеров, заполняя паузы между отдельными фрагментами текста. Особенно часто для заполнения пауз на письме используется эмотикон из нескольких скобок: *))))* или *=))*. Пишущий ставит подобные знаки также в момент, когда испытывает неприятные чувства — смущение, неловкость, стыд, страх, беспокойство. Ср. (25)–(26), а также (27–31):

(27) *Здравствуйте :)))) У меня нет зачета за прошлый семестр :((((Когда я могу к вам подойти?:))))*

(28) *<…> Вы, кажется, забыли прикрепить сам файл с литературой. :)* (неловко напоминать преподавателю)

(29) *Дорогая <…>, я УМЕЮ пользоваться Яндексом! :(((* (ответ оппонента на присланный полностью скачанный из Интернета диплом; преподавателю неловко и неприятно)

(30) *А Вы не могли бы прислать ту песенку на стихи G. Apolllinaire "Le pont Mirabeau"? Заранее большое спасибо)* (смущение)

(31) *Спасибо за список — я уже заказала все пособия) То, что группа несильная, меня не пугает, поскольку мне самой нужно будет многое по хожу вспомнить, может, так даже лучше) <…> Очень надеюсь быстро встроиться и нагнать группу :) Заранее Вам большое спасибо за такую возможность:)* (слово-паразит, своего рода дискурсивный маркер)

Многочисленные многоточия (часто не из трех, а из двух точек) посреди текста служат той же цели и практически приравниваются по функциям к эмотиконам из нескольких скобок:

(32) *<Имя Отчество>, высылаю весь диплом.. (на счет титульника и оглавления — знаю, что они не должны быть прономерованы, просто уже сил нет разбираться..и могло что-ниб поехать..но с этим я уже завтра разберусь…в целом, вроде все хорошо..113 стр. получилось!!!))*

В опросе, в котором принимало участие 38 преподавателей, в том числе спрашивалось: «Что раздражает или расстраивает Вас в студенческих письмах?» Многие назвали отсутствие прописных букв, абзацев и пунктуации, отсутствие обращения, приветствия и подписи (что порой ведет и вовсе к анонимности), отсутствие сопроводительного письма при пересылке файлов и подтверждения при их получении, обилие орфографических ошибок и опечаток, избыток эмотиконов. Некоторые отметили также фон с картинками, шрифты разного цвета, автоматические подписи, намеренное искажение орфографии. Употребление же разговорной и даже жаргонной лексики не вызвало нареканий. Не случайно пункты 1–5 из списка выше встречаются и в письмах студентов, и в письмах преподавателей.

Так что же тогда различает по тональности письма преподавателей и студентов? Многие информанты-преподаватели жаловались на «фамильярность, невежливость, запанибратство». Чтобы понять, что имеется в виду в данном контексте, нужно вспомнить о категории вежливости и различию по формальности, которое основано на характере социальных отношений между корреспондентами [Бергельсон 2007], [Кибрик 2009]. Ср. следующие примеры из писем студентов, обращенных к преподавателям:

(33) *<Имя Отчество>, да Вы просто метеор!:) Спасибо большое за быстрый ответ!*

(34) *Держитесь!!! :) Как вообще можно держать в голове такой поток информации?!!!...*

(35) *Я поняла Вас. Завтра я собираюсь появиться на занятии, тогда я с Вами и поговорю, что и как лучше. <…> Увидимся завтра.*

(36) *<Имя Отчество>! К сожалению, практически никто и нашей группы не остается на праздники в Москве, так что занятие, видимо, проводить не стоит. С уважением, группа <…>*

(37) *Добрый день! Я собираюсь рассказывать задачи 56, 19, 39. Ок?*
<Имя Фамилия>

Во всех письмах, откуда взяты эти цитаты, у студентов были самые добрые намерения: выразить свое восхищение, поблагодарить, сообщить об отмене занятия, чтобы преподаватель не приезжал зря, и т. д. Однако у преподавателей осталось впечатление невежливости и даже хамства, то есть налицо коммуникативная неудача. Почему? Потому что по своему статусу студенту не положено оценивать и подбадривать преподавателя (33–34), употреблять выражения *появиться на занятии* и *увидимся (35)*, судить о том, как правильно поступить (пример (36) был бы корректен, будь он в модальности предложения, а не суждения), ставить перед фактом и лишь после просить одобрения, тем более в разговорных выражениях (37).

В [de Siqueira, Herring 2009] исследовалась скорость переписки в системе *Skype* одного научного руководителя с четырьмя студентами. Было выявлено, что разговоры велись в четырех разных ритмах, в зависимости от студента. Таким образом, вопреки «статусной» гипотезе, именно преподаватель подстраивался под студентов, а не наоборот. Однако исследование вряд ли можно считать полным из-за недостаточной выборки: в эксперименте участвовал всего один преподаватель. Выборка также была некорректна: трое из четырех студентов, участвовавших в эксперименте, были иностранцами, то есть гостями в стране преподавателя (США), к тому же в разной степени владели языком переписки (английским). Почти все принимавшие участие в нашем опросе студенты отметили, что задавать тон в электронном диалоге должен преподаватель, а студент

под него подстроится, в частности, в вопросе допустимости употребления эмотиконов и жаргонных выражений. Преподаватели же согласились, что относятся положительно или нейтрально к сниженному регистру речи студентов лишь в том случае, если сам преподаватель первым перешел на менее формальный уровень общения.

Что же не нравится студентам в преподавательских письмах? Согласно результатам опроса, студентов больше всего расстраивает, когда преподаватель долго не отвечает на письма, а также не присылает подтверждений о получении каких-то важных текстов. Структура диалога таким образом нарушается, а слишком часто напоминать о себе студенту не позволяет его статус. Также многим студентам не нравится излишняя формальность и дистанция в преподавательских письмах: «Вы» с большой буквы, обращение «уважаемый» или «господа». В то же время некоторые преподаватели намеренно формализуют свою речь, боясь оскорбить студента отсутствием этих атрибутов вежливости и не подозревая, что студенту было бы приятнее без них. Происходит все та же коммуникативная неудача из-за нарушений категории «статусности», только уже в другую сторону.

5.2. Обращения, формулы приветствия и прощания

В *Таблице 1* и *Диаграмме 1* представлены наиболее часто встречающиеся в электронной переписке между преподавателями и студентами формулы приветствия и обращения.

Как видно из таблицы и диаграммы, 36,7 % преподавательских и 30 % студенческих писем вовсе не имеют стандартного для бумажных писем зачина — обращения или приветствия. В данную категорию попадают и не первые письма из серий-диалогов, и «пустые» письма-контейнеры, и письма, не содержащие обращения в силу выбранного автором стиля. В 47,1 % писем студенты обращаются к преподавателю по имени-отчеству, и, соответственно, в 22,2 % писем преподаватели называют по имени студентов. Еще приблизительно 16,2 % преподавательских писем тоже содержат обращение, но к группе (*Уважаемые студенты-филологи! Дорогая группа!*), которое тоже в некотором роде можно отнести к обращению по имени. Таким образом, прямое обращение к адресату используется студентами и преподавателями с близкой частотой (47,1 % и 38,4 %), но для студентов варианты ограничиваются вариантами *Здравствуйте / добрый день, X*, изредка *Уважаемый X* и вообще без приветствия, в то время как преподаватели в 16,2 % случаев используют еще форму с *Дорогой / Дорогие*. С обращением *Дорогой X* связаны самые противоречивые результаты опроса. Одни студенты не любят это обращение от преподавателя, воспринимая его как подчеркивание дистанции и даже иронию, для других, наоборот, это показатель хорошего отношения и особого расположения. В опросе несколько студентов отметили, что в свою очередь пишут преподавателю *Дорогой*, но в наших данных не встретилось ни одного такого примера.

Таблица 1

	Формулы приветствия, обращение	Студенты	Преподаватели
1	Приветствие и обращение отсутствуют	30%	36,7%
2	Здравствуйте, <Имя / И.О.>! <Имя / И.О.>, здравствуйте!	28,2%	10,8%
3	Здравствуйте! (без обращения)	10,6%	3,6%
4	<И.О.> или <Имя>	8,6%	8,4%
5	<Имя / И.О.>, добрый день (вечер, утро)! Добрый день (вечер, утро), <Имя / И.О.>!	7,2%	2,7%
6	Добрый день (вечер, утро, доброй ночи)! (без обращения)	4,1%	12%
7	Уважаемый (-ая) <Имя / И.О.>! Уважаемые студенты (бакалавры, пятикурсники и т.д.)	3,1%	3%
8	Дорогие студенты! (второкурсники, магистранты, коллеги и т.д.) Дорогая группа! Дорогой <Имя, И.О.>!		16,2%
9	Остальное	8,2%	6,6%
	Всего писем	298	332

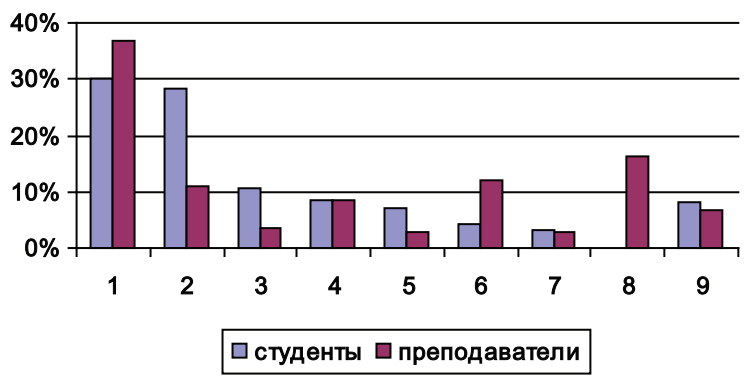


Диаграмма 1. Распределение формул приветствия и обращений в письмах студентов и преподавателей друг к другу[9]

Почти одинаковое количество (14,7% и 15,6%) студенческих и преподавательских писем содержат приветствие, но не обращение. Видимо, именно в стольких случаях студент не знает наверняка имени-отчества преподавателя, которому пишет, а преподаватель, в свою очередь, не знает, кому отвечает, или не знает, как лучше обратиться. Любопытно, что студенты при

[9] Цифрам по оси абсцисс соответствуют строки *Таблицы 1*, ось ординат показывает количество употреблений разных формул приветствия в процентах относительно общего числа студенческих или преподавательских писем соответственно.

этом отдают предпочтение форме «здравствуйте» (10,6% против 4,1%), а преподаватели — вариантам «добрый день», «добрый вечер» и т. д. (12% против 3,6%). Не включены в таблицу встречающиеся в нашем корпусе в единичных случаях обращения *Друзья, Господа, Граждане, Девочки, Дети мои, Привет!* от преподавателей и *Доброго времени суток!*, *здрасте!*, *hi!*, *Привет!* от студентов.

В *Таблице 2* представлены выбираемые студентами и преподавателями при переписке формулы прощания и подписи[10].

Таблица 2

	Формула прощания, подпись	**Студенты**	**Преподаватели**
1	Формула прощания и подпись отсутствуют	40,9%	26,8%
2	С уважением, <Имя Фамилия> (студенты такой-то группы, филологи и т. д.)	26,8%	5,1%
3	Имя	15,4%	6%
4	Имя Фамилия, И. Фамилия, И. О. Фамилия	7,4%	9,6%
5	И.О. И.Ф. Ф.И.О.		26,2% 12,9% 3%
6	Имя Отчество (полностью)		4,2%
8	Группа <…> (филологи, германисты и т. д.)	3,4%	
7	Спасибо	5,7%	1,2%
9	Всего доброго (хорошего, наилучшего)	2,3%	3%
11	Ваш (а)		6,6%
10	Удачи		4,2%
12	До встречи (до завтра, до пятницы и т. д.)		2,4%
13	Остальное	1,3%	3%
14	Всего писем[11]	298	332

Как видно из таблицы, 40,9% студенческих и 26,8% преподавательских писем не содержат ни подписи, ни формулы прощания. На втором месте (26,8%) среди студентов формула *С уважением*, не столь популярная, но все же использующаяся и среди преподавателей (5,1%). Студенты подписываются по имени или по имени и фамилии (22,8%), в то время как преподаватели предпочитают инициалы с преобладающим вариантом И.О. (26,2%), который, в отличие от варианта И.Ф. (12,9%), не используется ни в какой другой среде,

[10] Эти результаты не очень точно отражают реальное положение дел, поскольку подпись зачастую автоматически встраивается во все письма, и пишущий не всегда меняет ее в соответствии с тем, кому пишет.

[11] Многие письма содержат и прощальную формулу (*С уважением, Всего доброго*), и подпись, а в последней строке таблицы указано число писем в корпусе, поэтому сумма процентов выходит за 100.

кроме как в преподавательской или учительской. Похоже, подписываться инициалами И.О., а также полным именем и отчеством (без фамилии) — своеобразный профессиональный жаргон преподавателей. Статусно не маркированными оказались формулы *Всего доброго / хорошего / наилучшего* — они используются одинаково и студентами, и преподавателями, в отличии от вариантов *Удачи, До встречи* и *Ваш X* являющихся прерогативой преподавателей. В графу «остальное» вошли студенческая формула *Жду ответа / Буду ждать Ваш ответ* и преподавательские *Счастливо* и *Пока*.

Таким образом, выбор формы приветствия и прощания оказался более нейтральным к категории формальности (статусу), чем можно было предполагать *a priori*. Иными словами, студенты и преподаватели начинают и заканчивают свои письма к противоположной стороне в среднем одинаково. Исключениями, а значит, маркерами статуса, являются прощальная формула *С уважением* от студента, обращение *Дорогой(-ая, -ие) X(-ы)* от преподавателя, а также его подпись в виде инициалов, причем преимущественно <И.О.>.

Заключение

В работе переписка по электронной почте между преподавателями и студентами российских вузов рассмотрена с точки зрения поля, модуса и тональности. По содержанию (полю) письма подразделяются на письма-контейнеры, организационные письма и письма-по-существу. Письма второго типа по модусу оказываются ближе к устной речи, чем к письменной. Кроме того, организационные письма носят в себе черты телефонного разговора, а серии писем — черты устного диалога. Письма преподавателей и студентов не различаются по стилю дискурса (официальный, разговорный, жаргонный и т. д.). Но противопоставление по тональности тем не менее можно провести — оно лежит в области категории вежливости, характера социальных отношений между корреспондентами. Анализ формул приветствия и прощания подтверждает этот результат, хотя и показывает, что в целом, если судить по электронной переписке, социальные различия между преподавателями и студентами не так уж велики.

References

1. *Austin J.* Word as an Action [Slovo kak Deistvie]. Novoe v Zarubezhnoi Lingvistike, 17 : 22–129.
2. *Banks D.* 2001. Academic Mediation: the Functions of an Electronic Discussion List. ASp, la Revue du GERAS, 31–33 : 77–87.
3. *Baranov A. N., Kreidlin G. E.* 1997. Illocutive Forcing in the Dialogie Structure [Illokutivnoe Vynuzhdenie v Strukture Dialoga]. Voprosy Iazykoznaniia, 2 : 84–100.

4. *Bergel'son M. B.* 2007. Pragmatic and Socio-Cultural Motivation of Linguistic Form [Pragmaticheskaia I Sotsiokul'turnaia Motivirovannost' Iazykovoi Formy].
5. *Derks D., Bos A. E. R., von Grumbkow J.* 2007. Emoticons and Social Interaction on the Internet: the Importance of Social Context. Computer in Human Behavior, 23 : 842–849.
6. *Duran R. L., Kelly L., Keaten J. A.* 2005. College Faculty Use and Perceptions of Electronic Mail to Communicate with Students. Communication Quarterly, 53 (2) : 159–176.
7. *Frey A., Faul A., Iankelov P.* 2003. Student Perceptions of Web-assisted Teaching Strategies. Journal of Social Work Education, 39 (3) :443–457.
8. *Halliday M. A. K.* 1978. Language as Social Semiotic: The Social Interpretation of Language and Meaning.
9. *Hård af Segerstad Y.* 2000. Electronic Letters to the City Council: Factors Influencing the Composition of Email Messages. Human IT, 4 (1).
10. *Herring S. C.* 1996. Two Variants of an Electronic Message Schema. Computer-mediated Communication: Linguistic, Social and Cross-cultural Perspectives : 81–108.
11. *Hwang S. J. J.* 1998. Expository Discourse Schema for Scholarly Electronic Messages. LACUS Forum, 24.
12. *Karasik V. I.* 1992. The Language of the Social State [Iazyk Sotsial'nogo Statusa].
13. *Kibrik A. A.* 2009. Modus, Genre and Other Parameters of Discourse Classification [Modus, Zhanr I Drugie Parametry Klassifikatsii Diskursov]. Voprosy Iazykoznaniia : 2–21.
14. *Makarov M. L., Shkolovaia M. S.* 2006. Linguistic and Semiotic Aspects of E-mail Communication Identity Construction [Lingvisticheskie I Semioticheskie Aspekty Konstruirovaniia Identichnosti v Elektronnoi Kommunikatsii]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2006" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2006") : 364–369.
15. *de Siqueira A., Herring S. C.* 2009. Temporal Patterns in Student-Advisor Instant Messaging Exchanges: Individual Variation and Accommodation. Proceedings of the Forty-Second Hawai'i International Conference on System Sciences (HICSS-42).
16. *Zalizniak A. A., Mikaelian I. L.* 2006. Emailing as a Linguistic Object [Perepiska po Elektronnoi Pochte kak Lingvisticheskii Objekt]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2006" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2006") :157–162.
17. *Taylor M., Jowi D., Schreier H., Bertelsen D.* 2011. Students' Perceptions of E-Mail Interaction During Student-Professor Advising Sessions: The Pursuit of Interpersonal Goals. Journal of Computer-Mediated Communication,16 : 307–330.
18. *Walther J. B., D'Addario K. P.* 2001. The Impacts of Emoticons on Message Interpretation in Computer–Mediated Communication. Social Science Computer Review Fall, 19 (3) : 324–347.

19. *Punyanunt N. M.* 2004. Advisee-advisor Communication: An Exploratory Study Examining Interpersonal Communication Variables in the Graduate Advisee-Advisor Relationship. Communication Quarterly, 52 (3) : 224–236.
20. *Wrench J. S., Punyanunt-Carter N. M.* 2005. Advisee-Advisor Communication Two: The Influence of Verbal Aggression and Humor Assessment on Advisee Perceptions of Advisor Credibility and Affective Learning. Communication Research Reports, 22 (4) : 303–313.

ГРАММАТИЧЕСКОЕ ИССЛЕДОВАНИЕ НА ПОЛУРАЗМЕЧЕННОМ КОРПУСЕ ТЕКСТОВ (НА МАТЕРИАЛЕ НОМИНАЛИЗАЦИЙ В ОСЕТИНСКОМ ЯЗЫКЕ)

П. Гращенков (pavel.gra@gmail.com)
Институт Востоковедения, Москва, Россия

М. Ионов (max_ionov@mail.ru)

С. Малютина (i-am-stupid@list.ru)
МГУ, Москва, Россия

В настоящей работе мы предлагаем метод лингвистического исследования на полуразмеченном корпусе, предназначенный для изучения грамматической структуры тех языков, где создание полноценного корпуса текстов невозможно. В качестве примера применения нашего метода мы приводим исследование структуры номинализаций в осетинском языке. Грамматическое исследование было осуществлено в три основных этапа. Во-первых, было определено множество интересующих нас грамматических конструкций. Далее, была сформулирована гипотеза о вероятной грамматической структуре данных конструкций. Наконец, выдвинутая гипотеза была апробирована на корпусе текстов. Создание корпуса осуществлялось в два этапа. Во-первых, на основании выборки осетинских текстов была собрана значительная текстовая коллекция. Во-вторых, данный массив текстов был снабжен специальным средством поисковых запросов. В результате, наша исходная гипотеза подтвердилась, что позволило нам уточнить результаты проведенного ранее полевого исследования и сформулировать новые предположения о грамматическом устройстве номинализаций в осетинском языке.

Ключевые слова: номинализации, осетинский язык, полуразмеченный корпус, неполный корпус.

SEMI-TAGGED CORPORA METHOD EXEMPLIFIED WITH A STUDY OF OSSETIC NOMINALIZATION

P. Grashchenkov (pavel.gra@gmail.com)
Institute of Oriental Studies, Moscow, Russian Federation

M. Ionov (max_ionov@mail.ru)

S. Maliutina (i-am-stupid@list.ru)
Lomonosov Moscow State University, Moscow, Russia

We propose the method of Semi-Tagged Corpora (STC) for grammar research in languages that are not expected to have corpora in the nearest future. We exemplify this method with an STC study of internal structure of nominalization

in Ossetic. The research was implemented in three major steps: 1) a set of valid surface structures was established; 2) theoretical predictions were made; 3) the initial hypothesis was tested on the text corpora. The corpora were created in two steps. First we selected a signifi cant amount of texts available for Ossetic and merged them in a single text collection. Then we supplied the collection with specifi c search tools. The initial hypothesis was confi rmed that made our fi eld results more accurate and allows a further elaboration of the syntactic structure that we proposed for Ossetic nominalizations.

Key words: semi-tagged corpora, nominalization, ossetic naguage, Ossetic nominalization.

1. Introduction

Syntactic research is generally conducted via native speakers questioning. However when a speaker doesn't express clear preference for one surface structure in the set of possible structures, the questioning is not satisfactory. For instance, it was claimed in (Chelliah 2001, Brody1982, a. o.) that elicitation approach has quite limited scope and can not be applied to e. g. word order study.

Corpus-oriented researches (see Sinclair 1991 a.o.) were recently implemented on "major" languages like English (Biber et al. 1995) or Chinese (Huang 1994) and gave important output for the grammatical theory. But the enterprise of using corpora and quantitative study of minor or endangered languages seem strange at first. Indeed, languages like Ossetic seem not good candidates for corpus study. First, there are no corpora but only large text collections. Second, there are no electronic dictionaries or ready tag sets for them.

However, rich morphology of Ossetic allows to skip tagging and rely on affixation in corpus research. At the same time, Ossetic possesses a good collection of fiction and paper/magazine articles, sufficient for creation of a large text array.

We supplied our previous field study[1] of Ossetic syntax with a corpora study that favors up one of some initial hypothesis. We used a method of morphology-based search on the untagged corpora. Search results were subsequently filtered and tagged manually. We called this research strategy Semi-Tagged Corpora (STC) study. STC helped us to fill some theoretical gaps in syntactic structure analysis of Ossetic.

2. Linguistic object: Ossetic nominalization

Ossetic is an Iranian language with 0,5 mln of speakers. It has GenN, SOV word order and accusative case system. Cases are marked overtly except nominative and human direct object (unmarked). Non-human direct objects receive marking

[1] The original study of nominalization was provided during the MSU field research trips to Northen Ossetia in 2007–2010. We are very grateful to all our collegues and especially to the chiefs of the expedition, Sergei Tatevosov and Ekaterina Liutikova, for their assistance both in and outside linguistics.

phonologically identical to genitive. There are two nominalization strategies in Ossetic, *-yn* nominalization described here is more regular and productive one.

Two most prominent linguistic problems concerning nominalization in some particular language are the following. First — how many lexical and functional VP material receives nominal distribution. In particular — which arguments are involved in nominalization. Second — how DP structure influences nominalization, i. e. what are the way(s) of marking verbal argument(s), do they receive cases from a verb (accusative) or from nouns (genitive), etc. In Ossetic both these problems are relevant since Ossetic *-yn* forms are homonymous between nominalizations and infinitives. According to native speakers' judgments, both external and internal arguments are valid in the context of nominalization. Moreover, whereas the noun phrase displays strict left branching, the order of arguments in both simple predication and nominalization is quite flexible, see the Table 1. So, direct questioning of native speakers doesn't clarify which arguments (external, internal, both) are present on the argument list and what is the directionality of branching in nominalization.

The hypothesis that may help us to reveal the structure of Ossetic nominalization was reported in (Alexiadou 2004). According to Alexiadou's proposal, nominalizations, even if they allow different distribution and display distinct internal properties, are always merged[2] under the same structure. We can technically elaborate this proposal as follows: the syntactic material merged into enumeration is always the same, and what differs depending on context are phi-features[3] (see Chomsky 1999 and its developments). Differentiation of phi-features is induced by the external context where nominalization is merged. Every particular feature set forces specific internal syntactic configuration, see the Table 2.

Table 1. Constructions with nominalization attested during native speakers' questioning. External and internal arguments accepted under different orderings. Meaning: ***father's sharpening a scythe***[4]

fyd-y	sævæg	daw-yn-
father-GEN	scythe	sharp-*ING*
fyd-y	daw-yn-	sævæg
father-GEN	sharp-*ING*	scythe
daw-yn-	fyd-y	sævæg
sharp-*ING*	father-GEN	scythe
daw-yn-	sævæg	fyd-y
sharp-*ING*	scythe	father-GEN

[2] Merge is an operation that combines two items of the lexicon into a single unit with a label borrowed from one these items.

[3] Informally, phi-features are grammatical categories associated with particular nodes in syntactic structure, functional heads.

[4] In case if the DO is animate the patient will be marked with the genitive (=accusative) in both finite clause and nominalization, *fyd-y fyrt-y wyn-yn-* is a nominalization *father's seeing of his son*.

Table 2. Influence of the distribution of nominalization on its internal structure

1. Enumeration:
{Adjuncts, IntArg, Verb, D, (ExtArg, v,…)}

2. Nominalization merged as DP:
$[_{DP}\ [_{DP}\ Subj_{GEN}]\ \ldots\ XP\ \ldots\ [_{VP}Verb]\ \ldots\ D]$
$*[_{DP}\ [_{DP}\ Subj_{GEN}]\ \ldots\ [_{VP}Verb]\ \ldots$ **XP** … D]

3. Nominalization merged as infinitive:
$[_{DP}\ \ldots\ XP\ \ldots\ [_{VP}Verb]\ \ldots\ YP\ \ldots\ D]$
$*[_{DP}\ [_{DP}$ **Subj**] … $[_{VP}Verb]\ \ldots\ D]$

Then, two most prominent nominalization patterns are nominal and verbal ones. Merged under the postpositions and in noun phrases, nominalizations acquire all properties of noun phrases: they get able to assign genitive case to their subjects and should not exhibit word order permutation. Merged under modals and phase verbs they do not have their own subjects and exhibit word order dependency on the information structure as schematized in the Table 2.

Thus in case of Ossetic *-yn* nominalization we expect to observe the following distributional properties: (i) no difference in the number or marking of arguments in nominal and verbal nominalizations; (ii) differentiation in surface string ordering: strict left branching under the nominal external context and flexible ordering in verbal context. These two statements were chosen for testing by corpora method.

3. Creating corpora

Modern Ossetic has a status of a minor language (<0.5 million of speakers, the absolute majority of which reside on the North Caucasian Mountain) with a well-developed literary tradition. We collected and included into the corpora texts of modern Ossetic newspapers and writers of 20-th century with total volume of 1.3 million words. After that we supplied the text array with the search tools that allow to extract sentences including two words defined in the search query (with the regular expressions option) at some distance also specified in the query.

4. Extracting data and tagging results

Writing search queries, we relied on the rich morphology of Ossetic which made possible to select *-yn* nominalizations and distinguish between the nominal and verbal type of nominalizations.

Different case contexts of nominalizations of all verbs in the selected corpora provide about 20 thousand sentences. We chose eight of the most frequent verbs: *arazyn* ‘make’, *zuryn* ‘say’, *sæwyn* ‘go’, *hwydy kænyn* ‘think’, *maryn* ‘kill’, *særyn* ‘live’,

ahwyr kænyn 'study', *pajda kænyn* 'use'. To distinguish nominal uses from verbal ones we chose genitive forms of nominalizations as instances of the first type and contexts with the verbs 'start/begin', 'want' and 'need' as examples of the second type of representation, see examples in the Table 3. All such instances of the selected eight verbs provide about seven hundreds of contexts.

Then all these contexts were translated and tagged with respect to following properties: presence of subject, presence of direct object, directionality of branching of internal material (obliques and adjuncts considered as well).

Table 3. Corpora examples of nominal and infinitival contexts

1. Nominal:				
…iron	ævzag	ahwyr kænyn-y	raydayæn	etap…
Ossetic	language	study-*ING*	beginning	stage
the first stage of studying Ossetic				
2. Infinitival:				
…raidydta	ahwyr kænyn	matematikon	naukæ-tæ…	
he-started	study-*ING*	mathematical	science-PL	
he began studying mathematical sciences				

5. Evaluating results

The number of nominal contexts consists of 355 and verbal contexts — of 313 examples, 668 instances in total.

Concerning subjects, there were only 7 examples, all of them used in nominal contexts, see Figure 1. Paired t-test performed on the amount of subjects of each verb in nominal vs. verbal context revealed no significant difference between nominal and verbal contexts ($t(7) = 1{,}80$, $p > 0.1$).

Direct objects are met 291 times. Nominal contexts have 163 examples that represent 79% of 206 items of nominalizations of transitive verbs. Direct objects in verbal contexts are met 128 times which is 73% of 128 items, see Figure 2. Again, paired t-test performed on the amount of objects of each transitive verb in nominal vs. verbal context revealed no significant difference between nominal and verbal contexts ($t(5) = 0{,}34$, $p > 0.1$).

No nominalization with both subject and direct object has been attested.

Branching directionality is distributed as follows. Left branching is met in 98% of nominal and 65% of verbal contexts, Figure 3. Yates-corrected chi-square test revealed a significant difference between nominal and verbal context in the amount of examples with left vs. right branching ($p < 0.001$).

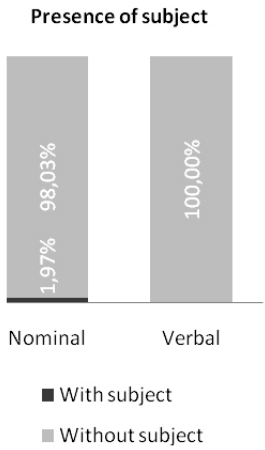


Figure 1. Subject DPs attested in nominal and infinitival contexts

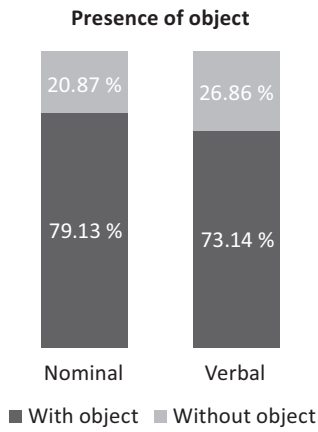


Figure 2. Direct object DPs in nominal and infinitival contexts

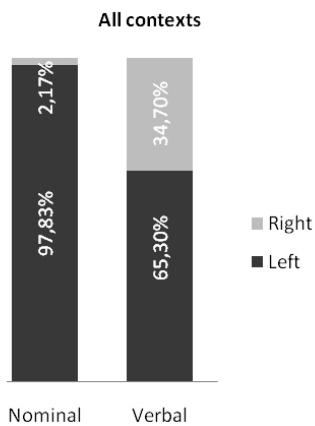


Figure 3. Word order directionality in nominal and infinitival contexts

6. Interpreting results

From the point of view of argument structure, two important observations can be done.

First, we can argue that both nominal and infinitival nominalizations lack subjects on their argument list. Attested 7 cases of subjects as well as artificial subject examples in the Table 1 should be addressed to as pragmatically introduced participants, not true arguments, cf. the traditional treatment of oblique agents. Genetive case can be assigned to such non-argumental DPs as a dummy case marker (see analysis

in Chomsky 1986 for English *of*). Verbs (both transitive and intransitive) other than those that we take for our study also exhibit less than 2% frequency of nominalized subjects. Based on such a low frequency, we argue that Ossetic nominalizations do not really have subject on their argument list.

Second, direct objects are equally frequent in nominal and infinitival nominalizations.

These two observations clearly show that the argument structure in both types of nominalizations is the same.

Concerning word order directionality, nominal contexts do not display any permutations — they are strictly left branching. At the same time more than one third of infinitival nominalizations display right branching. The explanation here is that nominalizations in nominal contexts (as well as in regular DPs) do not allow pragmatically driven scrambling. Infinitival nominalizations and simple clauses are not restricted in this option.

Thus branching directionality depends on phi-features "supplied" by external context, whereas other items that constitute nominalized structures are the same in different instances of nominalizations, see the Table 2. We can further speculate that only nominal phi-features create a phase, opaque for external syntactic processes but this statement comes beyond the scope of current research.

7. Conclusion

As we showed basing on our STG study of Ossetic, nominalizations in this language do not project external arguments. Their argument structure can include only direct objects (that may be marked genitive or nominative). Then, the internal structure of nominalization is a function of the context where it was used.

These results, that seem us quite interesting from the theoretical point of view, could hardly be achieved without quantitative corpora-based investigation of syntactic structure. And corpora creation for languages like Ossetic looks much more realistic under STC-methodology.

References

1. *Abney P. S.* 1987. The English Noun Phrase in its Sentential Aspect. Ph. D. Dissertation.
2. *Alexiadou Artemis.* 2004. Argument Structure in Nominals.
3. *Biber D., Johansson S., Leech G., Conrad S., Finegan E.* 1995. Longman Grammar of Spoken and Written English.
4. *Brody Jill.* 1982. Some Problems With the Concept of Basic Word Order. Linguistics, 22: 711–36.
5. *Chomskii, N.* 1999. Derivation by Phase. MIT Occational Papers in Linguistics. 18.

6. *Chomskii, N.* 1986. Knowledge of Language: Its Nature, Origin, and Use.
7. *Sinclair J.* 1991. Corpus Concordance and Collocation (Describing English Language).
8. *Chelliah Shobhana L.* 2001. The Role of Text Collection and Elicitation in Linguistic Fieldwork. Linguistic Fieldwork : 152–165.
9. *Huang Chu-Ren*. 1994. Corpus-based Study of Chinese: Preliminary Results. In Honor of William SY. Wang: Interdisciplinary Studies on Language and Language Change.

О МУЛЬТИМОДАЛЬНЫХ КЛАСТЕРАХ В УСТНОЙ РЕЧИ

Е. А. Гришина (rudi2007@yandex.ru)

Институт русского языка РАН, Москва, Россия

В статье вводится понятие мультимодального кластера (ММК), под которым понимается многокомпонентная единица устной речи, включающая в себя пары «смысл + жест», «смысл + фонетическое явление» (двойные ММК) или тройку «смысл + жест + фонетическое явление» (тройной ММК). Все компоненты ММК синхронизированы в речи, при этом фонетический и жестовый компонент ММК доступными им средствами передают связанный с ними смысловой компонент. Иными словами, ММК — это тесно связанное в речи сочетание языковых явлений разных модусов (собственно языкового, зрительного, звукового), которые в данной языковой культуре используются для стандартной передачи одного и того же смысла. В статье описываются разные случаи использования двойных и тройных ММК, характерных для устной русской речи.

Ключевые слова: кластер, мультимодальный кластер, устная речь, смысл, жест, фонетическое явление.

MULTIMODAL CLUSTERS IN SPOKEN RUSSIAN

E. A. Grishina (rudi2007@yandex.ru)

Russian Language Institute, Russian Academy of Sciences, Moscow, Russia

The paper introduces the notion of multimodal cluster (MMC). MMC is a multicomponent spoken unit, which includes diads “meaning + gesture”, “meaning + phonetic phenomenon” (double MMC) or triad “meaning + gesture + phonetic phenomenon” (triple MMC). All components of the same MMC are synchronized in the speech, gestural and phonetic components conveying the same idea as the semantic component (naturally, with available means). To put it another way, MMC is a combination of speech phenomena of different modi (semantic, visual, sound), which are closely connected in the spoken language, and roughly speaking mean the same, i. e. convey the same idea by their own means. The paper describes some examples of double and triple MMCs specifi c for the Spoken Russian.

Key words: cluster, multimodal cluster, spoken Russian, meaning, gesturem phonetic phenomenon.

1. Введение

Принципиальная ***мультимодальность*** устной речи, т. е. чрезвычайная важность для ее адекватного функционирования (и для ее понимания) разных каналов (модусов) информации, прежде всего, звукового и зрительного, привлекает в последние годы внимание все большего числа исследователей. Особенно активно эта идея развивается в рамках анализа повседневной жестикуляции, сопровождающей речепорождение. Ни в коей мере не претендуя на полноту, можно упомянуть работы А. Ченки, К. Мюллер, И. Миттельберг[1], в которых подробно анализируются формы взаимодействия речи и жестикуляции, координация или, напротив, рассогласованность жестикуляционного и собственно языкового рядов в устном высказывании[2].

В ходе создания устного подкорпуса Национального корпуса русского языка, и особенно Мультимедийного русского корпуса (МУРКО), нам приходилось сталкиваться с ситуациями, когда в устном высказывании в одних и тех же *смысловых* (в широком смысле) точках мы находили сочетание одних и тех же событий, относящихся к разным модусам речи, — к сопровождающей речь жестикуляции (зрительный ряд), к фонетическим — в широком смысле — особенностям речи (как к собственно фонетике, так и к интонации). Такое постоянство в сочетании столь разнородных признаков показалось нам не случайным и потребовало того или иного объяснения.

Для описания явлений такого рода мы предлагаем использовать термин ***мультимодальный кластер (ММК)***. Мы будем говорить, что в устной речи мы имеем дело с ММК в случае, если одно и то же смысловое событие (семантическое, прагматическое, синтаксическое, стилистическое) сопровождается в устной речи одним и тем же набором жестовых и/или фонетических событий, т. е. ММК — это связка, фасция, кластер разномодусных явлений (зрительных и/или звуковых), которые с частотой выше средней сопровождают одно и то же смысловое явление и образуют разномодусные тройки или пары, выступающие в устной речи как самостоятельные смысловые единицы.

Повышенная по сравнению со средними распределениями частота сочетаний разнородных явлений внутри мультимодального кластера позволяет нам считать совпадение именно такого набора событий неслучайным и ставит

---

* Исследование проведено при поддержке РФФИ, гранты 10-06-00151-а и 11-06-00030-а, и программ РАН РФ «Генезис и взаимодействие социальных, культурных и языковых общностей» и «Корпусная лингвистика».

1 [Cienki 2005], [Cienki&Müller 2008], [Müller 1998], [Mittelberg 2007].

2 Эти работы имеют непосредственный практический интерес, поскольку в рамках создания ECA (embodied computational agents), анимированных «посредников» между компьютером и человеком, в полный рост встала проблема воссоздания в анимационной среде естественной манеры каждодневного человеческого общения, — проблема, как оказалось, не решаемая без учета взаимодействий зрительного и звукового рядов в речи, т. е. без учета мультимодальности.

перед нами задачу объяснить, почему именно такие, а не иные явления входят в данный мультимодальный кластер.

Парные семантико-жестикуляционные ММК уже довольно давно являются предметом обширных и весьма продуктивных исследований — ср., например, работу [Richter 2010], в которой были исследованы жесты, сопровождающие те или иные грамматические явления в русском языке (например, круговое движение рукой для передачи значения итеративности в русских глаголах несовершенного вида), наши работы о вокальных жестах *А* и *О* [Гришина 2009, 2010], где значения этих русских слов, чрезвычайно характерных для устной речи, анализировались с опорой на сопровождающие их жесты. Огромное количество работ посвящено иконическим (изобразительным) жестам, которые сопровождают или заменяют слова определенной семантики.

Парные семантико-фонетические кластеры изучены в меньшей степени (если, конечно, оставить в стороне работы по звуковому символизму, которые во всем, что не касается звукоподражательных слов, требуют от придирчивого читателя слишком большого уровня доверия к авторам). Такая асимметрия связана, очевидно, с тем, что жест является двусторонним знаком, имеющим собственную семантику и внутреннюю форму, а следовательно, согласованность слова и жеста имеет общую территорию для сравнения. Звуковой ряд речи достаточно редко имеет прямой выход на значение, что, конечно, затрудняет интерпретацию фактов. Из интересных работ в этой области можно упомянуть, например, исследование [Кривнова 2007], в котором анализировалось значении вызова у гортанной смычки в русском языке.

Нам не известны работы не по *парным* соответствиям смысл — жест и смысл — звук, а по *тройкам* смысл — жест — звук. А между тем представляется, что такое сочетание может быть более продуктивным и нестандартным, чем попарные сопоставления, не говоря уже о том, что факт постоянного или частотного совпадения в потоке речи двух двусторонних знаков (слова и жеста) с некоторым звуковым явлением может повысить доказательность утверждений, касающихся семантизации фонетических явлений.

Данная работа ни в коей мере не представляет собой исчисления, систематизации или классификации ММК, характерных для устной русской речи. Предполагается дать вполне мозаичное описание отдельных явлений такого рода, встреченных нами к настоящему моменту. Некоторые из них мы уже отмечали ранее, некоторые описываем впервые. Таким образом, основной задачей настоящей работы является привлечение внимания к самому наличию не только двойных, но и тройных ММК в устной речи и приглашение к исследованию устного высказывания с этой точки зрения.

2. Двойной ММК: *щас/ща, вот/во, нет/не*

В работе [Гришина 2008] мы писали о том, что для устной русской речи характерно наличие у ряда частотных неизменяемых слов лексикализованных вариантов, которые, этимологически восходя к исходному слову, получают

и свои собственные, самостоятельные значения, отличные от значений базовых слов, и тем самым превращаются, в конечном итоге, в самостоятельные слова[3].

Интересно, что для трех частотных в русской устной речи слов — *щас* (регулярный стяженный вариант от *сейчас*), *вот* и *нет* (отрицательный ответ) — характерно наличие лексикализованных вариантов, которые образуются по одной и той же схеме: отпадает последняя согласная закрытого слога, при том, что каждое слово состоит ровно из одного этого закрытого слога: *щас* → *ща, вот* → *во, нет* → *не*. Мы считаем, что эта трансформация не случайна и представляет собой вполне закономерное событие.

Каждый из этих вариантов имеет свои собственные отношения с исходным словом. Так, *вот* и *во* достаточно сильно разошлись по значению и сфере функционирования (описанию их различий, собственно, и посвящена цитированная выше статья). *Щас* и *ща*, по-видимому, семантически идентичны. У *не*, как кажется, есть предпочитаемые контексты употребления, в частности, употребление варианта *не* в многократных (более чем трехкратных) повторах несколько отличается от употребления *нет*: *не* тяготеет к цельнооформленному повтору (связанному единым ударением), в то время как *нет* предпочитает иметь самостоятельное ударение на каждом повторяемом элементе.

Однако все эти три усеченных варианта объединяет одна особенность, а именно: все они относятся к неформальному, если не к просторечному регистру употребления языка, т. е. к тем сферам функционирования устной речи, где *ослаблен* контроль говорящего над соответствием его речи общепринятым официальным стандартам. На фоне этой общей особенности направление трансформации в вариантах по сравнению с исходным словом кажется уже неслучайным: при отсечении последней согласной закрытый слог превращается в открытый. Если иметь в виду, что 1) произнесение шумных согласных предполагает преодоление той или иной преграды в артикуляционном аппарате, а следовательно, предполагает большее напряжение, чем произнесение гласных, и 2) при произнесении согласных конфигурация артикуляционного аппарата в целом сложнее, чем при артикуляции гласных, то направление изменений в этих трех вариантах происходит по одной и той же схеме: напряжение + расслабление + напряжение → напряжение + расслабление. Тем самым, усеченные варианты предполагают завершение фонации на этапе более расслабленного состояния артикуляционного аппарата и меньшего контроля над ним.

Таким образом, в этом парном мультимодальном кластере мы имеем дело с соответствием фонетического ряда (расслабленность и пониженный контроль над артикуляцией) и стилистического ряда[4] (расслабленность и пониженный контроль над соответствием официальным нормам, характерные для просторечия).

---

[3] Такое направление развития характерно, в частности, для регулярных апокоп. Так, для некоторых контекстов характерно предпочтение апокопированных, а не полных вариантов, например, фраза *Прям, спешу и падаю!* звучит более естественно, чем *?Прямо, спешу и падаю!*, а проклятия *Чтоб ты сдох!, Чтоб ты жил на одну зарплату!*, по-видимому, в принципе могут существовать только в этом виде, т. е. варианты **Чтобы ты сдох!* **Чтобы ты жил на одну зарплату!*, скорее всего, просто невозможны.

[4] Если понимать смысловой ряд устного высказывания максимально широко, то ничто не может помешать нам трактовать стилистические характеристики как законные

3. Тройные ММК

3.1. Указательная частица *О*

В работе [Гришина 2009] мы анализировали разные значения вокального жеста *О* на основе сопровождающих этот вокальный жест телесных жестов. Было показано, что для одного из значений данного вокального жеста — указательной частицы *О* — ***доминантным***, т. е. наиболее частым сопровождающим жестом является указание пальцем. Этот указательный жест осуществляется в двух основных вариациях: 1) указательный палец, поднятый вверх, — жест, означающий, что в ходе беседы было озвучено (реже — собирается прозвучать) нечто важное с точки зрения говорящего; 2) указательный палец, направленный в сторону собеседника, — жест, означающий, что слушающий только что сделал или сказал что-то важное с точки зрения говорящего.

Одновременно было отмечено, что указательная частица *О* часто (примерно в 82% случаев) произносится с твердым приступом в начале, что предполагает относительно краткое (например, по сравнению с междометием *О*, выражающим удивление), *точечное* произнесение этой частицы.

В работе [Гришина 2008] было подробно проанализировано значение указательной частицы *О* по сравнению с *вот* и *во*. Нам представляется достаточно обоснованным полученный вывод, согласно которому основным значением указательного *О* является, собственно, не указание, а *фиксация* объекта, т. е. актуализация, констатация наличия того или иного объекта (вещественного или абстрактного) в окружающей действительности, в коммуникативном или тематическом пространстве диалога в момент использования частицы *О*. С помощью частицы *О* говорящий лишь фиксирует наличие существенного для себя объекта, а не указывает на него собеседнику (в отличие от *во* и в особенности *вот*, для которых существен именно момент указания): указательная частица *О* функционирует прежде всего в интересах говорящего, а *во* и *вот* — в интересах слушающего.

Представляется, что эти три компонента в осуществлении частицы *О* — семантический, жестовый и фонетический — представляют собой тройной ММК. Указательный жест пальцем, а не рукой действует как аналог иглы или любого иного острого предмета, который как бы пригвождает объект, фиксирует его в пространстве диалога. Твердый приступ и связанное с ним краткое звучание данной частицы иконически с помощью фонетических средств отображает **точку** фиксации. Значение же частицы сводится к фиксации, «обездвижению» некоторого объекта, который без этой фиксации прошел бы незамеченным мимо внимания аудитории.

компоненты смысла высказывания.

3.2. Итоговое *Да*

В работе [Гришина 2011а] нами были подробно рассмотрены разные случаи употребления ударного *да* в русском устном диалоге. Одним из значений *да* является т. н. итоговое *да*, с помощью которого *да*-говорящий подводит своего рода черту под своими размышлениями относительно какого-либо предмета.

Да в итоговом значении практически в обязательном порядке сопровождается специфическими фонетическими явлениями. Чаще всего это носовой призвук в начале произнесения, а также растяжка гласной. Эти фонетические признаки настолько специфичны и частотны, что даже нашли своей отражение на письме: итоговое *да* чаще всего обозначается как *нда* или *мда*, а также *да-а* или *да…* (многоточие стандартно обозначает растяжку гласной). Несколько реже встречается задержка рекурсии на [д].

С другой стороны, как было показано в работе [Гришина 2011б], посвященной исследованию закономерностей в изменении направления взгляда в устном диалоге, для итогового *да* характерно специальное поведение взгляда *да*-говорящего, отличающееся от поведения взгляда на *да* в иных значениях. Здесь для дальнейшего изложения существенны два параметра: 1) различение референтного (фиксированного) или нереферентного (расфокусированного) взгляда — говорящий фиксирует референтный взгляд на конкретном объекте (на любом физическом объекте или на собеседнике), нереферентный же взгляд свободен от конкретной фиксации, объект внимания в случае расфокусированного взгляда определить не удается; 2) в случае нереферентного взгляда — различение его направлений (вниз, перед собой, вдаль, в сторону). В результате самого общего обследования материала (более 400 контекстов употребления *да*) становятся очевидны две закономерности. Во-первых, для итогового *да* характерен расфокусированный (нереферентный) взгляд (72% против 27% в среднем) и нехарактерен фиксированный (референтный) взгляд (28% против 73% в среднем, см. Табл. 1); для остальных значений *да* соотношение обратное. Во-вторых, если рассматривать такой специфический взгляд, как *взгляд вдаль*, т. е. взгляд за пределы не только коммуникативного пространства, но и вообще за пределы сопредельной данному диалогу территории, то мы увидим, что *взгляд вдаль* в высшей степени характерен для итогового *да* (19% против 6% в среднем) и практически не свойствен остальным значениям *да*.

Таблица 1

Значения *да* Тип взгляда	**Итоговое *да***	**Остальные значения *да***	**Всего**
Нереферентный взгляд	**72%**	*18%*	27%
Фиксированный взгляд	*28%*	82%	73%
Взгляд вдаль	**19%**	*3%*	6%

Одновременно нужно упомянуть, что итоговое *да* сопровождается двумя разными типами жестикуляции. Первый, вполне ожидаемый в данном

случае, — это жесты, стандартно сопровождающие акт мысли, ситуацию сосредоточенного размышления: *взяться за подбородок, почесать затылок, сложить руки за спиной, чертить пальцем, нахмуриться*. Эти жесты естественно соотносятся с характерным для итогового *да* нереферентным взглядом: в ситуации размышления человек отгораживается от рассеивающих его внимание внешних объектов, в т. ч. и от собеседника, и старается не смотреть на них.

Второй тип жестов, объяснение которых уже не столь очевидно, — это жесты отстранения, дистанцирования. При этом дистанцирование может относиться как к собеседнику или некоторому внешнему объекту, так и к смысловой, тематической зоне высказывания. Это жесты *поднять брови, повести подбородком вбок, развести руками, отклониться назад, прищуриться* (т. е. трактовать предмет раздумий как находящийся далеко и, следовательно, плохо видный). С этим типом жестов естественно соотносится *взгляд вдаль*, который также помещает объект рассмотрения далеко, оценивая его как находящийся не только вне коммуникативного пространства, но и далеко за пределами сопредельной территории.

Представляется, что все описанные выше особенности итогового *да* хорошо укладываются в два ММК.

Первый кластер — двойной, семантико-жестикуляционный: *да* как итог (или, точнее, аккомпанемент) размышлений + жесты сосредоточенности, раздумий, а также нереферентный взгляд, отрешающий говорящего от объектов, способных помешать его сосредоточенности.

Второй кластер — тройной, добавляющий к семантике и жесту фонетическую составляющую. В этом кластере итоговое *да* является результатом размышлений, которые оценивают предмет размышлений как целое: очевидно, что именно охват ситуации в целом, во всех ее связях и аспектах, позволяет сделать правильные выводы. Но для того, чтобы появилась возможность оценить некоторый объект как целое, он должен располагаться от размышляющего субъекта на достаточном расстоянии, физическом или смысловом, и попадать, тем самым, целиком в «сектор осмотра». Следовательно, именно этот аспект семантики *да* естественно сопровождается жестами отстранения и дистанцирования от объекта размышлений, а также взглядом на него издалека (*взгляд вдаль* или *прищуриться*). И именно такая трактовка смысла итогового *да* объясняет его фонетические особенности (повторим — практически стопроцентно обязательные), а именно, — носовые призвуки и удлинение фонации (растяжка гласной или задержка рекурсии согласной): каждая из этих фонетических особенностей удлиняет итоговое *да*, оттягивает момент его завершения, что на фонетическом уровне дублирует отстранение говорящего от предмета его размышлений. Говорящий в этом случае словно делает фонетический «шаг назад» для улучшения осмотра.

3.3. Передразнивание

Цитирование, апелляция к чужому слову в тех или иных формах не просто часто встречается в устной речи, но составляет, по-видимому, ее обязательный

компонент, и может даже оцениваться как форма ее существования. Это является, очевидно, результатом непосредственного контакта в ходе устного диа- или полилога говорящего и слушающих, контакта, который требует от говорящего постоянного учета чужого слова, прежде всего, слова слушающего, при построении высказывания. В философском и общефилологическом аспектах тема чужого слова хорошо разработана исследователями, понятийный и операционный аппарат которых восходит к работам М. М. Бахтина (см., например, [Гоготишвили 2006]). Мы ни в коей мере не намерены рассматривать эту проблематику с теоретических позиций: в наши задачи входит анализ одного специфического случая реакции говорящего на высказывание слушающего, а именно, — анализ такой разновидности неодобрительного цитирования, как передразнивание. В этом случае, как представляется, использование понятия ММК будет достаточно продуктивным.

Анализ материала МУРКО показал, что речевые акты, содержащие передразнивание, регулярно включают в себя следующие компоненты.

1. **Повтор цитируемого фрагмента.** В наиболее стандартном случае повтор заключается не просто в однократном воспроизведении некоторого участка цитируемого высказывания, что естественно для цитирования в целом, а в двойном (чаще всего) или в многократном повторе этого участка:

(1) — *Ну кто ж за тебя будет бином Ньютона учить? — Опять за свое!* ***Бином-бином / бином…*** *Ну скока можно / ну?* (Друг мой, Колька!, 1961)

(2) *О! Как родная вошла! А ты /* ***не войдет / не войдет!*** (Операция «С Новым Годом!», 1996)

Лексическая редупликация при передразнивании сопровождается своеобразным интонационным движением, когда вторая часть редупликации повторяет интонационный контур первой части, но на более низком уровне тона[5]:

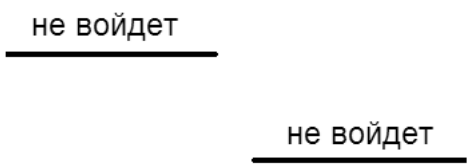


Рис. 1

Повтор, однако, может выглядеть и по-другому, а именно, — говорящий может повторять лексически не совпадающие между собой участки чужого высказывания, но при этом произносить их с одним и тем же интонационным контуром, расположенным на одном и том же уровне тона (этот тип повтора можно назвать ***интонационным*** — он используется говорящим, когда тот

[5] Для анализа интонационного движения на лексических мультипликациях данных пока недостаточно.

хочет уничижительно процитировать не одну, а две или больше зон цитируемого высказывания):

(3) *Новиков наговорил вам тут… **Хулиганить / избивать активных пионеров…** А кого мы избивали?* (Друг мой, Колька!, 1961)

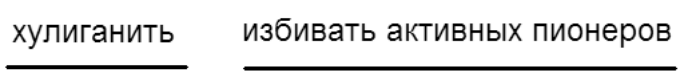


Рис. 2

2. Передразнивание очень часто сопровождается **ненатуральным** для данного говорящего **уровнем тона**, причем это не локальный, точечный выброс за пределы нормы, а достаточно длительное ненатуральное звучание. Обычно голос поднимается выше стандартного уровня, но встречаются случаи и понижения тона:

(4) *Я его сегодня встретил. Говорит / **как там наш Лёва? Передайте ему пожалуйста / я хочу чтоб он в моем спектакле Ленина сыграл.*** (Операция «С Новым Годом!», 1996)

(5) *— Какие деньги? — **Какие деньги…** Обыкновенные!* (Друг мой, Колька!, 1961)

3. Для передразнивания характерны **жесты *скривиться*** и ***трясти головой***.

(6) *— А чего ж она? — Чего… Мамку вызовут в школу / будешь знать /* [кривится] ***чего***. (Друг мой, Колька!, 1961)

(7) *— Но тот тоже сказал пароль «черт побери»!* — [кривится, трясет головой] ***Черт побери, черт побери…*** (Бриллиантовая рука, 1968)

Все эти три ряда явлений могут быть использованы для передачи идеи передразнивания как поодиночке, так и все вместе в одном высказывании. С нашей точки зрения, они представляют собой ММК, т. е. их сочетание для передачи речевой ситуации передразнивания не случайно, а закономерно, поскольку каждое из этих явлений отражает сущностные свойства передразнивания, и, кроме того, связано с другим явлением, относящимся к иному модусу.

1. **Повтор + жест *трясти головой*.** Естественно, первый же вопрос, который возникает при анализе материала, содержащего речевое действие передразнивания, — почему передразнивание так тесно связано с повтором?

Для ответа на этот вопрос следует сформулировать, чем, собственно, является передразнивание? Передразнивание — это неодобрительное цитирование чужой речи с одновременным понижением этой чужой речи в ранге, в статусе.

Каким образом можно понизить статус чужой речи?

Прежде всего, можно подчеркнуть, что чужая речь бессодержательна. Бессодержательность речи выражается в том, что слов говорится слишком много, а реально передаваемой новой информации в этой речи слишком мало по сравнению с количеством слов. Следовательно, если какое-либо высказывание оценивается как многословное, то тем самым его статус понижается. Повтор чужой реплики или ее части (лексический или интонационный — все равно) как раз и символизирует, что цитируемое лицо произносит слишком много слов и при этом передает слишком мало полезной информации. Таким образом, цитация с повтором автоматически понижает ранг цитируемого высказывания, трактуя его как многословное и, соответственно, бессодержательное.

Почему речевой акт передразнивания часто сопровождается трясением головой? По всей вероятности, потому, что трясение головой имитирует вынужденные движения головы человека в момент говорения, и, следовательно, в момент произнесения цитируемой реплики. Таким образом, трясение головой передает ту же самую идею, что и повтор, но на уровне жестов: цитирующий пародирует физический акт произнесения цитируемой реплики, многократно и утрированно двигая головой (чаще всего — подбородком из стороны в сторону). Таким образом, на уровне жеста проводится та же идея, которая передается лексическим или интонационным повтором: цитируемый автор говорит слишком много и при этом передает слишком мало информации, следовательно, его речь бессодержательна, следовательно, ее можно не принимать во внимание ввиду ее низкого статуса.

2. **Ненатурный уровень тона + жест *скривиться***. Вторым способом понизить ранг источника цитаты является утверждение о том, что цитируемая речь неверна (т. е. сказанное не соответствует действительности) или неуместна (сказанное не соответствует конситуации речи). Несоответствие речи действительному положению вещей означает, что эта речь отражает действительность в испорченном, изуродованном виде, выступая по отношению к ней в качестве кривого зеркала. Именно эта идея порчи реального положения дел посредством неверного высказывания о нем и передается «уродованием», «порчей» цитируемой речи с помощью ненатурального уровня тона и жеста *скривиться*. Таким образом, цитируемое лицо произносит свое высказывание естественно, считая, что оно органично соответствует передаваемому содержанию или ситуации. Передразнивающий же субъект цитирует чужое высказывание, но одновременно портит его звучание (ненатуральный уровень тона) или дефектно отображает лицо цитируемого в момент произнесения им реплики (жест *скривиться*), воплощая на зрительном уровне идею несоответствия цитируемого высказывания реальному положению дел или ситуации и, соответственно, его низкий статус.

Следует отметить, что передразнивание — социологически маркированный речевой акт, т. е. он возможен только в том случае, 1) если передразнивающий субъект находится на одном уровне иерархии с передразниваемым или выше его, 2) если социальная ситуация общения воспринимается как неофициальная. В случае, если эти условия не выполняются, мы сталкиваемся

не с передразниваем, а с неодобрительным цитированием. При неодобрительном цитировании не используются жестовые и интонационные составляющие кластера передразнивания, т. е. неодобрительное цитирование сводится лишь к лексическому или интонационному повтору:

(8) [Директор на педсовете] *Да… Как это просто у нас получается.* ***Исключить… Поставить вопрос…*** (Друг мой, Колька!, 1961)

(9) [Отрядный вожатый] — *Какое общество?* — [Старшая пионервожатая] ***Какое общество… какое общество…*** *Вот вы разберитесь / а потом судите.* (Друг мой, Колька!, 1961)

Таким образом, повтор является самым нейтральным и социально приемлемым способом понизить ранг цитируемой речи. Остальные же составляющие рассматриваемого кластера (трясение головой, скривившееся лицо, ненатуральность уровня тона) переводят неодобрительное цитирование в разряд передразнивания и понижают статус цитируемой реплики существенно сильнее[6].

4. Вместо заключения: иконичность и компенсация

Во всех рассмотренных в предыдущем изложении ММК отношение, которое связывает компоненты кластеров между собой, можно охарактеризовать как иконичность, т. е. повтор, имитация одного и того же значения средствами, относящимися к разным модусам устного высказывания:

1. расслабленность произнесения и отсутствие контроля за соответствием речи официальным нормам, характерные для просторечия, имитируются расслабленностью произнесения открытого слога в *ща*, *во* и *не* по сравнению с литературными *щас, вот* и *нет*, представляющими собой закрытые слоги;
2. точка в конситуации общения, которая фиксируется говорящим как важная, имитируется указательным жестом, который осуществляется пальцем (а не рукой, к примеру), поскольку кончик указательного пальца на жестовом уровне отображает фиксируемую смысловую точку; на фонетическом уровне эта же точка отображается «точечным» звучанием указательной частицы *О*, которое обеспечивается твердым приступом в начале ее произнесения;

[6] Следует отметить, что, по нашим данным, жестовые и фонетические составляющие мультимодального кластера передразнивания имеют гендерные ограничения: женщинам не свойственны ни жесты *скривиться* и *трясти головой*, ни ненатуральный тон речи. По-видимому, это связано с наличием внутренних запретов на уродование своего голоса и внешнего вида, характерных для женщин. Таким образом, женщинам доступно передразнивание в основном только в форме повторов. Впрочем, для уверенных утверждений материала пока недостаточно.

3. сосредоточенность говорящего на итоге своих размышлений, характерная для высказываний, содержащих итоговое *да*, на жестовом уровне имитируется жестами сосредоточенности, а на уровне грамматики взгляда — выходом *да*-говорящего из коммуникативного пространства и нереферентным (несфокусированным) взглядом, которые позволяют говорящему не отвлекаться на несущественные для его размышлений объекты (прежде всего — на слушающего) и полностью сосредоточиться на подведении итогов; стремление рассмотреть предмет размышлений в целом, как бы издалека, чтобы подвести правильный итог, имитируется на жестовом уровне взглядом вдаль, далеко за пределы коммуникативного пространства, и прищуром, функционирующим аналогичным образом; на фонетическом же уровне итоговое *да* характеризуется растянутым звучанием, и удлинение *времени* звучания имитирует увеличение расстояния (*пространства*) до предмета размышлений;
4. многословие и бессодержательность цитируемого высказывания имитируются лексическим или интонационным повтором, а на жестовом уровне — трясением головы, пародирующим вынужденные движения головы говорящего в момент говорения; ложность или неуместность цитируемой речи на интонационном уровне передается ненатуральным уровнем тона, а на жестовом уровне — гримасой; все эти элементы, организованные в ММК, составляют стандартный речевой акт передразнивания.

Рассмотрим еще один пример. Он касается взаимоотношений взгляда и эмфазы внутри фразы[7]. Далее мы будем различать *эмфазу* (фонетически, интонационно и акцентологически выделенное фонетическое слово), *фон* (одно или несколько идущих подряд фонетических слов, не являющихся эмфазой), *стандартную зону темы* (начало фразы, т. е. первое фонетическое слово), *стандартную зону ремы* (конец фразы, последнее фонетическое слово), *независимую* позицию для эмфазы или фона (часть фразы, не являющуюся ее началом или концом).

Анализ достаточно обширного материала показывает, что в независимой позиции поведение взгляда на эмфазе и фоне достаточно предсказуемо[8]:

Таблица 2

Положение взгляда	Независимая эмфаза	Независимый фон
Внутри зоны коммуникации	+	0
Вне зоны коммуникации	–	0

[7] Подчеркнем, что мы рассматриваем взгляд и эмфазу именно внутри фразы, а не внутри реплики: на границах реплики взгляд, по-видимому, зависит от наличия или отсутствия эмфазы в существенно меньшей степени, чем на границах фразы, и управляется иными закономерностями, связанными с очередностью реплик (см. [Гришина 2011б]).

[8] Далее + означает отклонение от средних распределений в бóльшую сторону, – (минус) — отклонение в меньшую сторону, 0 — незначимость параметра, т. е. свободное глазное поведение говорящего, не связанное с наличием/отсутствием эмфазы; положение взгляда внутри зоны коммуникации = взгляд на собеседника, вне зоны коммуникации = взгляд не на собеседника.

Как видим, на независимом фоне, т. е. на той части фразы, которая не совпадает с ее границами и не подпадает под эмфазу, взгляд говорящего свободен находиться как внутри, так и вне коммуникативного пространства. Что касается независимой эмфазы, то она заметно чаще среднего сопровождается взглядом говорящего на собеседника. Очевидно, что в случае независимой эмфазы мы имеем дело с двойным ММК: то фонетическое слово, которому говорящий придает особое значение и, соответственно, выделяет его фонетически, интонационно и акцентологически, сопровождается взглядом говорящего на собеседника. Этот взгляд так же привлекает внимание слушающего к выделенному слову, как и звуковое его подчеркивание. Таким образом, взгляд в данном случае имитирует, иконически отображает на телесном уровне фонетический аспект эмфазы.

Если экстраполировать данные Табл. 2 на границы фразы, то на границах фразы мы должны ожидать следующих распределений (см. Табл. 3):

Таблица 3

Положение взгляда	**Эмфаза**		**Фон**	
	Начало фразы	Конец фразы	Начало фразы	Конец фразы
Внутри зоны коммуникации	+	+	0	0
Вне зоны коммуникации	–	–	0	0

Реальные же распределения отличаются от ожидаемых (см. Табл. 4):

Таблица 4

Положение взгляда	**Эмфаза**		**Фон**	
	Начало фразы	Конец фразы	Начало фразы	Конец фразы
Внутри зоны коммуникации	–	0	0	+
Вне зоны коммуникации	+	0	0	–
Номер столбца	1	2	3	4

Как видим, три столбца Табл. 4 (из четырех возможных) показывают нам отклонение от ожидаемого поведения взгляда.

Значит ли это, что ММК взгляд + эмфаза, который мы наблюдали на независимой эмфазе внутри фразы, распадается и перестает функционировать на ее границах? Это, по-видимому, одно из возможных решений, но нам показалось трудным, почти невозможным найти обоснование для такого внезапного разрушения кластера на границах фразы. Кроме того, простая констатация распадения кластера не дает нам объяснения реальному поведению взгляда на границах фразы.

Приведенные данные, однако, могут быть объяснены иначе: внутри мультимодальных кластеров отношения, которые могут связывать между собой компоненты кластера, относящиеся к разным модусам, не исчерпываются прямой иконичностью.

Прежде всего, обратим внимание на то, что когда эмфаза и фон попадают в свои стандартные позиции (эмфаза — в конец фразы, т. е. в стандартную зону ремы, а фон — в начало фразы, т. е. в стандартную зону темы), распределение взгляда между положением внутри и вне зоны коммуникации в количественном отношении не отличаются от средних распределений, т. е. взгляд говорящего в этих ситуациях свободен (столбцы 2 и 3 Табл. 4).

В случае же, если эмфаза или фон отмечены в нестандартных для себя положениях (эмфаза — в начале фразы, а фон — в конце, столбцы 1, 4 в Табл. 4), взгляд ведет себя по-разному. Если эмфаза приходится на начало фразы, взгляд предпочитает положение вне зоны коммуникации. Если фон приходится на конец фразы, взгляд предпочитает положение внутри зоны коммуникации. Таким образом, можно сказать, что взгляд компенсирует нестандартное сочетание эмфазы/фона и их позиций во фразе: если в начале фразы наблюдается избыток эмфазы по сравнению с нормой, то положение взгляда вне зоны коммуникации нивелирует этот избыток своим отсутствием в зоне коммуникации; если же в конце фразы наблюдается недостаток эмфазы по сравнению с нормой, то взгляд компенсирует этот недостаток своим нахождением внутри зоны коммуникации.

Это означает, очевидно, что положение взгляда внутри зоны коммуникации и эмфаза — явления одного плана, хотя и принадлежащие к разным модусам устного высказывания (что мы и наблюдали в случае независимой эмфазы). Однако на границах фразы, в случае нестандартной с точки зрения коммуникативного членения предложения конфигурации, прямая иконичность разномодусных явлений заменяется их компенсационным поведением — взгляд компенсирует избыток и недостаток эмфазы по сравнению с нормой, т. е. эмфаза и взгляд ведут себя, как жидкость в сообщающихся сосудах, участвуя в игре с нулевой суммой. Очевидно, что такое компенсационное поведение возможно только в том случае, если два явления связаны иконически, т. е. выполняют одну и ту же задачу, находясь в разных модусах устной речи, однако иконичность в этом случае претерпевает достаточно необычную трансформацию.

В заключение следует отметить, что в рамках настоящей работы остался без ответа вопрос о том, имеет ли фонетический компонент ММК самостоятельное значение, или возможность семантизировать его появляется только внутри данного кластера, с опорой на такие двусторонние знаки, как лексика/синтаксис и жесты. Для того, чтобы фонетический компонент мог быть расценен как семантически самостоятельный, он должен а) встречаться и б) сохранять одно и то же значение внутри разных ММК. Для каких бы то ни было уверенных утверждений материала пока чрезвычайно мало, вернее, попросту нет. Мы на сегодняшний день склоняемся к мысли, что фонетический компонент внутри кластера семантизуется посредством двусторонних компонентов.

В упомянутой выше работе [Кривнова 2007: 348] было замечено, что «неформальное отрицание типа *«не-а»* [точно так же, как неформальное согласие *да*ˀ — *Е. Г.*] произнесенное с ларингализацией, может обладать оттенком вызова, противопоставления мнения говорящего мнению собеседника, размежевания и нарочитого разграничения и рассогласования их позиций». С другой стороны, на нашем материале было замечено, что такая же ларингализация может быть использована как иконическое отображение свертывания фразы до начальной частицы, типа: *Да какая разница / наплевать! → Да / наплевать! → Даˀ / наплевать!, Ну что тут скажешь / тогда другое дело! → Ну / тогда другое дело! → Нуˀ / тогда другое дело!* (см. [Гришина 2010]; иконичность состоит в том, что внезапно заканчивается, будучи свернутой, фраза, которая должна была бы продолжаться дальше, и гортанная смычка концентрирует в себе всю неизрасходованную энергию несказанного, осуществляя что-то типа скрипа тормозов при экстренном торможении). Как видим, в разных контекстах одно и то же фонетическое явление может иметь разные значения, очевидно не сводимые друг к другу. Однако для более уверенных утверждений материала совершенно недостаточно.

References

1. *Cienki.* 2005. Image Schemas and Gesture. From Perception to Meaning: Image Schemas in Cognitive Linguistics, 29 : 421–442
2. *Cienki A., Müller C.* 2008. Metaphor, Gesture, and Thought. The Cambridge Handbook of Metaphor and Thought : 483–501
3. *Gogotishvili L. A.* 2006. Indirect Speaking [Nepriamoe Govorenie].
4. *Grishina E. A.* 2008. The Particle 'VOT': Variants Used in Spontaneous Speech [CHastitsa 'VOT': Varianty, Ispol'zuemye v Neprinuzhdennoi Rechi]. Instrumentarii Rusistiki: Korpusnye Podkhody (Slavica Helsingiensia 34) : 63–91 (available at: http://docs.google.com/View?id=df52fjjj_9fntn6kxq)
5. *Grishina E. A.* 2009. On the Word and Gesture Correlation Problem (Vocal Gesture in the Oral Speech) [K Voprosu o Cootnoshenii Slova I Zhesta (Vokal'nyi Zhest v Ustnoi Rechi)]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2009" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2009"), 8 (15) : 80–90 (available at: http://www.dialog-21.ru/dialog2009/materials/html/14.htm)
6. *Grishina E. A.* 2010. Vocal Gesture in the Oral Speech [Vokal'nyi Zhest v Ustnoi Rechi]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2010" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2010") : 102–112, available at: http://www.dialog-21.ru/dialog2010/materials/html/17.htm
7. *Grishina E. A.* 2011. 'DA' in Russian Oral Dialogue ['DA' v Russkom Ustnom Dialoge]. Russian Linguistics, 2(35).

8. *Grishina E. A.* 2011. Grammatics of Look (Look Direction as a Linguistic Factor) [Grammatika Vzgliada (Napravlenie Vzgliada kak Lingvisticheskii Faktor)]. Filologiia, 1(01) : 50–71.
9. *Krivnova O. F.* 2007. Phemonen of Laryngalization in Russian Speech [Iavlenie Laringolizatsii v Russkoi Rechi]. Russkii Iazyk: Istoricheskie Sud'by I Sovremennost'. III Mezhdunarodnyi Kongress Issledovatelei Russkogo Iazyka (Russian Language: Historic Destinies and the Modernity. III International Congress of Russian Language Researchers): 348 (available at: http://www.philol.msu.ru/~rlc2007/abstracts/?sectionid=12)
10. *Mittelberg I.* 2007. Methodology for Multimodality: One Way of Working with Speech and Gesture Data. Methods in Cognitive Linguistics : 225–248.
11. *Müller C.* 1998. Redegebegleitende Gesten: Kulturgeschichte-Theorie-Sprachvergleich
12. *Richter N.* 2010. The Gestural Realization of Some Grammatical Features in Russian. Gesture: Evolution, Brain, and Linguistic Structures. 4th Conference of the International Society for Gesture Studies : 245

ИССЛЕДОВАНИЕ СТРУКТУРЫ НОВОСТНОГО ТЕКСТА КАК ПОСЛЕДОВАТЕЛЬНОСТИ СВЯЗНЫХ СЕГМЕНТОВ

Е. В. Ягунова (iagounova.elena@gmail.com)

Л. М. Пивоварова (lidia.pivovarova@gmail.com)

Санкт-Петербургский государственный университет,
Санкт-Петербург, Россия

Мы рассматриваем сочетания двух и более лексических единиц, которые выделяются нами из текста на основании статистических критериев и экспериментов с информантами с учетом контекстов различного типа. Полученные списки связанных сегментов эксплицируют разные информационные структуры одного и того же текста.

Ключевые слова: списки, информационные структуры, списки сегментов, связные сегменты.

A STUDY OF THE NEWS TEXT STRUCTURE AS A CONSEQUENCE OF CONNECTED SEGMENTS

E. V. Iagunova (iagounova.elena@gmail.com)

L. M. Pivovarova (lidia.pivovarova@gmail.com)

Saint-Petersburg State University, Saint-Petersburg,
Russian Federation

The main object of this study is connected segments (collocations, compound nominations, predicative constructions, multiword expressions, etc.) extracted from the text by different statistical measures and during experiments with native speakers. This paper deals with news texts: i) 2010 news from lenta.ru (40000 texts, 9.5 million tokens); ii) a small highly homogeneous corpus that deals with some particular event: Schwarzenegger in Moscow (360 texts,110 thousand tokens) and The appointment of Sobyanin (660 texts,170 thousand tokens); iii) three individual texts about Schwarzenegger and two individual texts about Sobyanin. These texts are part

of both the small homogeneous corpus and the large news corpus. In this paper we use an open-source “Cosegment” system (http://donelaitis.vdu.lt/~vidas/tools. htm). The program cuts the text into strongly connected segments depending on the corpus. We study different types of context using overlapping corpora as the input of the system. We also compare result based on the whole corpus and on individual texts from this corpus. During the experiments with native speakers we ask 18 students to put a number from 0 to 5 between every two words in the text. 5 means that these two words are strongly connected, 0 that there is no connection at all. Then we use a cutoff 3.7 to divide a text into connected segments. Our results are the following: i) Longer connected segments are found in the more homogeneous corpus; ii) Frequent connected segments in highly homogeneous corpora (as opposed to lenta. ru corpus) are mostly predicative constructions; iii) The computer processing data are very close to the native speakers’ data; iv) Native speakers tend to extract longer segments; they also prefer predicative constructions to collocations.

Key words: words connection, connected segments, information structures, context.

1. Введение

Эта работа посвящена экспериментальному исследованию структуры текста в духе принципиальной неединственности структурированности текста. Невозможность построения единственной структуры текста, отвечающей всем возможным видам анализа, уже рассматривалась ранее (напр., Падучева 2001: 109–112; Ягунова 2008).

На предыдущих этапах исследования (см. Ягунова 2008) выборка анализируемых текстов была ограничена возможностями экспериментов с информантами, т.е. объектом исследования становились отдельные тексты. Сейчас мы пытаемся реализовать следующий виток, когда объектом исследования становятся большая текстовая коллекция объемом в миллионы словоупотреблений и тематически однородные кластеры (подколекции). В результате различных вычислительных экспериментов на основе таких коллекций мы получаем данные, с одной стороны, позволяющие соотнести особенности структуры двух разных объектов (коллекции vs. единичного текста), с другой — определить интересующие нас типы текстов (структур текстов) и, тем самым, сузить материал для экспериментальной работы с информантами. В результате мы имеем возможность наиболее тщательно исследовать роль контекста: большой коллекции текстов → тематически однородной подколлекции текстов (сюжет или кластер) → единичного текста и → минимального синтаксического контекста (подробнее см. (Ягунова 2008; Ягунова, Пивоварова 2010)). Мы в своем исследовании языка и речи идем от реализации, от имеющегося в нашем распоряжении материала.

Мы рассматриваем **все** связанные сочетания двух и более лексических единиц, которые выделяются нами из текста на основании статистических критериев и/или экспериментов с информантами. Выделяемые таким образом единицы представляют собой неоднородное множество с точки зрения соотнесенности

со словарем и/или грамматикой, номинативностью и/или предикативностью. Возникает задача интерпретации выделенных единиц. Если подойти к этой задаче не столько с точки зрения четкого разбиения на классы, но выстраивания гибкой шкалы, то типовые, или ядерные, коллокации являются сложными номинациями (единицами словаря и парадигматическими единицами). Типовые, или ядерные, конструкции находятся на другом конце шкалы и характеризуются высокой предикативностью и синтагматичностью (см. подробнее в (Ягунова, Пивоварова 2011)).

Как известно, в процедурах обработки текста происходит максимальная опора на контекст. Причем понятие «контекст» также рассматриваться в разных смыслах. Для нас контекст предполагает широкое понимание:

- минимальный контекст, в котором реализуются лексические и морфолого-синтаксические явления;
- текстовый контекст, включающий в себя фрагменты текста вплоть до текста целиком;
- контекст, предполагающий учет текстов определенного типа (заданного функционального стиля, отобранной коллекции текстов и т. д.)

Степень связанности неоднословной единицы и закономерности ее появления в тексте, по всей видимости, описываются вероятностной моделью; оценки могут быть получены лишь на основании статистических данных. Причем статистические характеристики должны описывать данные в зависимости от перечисленных выше типов контекста, т. е. контекст должен учитываться как один из параметров модели.

Таким образом, неоднословные связанные сегменты выступают, прежде всего, как структурные составляющие текста или однородных коллекций (например, сюжетов). Анализ этих структурных составляющих позволил исследовать структуру текста и/или текстов. Единицы и контекст(-ы) анализировались во взаимодействии: главным образом, контекст (равно как и коммуникативная задача) определяют выбор единиц анализа. Тематически однородная коллекция (сюжет) изучалась методами лингвистики текста (дискурса).

Основной целью работы является оценка исследовательского потенциала предлагаемого метода в изучении теоретических аспектов лингвистики текста (дискурса).

Нами оценивались следующие данные:

- данные, полученные в ходе вычислительных экспериментов:
  - список наиболее связанных n-грамм по коллекции;
  - список наиболее связанных n-грамм по подколлекции (подколлекция является тематически более однородной, чем исходная коллекция);
  - отдельные тексты, представленные в виде последовательности связанных сочетаний («сегментов» в терминологии автора программы);
- отдельные тексты, представленные в виде последовательности связанных сочетаний, полученных в ходе эксперимента с информантами.

В ходе предварительного анализа пилотного эксперимента были сформулированы некоторые **гипотезы**, которые мы проверяем в ходе данной работы:

- с увеличением степени однородности (коллекция → однородная коллекция → текст) характерными становятся более длинные n-граммы;
- с увеличением степени однородности (коллекция → однородная коллекция → текст) увеличивается число конструкций (в соотношении конструкция vs. типовая коллокация), увеличивается число предикативных сочетаний;
- набор связанных сочетаний, подсчитанных для каждого текста отдельно в ходе вычислительного эксперимента, сходен с набором сочетаний, полученных в ходе экспериментов с информантами,
- набор связанных сочетаний, выделенный в ходе экспериментов с информантами, содержит несколько больше предикативных сочетаний, чем набор связанных сочетаний, сформированный в ходе вычислительного эксперимента.

2. Методика. Гипотезы

Данное исследование предполагает сочетание вычислительного эксперимента и эксперимента с информантами. В ходе вычислительного эксперимента меры совместной встречаемости определяется на основании видоизмененной меры Дайса (Dice) (Daudaravicius 2010a):

$$Dice'(x, y) = \log_2 \left(\frac{2 * f(x, y)}{f(x) + f(y)} \right),$$

где f(x) и f(y) — частота встречаемости слов x и y в коллекции, а f(x,y) — частота совместной встречаемости слов x и y.

Процесс вычислительного эксперимента можно коротко описать следующим алгоритмом. Сначала для всех пар слов по всей коллекции считается коэффициент Дайса. Затем для каждого конкретного текста, представляющего собой цепочку слов или, вернее, цепочку пересекающихся пар (слово *x* с предшествующим словом и слово *x* с последующим словом), осуществляется «сборка» связанных сегментов. При последовательном прохождении от слова к слову в каждом тексте уже известны соответствующие значения меры Дайса для всех пересекающихся пар. На основании значений этой статистической меры слова объединяются в связанные группы с учетом ближайшего контекста (принимается решение о том, надо ли присоединить текущее слово к предыдущему). Слово не присоединяется к предыдущему, если значение коэффициента Дайса для данной пары ниже порогового, или если оно ниже, чем среднее арифметическое того же коэффициента для левой и правой пары. Во всех остальных случаях слово присоединяется. Связанный сегмент может включать не более семи слов. В результате такого вычислительного эксперимента мы получаем набор связанных сочетаний, подсчитанных для каждого текста отдельно, а затем объединенный в некое подобие частотного словаря связанных сочетаний. Программа, реализующая этот алгоритм, доступна для скачивания с сайта ее создателя: http://donelaitis.vdu.lt/~vidas/tools.htm.

В ходе интерпретации мы исходили из того, что используемая мера выделяет связанные сегменты, характеризующиеся информационной ценностью

на материале однородной коллекции текстов (ср. Daudaravičius 2010б; Daudaravičius, Marcinkevičienė 2004). Свое предположение мы проверили через сопоставление с результатами, полученными с помощью стандартных статистических мер MI и t-score, и с ключевыми словами, выделяемыми на основании коэффициента важности tf-idf (этот коэффициент позволяет оценить степень важности слова по отношению к той или иной коллекции (подколлекции)). Выдвинутое предположение об информационной значимости связанных сегментов, выделяемых с помощью меры Дайса на материале тематически однородной коллекций текстов, подтверждается в ходе предыдущих исследований с использованием меры MI (напр., Ягунова, Пивоварова 2010; Ягунова, Пивоварова 2011). При рассмотрении указанных сегментов в рамках единичных текстов (по результатам вычислительного эксперимента и эксперимента с информантами) будем называть их значимыми структурными составляющими текста (значимыми для анализа текстов).

Материалом послужили тексты и/или коллекции:

- коллекции
  - Тексты портала Лента.ру за 2010 год — 40 000 текстов общим объемом около 9,5 млн. токенов (т. е. словоупотреблений и знаков препинания);
- два сюжета (или кластера), т. е. две небольших коллекции тематически однородных текстов, полученных с помощью ресурса «Галактика Зум»[1]:
  - приезд А. Шварцнеггера в Москву — 360 текстов, около110 тыс. токенов,
  - назначение С. Собянина — 660 текстов, 170 тыс. токенов,

  все тексты кластеров берутся из новостного потока, они близки по времени появления и посвящены одному событию;
- три текста о А. Шварцнеггере (из Лента.ру, РИАН, Газета.ру) и два текста о Собянине (Лента.ру, РИАН). Эти тексты использовались в вычислительных экспериментах (с соответствующим кластером и коллекцией данного информационного источника за 2010 год в роли двух разных контекстов) и в эксперименте с информантами.

Выбор конкретных новостных текстов и сюжетов (кластеров), т. е. подколлекций, состоящих из максимально тематически однородных текстов, определялся следующими соображениями. Материал должен был обладать сравнительно четкой и простой синтаксико-семантической структурой. Отбирались кластеры сравнительно большого объема с информационно значимым сюжетом (по субъективной оценке), имеющие четко выстроенный сюжет (основное действующее лицо (или лица), основное действие, сопровождающие действующие лица и/или организации, сопровождающие действия, время, место и т. д.).

В эксперименте с информантами — эксперименте по шкалированию — приняло участие 18 студентов СПбГУ, получающих гуманитарное образование[2]. Эксперимент с информантами представлял собой оценку связности между

---

[1] Этот материал любезно предоставлен нам Александром Антоновым и Станиславом Баглеем, Галактика-Zoom: galaktika-zoom.ru, http://www.webground.su

[2] Пользуясь случаем, хотим поблагодарить Галину Доброву за помощь в проведении эксперимента.

текстоформами (пробельными словами) в шкале от 0 до 5, где 5 — соответствовало максимальной, а 0 — минимальной степени связности. В инструкции информанту предлагалось оценить «*степень связности между словами или словом и знаком препинания в шкале от 0 до 5 баллов. «0» соответствует минимальной силе связности, а «5» — максимальной силе связности. Проставьте эти баллы (от 0 до 5) во ВСЕ позиции, между ВСЕМИ словами и/или словами и знаками препинания*». Информантам отдельно не объяснялся принцип оценки связности, они должны были действовать, опираясь на интуитивные представления о связности и, конечно, на свою текстовую базу знаний. Экспериментатор не навязывает информанту предпочтение, например, синтаксического или лексико-семантического подхода, однако полученные данные позволяют судить о том, что информанты в целом справляются с поставленной задачей. Усредненные данные по группе информантов не менее 18 человек, представили непротиворечивую оценку степени связности между словами. На основании этих данных можно выстраивать сколь угодно длинные цепочки слов в соответствии с устанавливаемым пороговым значением связности. Эмпирически мы подобрали пороговое значение, равное 3,7 баллам. Если полученное число было больше, чем 3,7, пару слов рассматривали как связную, если меньше — как не связную.

Носитель языка имеет интуитивные представления о неслучайно встречающихся сочетаниях слов: текстовые базы по текстам разных функциональных стилей, по текстам разных тематик или по текстам, посвященным определенной теме. На основании этого знания адресат воспринимает каждый конкретный текст как непротиворечащий некоторой текстовой базе адресата (в качестве ее аналога при вычислительном эксперименте выступают коллекции и подколлекции текстов разной степени однородности). Тематически однородные кластеры представляли достаточно обсуждаемые события, поэтому нельзя было предположить, что информанты не знакомы с этими темами. Эксперимент проводился примерно через месяц после описываемых событий, так что эти темы не могли быть забыты.

3. Предварительные результаты

Наибольший интерес в данном докладе представлял анализ данных, полученных на материале кластеров для словоформ[3]. При интерпретации данных по рассматриваемым сюжетам мы опирались на данные, полученные на материале двух сюжетов и пяти указанных текстов, однако для иллюстрации возможностей предлагаемого метода в статье приведены результаты только двух текстов: одного текста о А. Шварценергере и одного текста о С. Собянине из «Лента.ру» 2010 года[4].

---

[3] Анализ связности для лексем (лемм) также был нами произведен, но эта тема не является предметом данной статьи. Лемматизация текстов была произведена при помощи свободно распространяемого программного обеспечения АОТ (www.aot.ru), адаптированного под наши задачи В. В. Бочаровым.

[4] В статье мы ограничиваемся новостными текстами, однако при интерпретации данных частично учитывались также результаты, полученные на материале научных

В табл. 1 представлены данные вычислительного эксперимента и эксперимента с информантами на материале сюжета и текста о А. Шварценеггере. В таблице представлены сегменты, состоящие не менее чем из трех текстоформ (слов, разделителем между которыми служат пробелы и/или знаки препинания). Полужирным шрифтом выделены те сегменты или их фрагменты, которые присутствуют как в списке, полученном в ходе вычислительного эксперимента, так и в эксперименте с информантами. В графу «Сюжет о Шварценеггере (однородная коллекция)» попала верхушка наиболее частотных связанных сегментов, упорядоченных по частоте, остальные графы (наборы) представлены в табл. 1 полностью.

Предложенная нами методика учитывает различные виды контекстов: «тематический» (сюжет) и «стилистический» (Лента.ру) (см. табл. 1). В «стилистическом» контексте существенными оказывались характерные для СМИ конструкции и обороты (например, *в настоящее время*, *со ссылкой на*), из которых нельзя сделать выводы о конкретном содержании текстов, но можно составить общее впечатление об их стилистической направленности (см. табл. 1). В «тематическом» контексте наиболее значимыми оказывались сложные номинации (*глобальное инновационное партнерство*) и предикативные конструкции, описывающие ситуацию (*только что приземлился*) (см. табл. 1). Структурные составляющие сюжета дали более полное и объективное представление о сюжете, чем структурные составляющие единичного текста. Информанты в целом выделяли более длинные сегменты, чем программа. Информанты были нацелены на описание ситуаций, они выделяли большее число предикативных сочетаний — длинные конструкции в целом более типичны, чем длинные коллокации.

Таблица 1. Связанные сегменты, состоящие не менее чем из трех текстоформ

Вычислительный эксперимент			**Эксперимент с информантами, единичный текст о А. Шварценеггера**
Коллекция (Лента.ру 2010 г)	**Сюжет о Шварценеггере (однородная коллекция)**	**Единичный текст о А. Шварценеггера**	
тем не менее	глобальное инновационное партнерство	только что приземлился	Губернатор Калифорнии Арнольд Шварценеггер
в связи с	представителей ведущих компаний	**могу дождаться встречи**	прилетел в Москву.
в 2009 году	с губернатором калифорнии	вскоре после этого	в российскую столицу
то же время	могу дождаться встречи	**ответил** калифорнийскому губернатору	Не **могу дождаться встречи** с президентом Медведевым

текстов (тематически однородная коллекция материалов конференции «Корпусная лингвистика» и 4 текста из этой коллекции).

Вычислительный эксперимент			**Эксперимент с информантами, единичный текст о А. Шварценеггера**
Коллекция (Лента. ру 2010 г)	**Сюжет о Шварценеггере (однородная коллекция)**	**Единичный текст о А. Шварценеггера**	
в настоящее время	во главе делегации	англоязычная версия твита	российский президент Дмитрий Медведев **ответил**
со ссылкой на	создать настоящий технологический бум	ответил ему взаимностью	в своем микроблоге
возбуждено уголовное дело	сфере высоких технологий	это же время	добро пожаловать в Москву
по сравнению с	только что приземлился		Жду встречи с вами
в 2008 году	тогда вам сказал		Медведев добавил микроблог
и т. д.	которые занимаются инновационными разработками		с делегацией представителей
	их российскими партнерами		он встретится с российскими министрами
	российская венчурная компания		во время посещения Медведевым
	стать мэром москвы		российский президент завел себе
	Global Technology Symposium		
	главами американских инвестиционных компаний		
	видение дальнейшего развития		
	Silicon Valley Bank		
	пост мэра москвы		
	самых разных событий происходит		
	июне этого года		
	после непродолжительной беседы		
	и т. д.		

Число пересекающихся длинных связанных сегментов, выделяемых программой и информантами, в существенной степени зависит от типа текста. Для

более динамичных сюжетов и текстов (включающих описание последовательности событий) число пересечений меньше, для более статичных — больше[5]. Это один из параметров, позволяющих оценить структуру единичного текста и текстов сюжета в целом. Нам кажется неправильным рассматривать процент совпадений между программой и информантами как меру оценки качества работы программы, поскольку информанты ничего не знали о конечных целях исследования. Оценка работы самой программы производилась ранее ее автором. В частности, в (Daudaravičius 2010b) показано, что для задачи выделения ключевых слов использование алгоритма сегментации дает улучшение F-меры на 17–27 % в зависимости от данных.

Набор длинных связанных сегментов, выделяемых информантами, на наш взгляд, может считаться самоценным для анализа структуры текста, т. к. вполне вероятно, что они отражают расстановку структурных составляющих текста, важных для восприятия (ср. идею о том, что при восприятии адресат стремится оперировать наиболее крупными оперативными единицами, напр., Грановская 1974). Продемонстрируем это на примере текста, в котором длинные связанные сегменты интерпретировались в духе гештальтпсихологии в качестве фигуры (они выделены полужирным шрифтом), а все остальные фрагменты текста рассматриваются как фон (выделены зачеркнутым шрифтом):

Губернатор Калифорнии Арнольд Шварценеггер ~~10 октября~~ **прилетел в Москву.** / ~~После прибытия~~ **в российскую столицу** ~~он сделал в своем микроблоге на Twitter соответствующую запись (Только что приземлился в Москве. Прекрасный день.~~ **Не могу дождаться встречи с президентом Медведевым**), ~~а также разместил фотографию, сделанную по дороге из аэропорта.~~

~~Вскоре после этого~~ **российский президент Дмитрий Медведев ответил** ~~калифорнийскому губернатору~~ **в своем микроблоге**: ~~@Schwarzenegger~~, **добро пожаловать в Москву**. ~~Англоязычная версия твита Медведева также содержала слова~~ «**Жду встречи с вами** ~~и вашей делегацией в @skolkovo».~~

~~Кроме того,~~ **Медведев добавил микроблог** ~~Шварценеггера в друзья. Губернатор Калифорнии ответил ему взаимностью.~~

~~Как сообщает РИА Новости, Шварценеггер приехал в Россию~~ **с делегацией представителей** ~~венчурных фондов и инновационных компаний Кремниевой долины. Планируется, что помимо президента Медведева~~, **он встретится с российскими министрами**.

~~Президент России и губернатор Калифорнии~~ **в этом году** ~~уже встречались —~~ **это произошло в июне / во время посещения Медведевым** ~~США. В это же время~~ **российский президент завел себе** ~~микроблог~~.

Набор двухсловных связанных сегментов (полученных в эксперименте с информантами), конечно, имел информационную ценность, однако количество предикативных сочетаний в нем минимально (см. табл. 2). Объединение набора двухсловных и длинных связанных сегментов «улучшает» понимание значимости визита А. Шварцнеггера для развития высоких технологий (см.

[5] По нашим предварительным данным, для научных текстов такого рода пересечений гораздо больше, чем для новостных текстов.

табл. 2), а насколько эта составляющая важна — решать адресату, т. е. тому, кто анализирует и понимает этот текст. Возможно, причина невыделения сегментов, несущих такую информацию, в том, что большинство информантов — гуманитарии, однако структура рассматриваемых текстов как минимум позволяет прочтение, в котором «*развитие высоких технологий*» является второстепенным фактом.

На материале результатов вычислительных экспериментов картина более неоднозначная. Если для кластера в целом длинные связанные сегменты информативны, то в случае единичного текста в указанном примере длинных связанных сегментов мало, мы не можем извлечь ценную информацию (понять текст) из их набора. Рассмотрим набор связанных сегментов, состоящих из 2 текстоформ, полученных в ходе вычислительного эксперимента с этим текстом (см. табл. 2). Среди них много информационно значимых единиц для описания сюжета (наименования персон, организаций, места и времени), более того — среди них неожиданно много предикативных единиц (основные выделены курсивом в табл.2). Среди них встречались цепочки (здесь и далее знак «/» показывает границу между связанными сегментами), например, ***Губернатор Калифорнии / Арнольд Шварценеггер,*** ~~*Как сообщает / РИА Новости,*~~ ~~*Шварценеггер приехал / в Россию*~~. При сопоставлении материалов экспериментов с информантами и вычислительного эксперимента фигура и фон — связанные сегменты, выступающие в качестве фигуры и/или фона, — могли меняться или оставаться прежними (здесь и далее полужирный и зачеркнутый шрифт соответствует выше описанному анализу результатов эксперимента с информантами). Кроме того, часто встречались случаи опущения однозначно восстановимого предлога (в примере такие предлоги заключены в скобки), напр., ~~*После прибытия*~~ */* ***(в) российскую столицу*** */* ~~*он сделал / (в) своем микроблоге,*~~ ***он встретится / (с) российскими министрами***. Объединение двухсловных и длинных связанных сегментов — по результатам вычислительного эксперимента — дало достаточно полное представление о структуре текста, необходимой для извлечения смысла.

Почему, если рассматривать каждый из текстов из кластера про Шварцнеггера, то длинных связанных сегментов, полученных в результате вычислительного эксперимента, практически никогда не оказывается достаточно для анализа информационной структуры этого текста? Почему для этого материала столь велико различие между набором длинных связанных сегментов, полученных в результате эксперимента с информантами и вычислительного эксперимента?

Одна из основных причин лежит в особенностях структуры анализируемого в примере текста. Телетайпный, отрывочный стиль написания большинства текстов кластера про А. Шварцнеггера (возможно, обыгрывающий общение в твиттере) характеризуется короткими структурами и навязывает короткие связанные сегменты. Характеристику анализируемого текста можно дополнить отсутствием четко выраженной композиционной структуры сюжета. Выбор примера — и сюжета, и текста как его наиболее яркого представителя — обусловил резкое различие между результатами эксперимента с информантами и вычислительного эксперимента.

Таблица 2. Связанные сегменты, состоящие из 2 текстоформ, полученные в ходе вычислительного эксперимента

губернатор Калифорнии	из аэропорта	венчурных фондов
Арнольд Шварценеггер	российский президент	инновационных компаний
10 октября	Дмитрий Медведев	
в Москву	своем микроблоге	Кремниевой долины
после прибытия	*добро пожаловать*	, что
российскую столицу	в Москву	президента Медведева
он сделал	*содержала слова*	*он встретится*
своем микроблоге	*жду встречи*	российскими министрами
соответствующую запись	вашей делегацией	
	кроме того	президент России
в Москве	добавил микроблог	губернатор Калифорнии
прекрасный день	в друзья	
президентом Медведевым	губернатор Калифорнии	этом году
	как сообщает	*уже встречались*
а также	РИА Новости	во время
разместил фотографию	*Шварценеггер приехал*	российский президент
сделанную по	в Россию	*завел себе*

Таблица 3. Связанные сегменты из текста про С. Собянина, состоящие не менее, чем из 3 текстоформ[6]

Кластер про С. Собянина (однородная коллекция)	**Вычислительный эксперимент**	**Эксперимент с информантами**
на пост мэра	**Московской городской думы**	Сергей Собянин утвержден
Московской городской думы	**проголосовали 32 депутата**	на посту мэра Москвы
проголосовали 32 депутата	участвовали 34 человека	**Московской городской думы**
тот же день	присяга нового мэра	**проголосовали 32 депутата**
губернатор Нижегородской области	тот же день	против высказались двое
нового мэра Москвы	**Как сообщалось ранее**	голосование в Мосгордуме
из 35 депутатов	**18 : 00**	**Как сообщалось ранее**
инаугурация нового мэра	избрании нового градоначальника	торжественное мероприятие планируется провести

[6] Полужирным шрифтом выделены те сегменты или их фрагменты, которые присутствуют в списках, полученных как в ходе вычислительного эксперимента, так и эксперимента с информантами.

Кластер про С. Собянина (однородная коллекция)	**Вычислительный эксперимент**	**Эксперимент с информантами**
центральном Федеральном округе	руководивший исполнительной властью	в **18:00**
кандидатуру Сергея Собянина	9 октября партия	21 октября 2010 года
на посту мэра	представила президенту четыре кандидатуры	нового градоначальника Москвы
добросовестно исполнять возложенные	список единороссов попали	исполнительной властью столицы
благополучию его жителей	**губернатор Нижегородской области**	с утратой доверия президента
участвовали 34 человека	прошлом — вице-мэр	Соответствующий указ Дмитрия Медведева
губернатором Тюменской области	**исполняющая обязанности вице-мэра**	на пост мэра Москвы
остановил свой выбор	остановил свой выбор	**губернатор Нижегородской области**
по его словам	после этого фракция	**исполняющая обязанности вице-мэра** Москвы
присяга нового мэра	из 35 мест	президент Медведев объявил
Московская городская дума	органах власти начался	аппарата правительства РФ
руководивший исполнительной властью	**городе Когалым Ханты-мансийский округа**	пообещала поддержать выбор Дмитрия Медведева
9 октября партия	**ответственные государственные посты**	в **городе Когалым Ханты-Мансийский округа**
избрании нового градоначальника	губернатором Тюменской области	в разные годы
до 2008 года	до 2008 года	занимал **ответственные государственные посты**
из 35 мест		
органах власти начался		
ответственные государственные посты		

В качестве контрпримера приведем кластер текстов о С. Собянине. В таблице 3 в первом столбце представлена верхушка наиболее частотных связанных сегментов из кластера, упорядоченных по частоте; во втором и третьем столбце результаты вычислительного эксперимента и эксперимента с информантами

на материале конкретного текста из этого кластера[7], также принадлежащего источнику Лента.ру. Наблюдается значительное сходство между наборами длинных связанных сегментов, полученных в результате эксперимента с информантами и вычислительного эксперимента. Длинные связанные сегменты, полученные в результате эксперимента с информантами, рассмотрим в силу нашего допущения как достаточные для анализа (понимания) текста.

Длинные связанные сегменты, полученные в результате вычислительного эксперимента, обладают, главным образом, одним «недостатком»: в их состав не попадают наименования персон, действующих лиц этого сюжета. Если бы мы добавили к этому набору набор двухсловных связанных сегментов или наименования персон (с элементами Ф.И.О.), то вся информация, необходимая для восстановления текста, присутствовала бы в объединенном наборе. Для рассматриваемого текста набор двухсловных связанных сегментов с элементами ФИО следующий: *Собянин утвержден, Сергей Собянин, за Собянина, Юрий Лужков, Дмитрия Медведева, помимо Собянина, Игорь Левитин, соратник Лужкова, Валерий Шанцев, Людмила Швецова, Медведев объявил, Сергее Собянине, Дмитрия Медведева, избрать Собянина, Сергей Собянин, Владимира Путина, Дмитрия Медведева, Владимира Путина*.

Полученные в наших экспериментах данные можно и нужно интерпретировать в духе исследования принципиальной неединственности структурированности текста (напр., Ягунова 2008) — в данном случае, прежде всего, информационной структурированности. Результаты вычислительного эксперимента и эксперимента с информантами эксплицируют разные информационные структуры одного и того же текста: разные варианты извлечения информации в соответствии с намерениями и возможностями адресата. Адресат (носитель языка или автомат) выделяет важные вехи в тексте на основании коммуникативной ситуации, собственных целей и задач. Разные возможности и задачи соответствуют разным коллекциям (в соответствии тематической областью коллекции и/или разной степенью однородности) или разным базам знаний информантов (степени компетентности информантов).

4. Заключение

Полученные данные не противоречат выдвинутым гипотезам. Рассматриваемая методика предоставляет исследователю возможность анализировать информационную структуру текстов, варьируя, как минимум, варианты коллекций и подколлекций и, исследуя, таким образом, тексты разных функциональных стилей (новостные, научные, официально-деловые), разных жанров, разной тематики.

В заключение хотим перечислить те сопоставительные результаты, которые не вошли в формат статьи, но, надеемся, смогут быть предметом обсуждения на конференции «Диалог 2011». Речь идет о двух видах сопоставления:

[7] Объем текста — 273 слова.

- с результатами, полученными на этом же материале с помощью статистических мер MI и t-score[8], а также с ключевыми словами, выделяемыми на основании коэффициента TF-IDF[9],
- с результатами, полученными по полностью аналогичной методике, на основе 4 текстов из материалов конференции «Корпусная лингвистика» (в контексте коллекции материалов «Корпусная лингвистика» за 2004–2008 годы и коллекции трудов конференции Диалог за 2003–2009 годы).

Настоящий этап исследования был посвящен экспериментальному исследованию теоретических аспектов лингвистики текста (дискурса). Надеемся, что в результате следующего этапа нам удастся получить те данные, которые будут полезны для построения модели понимания текста адресатом и для решения технологических вопросов анализа текстов и текстовых коллекций.

References

1. *Comptunig* Resourse, available at: http://donelaitis.vdu.lt/~vidas/tools.htm
2. *Daudaravicius V.* 2010. Automatic Identification of Lexical Units. Computational Linguistics and Intelligent text processing CICling-2009.
3. *Daudaravičius V.* 2010. The Influence of Collocation Segmentation and Top 10 Items to Keyword Assignment Performance. Proceedings of Computational Linguistics and Intelligent text processing CICling-2010 : 648–660.
4. *Daudaravičius V., Marcinkevičienė R.* 2004. Gravity Counts for the Boundaries of Collocations. International Journal of Corpus Linguistics, 9 (2).
5. *Granovskaia R. M.* 1974. Memory Model Perception [Vospriiatie Modeli Pamiati].
6. *Iagunova E. V.* 2008. Variability of the Strategies of Sounding Text Perception [Variativnost' Strategii Vospriiatiia Zvuchashchego Teksta].

---

[8] Меры MI и t-score содержательно различны. MI наилучшим образом позволяет определять наименования объектов (персон, организаций, географические наименования), термины и другие сложные номинации. Мера t-score, напротив, обычно позволяет выделять частотные составные слова (служебные и дискурсивные слова) и частотные конструкции (напр., *по словам*, *со ссылкой*, *сообщает РИА*) (ср. Ягунова, Пивоварова 2010). Однако для тематически однородных коллекций мера t-score выделяет информативно значимые сложные номинации, которые характеризуют коллекцию в целом и присутствуют (почти) во всех текстах коллекции (Пивоварова, Ягунова 2010; Ягунова, Пивоварова 2011).

[9] Для примера возможности сопоставления информационной важности слов текста (на примере текста про С. Собянина по отношению к коллекции лента.ру за 2010г.) приведем топ слов с наибольшими значениями TF-IDF: *Собянин, мэр, Москва, заседание, Медведев, вице-мэр, мосгордума, пост, избрание, градоначальник, 2001–2005, октябрь, Когалым, Лужков, кандидатура, Дмитрий, вице-премьер, 32, утвердить, Сергей, голосование, губернатор, аппарат, столичный, единый, четверг, Швецов, президент, Ханты-мансийский, Шанцев, Путин, новое, Левитин, тюменский, присяга, депутат, минтранс, тайный, утрата, выбор, 21, занимать, единорос.*

7. *Iagunova E. V., Pivovarova L. M.* 2010. The Nature of Collocations in Russian Language. Experimetn of Automatic Extraction and Classification on the Material of New Texts [Priroda Kollokatsii v Russkom Iazyke. Opyt Avtomaticheskogo Izvlecheiia I Klassifikatsii na Materiale Novostnykh Tekstov]. NTI, 2 (6).
8. *Iagunova E. V., Pivovarova L. M.* 2011. From Collocations to Constructions [Ot Kollokatsii k Konstruktsiiam]. Russkii Iazyk: Konstruktsionnye I Leksiko-Semanticheskie Podkhody.
9. *Manning C., Schutze H.* 2002. Collocations. Foundations of Statictical Natural Language Processing :151–189
10. *Paducheva E. V.* 2001. Statement and its Correlation with the Reality [Vyskazyvanie I ego Sootnesennost' s Real'nost'iu].
11. *Pivovarova L. M., Iagunova E. V.* 2010. Extraction and Classification of Terminological Collocations on the Material of Linguistic Scientific Texts (Preliminary Observations) [Izvlechenie I Klassifikatsiia Terminologicheskikh Kollokatsii na Materiale Lingvisticheskikh Nauchnykh Tekstov (Predvaritel'nye Nabliudeniia)]. Materialy Simpoziuma "Terminologiia I Znanie" (Proc. of Symposium "Terminology and Knowledge").
12. *Stubbs M.* 1995. Collocations and Semantic Profiles: On the Case of the Trouble with Quantitative Studies. Functions of Language, 2 (11) : 23–55.

ПРИНЦИПЫ ВЫБОРА СЛОВ-НОСИТЕЛЕЙ АКЦЕНТНЫХ ПИКОВ В РУССКОМ ЯЗЫКЕ

Т. Е. Янко (tanya_yanko@list.ru)

Институт Лингвистики РАН, Москва, Россия

Тональные акценты, несущие значения темы, ремы, контраста и незавершенности текста, накладываются на сегментный материал не случайным образом, а в связи с определенными словоформами-акцентоносителями. Базовый принцип выбора акцентоносителя контролируется синтаксическими иерархиями. Однако этот принцип не единственный. На разных уровнях построения предложения и текста действуют периферийные принципы, демонстрирующие отклонения от базовых приоритетов.

Ключевые слова: акцент, акцентный пик, тональный акцент, акцентоноситель, слова-носители.

ACCENT PLACEMENT PRINCIPLES IN RUSSIAN[1]

T. E. Ianko (tanya_yanko@list.ru)

Institute for Linguistics, Russian Academy of Sciences, Moscow, Russian Federation

The basic constituents of intonation structure are pitch accents. Pitch accents designate topic-focus distinctions, contrast, and discourse structure. The question arises as to what phonetic words the accents are placed on. This paper gives an account of various accent placement principles in modern Russian.

Key words: accent, accent pitch, tonal accent, accent placement.

[1] This work was supported by the program ОИФН РАН «Текст во взаимодействии с социокультурной средой: уровни историко-литературной и лингвистической интерпретации», раздел VI. Текст в социокультурном и языковом пространствах РФ, проект «Универсальные и идиоэтнические стратегии продуцирования и интерпретации текста».

The basic constituents of intonation structure are pitch-accents. They designate topic-focus distinctions, contrast, and discourse structure. Pitch-accents are placed on the segmental material not in a random way but as properties of the phonetic words that they fall on [Bolinger 1958, 1961; Halliday 1963, 1967a, b; Steedman 2007]. There is a considerable body of literature that assumes that the relevant notions of accent placement are information structure, focus, contrast, or emphasis, i. e. the words bearing pitch-accents refer to either new, or emphatic, or contrasted information (cf. most recent investigations [Jaeger, Wagner (in press); Steedman 2007; Kadmon 2009], and references cited therein[2]). However, the segmental material referring to either new, or contrastive, or emphatic information is not restricted to one phonetic word. It may rather have complex syntactic structure. The question would then arise as to how the words bearing pitch-accents are selected. Many theories of information structure and accent placement proposed particular sets of syntactic priorities accounting for the selection of the accent-bearer in syntactic structures, such as the priority of the object to the verb ([Schwarzschild 1999]), and the subject to the verb ([Halliday 1967b: 208; Enkvist 1979; Schwarzschild 1999]). But nobody proposed a full hierarchy of syntactic constituents relevant for accent placement. Moreover, the parameters relevant for the accent-bearers selection are not limited to information structure and syntactic hierarchies. Modern Russian displays a number of accent placement principles for which, apart from the information and syntactic structure, some lexical, illocutionary and discourse parameters are also significant. Some of the principles are presumably valid not only for Russian but for a wide variety of languages with no lexical pitch-accents. This paper gives a concise account of accent placement principles in Russian. The formulation of accent placement principles is aimed at being used in oral speech synthesizers.

1. Basic principle

The basic principle of accent placement in topics, foci, constituents of questions and imperatives is in the first place regulated by 1) the given-new distinction [Chafe 1976] (or activation, on the terminology of [Dryer 1996]), and 2) the syntactic structure (the argument structure of the sentence predicate and the sentence syntactic type).

The ‘given-new’ parameter affects the accent placement in such a way that the items referring to activated (in the mind of the hearer) information should be eliminated from the set of items of which the focus (= the rheme, on Mathesius terminology) accent-bearer is to be selected. For example, the replies in (1) and in (2) with the

[2] One of the most recent accounts of accent placement given in paper presentation [Kadmon 2009] is based on the notion of 'recoverable' harking back to Kuno's 'predictable' ([Kuno 1978: 282–283]), Halliday's 'recoverable' ([Halliday 1967b: 204]), and Prince's given$_p$ ([Prince 1981] “which is the informational status of a word depending not just on preceding context, but also on its relation to its own utterance”. The author also claims that accent placement can be interpreted “at the level of the word alone, without recursive projection or interpretation of a syntactic feature” [Kadmon 2009].

accents on *дети.Nom*[3] *and on едят.*V_{fin} respectively demonstrate the priority of the new information to the activated one in accent-bearer selecting. In sentence (2) the accent-bearer of the reply is the verb because the subject *дети* is activated in the question. (The accent-bearers in examples below are boldfaced.)

(1) — *Чем ты расстроен? — **Дети** плохо едят;*

(2) — *Чем тебя дети расстроили? — Дети плохо **едят**.*

Topics can also have accented words carrying specific "topical" pitch-accents. The accent placement in topics (themes) follows similar accent placement principles as the foci. The differences between the focal and the topical accent placement are discussed in [Янко 2008: 60–72].

The syntactic parameter of accent placement is represented by a variety of hierarchies which are determined by the argument structure of the predicate and the sentence structure [Янко 1991; 2008: 43–73]. These hierarchies are 1) the Basic Hierarchy of the predicate and its arguments, and 2) the local hierarchies affecting accent placement in non-terminal arguments (or the arguments which have complements or adjuncts).

1.1. Basic hierarchy

The basic syntactic hierarchy ranges the sentence components according to the predicate-argument structure. Thus the indirect object in all-new[4] sentences is prior to the direct object and the subject; the subject is prior to the finite verb, the object No. N+1 is prior to the object No. N, the arguments are prior to the adverbial modifiers, cf. the Hierarchy on Figure 1.

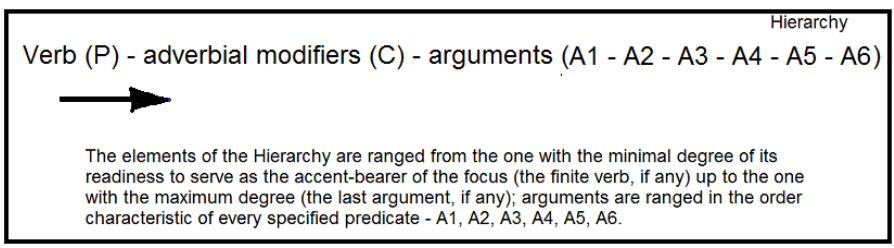


Figure 1. The Basic Syntactic Hierarchy controlling the accent placement in a sentence

[3] The following abbreviations are adopted in this paper: ACC — Accusative, INSTR — Instrumental, NOM — Nominative, NP — Noun Phrase, V_{fin} — finite verb.

[4] About all-new (or thetic, on the terminology of V. Mathesius) cf. [Баранов, Кобозева 1983; Sasse 1987; Николаева 1996: 53], and literature cited therein.

Thus the replies in (3) and in (4) with the accents on *дети.NOM* (A1) and on *кашу.ACC* (A2) respectively exemplify the priority of the arguments to the predicate, and the object to the subject.

(3) — *В чем дело?* — ***Дети*** (A1) *плохо едят.*

(4) — *В чем дело?* — *Дети* ***кашу*** (A2) *плохо едят.*

The second sentences in examples (5)–(10) exemplify a number of priorities determined by the Basic Hierarchy:

(5) — *Чем ты расстроен*? — *Бабушка* (A1) *в дороге* (C) ***очки*** (A2) *сломала* (P);

(6) — *Почему пусто в отделе*? — *Директор* (A1) *пять человек* (A2) *в Москву* (A4) *для обмена* ***опытом*** (A5) *командировал* (P);

(7) — *Ты куда*? — ***Обедать*** **(**A2) *пора* (P);

(8) — *В чем дело*?> — ***Денег*** (A) *нет* (P);

(9) *Худо, брат, жить в Париже.* ***Есть*** (A) *нечего* (P) (Pushkin);

(10) *Мальчики платья не носят*. ***На горшок*** (A2) *неудобно* (P) (from children's speech).

In addition to activation and syntax, some other parameters may take effect in accent placement. For instance, the parameter of idiomatic filling the argument position may produce an accent placement distinct from the one in a formally identical but non-idiomatic phrase. Thus sentence (11) with the idiom *входить в силу* 'come into effect' (lit. 'come into power') has the subject *закон.ACC* 'law' as the accent-bearer. Here, the "privileged" accent-bearer licensed by the basic Hierarchy is the object, while within the idiomatic phrase *входить в силу* the desemantisized object concedes its right of the accent-bearer to the subject which is previous to the object in the Hierarchy.

(11) — *В чем дело? — Новый* ***закон*** (A1) *в силу* (A2) *вошел.*

Thus example (11) shows that idiomatic expression may violate the main principle of accent placement.

Accent placement is also affected by the Animacy hierarchy in such a way that the non-prototypical argument expression may change the primary priority of elements within the Basic Hierarchy. For instance, an animate — non-prototypical — object may give its right of the accent-bearer to an inanimate — non-prototypical — subject. For example, the pair of sentences (12) with the prototypical accent-bearer the object *мышку.ACC* and (13) with the non-prototypical accent-bearer the subject

совесть.NOM exemplify the derived priority of the non-prototypical subject to the non-prototypical object displayed by sentence (13).

(12) — *В чем дело? — Кошка* (A1) ***мышку*** (A2) *мучает.*

(13) — *В чем дело? — Кошку* (A2) ***совесть*** (A1) *мучает.*

Finally, the accent placement may also be affected by the parameter of syntactical complexity of a focus which can consist of more than one phrase having no dominating category in the immediate constituents tree. For instance, in (14) the reply with the focus *Хартман в 1882 году* consists of the two NPs *Хартман.NOM* and *в 1882 году* (lit. 'in the 1882th year'). These NPs have no shared immediate dominating node. (The focus in (14) below is underlined.)

(14) — *Кто еще занимался этой проблемой? — Этой проблемой занимался Хартман в 1882 году.*

In (14), there are two accents within the focus — the sentence-final focal accent on the accent-bearer of the constituent *в 1882 году* and the non-final accent carried by the constituent *Хартман* [Янко 2008: 50].

1.2. Accent placement in coordinate phrases and in phrases with complements or adjuncts

If the selected element — be it a subject, an object, or an adverbial modifier — includes coordinate phrases, complements or adjuncts (dependent elements), additional local hierarchies allow for selecting *Ванечка* in (15) *Танечка и* ***Ванечка***, *Иванов* in (16) *Вася* ***Иванов***, *гостей* in (17) *ждем* ***гостей*** *дорогих*, and *белье* in (18) *грязное* ***белье*** *стирать*, cf. [Ковтунова 1976: 146, Русская грамматика 1982, II: 203–206; Светозарова 1993; Кодзасов 1996: 202]:

(15) *Танечка и* ***Ванечка***;

(16) *Вася* ***Иванов***;

(17) *ждем* ***гостей*** *дорогих*;

(18) *грязное* ***белье*** *стирать.*

Accent placement in English compound NPs is discussed in [Zwicky 1986].

The rules based on the local hierarchies are employed recursively until the terminal node is obtained.

1.3. Accent placement of contrast

Accent placement in sentences which include contrastive or emphatic components demands a specific consideration. In sentence (19) the accent-bearer of the focus is the word *Вася*, while in (20) with no contrast the accent-bearer is the family name *Иванов*. Similarly, in (21) the accent-bearer is the emphatically stressed verb *не хотелось* 'did not want', while in (22) the accent-bearer of a focus is the object *берег*.

(19) *Это* ***Вася*** *Иванов, а не Ваня Иванов.*

(20) *Это Вася* ***Иванов****.*

(21) *Не* ***хотелось*** *мне переходить на другой берег.*

(22) *Мне не хотелось переходить на другой* ***берег****.*

The contrastive and emphatic foci (and topics) are regarded here as non-violating the Basic Principle because they change not the syntactic principle of the accent-bearer selection, but only the boundaries of the foci (or the topics). For instance, in (19) the contrastive focus is the word *Вася*, and therefore it is accented. Whereas the first occurrence of the word *Иванов* here belongs to the topic as being activated in the mind of the hearer, and it therefore remains unstressed. In sentence (20), however, the focus is *Вася Иванов* with the accented word *Иванов* in full accord with the local hierarchy viewed in Section 1.2. Consequently, both cases totally agree with the Basic Principle as presented in Sections 1.1 and 1.2. Similarly, in (21) the emphatic focus is the prosodic group *не хотелось*, and thus the verb *хотелось* is stressed. In sentence (22), however, the focus is the verb phrase *не хотелось переходить на другой берег* 'did not want to pass to the opposite bank', therefore the accent-bearer is the word *берег* 'river bank'. This accent placement also strictly follows the Basic Principle. About accent-bearers in contrastive components of sentences cf. [Янко 2008: 58].

The Basic Principle of accent placement is the main but not the only accent placement principle relevant for Russian. A question arises as to whether there are accent placement types deviating from the basic priorities. At various discourse levels some peripheral principles are taking effect. These are:

- the principle based on the linear order of the words in a sentence;
- the principle based on accenting the illocutionary markers in a variety of specific speech act types;
- the principle relevant for the Russian colloquial speech which employs an additional (unspecified by the Basic Principle) accent-bearer to designate text incompleteness, i. e. the meaning of 'to be continued';
- accent placement types relevant for various cultural traditions, such as chanting, praying, verse reading, and begging.

  These principles are considered in Sections 2–5 respectively.

2. Linear principle

The linear principle is employed in more than one-word Russian vocatives, imperatives and exclamations which are composed with additional illocutionary meanings, such as gentle reproaching, or persuading or, on the contrary, with insistent urging to do something, indignation, or anger. This principle is based on positioning the frequency peak sentence-initially, or, on the contrary, sentence-finally, irrespective of syntactic priorities. For instance, when addressing a person, close to the speaker either psychologically or in space, the accent moves to the "left". Consider sentences (23) with the accent on *Марья* and (24) with the accent on *Ивановна* respectfully:

(23) ***Марья*** *Ивановна! Ну как же это вы так неосторожно?!*

(24) *Меня зовут Марья* ***Ивановна.***

Here, in vocative sentence (23) which expresses a sympathetic reproaching the accent-bearer in the name *Марья Ивановна* is the first name *Марья*. However, in sentence (24) which is a simple statement with no any attendant illocutionary meanings the accent-bearer is the sentence-final patronymic name *Ивановна*, as the Basic Principle demands.

Similarly, in (25) which designates a gentle reproach the accent-bearers are the sentence-initial words *ваша* and *как*, while in the corresponding sentences (26) which is a simple statement and (27) which expresses indignation the accent-bearers are the words *честь* (cf. (26)) and *так* (cf. (27)) respectively:

(25) ***Ваша*** *честь! Ну* ***как*** *же так?! Мы же осуждаем невиновного!*

(26) *На карту поставлена ваша* ***честь.***

(27) *Как же* ***так****?! Это возмутительно!*.

In sentence (28) which expresses gentle persuading the accent-bearers are the sentence-initial words the adjective *молодой* and the imperative *купите*:

(28) ***Молодой*** *человек!* ***Купите*** *букетик!*

Whereas in a more persistent turning a looker to buy something the accent-bearer is the sentence-final object:

(29) *Купите* ***букетик***; *Гоните* ***рублики***.

The "left shift" in Russian vocatives was examined in [Кузьмичева 1964; Светозарова 1993]. In English and German vocatives, a similar "left" shift is inapplicable:

(30) **Your* *Honour!*; **Young* *man!*; **Mister* *Johnson!*; **Herr* *Janzen!*; **Frau* *Müller!*; **Doktor* *Kozak!*; **Liebe* *Kollegen!*; **Marie-Luise!*.

About accent-bearers in German vocatives cf. [Палько 2009]. Thus the "left" shift in vocatives is a specific Russian typological feature.

A symmetric "right" shift is also applicable to speech acts occurring within either space or psychological distance between the interlocutors. However, the result of it generally coincides with the accent placement governed by the basic rules, because in statistical majority of cases the accent-bearer, irrespective of possible extraneous meanings (either space remoteness, or anger, or indignation, or sharp rebuke), the accent-bearer remains sentence- final. Nevertheless, the minimal pair (31) proves that a remote call follows not the Basic Principle, but the linear one:

(31) a. *Дорогие* ***гости-и!*** *К столу-у!* vs. b. *Гости* ***дорогие-е!*** *К столу-у!*

In (31b), the accented sentence-final adjective *дорогие* demonstrates that the accent placement here follows the linear principle because the "syntactic" (or the "basic") accent-bearer in (31b) would be not the word *дорогие* but the word *гости*.

In sum, the linear accent placement principle is employed in Russian to designate the distance between the interlocutors. Within the short distance communication the accent-bearer is sentence-initial, whereas within the remote distance communication the accent bearer is sentence-final.

3. "Illocutionary" principle

The "illocutionary" principle presumes that in certain specific illocutions (dreaming, being puzzled, recollecting, urgent requests) the accent placement can follow individual rules, cf. sentence (32) with the accent on the object *бутербродик.ACC* and sentence (33) with the "basic" accent on the adjunct *колбаской.INSTR*:

(32) *Вот бы нам сейчас дали* ***бутерброо-одик*** *с колбаской!*

(33) *Мама дала мне с собой в школу бутербродик с* ***колбаской***.

The accent placement in (32) obviously violates the Basic rules.

The rules of individual accent placement are based on 1) the taxonomy of illocutionary meanings, and 2) the syntactic taxonomy of noun groups. The taxonomy of speech acts and noun groups specify the types of sentences demanding a shift. For instance, sentence (32) displays the accent shift within the object NP from *колбаской* to *бутербродик* because (32) expresses a specific state of dreaming. Whereas in (33) that is just a statement not combined with any possible attendant illocutionary meanings the accent-bearer within the object (which, in its turn, has a syntactic structure of a noun group with an adjunct) is the word *колбаской*. It is selected in full accord with the Basic Principle.

Thus dreams and recollections are characterized by the following accent shifts. In an NP which itself is a privileged accent-bearer in terms of the Basic Hierarchy and has dependent nodes the accent moves from 1) the family name to the first name (cf. examples (34) where (34a) is a sentence of dreaming, while (34b) is an ordinary statement) and 2) from the complement (or the adjunct) to the head (cf. examples (35)):

(34) a. *Вот бы к нам сейчас сюда* ***Ва-асю*** *Иванова!* — b. *Я сейчас приведу сюда Васю* ***Иванова***;

(35) a. *Вот бы сейчас* ***пирожко-ов*** *с капустой!* — b. *Мама дала нам с собой пирожков с* ***капустой***.

However, there are no any expected shifts from the noun to the adjective (cf. (36)), from the patronymic to the name (cf. (37)), from the name to title (cf. (38)), and from the second coordinate group to the first coordinate group (cf. (39)):

(36) **Вот бы сейчас* ***квашеной*** *капустки!*

(37) **Вот бы к нам сейчас сюда* ***Виктора*** *Ивановича!*

(38) **Вот бы к нам сейчас сюда* ***сержанта*** *Иванова!*;**Вот бы к нам сейчас сюда* ***почтальона*** *Печкина!*

(39) **Вот бы к нам сейчас сюда* ***Танечку*** *и Ванечку!*

The accent-bearers here are the same as those selected by the Basic Principle:

(40) *Вот бы сейчас квашеной* ***капустки!***; *Вот бы к нам сейчас сюда Виктора* ***Ивановича*** */ сержанта* ***Иванова!*** */ почтальона* ***Печкина!*** */ Танечку и* ***Ванечку!***

Accent shifts are also displayed by sentences of identification:

(41) *Да это ж* ***Вася*** *Иванов!*; *Алё, здраствуйте, это* ***Вася*** *Иванов, вы меня помните?*; *Еще ж* ***Васи*** *Иванова не хватает. Он придет?*

In sentences with illocutionary markers *какой*, *когда*, *куда* designating hesitation, distrust, or being puzzled the accent-bearer is the illocutionary word:

(42) *И* ***куда-а*** *только он запропастился, не знаю!*; *И* ***кака-ая*** *еще может быть прогулка в такую погоду, не представляю себе?!*

The accent placement in the sentences where the words *какой*, *когда*, *куда* are complementizers follows the basic rules:

(43) *Не знаю, куда он* ***запропастился***; *Не представляю себе, какая прогулка может быть в такую* ***погоду***.

In urgent requests the accent moves from the object to the imperative:

(44) *Ну* ***купи*** *мне эту куклу, ну пожалуйста*; ***Зайдите*** *к хирургу*; *Не* ***забывайте*** *свои вещи*.

Cf. the basic selection in (45):

(45) *Купи мне, пожалуйста, эту* ***куклу***; *Зайдите к* ***хирургу***; *Не забывайте свои* ***вещи***.

Thus in dreams, recollections, hesitations, identifications, and insistent requests the accent placement may differ from the basic one. The accent placement principles in such sentences are based on the speech acts taxonomy and the syntactic classification of noun groups.

4. Text principle

(46) A sentence as an element of a coherent discourse can conclude specific markers designating text incompleteness. For example, in sentence (46) the fall on the object *пиджак.ACC* is a focus marker, whereas the rise on the verb *снял.*V_{fin} designates text incompleteness, cf. Figure 2. (The rises and falls of frequency in examples below are marked by up and down arrows respectively. The arrows are placed after the words carrying pitch-accents. The context of the sentences in question is enclosed in angle brackets)

(47) *Я тогда* ***пиджак***↘ ***снял***↗, <*на* ***почту***↘ *поскорее* ***побежал***↗, *жене телеграмму о снижении цен на* ***фрукты***↘ ***дал***↗. *Потому что личный покой прежде всего.*

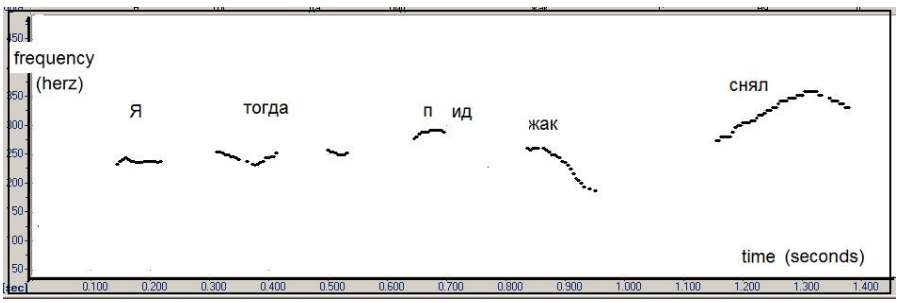


Figure 2. Frequency tracing of sentence (46)

Thus the accent-bearer of the focus the object *пиджак* follows the Basic Principle. Whereas the rise on the finite verb can be accounted for only by positing an additional and autonomous accent-bearer to designate text incompleteness. The finite verb as a marker of text incompleteness is positioned sentence-finally and its selection is not governed by the basic rules. In English this type of accent placement is obviously barred, in Russian, however, it is highly frequent in colloquial speech:

(48) *Я из* ***комнаты↘ выхожу↗****, когда вхожу, она уже наполовину пустая; И когда* ***обратно↘*** *уже* ***бежали↗****, ммм сейчас… ; И вот во сне меня какое-то чувство* ***страха↘ охватило↗****…*

5. Culturally bound principles

Intonation of begging, praying, verse-reading is characterized by pitch-accents and accent placement principles distinct from the basic ones. For instance, the Russian traditional Orthodox liturgical reading has specific intonation that differs from that of everyday speech or from that found in other liturgical traditions[5]. A number of questions then arise:

- what does such intonation express?
- what type of pitch-accents does it have?
- what words do the pitch-accents fall on?

The analysis below is based on recordings of the Morning and Evening cycles of prayers recited by contemporary Russian priests in Church Slavonic. I argue that the Russian liturgical intonation does not express illocutionary or any other language-specific distinctions. It only serves to divide a prayer into lines and a series of prayers into single prayers. The marker of a line is a rise on the tonic syllable of the initial phonetic group followed by high and level groups up to the end of the line. Thus in line (48) the accent on *ангельскую* marks the onset of the line, while the default accent-bearer *песнь*, "inherited" from the Basic Principle, can remain unaccented, cf. Figure 3.

(49) *…и* ***ангельскую*** *песнь вопием Ти, Сильне…*

The marker of a prayer is a similar rise and longer duration on the terminal phonetic group of a prayer. Thus in line (49) the accent on *лукаваго* marks the end of the prayer (while the accent on *избави* designates the onset of the line), cf. Figure 4.

(50) *…но* ***избави*** *насъ от* ***лука-ава-аго-о-о***.

[5] The Russian traditional Orthodox liturgical prosody has been thoroughly investigated in [Прохватилова 1999]. In my paper I am only attempting to answer the question what are the words carrying pitch accents in this reading.

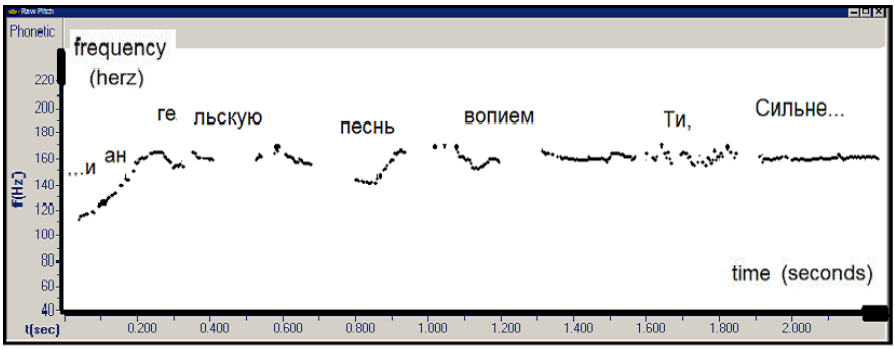


Figure 3. Frequency tracing of example (48)

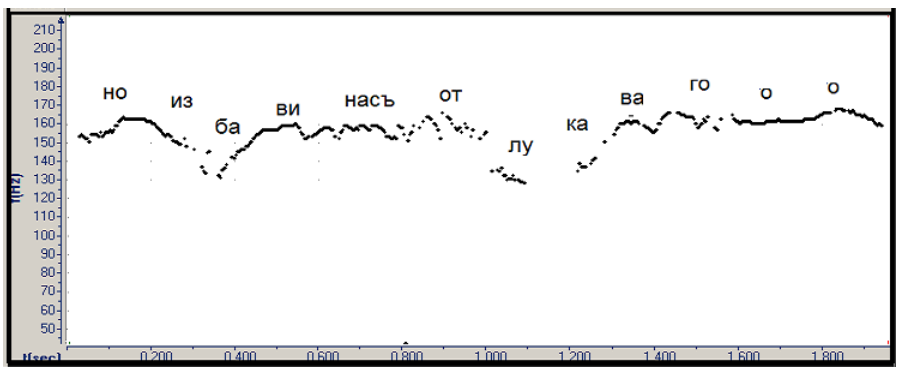


Figure 4. Frequency tracing of example (49)

Thus the Russian liturgical reading has a highly concise system of intonational text-segmenting, whereas many other liturgical traditions preserve all language-specific distinctions.

6. Conclusion

The account of accent placement principles proposed here shows that the Basic Principle is not the only one in oral discourse: within the constraints imposed by the context, the accent placement can be also governed by a diversity of peripheral intonational strategies.

References

1. *Baranov A. N., Kobozeva I. M.* 1983. Semantics of General Questions in Russian [Semantika Obshchikh Voprosov v Russkom Iazyke (Kategoriia Ustanovki)]. Izvestiia Akademii Mauk SSSR. Seriia Literatura I Iazyk, 42 (7) : 263–275.
2. *Bolinger D.* 1958. A Theory of Pitch Accent in English. Word, 14 : 109–149.
3. *Bolinger D.* 1961. Contrastive Accent and Contrastive Stress. Language, 37 : 83–96.
4. *Chafe W.* 1976. Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View. Subject and Topic.
5. *Dryer M. S.* 1996. Focus, Pragmatic Pressupposition, and Activated Propositions. Journal of pragmatics, 26.
6. *Enkvist N. E.* 1979. Marked Focus: Functions and Constraints. Stuides in English Linguistics for Randolph Quirk : 134–152.
7. *Halliday M.* 1963. The Tones of English. Archivum Linguisticum, 15 (1).
8. *Halliday M.* 1967. Intonation and Grammar in British English.
9. *Halliday M.* 1967. Notes on transitivity and theme in English . Part 2 . Journal of Linguistics, 3 : 199–244 .
10. *Ianko T. E.* 1991. Communicative Structure with No-ingerent Theme [Kommunikativnaia Struktura s Neingerentnoi Temoi]. Nauchno-Tekhnicheskaia Formatsiia, 2 (7).
11. *Ianko T. E.* 2008. Intonational Strategies of Russian Speech in Comparative Aspect [Intonatsionnye Strategii Russkoi Rechi v Sopostavitel'nom Aspekte].
12. *Jaeger F., Wagner M.* When Warriors Mourn Longer. Testing Some Phonetic Predictions of Current Focus Theories, available at: web.mit.edu/~chael/www/JaegerWagner03_SemFest.pdf
13. *Kadmon N.* 2009. Some Theories of the Interpretation of Accent Placement.
14. *Kovtunova I. I.* 1976. Modern Russian Language. Word Order and Actual Sentence Segmentation [Sovremennyi Russkoii Iazyk. Poriadok Slov I Aktual'noe Chlenenie Predlozhenii].
15. *Kodzasov S. V.* 1996. Phrase Accentuation Laws [Zakony Frazovoi Aktsentuatsii]. Prosodicheskii Stroi Russkoi Rechi.
16. *Kuz'micheva V. K.* 1964. Address Intonation in Modern Literary Russian Language [Intonatsiia Obrashchenii v Sovremennom Russkom Literaturnom Iazyke].
17. *Nikolaeva T. M.* 1996. Balkan Prosody. Word – Statement – Text [Prosodiia Balkan. Slovo – Vyskazyvanie – Tekst].
18. *Pal'ko M. L.* 2009. Addresses Prosody in German Language in Comparison with Russian [Prosodiia Obrashchenii v Nemetskom Iazyke v Sopostavlenii s Russkim]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2009" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2009").
19. *Prince E. F.* 1981. Toward a Taxonomy of Given-new Information. Radical pragmatics : 223–255.
20. *Prokhvatilova O. A.* 1999. Orthodox Sermon and Prayer as Modern Speech Phenomenon [Pravoslavnaia Propoved' I Molitva kak Fenomen Sovremennoi Ustnoi Zvuchashchei Rechi].

21. *Russian* Grammar, 2. 1982.
22. *Sasse H. J.* 1987. The Thetic/Categorical Distinction Revisited. Linguistics, 25 (3) : 511–580.
23. *Schwarzschild R.* 1999. Givenness, AvoidF and other Constraints on the Placement of Accent. Natural Language Semantics, 7 : 141–177.
24. *Steedman M.* 2007. Information Structural Semantics for English Intonation. Topic and Focus: Cross-Linguistic Perspectives on Meaning and Intonation.
25. *Svetozarova N. D.* 1993. Accent-Rhythmical Innovations in Russian Spontaneous Speech [Aktsentno-Ritmicheskie Innovatsii v Russkoi Spontannoi Rechi]. Problemy Fonetiki, I.
26. *Zwicky A. M.* 1986. Forestress and afterstress. OSU WPL, 32 : 46–62.

КАК РАЗНЫЕ ЯЗЫКИ КЛАССИФИЦИРУЮТ ПРЕДМЕТЫ БЫТА[1]

Б. Л. Иомдин (iomdin@ruslang.ru)
Институт Русского Языка, Москва, Россия

А. Ч. Пиперски (apiperski@gmail.com)
МГУ, Москва, Россия

М. М. Руссо (apiperski@gmail.com)
Институт Лингвистики, Москва, Россия

А. А. Сомин (somin@tut.by)
РГГУ, Москва, Россия

Исследуются классификации бытовых предметов на материале более 40 языков. Показано, что большинство классов являются «скрытыми» — не имеют нейтральных общепринятых названий (ср. офиц. предметы личной гигиены и разг. мыльно-рыльное). Кроме того, наборы и состав классов в разных языках существенно различаются.

Ключевые слова: быт, бытовая лексика, бытовые предметы, классификация.

[1] Under partial financial support by Russian Foundation for Humanities (Project No. 10-04-00273a), Fundamental Research Program of History and Philology Branch of the Russian Academy of Sciences, and a President grant for leading scientific schools of Russia (No. NSh-4019.2010.6).

HOW DIFFERENT LANGUAGES CATEGORIZE EVERYDAY ITEMS

B. L. Iomdin (iomdin@ruslang.ru)

V. V. Vinogradov Russian Language Institute,
Russian Academy of Sciences, Moscow, Russian Federation

A. Ch. Piperski (apiperski@gmail.com)

M. V. Lomonosov Moscow State University, Moscow,
Russian Federation

M. M. Russo (rousseau@mail.ru)

Institute for Linguistics, Russian Academy of Sciences, Moscow,
Russian Federation

A. A. Somin (somin@tut.by)

Russian State University for the Humanities, Moscow,
Russian Federation

Classifications of everyday items (category words for clothing, stationery, personal hygiene, beauty products etc.) are studied. A survey of 41 languages was performed. Several results are reported, in particular:

1. Speakers of some languages provide generic terms relatively easy, while for speakers of other languages it is often diffi cult to perform this task.
2. Some items (such as keys, ear plugs, umbrellas) are virtually unclassifi able in all languages.
3. All languages have covert classes without well-established names (such as personal hygiene or data storage), and people either resort to awkward offi cial phrases like Russian предметы личной гигиены or highly colloquial occasional words like Russian мыльно-рыльное. For items belonging to such classes, high variation of category words was observed.
4. Classes existing in several languages often overlap and include different items. So, посуда in Russian corresponds to dishes, cookware and cutlery in English.

Possible areas of further research are discussed, including studies of language acquisition and bilingualism and comparisons with folk biology and folksonomies.

Key words: everyday life, everyday life vocabulary, everyday life objects, classification.

Introduction

The idea of the present study was born at our academic seminar devoted to developing an explanatory and encyclopedic thesaurus of Russian everyday life terminology. In Iomdin 2009, 2010, 2011 this lexicon has been shown to be treated very differently in dictionaries, industrial standards, and usage; uniform lexicographic definitions of such words are very difficult to produce. The thesaurus is being developed by a group of researchers led by Boris L. Iomdin. The group members helped to perform the study at all stages (organizing the survey, recruiting participants, collecting and discussing the results). We would like to especially thank those members who made many valuable contributions: Anna Kadykova, Anastasiya Lopukhina, Varvara Matissen-Rozhkova, Pavel Vasilyev, Fedor Vinokurov, and Anna Vybornova2.

1. Classification

We conceive our dictionary as a thesaurus where similar objects are grouped together, which allows an easy search for information on a certain object or group of objects. However, when trying to classify the lexicon, we were faced with problems of different kinds.

1.1. Unclassifiable items. Certain items simply defied any reasonable categorization. These included *ваза* 'vase', *веер* 'hand fan', *зажигалка* 'lighter', *ключ* 'key', *открытка* 'postcard', *очки* 'glasses', *носовой платок* 'handkerchief', *полотенце* 'towel', *штора* 'blind, curtain', etc.

1.2. Covert classes. Several classes that obviously exist in speakers' minds do not have natural names. For example, most travelers pack their toothbrush, toothpaste, soap, shampoo, sponge etc. together, but no good Russian word exists for this class. If asked, or urged, to use a superordinate, people either resort to awkward official phrases like *предметы личной гигиены* 'personal hygiene items', or highly colloquial occasional words like *умывалки* [from умываться 'to wash oneself']. In fact, most superordinates prevail either in official documents (e. g. *парфюмерия* 'perfumery', *бытовая химия* 'household chemistry', *писчебумажные принадлежности* 'stationery supplies', *чулочно-носочные изделия* 'hosiery'), or in colloquial texts such as blogs (*мыльно-рыльное* ≈ 'soap and stuff' / 'phiz wash', *косметика* 'make-up',

[2] We would also like to thank Julia Khaleeva who calculated all statistics for us; Vladimir Belikov and Aleksandrs Berdicevskis who made valuable comments; Elena Muravenko, Elizaveta Kushnir and Hugo Dobbs who promoted the survey among speakers of various languages; Anastasia Zaytseva who commented on Japanese; professors and students from the Slavistic Institute of Karl-Franzens-Universität (Graz, Austria) and the Department of Foreign Languages of University of Bergen (Norway); subscribers of http://community.livejournal.com/by_mova; and everyone who submitted answers for our survey.

аксессуары ‘accessories’, *шмотки* ‘duds’, *прибамбасы* ‘gismos’)[3]. Consider two examples covering similar topics, where *игрушки* ‘toys’ is the only item named in the same way:

(1) Согласно договору о патронате воспитателю перечисляются заработная плата и денежные средства на содержание ребенка (питание, приобретение предметов хозяйственного обихода, личной гигиены, медикаментов, канцелярских товаров, игрушек и др.) (Russian Tax Courier, 2008, No.13–14)

[≈ ‘According to the patronage agreement, salary and money for upkeeping the child (nutrition, purchase of household objects, medicaments, stationery, toys, etc.) is transferred to the tutor’s account]

(2) 2000 руб единовременно на весь год — родит. комитет — вода, рыльномыльное, игрушки, подарки-поздравлялки, канцелярка и т. п. (http://www.mamask.ru/forum/index.php?topic=11641.0)

[≈ ‘2000 roubles for the whole year as flat payment by the parent org: water, soap and stuff, toys, prezzies and gz, office stuff, etc.’, a highly colloquial forum message].

1.3. Vague classes. Some other classes with relatively established names are too fuzzy: for instance, Russian *галантерея* ≈ ‘haberdashery’ for different speakers might refer to handkerchiefs, ties, gloves, belts, bags, purses, threads, needles, pins, umbrellas, combs, hair rollers, beads, costume jewellery, mirrors, clothes hangers, etc. Similar phenomena were discussed in semantic literature in the 1970s, consider e. g. Kempton 1978. Cruse 1995 reports on a study where some 200 American college students were asked to estimate sixty household items as good or bad examples of furniture.

For some classes, standard dictionaries often provide vague, hardly translatable explications of doubtful usability, e. g. *ширпотреб* ‘Товары широкого спроса и массового производства’ [‘mass demand and mass production goods’], *утварь* ‘Совокупность предметов, необходимых в обиходе, в какой-л. области жизни’ [‘a range of items needed in common use, in one of life spheres’]; *аксессуары* ‘1. Мелкие предметы сценической обстановки, бутафория. 2. Принадлежность чего-л.; сопутствующие предметы’ [‘1. Small items of stage set, props. 2. Accessories of something, accompanying items’] (Kuznetsov 1998) (see below for more on Accessories).

Considering all this, we decided to investigate into the subject with the purpose to find out whether unclassifiable items, covert and vague classes are universal or language specific. Semantics of category words was studied a lot (see e. g. Wierzbicka 1985 or Taylor 1995); some studies are also under way in the domain of natural

[3] Data obtained via several searches in Russian blogs (blogs.yandex.ru) and in the ConsultantPlus juridical information system (base.consultant.ru).

ontologies, cf. Mihatsch 2007. However, this topic has not received much attention cross-linguistically. With this aim in view, we launched a survey, to be outlined below.

2. Survey

Under http://www.lingling.ru/useful/pics-survey-en.php, we posted 33 clickable images depicting the following items: suitcase, pot, notepad, toothbrush, receipts, toy blocks, eraser, sock, glasses, pencil, blanket, passport, gloves, tacks, umbrella, ruler, make-up bag, ear plugs, handkerchief, CD, vase, barrette, charger, keys, spoon, soap, slippers, teapot, lipstick, table cloth, high heels, comb, glass. The following task was given:

Please add two headings for each image according to the examples below:
[image of a chair] chair furniture
[image of a bed] bed furniture
[image of an iron] iron appliances
Write in your native language. Choose words that you use yourself when speaking with your family members. If you find it difficult to add headings to some of the pictures, leave those fields blank.

563 participants aged 12 to 64 (mean age 30) submitted their results in 41 languages: Albanian, Arabic, Azerbaijani, Belorussian, Bulu, Catala, Chinese, Croatian, Czech, Dutch, English, Finnish, French, Georgian, German, Greek, Hebrew, Hindi, Hungarian, Icelandic, Indonesian, Italian, Japanese, Korean, Latvian, Lithuanian, Luxembourgish, Norwegian, Occitan, Persian, Polish, Portuguese, Romanian, Russian, Serbian, Slovenian, Spanish, Swedish, Tagalog, Turkish, Ukrainian.

3. Results

3.1. Categorization difficulties in different languages. Our survey allowed making preliminary cross-linguistic observations on how easily speakers of a given language can use generic terms. The subjects were instructed to leave a field blank if they had difficulties filling it in. If speakers of language L_1 fail to provide generic terms more often than speakers of language L_2, then L_2 is probably richer in generic terms than L_1 and uses them more frequently.

However, when analyzing the data, certain precautions had to be taken. First of all, samples must of course be rather large: it is not enough to count the mean number of blank fields in the responses of e. g. three speakers of a language. Second, it is well understandable that not all participants of the survey invest as much zeal and enthusiasm into this work as the researchers would wish. Each response had to fulfil two criteria in order to be counted: (1) all 33 specific terms should be provided; (2) at least 10 generic terms should be provided (in other words, no more than 23 gaps are allowed).

This left us with representative samples (≥ 15 responses) for five languages: Belorussian, English, German, Norwegian, and Russian. Even though this sample is obviously ill-balanced genealogically as well as geographically, we can see that the mean number of gaps varies significantly even within this sample:

Language	Responses (total)	Responses (suitable)	Mean number of gaps
Norwegian	21	18	8.2
German	44	37	5.1
Russian	230	188	4.2
Belorussian	32	17	3.2
English	21	18	1.7

This shows that speakers of Norwegian had the greatest difficulty finding generic terms, while speakers of English had the fewest problems with this[4]. The fact that such closely related and similar languages as Russian and Belorussian pattern similarly in respect to the number of gaps (they occupy neighbouring rows in the table above) supports the assumption that the amount of gaps in generic terms is not arbitrary but constitutes an important characteristic of a language[5].

One of the important consequences of this fact is that generic terms cannot serve as markers of linguistic identity because of high fluctuation in their use. For example, it is shown elsewhere in this volume (Piperski 2011) that Serbian and Croatian speakers make less notice of differences between their languages that concern generic terms than of differences in specific terms even though the sociolinguistic situation in the Balkans favours language awareness.

Furthermore, as we will show below, languages vary considerably in how much speakers agree when using generic terms for the same items.

3.2. Unclassifiable items. In total, 11 items were not classified at all by 20% or more respondents: keys (48.4%), ear plugs (42.4%), table cloth (32.1%), umbrella (30.8%), glasses (29.5%), handkerchief (25.5%), make-up bag (24.8%), CD (24.0%), comb (23.5%), receipts (22.4%), barrette (21.5%). The answers of respondents who did submit generic terms for these items exhibit great variation. E.g. in Russian, 230 respondents submitted 63 different unique superordinates for ear plugs (with a maximum of 7 answers (3.0%) reached in vague *личные вещи* 'personal belongings') and 53 for keys (a maximum of 11 answers (4.8%) reached in helpless *ключи* 'keys'), while

[4] The hypothesis that English develops category words more easily than other languages could be supported by the following funny observation. In current English usage, suffixes *-wear* and *-ware* tend to be mixed: Google search yields thousands of occurences for *footware*, *eyeware*, *outerware* etc. as well as *cookwear*, *silverwear*, *glasswear* etc. This might mean that speakers of English start to consider this a single suffix with a general meaning 'category of artifacts'.

[5] Of course we should bear in mind that all Belorussian speakers live in Russian language environment, so that they are inevitably strongly influenced by Russian.

e. g. for blocks, only 8 different superordinates were offered, other than *игрушки* 'toys' used by 207 (90.0%) respondents. For some of these items, these numbers were comparable in all languages, but there are exceptions.

The umbrella was classified by more than half of Japanese speakers as 雨具 [*amagu*]; web search confirms that this category is indeed well established in the Japanese language and includes umbrellas, raincoats, rubber boots, tents etc (see e. g. http://shopping.yahoo.co.jp/category/2585). *Rain*()*gear* in English was used by 25% respondents, *Regenschutz* in German by 17.5% respondents, *regnutstyr/regntøy* in Norwegian by 10.5% respondents each. According to the submitted results, this category is virtually non-existent in other languages participating in the survey: e. g. in Russian only one respondent (0.5%) used *защита от дождя* 'rain protection' and another one used *средство от дождя* 'aid against rain'.

An interesting tendency we observed in many languages is categorizing unclassifiable items like these (but not only them!) under a special extremely nebulous class called Accessories. This word was borrowed into and widely used in almost half of the languages studied, and is the most frequent category word used by the survey participants[6].

3.3. Well-established classes. Surprisingly, only one class appears to have a distinct name in most languages, namely Toys. For almost all languages, there is one word with this meaning that gets more than 60% of answers (and usually much more, cf. *игрушки* (91%) in Russian mentioned in the previous subsection) and little variation. An interesting example is presented by the Documents class: it appears to be well-established in most languages, but not quite so in some others (mostly Germanic ones[7]). For passport, we got *дакументы* in Belorussian (100%), *dokumenty* in Polish (100%), *документы* in Russian (92%), *documentos* in Spanish (67%), etc. However, in Norwegian, the best result was *reisedokument* (21%), with 17% of respondents who couldn't provide any answer and the rest using various other words (*dokument, identifikasjon, identifikasjonspapir, identitetsbevis, identitetspapir, legitimasjon, reisepapirer*). Similar situation happened in Dutch: *document* (25%) and many other answers (*reisdocument, identiteitspapieren, indentificatiebewijs, officiële papieren, paperassen, reisbenodigdheden*), English, Swedish and (outside the Germanic branch) also Japanese, where 36% of respondents could not come up with any answer and different words were offered by the rest.

We also discovered some language-specific classes. These include Raingear in Japanese described above and Luggage in English and Polish. For the suitcase, 66% used *bagaż* in Polish and 65% used *luggage* in English, while the next closest result was 32% for German *Gepäck*, with similar or lower numbers for other languages. 22% of Russians did not come up with any answer, and the leading one was *сумки* (29%).

---

[6] Note also the Tagalog word *gamit*, which was used by our surveyees as the superordinate for a considerable number of different objects. Cf. "The term *gamit* means several things. Its definition as a Filipino word is legion. In Tagalog colloquial term, it means an object that has several utilitarian purposes or simply a utilitarian object with specific usage in a particular space and time" (Ruston Ocampo Banal Jr. Gamit: subjectifying objectivity).

[7] One of the striking results of our survey that requires much more data to be confirmed is that cognate languages often use similar categorization patterns even when words they choose are not related.

3.4. Covert classes. Several other classes have no good names in most analyzed languages. These are: (a) personal hygiene, (b) appliances, (c) stationery, and (d) data storage. For all items belonging to these classes, the variation (calculated as the total number of different answers in all languages divided by the number of non-empty answers) was twice as high as for the well-established class of toys. Most answers here are compounds or word combinations[8], and no answer was given by 50 % of respondents or more. The highest results for all languages in these groups are (a) *Hygieneartikel* in German (45 %), (b) 電化製品 [*denkaseihin*] in Japanese (45 %), (c) *skrivesaker* in Norwegian (42 %), and (d) *носьбіт інфармацыі* (32 %) in Belorussian. Interestingly, in many cases there is one leading word or root (depending on the morphological structure of the language) which occurs in various compounds or word combinations. E.g. generic names for items from group (c) in Russian mostly contain the root *канц-*, and counting together all answers containing this root (*канцелярские товары, канцтовары, канцелярские принадлежности, канцелярская продукция, канцелярские мелочи, канцелярские предметы, канцелярское изделие, канцелярия, канцелярка, канцелярщина*, etc.) we get as much as 81 %. The same applies to roots *toilet-* and *hygien-* for group (a) or to *technic-* and *electro-* for group (b), in various forms depending on the language.

3.5. Overlapping classes. Some classes exist in many languages but include different items. Let us give two examples. For four images that are grouped together as *посуда* in Russian, different languages have several classes. Cf. the summarization table, characteristic for lexical typology (cf. Hjelmslev 1957, Haspelmath 2001, Koch 2005), which only uses data from languages where 50 % or more respondents agree on certain generic term:

Item	Russian	Belorussian	English	Norwegian	German	Japanese	Arabic
Pot	*посуда*	*посуд*	*cookware*	*kjøkkenutstyr*	?	調理器具 [chorikigu]	أدوات (الـ) مطبخ ['adwa:t (al-)maṭbaḫ]
Teapot			?		*Geschirr*	食器 [shokki]	?
Wineglass				?			
Spoon				*bestikk*	*Besteck*		

Russian word *посуда* is indeed rarely translated as a similar generic term in English: interpreters use various strategies to avoid direct translation. Cf. examples from parallel corpora:

[8] According to Mihatsch 2007, superordinates are often morphologically more complex than subordinates; our data clearly support this hypothesis.

(3) Проходивший в это время по коридору старший доктор, услыхав звон разбитой посуды и увидав выбежавшую раскрасневшуюся Маслову, сердито крикнул на нее (Л. Н. Толстой, Воскресение).

(4) The head doctor, who was passing at that moment, heard the sound of breaking glass, and saw Maslova run out, quite red, and shouted to her (Lev Tolstoï, Resurrection, translated by Louise Maude).

(5) Хохлушка в платке внесла поднос с посудой, потом самовар (А. П. Чехов, Красавица).

(6) A Little Russian peasant woman in a kerchief brought in a tray of tea-things, then the samovar (Anton Chekhov, The Beauties, translated by Constance Garnett).

A less clear situation takes place with the Clothing class. In Russian, it distinctly falls into two subclasses: *одежда* 'garments' and *обувь* 'shoes'. In most other languages, the respondents disagree as to which of the four items in the survey fall into which class. The following table summarizes the results (words in brackets correspond to answers that received 40 % to 50 % votes).

Item	Russian	Belorussian	English	Norwegian	German	Japanese	French
High heels	*обувь*	*абутак*	?	?	*Schuhe*	靴 [kutsu]	?
Slippers			(*footwear*)	(*sko*)	(*Schuhe*)	?	*vêtements*
Socks	*одежда*	(*вопратка / адзенне*)[9]	(*clothing*)	*klær*	*Kleidung*	衣類 [irui]	
Gloves		?		?	?	?	?

Google search results[10] seem to confirm that *clothes/clothing* and *shoes/footwear* do not constitute same-level classes in English, as do *одежда* and *обувь* in Russian:

[9] In Belorussian, two different superordinates are used for socks, both of which have received more than 40 % votes. *Вопратка* is explained in dictionaries as outerwear, but is used to describe all clothing as can be seen from the survey and confirmed by web searches. This could have happened either under the influence of Russian that does not have a separate word for outerwear (only word combinations like *верхняя одежда*) and does not specially name it unless needed, using the neutral word *одежда* for all types of clothing (so *вопратка* develops the same meaning as *одежда* and starts competing with *адзенне*), or by following the tendency of lexical differentiation of closely related Russian and Belorussian languages (the word *адзенне* is akin to the Russian *одежда*).

[10] Data obtained on January 31, 2011. The total number of hits given by Google varies considerably and may only serve as a very approximate estimate.

	Google hits
clothing site:uk	48,500,000
shoes site:uk	47,800,000
footwear site:uk	2,190,000
“clothing and shoes” site:uk	33,400
“clothing and footwear” site:uk	105,000
“clothing such as shoes” site:uk	2,050
одежда site:ru	21,700,000
обувь site:ru	14,800,000
“одежда и обувь” site:ru	16,700,000

The same might be the case in Arabic, where ملابس [mala:bis] ‘clothing’ is used much more frequently than أحذية [’aḥḏi:ya] ‘shoes’. In the survey, no Arabic speakers used *’aḥḏi:ya*, and two of them even referred high hills to *mala:bis*.

4. Possible areas of further research

4.1. Language acquisition and bilingualism studies. Superordinate categories play an important role in language acquisition. Reportedly small children master well enough many category names that are well established in a language, including those of everyday objects they use. This might be a way of finding out which categories play a key role in a language. Consider e. g. a characteristic quotation about Russian children: “К 3 годам среди родовых наименований появляются более «книжные» термины: фрукты, овощи, животные, посуда, насекомые, обувь, одежда, транспорт и т. п. <…> В речи ребенка появляются конструкции, <…> соотносящие видовое и родовое <…>: Кастрюля — это посуда. Чашка — это посуда” (Yeliseeva 2006) [‘3-year-olds start using more “bookish” generic terms: fruits, vegetables, animals, dishes, insects, footwear, clothing, transport, etc. The child starts producing constructions correlating specific and generic terms: Pot is dishes. Cup is dishes’]. It is clear from the awkward translation we provided that Russian-speaking kids learn other hierarchies that English-speaking ones. Bilinguals are especially interesting here, since they might mix up different classification strategies (see e. g. Malt & Pavlenko 2009 who report a study of English-Russian bilinguals naming cups, mugs and glasses of different types).

4.2. Folk biology. Further research of everyday items classification in different languages might use the experience of folk biology, which studies linguistic classifications of animals and plants (Berlin 1992, Atran 1990). It also describes covert categories that have no special names in languages but apparently exist in speakers’ minds. Often these even include the highest taxa, which are animals and plants (Berlin 1973: 266–267). Latin started to use *plantae* for all plants only in the 13th century, English and French accepted this term only in the 16th century (Kupriyanov 2005: 14). It is suggested that generic terms for animals and plants appear when a language becomes a written one (Slaughter 1982). Covert classes in folk biology and in everyday

items are evidently similar. In folk biology, several techniques for revealing covert classes through speaker surveys have been developed (cf. Hays 1976), which could be used in deeper studies of everyday items classification. Probably their names appear in professional sublanguages before progressing into standard language and then into colloquial speech; this is subject to further investigation.

Scholars of folk biology believe that folk taxa in world languages are organized into a hierarchical system of levels, or ranks: folk kingdom (e. g. animal, plant), life form (e. g. bug, fish, bird, mammal/animal, tree, herb/grass, bush), folk species (gnat, shark, robin, dog, oak, clover, holly), folk specific (poodle, white oak), and folk varietal (Berlin 1992: 15–25). The levels are thought to be universal, unlike the taxa. Folk species can unite into folk series: chains of species that look similar to the speakers. These chains only rarely have names (see Merkulova 1967) and obey several universal tendencies. For instance, longer series less frequently get names. This might be explained by the fact that speakers cease to consider remote elements of a long chain as similar enough (Kupriyanov 2005:15). Such phenomena might be present in our case, too. S. Atran (Atran 1990) mentions that linguistic classifications of artifacts provide much more freedom for intersecting classes and several alternative groupings: e. g. a piano could be considered a musical instrument or a piece of furniture. However this might vary for different items and in different languages.

4.3. Folksonomics. This is another domain thoroughly investigated in recent years. It studies classification emerging from the collective action of users who tag resources with an unrestricted set of key terms, such as flickr.com (Veres 2006). Since on many websites like these objects of everyday use are discussed and tagged, it would be interesting to compare these tag sets with the categories we describe.

References

1. *Atran S.* 1990. Cognitive Foundations of Natural History.
2. *Berlin B.* 1973. The Relation of Folk Systematics to Biological Classifications and Nomenclature. Annual Review of Systematics and Ecology, 4 : 259–271.
3. *Berlin B.* 1992. Ethnobiological Classification.
4. *Cruse D. A.* 1995. Lexical Semantics.
5. *Eliseeva M. B.* 2006.On Lexical Development of a Child of Early Age [O Leksicheskom Razvitii Rebenka Rannego Vozrasta]. Logoped v Detskom Sadu,1 (10).
6. *Haspelmath M.* 2001. Typologie des Langues et les Universaux Linguistiques. Manuel International.
7. *Hays T. E.* 1976. An Empirical Method for the Identification of Covert Categories in Ethnobiology. American Ethnologist, 3 (3) : 489–507.
8. *Hjelmslev L.* 1970. Sémantique Structurale. Essais linguistiques : 96–112.
9. *Iomdin B. L.* 2009. Everyday life Vocabulary. Search of Standard [Terminologiia Byta. Poiski Normy]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2009" (Computational Linguistics

and Intelligent Technologies: Proceedings of the International Conference “Dialog 2009”), 8 (15) : 127–135.
10. *Iomdin B. L.* 2010.Russian Everyday Object Vocabulary: Ontology and Description [Russkaia Bytovaia Predmetnaia Leksika: Ontologiia I Opisanie]. Trudy 33 Konferentsii Molodykh Uchenykh I Spetsialistov IPPI RAN “Informatsionnye Tekhnologii I Sistemy (Proc. of the 33 Conference "Information Technologies and Systems"), avaiable at: http://www.itas-proceedings.iitp.ru/pdf/1569326461.pdf. (In Russian).
11. *Iomdin B. L.* 2011. Materials for Everyday Terminology Dictionary. ‘JERSEY’: An Example of Dictionary Paragraph [Materialy k Slovariu-tezaurusu Bytovoi Terminologii. ‘SVITER’: Obrazets Slovarnoi Stat’i]. “Slovo I Iazyk”. Sbornik k 80-letnemu Iubileiu Akademika Iu.D. Apresiana : 394–408.
12. *Kempton W.* Category Grading and Taxonomic Relations: a Mug is a Sort of a Cup. American Ethnologist, 5: 44–65.
13. *Koch P.* 2005. Aspects Cognitifs d’une Typologie Lexicale Synchronique. Les Hiérarchies Conceptuelles en Français et dans d’Autres Langues. Langue française, 145 : 11–33.
14. *Koptjevskaja-Tamm M., Vanhove M., Koch P.* 2007. Typological Approaches to Lexical Semantics. Linguistic Typology, 11 (1) : 159–185.
15. *Kupriianov A. V.* 2005. Prehistory of Biologic Systematics [Predystoriia Biologicheskoi Sistematiki].
16. *Kuznetsov S. A.* 1998. Large Explicative Dictionary of Russian Language [Bol’shoi Tolkovyi Slovar’ Russkogo Iazyka].
17. *Malt B. C., Pavlenko A.* Kitchen Russian: First-language Object Naming by Russian-English Bilinguals. Proceedings of the 31th Annual Conference of the Cognitive Science Society.
18. *Merkulova V. A.* 1967. Sketches on Russian Popular Plants Nomenclature [Ocherki po Russkoi Narodnoi Nomenklature Rastenii].
19. *Mihatsch W.* Taxonomic and Meronomic Superordinates with Nominal Coding. Ontolinguistics. How Ontological Status Shapes the Linguistic Coding of Concepts : 359–378.
20. *Piperski A.* 2011. Generic Terms in Everyday Vocabulary as a Sphere of Subtle differences between Serbian and Croatian. Komp’iuternaia Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii “Dialog 2011” (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference “Dialog 2011”).
21. *Slaughter M. M.* 1982. Universal Languages and Scientific Taxonomy in the Seventeenth Century.
22. *Taylor J. R.* 1995. Linguistic Categorization: Prototypes. Linguistic Theory : 43.
23. *Veres C.* 2006. The Language of Folksonomies: What Tags Reveal About User Classification. Natural Language Processing and Information Systems, 3999 : 58–69.
24. *Wierzbicka, A.* 1985. Lexicography and Conceptual Analysis.

ГОВОРЯЩИЙ «ЭТАП». ОПЫТ ИСПОЛЬЗОВАНИЯ СИНТАКСИЧЕСКОГО АНАЛИЗАТОРА СИСТЕМЫ ЭТАП В РУССКОМ РЕЧЕВОМ СИНТЕЗЕ[1]

Л. Л. Иомдин (iomdin@iitp.ru)

Институт проблем передачи информации РАН им. А. А. Харкевича, Москва, Россия

Б. М. Лобанов (lobanov@newman.bas-net.by)

Ю. С. Гецевич (mix1122@gmail.com)

Объединенный институт проблем информатики НАН Беларуси, Минск, Беларусь

Излагаются результаты работы по созданию экспериментальной гибридной системы синтеза русской речи, использующей в качестве промежуточного этапа поверхностно-синтаксический анализ читаемого текста. Синтаксическая структура предложения в виде размеченного дерева зависимостей, формируемая в ходе синтаксического анализа, обеспечивает лучшие качественные характеристики звучащей речи по сравнению с классической системой речевого синтеза, не учитывающей в явной форме информации о связях слов в предложении.

Ключевые слова: синтез речи, речевой синтез, синтаксический анализ, звучащая речь, система синтеза речи.

[1] Авторы благодарят Российский фонд фундаментальных исследований и Белорусский республиканский фонд фундаментальных исследований, поддержавшие данную работу грантами РФФИ (№ 10-07-90001-Бел) и БРФФИ (№ Ф10Р-006) в рамках программы совместных исследований России и Беларуси. Авторы выражают также искреннюю благодарность активным участникам проекта В. Г. Сизову, осуществившему программную интеграцию синтаксического анализатора с речевым синтезатором, и О. Ю. Подлесской, в задачу которой входит акцентуирование большого морфологического словаря интегрированной системы: эта работа в настоящее время подходит к концу.

THE TALKING ETAP. USING THE ETAP PARSER IN RUSSIAN SPEECH SYNTHESIS

L. L. Iomdin (iomdin@iitp.ru)

A. A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russian Federation

B. M. Lobanov (lobanov@newman.bas-net.by)

Iu. S. Getsevich (mix1122@gmail.com)

United Institute of Informatics Problems, National Academy of Sciences of Belarus, Minsk, Belarus

The paper presents an attempt to create an experimental hybrid system of Russian speech synthesis, which makes use of surface-syntactic analysis of the text to be read. The syntactic structure of the sentence, a labeled dependency tree formed by the parser, provides better speech parameters as compared to the classical system of speech synthesis, which does not take explicit account of the information on how words are related in a sentence. The hybrid algorithm works as follows: the text to be read is sent to the parser of the ETAP-3 linguistic processor sentence by sentence; the ready syntactic structure of each sentence is treated by a number of specially designed rules that mark certain elements of the sentence as prosodically salient, specifying several element types like sentence head, last element of noun phrase etc. The Multiphone speech synthesis module uses this information to produce intrasentential pauses and emphasize certain words or word groups.

Key words: speech synthesis, syntactic analysis, hybrid system, system of speech synthesis

1. Вводные замечания

Данное исследование продолжает работу, начатую два года назад коллективами Лаборатории компьютерной лингвистики ИППИ РАН им. А. А. Харкевича и Лаборатории распознавания и синтеза речи Объединенного института проблем информатики НАН Беларуси и направленную на создание интегрированной системы русского речевого синтеза, в которой просодические и интонационные и акцентные характеристики генерируемого текста формируются с учетом информации о синтаксической структуре читаемого предложения.

В работе [1] были изложены результаты первых экспериментов в этом направлении. Эти эксперименты сводились к тому, что подлежащий синтезу текст пропускался через синтаксический анализатор многоцелевого лингвистического процессора ЭТАП-3 [2, 3], который формировал информацию об эмфатически выделенных элементах предложения.

В настоящем проекте речь идет о значительно более масштабном участии системы ЭТАП-3 в системе русского речевого синтеза, разработанного Б. М. Лобановым и его коллегами [4]. Синтезируемый текст в нормальной орфографической записи подвергается полному синтаксическому анализу, осуществляемому парсером ЭТАП-3, который (1) членит текст на отдельные предложения, (2) для каждого предложения строит его древесную синтаксическую структуру (СинтС), (3) с помощью специальных правил, применяемых к готовой синтаксической структуре, устанавливает границы речевых синтагм предложения и его эмфатически выделенные элементы. Система Мультифон обрабатывает эту информацию и определяет длительность пауз между синтагмами в зависимости от их синтаксического типа (на принципах, изложенных, в частности, в [5]). Попутно в формирующейся гибридной системе речевого синтеза успешно решается критическая для такой системы задача снятия омографии словоформ, которые различаются ударением и/или противопоставлением букв *е* и *ё*.

2. Синтаксический анализатор системы ЭТАП-3: современное состояние

Синтаксический анализатор (парсер) системы ЭТАП-3 используется в различных приложениях, разрабатываемых в Лаборатории компьютерной лингвистики ИППИ РАН, в том числе в системе машинного перевода с русского языка на английский, в системе синонимического перифразирования, для построения синтаксически размеченного корпуса русского языка СинТагРус [6, 7], а также, в последнее время, в целях создания онтологии для автоматической обработки текстов [8].

Этот парсер в значительной мере основан на лингвистической теории «Смысл ⇔ Текст» И. А. Мельчука. Для каждого предложения письменного текста он строит его синтаксическую структуру (СинтС) (в терминах теории «Смысл ⇔ Текст» — поверхностно-синтаксическую структуру) в виде дерева зависимостей. В дереве СинтС любого предложения имеется единственная вершина, которой непосредственно или опосредованно подчиняются все остальные узлы. Каждый узел такого дерева соответствует одному слову предложения (или некоторому словосочетанию, по тем или иным причинам трактуемому как слово, такому как *несмотря, по меньшей мере, во что бы то ни стало* и т. п.), а его дуги помечены именами синтаксических отношений (СинтО). Имена СинтО эксплицируют различные типы синтаксических связей между словами; в современной версии парсера используется 65–70 различных СинтО. Например, связь между глагольным сказуемым в качестве вершины и именным подлежащим при нем в качестве зависимого члена (*старик* ← *получил*) представляется **предикативным СинтО**; связь между предикатным словом и первым дополнением при нём (*получил* → *письмо, получение* → *письма*) представляется **1-ым комплетивным СинтО**; связь между существительным и определяющим его прилагательным (*заказное* ← *письмо*) оформляется **определительным СинтО**, связь между глаголом и наречным обстоятельством (*неожиданно* ← *получил*) задаётся

обстоятельственным СинтО, а аналитические формы слов, рассматриваемые как синтаксические конструкции, оформляются с помощью **аналитического СинтО** (*получил → бы, более ← интересный, будет → работать*)**.**

Дерево СинтС предложения, генерируемое парсером ЭТАП-3, является упорядоченным — оно сохраняет информацию о порядке следования слов в предложении, который имел место в его исходной форме.

Алгоритм синтаксического анализа обращается к лингвистическим ресурсам двух основных типов: набору бинарных синтаксических правил, или синтагм[2], и так называемому комбинаторному словарю, содержащему богатую и разнообразную информацию о каждом входящем в него слове. Парсер работает пофразно и может функционировать в нескольких режимах, в частности, 1) в полностью автоматическом режиме, применяемом по умолчанию: в этом случае для каждого предложения строится ровно одна СинтС; 2) в режиме множественного анализа, когда пользователь может потребовать от системы построить для неоднозначного предложения несколько СинтС или даже все возможные СинтС; 3) в интерактивном режиме, когда в определенных точках алгоритма парсер, встретив неоднозначную лексическую единицу или омонимичную синтаксическую конструкцию, предлагает пользователю выбрать ту или иную морфологическую, лексическую и/или синтаксическую интерпретацию элементов предложения и тем самым направить работу по некоторому конкретному пути.

Система ЭТАП-3 в целом и ее синтаксический анализатор рассчитаны в первую очередь на тексты нейтрально-деловой прозы. Это, в частности, означает, что в составе некоторых приложений (в первую очередь, в машинном переводе) её нецелесообразно применять к стилистически окрашенному материалу, к авторской художественной прозе, поэзии или к разговорной речи. Однако, как показали наши эксперименты, в рамках рассматриваемой здесь задачи синтеза звучащей речи парсер ЭТАП-3 вполне применим для художественной прозы и публицистики: хотя СинтС предложений, образующих такого рода тексты, могут содержать ошибки, неприемлемые в задачах, требующих глубокой семантической переработки, эти ошибки не критичны для речевого синтеза, поскольку, как правило, информация о границах фонетических синтагм и эмфатически выделенных элементах предложения передаётся верно.

Современная версия русского парсера ЭТАП-3 характеризуется существенно лучшей производительностью и более высокой надёжностью по сравнению с предшествующими его вариантами благодаря включению в парсер достаточно развитого статистического компонента, основанного на материале синтаксически размеченного корпуса СинТагРус. Эти факторы оказываются весьма важными для разрабатываемой гибридной системы речевого синтеза.

[2] Тем самым термин «синтагма» используется здесь иначе, чем это принято в литературе, посвященной автоматической обработке устной речи (в том числе и в настоящей статье).

3. Интерфейс «ЭТАП — МУЛЬТИФОН» и используемые правила

Разработанная в Лаборатории распознавания и синтеза речи ОИПИ НАН Беларуси система речевого синтеза «Мультифон» в целом хорошо справляется с членением и просодическим оформлением фонетических синтагм, идентифицируемых в тексте в первую очередь знаками препинания. Однако при чтении развёрнутых предложений с минимальным количеством знаков препинания (а встречаемость таких предложений в текстах весьма высока, см. [9]) система даёт ощутимые сбои, поскольку глубина их синтаксического анализа оказывается недостаточной. Эти сбои в значительной мере удаётся устранить, если прибегнуть к полному синтаксическому анализу предложения, осуществляемому парсером ЭТАП-3.

На рис. 1 представлена схема, реализующая интерфейс взаимодействия систем ЭТАП-3 и МУЛЬТИФОН. Интеграция двух систем осуществляется через стандартный способ взаимодействия SAPI 5.1 (The Speech Application Programming Interface), который предназначен для стыковки программ синтеза речи с другими программами, работающими в операционной среде Windows.

Прежде чем поступить на вход «Мультифона», входной текст в нормальной орфографической записи подвергается синтаксическому анализу, осуществляемому парсером ЭТАП-3. Специально сконструированный блок правил, применяемых к построенной парсером синтаксической структуре каждого предложения, формирует информацию о его просодически значащих элементах. Размеченный таким образом текст передаётся через SAPI в синтактико-просодический препроцессор (СПП), который на основе этой информации осуществляет членение предложений на фонетические синтагмы, определяет длительность пауз между ними и устанавливает интонационный тип полученных синтагм.

К настоящему времени как в постпроцессоре парсера ЭТАПа-3, так и в СПП задействован ограниченный массив фонетических правил, которые носят достаточно общий характер. Этот массив пока далеко не полон и будет совершенствоваться в дальнейшем. Тем не менее даже небольшое число синтаксических правил, применяемых при синтезе речи, даёт обнадёживающие результаты.

Так, для определения положения границ фонетических синтагм были использованы следующие типы синтаксических элементов предложения:

1) абсолютная вершина предложения;
2) вершины всех частей сложносочиненного предложения;
3) вершины всех придаточных предложений;
4) самые правые субстантивные элементы группы подлежащего, дополнения или обстоятельства при вершинах, перечисленных в пп. 1–3;
5) самый правый субстантивный элемент первой именной подгруппы в группах, перечисленных в п. 4;
6) отдельные классы лексических единиц и конкретные лексические единицы, стоящие в определенной позиции, такие как

наречия-детерминанты в начале предложения типа *вовремя, наверняка, непременно,* числительные и количественные существительные типа *миллион, количество, часть* и пр).

Дополнительно опытным путём были определены также предпочтительные значения длительности межсинтагменной паузы и интонационного типа каждой из полученных синтагм в зависимости от используемых синтаксических типов элементов.

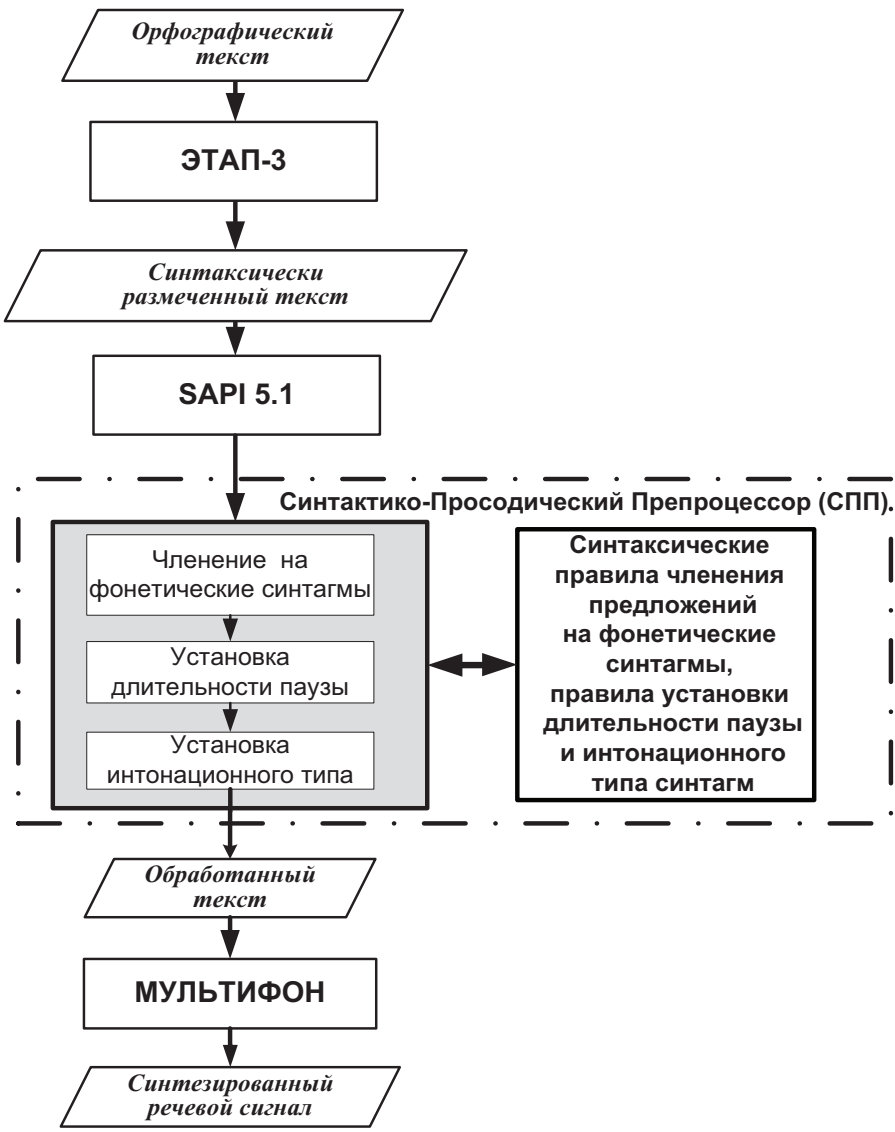


Рис. 1. Схема интегрированной системы «ЭТАП-3 — МУЛЬТИФОН»

Эффективность разработанных правил хорошо видна из приводимых ниже примеров обработки небольшого отрывка текста (7 предложений) из книги Генри Форда «Моя жизнь, мои достижения». В приведенных примерах (1)–(7) после каждого из анализируемых входных предложений представлен обработанный интерфейсом «ЭТАП — МУЛЬТИФОН» выходной текст, размеченный на фонетические синтагмы и поступающий на вход системы «Мультифон». В конце каждой синтагмы присутствует один из интонационных знаков: C — интонация незавершённости, P — интонация завершённости и Q — интонация вопроса. Каждый знак сопровождается цифровым индексом, по которому «Мультифон» выбирает один из подтипов интонационного типа (C, P или Q) и соответствующую этому подтипу длительность межсинтагменной паузы. Кроме того, каждая синтагма разбивается на акцентные единицы (знак [/]), внутри которых выделяются фонетические слова с сильными, или главными акцентами (знак +), и со слабыми, или побочными, акцентами (знак =). Служебные и значимые слова в синтагмах объединяются в фонетические слова посредством твёрдого знака (Ъ).

(1) Если бы имелось средство сэкономить время на 10% или повысить результаты на 10%, то неприменение этого средства означало бы десятипроцентный налог на все производство.

```
1  е=слиЪбы име+лось/сре+дство/                                  C3
2  сэконо+мить/вре+мя/                                           C3_1
3  наЪде=сять проце+нтов/                                        C2
4  и=ли повы+сить/результа+ты/                                   C3_2
5  наЪде=сять проце+нтов/                                        C7
6  то= непримене=ние э=того сре+дства/                           C3
7  означа+лоЪбы/десятипроце=нтный нало+г/наЪвсё= произво+дство/  P4
```

(2) Если, скажем, время одного человека стоит 50 центов в час, то десятипроцентная экономия составит лишний заработок в пять центов.

```
1  е=сли ска+жем/вре+мя/одного= челове+ка/                       C3_1
2  сто+ит/пядеся=т це+нтов/вЬча+с/                               C7
3  то= десятипроце=нтная эконо+мия/                              C3_2
4  соста+вит/ли=шний за+работок/                                 C01
5  вЬпя+ть/це+нтов/                                              P4_1
```

(3) Если бы владелец небоскреба мог увеличить свой доход на десять процентов, он отдал бы охотно половину этого добавочного дохода только для того, чтобы узнать это средство

```
1  е=слиЪбы владе+лец/небоскрё+ба/                               C3
2  мо+г/увели+чить/сво=й дохо+д/                                 C3_1
3  наЪде=сять проце+нтов/                                        C3_2
4  о=н отда+лЪбы/охо+тно/                                        C02
5  полови+ну/э=того доба=вочного дохо+да/                        C3
6  то=лько для того+/                                            C9
7  што=бы узна+ть/э=то сре+дство/                                P9
```

(4) Почему он построил себе небоскреб?

```
1  почему+/                                  Q1_1
2  о=н постро+ил/себе= небоскрёб/            Q2
```

(5) Потому что научно доказано, что известные строительные материалы, примененные известным образом, дают известную экономию пространства и увеличивают наемную плату.

```
1  потому= што= нау+чно/дока+зано/           C8
2  што= изве=стные строи=тельные материа+лы/ C10
3  применё+нные/изве=стным о+бразом/         C3_1
4  даю+т/изве=стную эконо+мию/простра+нства/ C1
5  и= увели+чивают/наё+мную/пла+ту/          P4_2
```

(6) Тридцатиэтажное здание не требует больше фундамента и земли, чем пятиэтажное.

```
1  тридцатиэта=жное зда+ние/                 C3_2
2  неЬтре+бует/бо+льше/фунда+мента/          C1
3  и= земли+/                                C8
4  че=м пятиэта+жное/                        P8
```

(7) Следование старомодному способу постройки стоит владельцу пятиэтажного здания годового дохода с двадцати пяти этажей.

```
1  сле=дование старомо=дному спо+собу/постро+йки/ C3
2  сто+ит/владе+льцу/пятиэта=жного зда+ния/       C3_1
3  годово=го дохо+да/                             C3_2
4  сЬдвадцати+/пяти+/этаже+й/                     P6
```

Нетрудно убедиться в том, что звучащий текст, синтезированный из приведенных выше синтагм, адекватно передаёт как межсинтагменные паузы, так и эмфатически выделенные слова.

Для сравнения приведём примеры разбиения на синтагмы предложения (7), осуществляемого синтезатором ЭТАП-Мультифон (7а), Мультифон (7б), а также двумя доступными в Интернете синтезаторами русской речи «Катерина» компании ScanSoft (7в) и «Алёна» группы компаний Acapela (7г).

(7а) Следование старомодному способу постройки // стоит владельцу пятиэтажного здания // годового дохода // с двадцати пяти этажей.
(7б) Следование старомодному способу постройки // стоит владельцу // пятиэтажного здания годового дохода // с двадцати пяти этажей.
(7в) Следование // старомодному // способу постройки // стоит владельцу // пятиэтажного здания // годового дохода // с двадцати пяти этажей.
(7г) Следование // старомодному // способу постройки // стоит владельцу // пятиэтажного здания // годового дохода // с двадцати пяти этажей.

Пример (7а), на наш взгляд, демонстрирует наилучший способ синтагматического членения. Кроме того, как видно из примеров (7в) и (7г),

способы разбиения на синтагмы синтезаторами «Катерина» и «Алёна» оказались идентичными.

Идея применения синтаксического анализа текста к задаче синтеза звучащей речи высказывалась и раньше, хотя и нечасто. Так, в работе Ф. Кёна и соавт. [10] высказывалось предположение, что характер синтаксической структуры фразы и ее интонационный рисунок (включая межсинтагменные паузы) взаимосвязаны. Авторы провели небольшой эксперимент на материале английского языка, который это подтвердил. Дж. Тауберер [11] показал на материале достаточно объемного корпусного исследования звучащей английской речи, что между синтаксической структурой предложения и внутрисентенциальными паузами имеется безусловная корреляция. Наконец, в недавнем исследовании Ф. Кампилло Диаса и соавт. [12], выполненном на материале галисийского языка, обнаружена несомненная корреляция синтаксической структуры, интонационного контура и паузации.

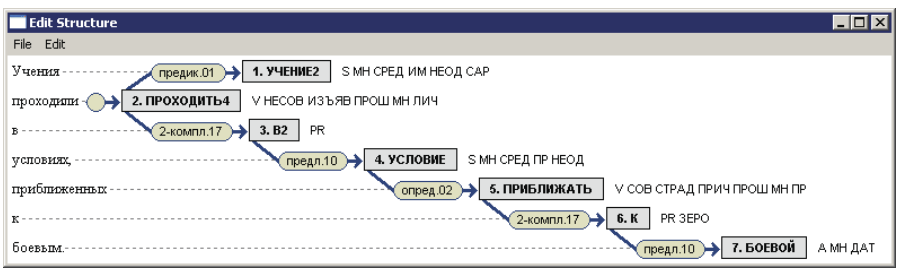
Насколько известно авторам, попытки непосредственной интеграции синтаксического анализатора в систему речевого синтеза до сих пор не предпринимались ни для какого языка.

4. Разрешение омографии

Использование синтаксического анализатора ЭТАП-3 в интегрированной системе речевого синтеза в значительной степени снимает проблему правильной передачи омографичных словоформ текста, которые различаются ударением и/или буквами *е* и *ё*. Дело в том, что в подавляющем большинстве случаев такие словоформы в результате синтаксического разбора получают однозначную лексико-морфологическую интерпретацию, исключающую необходимость применения каких-либо эвристических правил выбора вариантов произнесения. Так, построенная анализатором СинтС предложения

(8) Учения проходили в условиях, приближенных к боевым.

имеет вид

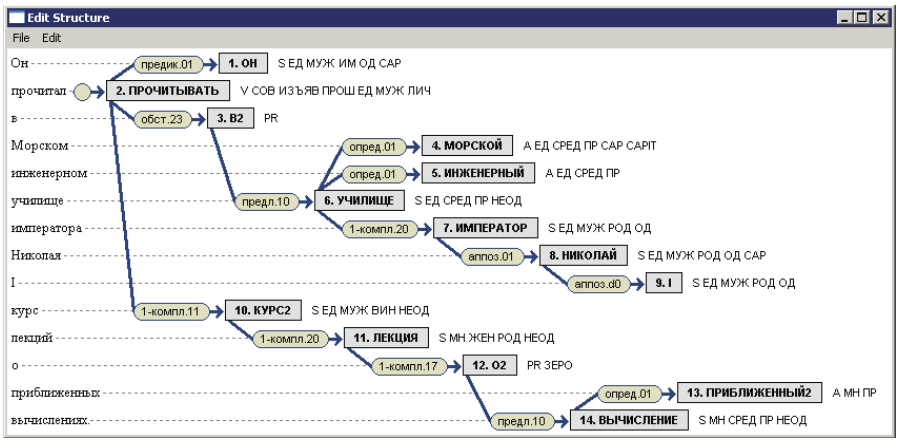


Как легко увидеть, словоформа *приближенных* интерпретируется ЭТАП-ом как страдательное причастие совершенного вида, прошедшего времени, множественного числа и предложного падежа от глагола ПРИБЛИЖАТЬ, что соответствует произносительному варианту *приблúженных*.

В то же время словоформа *приближенных* в предложении

(9) Он прочитал в Морском инженерном училище императора Николая I курс лекций о приближенных вычислениях.

интерпретируется как прилагательное ПРИБЛИЖЁННЫЙ во множественном числе и предложном падеже:

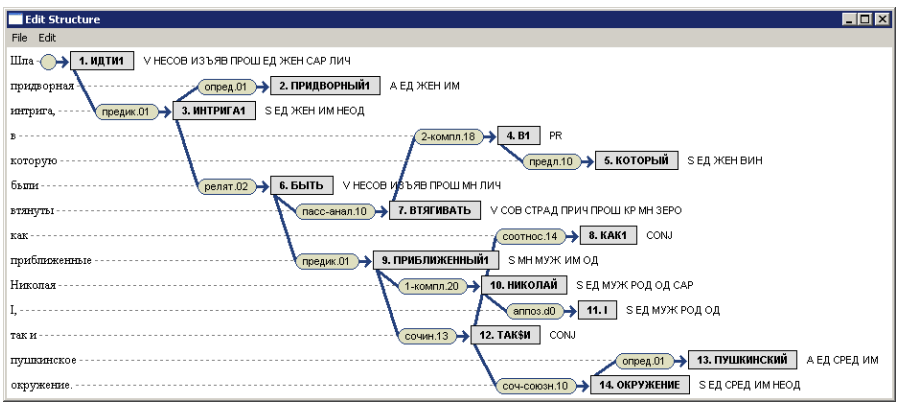


что соответствует произносительному варианту *приближённых*.

Наконец, во фразе

(10) Шла придворная интрига, в которую были втянуты как приближенные Николая I, так и пушкинское окружение,

СинтС которого имеет вид



словоформа *приближенные* интерпретируется как существительное и, разумеется, произносится как *приближённые*.

Очевидно, что и синтаксический анализ не обеспечивает стопроцентного разрешения омографии: чтобы правильно выбрать интерпретацию словоформы типа *замок,* в общем случае необходимо обращение к глубокой семантике. Тем не менее и в таких ситуациях применение синтаксического анализатора нередко приводит к хорошим результатам, особенно с учетом того факта, что при разрешении лексической неоднозначности в современной версии ЭТАП-3 задействуется статистическая информация из размеченного корпуса, в том числе и информация о встречающихся там словосочетаниях (ср. *за́мки Луары* и *дверные замки́*).

5. Использование модуля русского речевого синтеза в других задачах автоматической обработки текстов

Разрабатываемая интегрированная система речевого синтеза «ЭТАП-Мультифон» может использоваться не только для решения задачи выразительного чтения готового русского текста. В частности, она может быть применена и для озвучивания результата машинного перевода в системах, в которых русский язык является выходным, а также в любых других задачах, где необходимо или желательно произнесение сформированного компьютерной системой русского текста (перифразирование, общение с компьютером на естественном языке и т. д.). Первые эксперименты с фонетическим синтезом текста в составе системы англо-русского перевода ЭТАП-3 дали обнадеживающие результаты. В этих экспериментах правила идентификации значимых элементов предложения применяются не к готовой СинтС предложения русского текста, а к промежуточному результату работы перевода (непосредственно перед этапом морфологического синтеза).

Отметим в заключение, что ограниченный масштаб проведенных нами экспериментов не позволяет пока провести полноценную статистически обусловленную оценку полученных результатов. Эта задача предстоит авторам в ближайшем будущем.

References

1. *Apresian Iu. D., Boguslavskii I. M., Iomdin L. L.* et al. 1992. Linguistic Processor for Complex Information Systems [Lingvisticheskii Protsessor dla Slozhnykh Informatsionnykh Sistem].
2. *Apresian Iu., Boguslavskii I., Iomdin L., Lazurskii A., Sannikov V., Sizov V., Tsinman L.* 2003. ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT. MTT 2003, First International Conference on Meaning — Text Theory (June 16–18 2003) : 279–288.
3. *Boguslavskii I. M., Iomdin L. L., Valeev D. R., Sizov V. G.* 2008. Syntactic Parser of the System ETAP and its Evaluation with Parsed Russian Corpus [Sintaksicheskii Analizator Sistemy ETAP i ego Otsenka c pomoshch'iu Gluboko

Razmechennogo Korpusa Russkikh Tekstov]. Trudy Mezhdunarodnoi Konferentsii "Korpusnaia Lingvistika 2008" (Proc. of International Conference "Corpus Linguistics 2008") : 56–74.

4. *Boguslavskii I., Iomdin L., Timoshenko S., Frolova T.* 2009. Development of the Russian Tagged Corpus with Lexical and Functional Annotation. Metalanguage and Encoding Scheme Design for Digital Lexicography. MONDILEX Third Open Workshop. Proceedings : 83–90.
5. *Boguslavskii I., Iomdin L., Sizov V., Tsinman L., Timoshenko S.* 2010. Interfacing the Lexicon and the Ontology in a Semantic Analyzer. COLING 2010. Proceedings of the 6th Workshop on Ontologies and Lexical Resources (Ontolex 2010) : 67–76.
6. *Campillo Díaz F., van Santen J., Rodríguez Banga E.* 2009. Integrating Phrasing and Intonation Modelling Using Syntactic and Morphosyntactic Information. Speech Communication, 51 (5) : 452–465.
7. *Iomdin L. L., Lobanov B. M.* 2009. Syntactic Correlatives of Prosodically Marked Sentence Elements [Sitaksicheskie Korreliaty Prosodicheski Markirovannykh Elementov Predlozheniia]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2009" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2009"), 8 (15) : 136–142.
8. *Koehn P., Abney S., Hirschberg J., Collins M.* 2000. Improving Intonational Phrasing with Syntactic Information. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 3 : 1289–1290.
9. *Lobanov B. M.* 2008. Syntactic Syntagms Text Segmentation Algorythm for Speech Synthesis [Algoritm Segmentatsii Teksta na Sintaskicheskie Sintagmy dlia Sinteza Rechi]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2008" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2008") : 323–329.
10. *Lobanov B. M.* 2010. Punctuation Structure of Fiction Texts and its Role for Expressive Text to Speech Synthesis [Punktuatsionnaia Struktura Khudozhestvennykh Proizvedenii i ee Rol v Sinteze Vyrazitel'noi Rechi po Tekstu]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2010" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2010") : 330–338.
11. *Lobanov B. M., Tsirul'nik L.I.* 2008. Computational Synthesis and Speech Cloning [Komp'iuternyi Sintez i Klonirovanie Rechi].
12. *Tauberer J.* 2008. Predicting Intrasentential Pauses: Is Syntactic Structure Useful? Proceedings of the Speech Prosody 2008 Conference : 405–408.

НЕКОТОРЫЕ МЕТОДЫ ОЧИСТКИ СЛОВАРЯ ЗАПРОСОВ ПОИСКА

М. П. Карпенко (m.karpenko@rambler-co.ru)

С. В. Протасов (s.protasov@rambler-co.ru)

Рамблер Интернет Холдинг

В докладе рассматривается система очистки статистической модели языка. С помощью нескольких этапов очистки система удаляет из модели данные, содержащие опечатки. Описаны эксперименты и различные методики для подавления данных, содержащих опечатки. Результаты экспериментов показывают рост эффективности работы системы коррекции ошибок поиска Рамблера.

Ключевые слова: запрос, поиск, Рамблер, очистка, опечатки.

SOME METHODS FOR LANGUAGE MODEL PRUNING

M. P. Karpenko (m.karpenko@rambler-co.ru)

S. V. Protasov (s.protasov@rambler-co.ru)

Rambler Internet Holding

This paper describes a pruning system of statistical language models. We present a method for pruning the internal vocabulary which is made completely automatically based on users requests and texts drawn from the Internet. The spellchecker system is one of the components of a search engine, and uses this dictionary for language modeling. The described methods can signifi cantly reduce the size of a language model, and open the possibility to improve spellchecker quality. Experimental results show an improvement in the effi ciency of the spellchecker. In our tests the pruning method removed 48 % of the language model without sacrifi cing the quality (in fact, the quality went up 2.7 %). This reduction resulted in the speed increase by 87 %. Pruning the model allows using a greater volume of query logs in the scenario when the amount of available RAM is fixed. This in turn can improve the quality of the spellchecker.

Key words: query, search, Rambler, pruning, misprint.

1. Введение

В данном докладе мы рассматриваем метод очистки внутреннего словаря, который составляется полностью автоматически на основе запросов пользователей и текстов, извлеченных из сайтов в сети Интернет. Система коррекции ошибок (далее «опечаточник») в запросах пользователей к поисковой системе Rambler является одной из компонент поисковой системы и использует данный словарь для моделирования языка запросов пользователей.

Одно из первых упоминаний о проблематике исправления орфографических ошибок можно найти в работе Дамеро [7]. В ней коррекция ошибок предполагает поиск проверяемого слова в эталонном словаре и в случае, если слово отсутствует в словаре, то предлагаются близкие варианты. В случае поиска по Web для коррекции ошибок в запросах вариант ручной проверки малоэффективен, поэтому используется статистическая языковая модель на основе запросов пользователей [3][4]. Питер Норвиг в своей статье «How to Write a Spelling Corrector»[1] приводит пример такого подхода и отмечает, что данный подход прост (28 строчек кода), но недостаточно эффективен. Отмечается, что если данные содержат опечатки, то опечаточник не будет их исправлять. Эрик Брилл и Роберт Мур в своей статье [2] предлагают решить эту проблему с помощью более сложной модели ошибок. Благодаря усложнению модели ошибок опечаточники правильно исправляют «thay» на «they», которое имеет меньший рейтинг чем «that». Модель ошибок, зачастую, основывается на понятии расстояния, введенным Левенштейном[11], и использует функцию вычисления расстояния между строками на основе алгоритма Вагнера и Фишера [8]. Варианты с использованием модели ошибок описаны в работах Цобеля и Дарта[9][10]. В них описывается, что модель ошибок учитывает несколько параметров: статистические данные о реальных опечатках, фонетическую близость слов, а так же близость на клавиатуре. Они также проводили сравнение алгоритмов анализа строк по произношению, и отметили, что вариант Soundex[1], плохо подходит для общей задачи коррекции опечаток. Помимо модели ошибок хорошим средством повышения качества работы опечаточника является использование контекста, о чем отмечают Маерс и Дамерау в своей статье [12]. Суть метода заключается в анализе слов используемых по-соседству с проверяемым. Но для построения контекстной расширенной модели требуются значительные объемы данных.

Несмотря на наличие расширенной модели языка и достаточно полной и точной модели ошибок, остаются слова с опечаткой, которые не могут исправиться, так как их рейтинг в модели языка слишком велик и модель ошибок не может это компенсировать. А так же слова с более чем одной ошибкой, которые часто исправляются на слова с одной опечаткой. С точки зрения быстродействия сложно и малоэффективно создавать модель ошибок, которая бы покрывала эти проблемные места. Поэтому далее мы рассматриваем технологии предварительной очистки модели языка, которая не работает в режиме онлайн, а используется при подготовке исходных данных.

---

[1] http://ru.wikipedia.org/wiki/Soundex

Отдельной обширной темой являются различные способы сглаживания моделей языка: Good-Turing, Back-Off, интерполяция через удаление. Вопросы сглаживания языковых моделей не рассматриваются в данной статье и информацию о них можно найти с трудах Джефри Сампсона [18] и Манина и Шютце[19]. В данной статье мы описываем вариант повышения качества не через сглаживание, а через очистку модели языка от слов и словосочетаний, содержащих опечатки.

Для обучения (тренировки) модели используется статистика запросов пользователей, из которой извлекается языковая модель и модель ошибок. Пользователи поисковой системы совершают опечатку примерно в 15 процентах случаев (среди половины случаев, когда они набирают запрос вручную). Более половины этих опечаток наша система исправляет успешно. Однако, только в 20 процентах случаев, пользователи обращают внимание на заведомо правильное исправление и «кликают» на него. Можно выделить следующие типы опечаток:

1. Опечатки в отдельных словах
   - удаление (агенство вместо аген**т**ство) (8% случаев)
   - перестановка (ко**к**ши вместо кошки) (4% случаев)
   - вставка (пар**р**аллельный вместо параллельный) (4% случаев)
   - замена (однокласс**в**ики вместо одноклассники) (80% случаев)
   - вставка пробела как частный случай вставки (апель син вместо апельсин)
2. Опечатки в последовательностях — склейка слов (недвижимостьв москве вместо недвижимость в москве)
3. Смысловые опечатки («купить акации» вместо «купить акции»)
4. Раскладка клавиатуры (ckjy вместо слон)
5. Латиница (varezhka вместо варежка).

От 80% до 90% всех опечаток отстоят от оригинала на одно изменение символа.

Опечаточник в работе использует два основных компонента:

- Языковая модель. Состоит из униграммной и биграммной моделей.
- Модель ошибок. Модель ошибок представляет из себя модуль, который на основе статистики предсказывает вероятность той или иной опечатки. Например, замена «а» на «о» более вероятна, чем «а» на «б».

Пользовательские опечатки являются источником шума в статистических моделях и приводят к падению точности исправлений, чрезмерному потреблению оперативной памяти серверов и падению скорости работы опечаточника.

В данном докладе мы рассмотрим итерационный алгоритм очистки, который по сути является разновидностью EM-алгоритма [13][15]. В каждом шаге мы уточняем рейтинги слов, добиваясь занижения рейтинга у слов с ошибкой и повышая рейтинги слов без ошибок.

Более полная информация об опечатках позволяет более точно сконструировать не только языковую модель, но и модель ошибок.

Кроме корпуса запросов, мы используем дополнительные сигналы, такие как — было ли нажато исправление, а также дистрибутивная схожесть между опечаткой и кандидатом на исправление. [4].

В качестве метрик качества работы опечаточника мы используем размеченный корпус опечаток, проверенный вручную. Он состоит из более чем 20 000 уникальных запросов пользователей в течение выбранного дня к поисковой системе Rambler. Данный тестовый корпус позволяет примерно оценить изменения характеристик опечаточника: качества (точность, полнота) и скорости работы. Ключевым показателем для нас является информация о количестве кликов и показов опечаточника при контрольном эксперименте на части аудитории.

2. Цели и задачи

Опечатки в данных для обучения создают много проблем. Во-первых, 60%-90% объема памяти занимаемого моделью языка — это опечатки. Во-вторых, становится сложнее исправлять популярные опечатки. Например, слово с популярной опечаткой уже есть в словаре и имеет рейтинг, ориентируясь на который, мы ошибочно предполагаем, что оно правильно написано. В-третьих, сложнее исправлять запросы с более чем одной опечаткой.

Цель очистки модели языка, состоит в том, чтобы улучшить показатели системы:

- Качество работы системы. Увеличивающийся CTR при незначительном уменьшении полноты.
- Скорость работы. Очистка позволит подбирать меньше кандидатов на исправление и это значительно ускорит работу системы.
- Объем языковой модели. Увеличение объема модели ограничивается быстродействием и объёмом используемой памяти для хранения модели. Чем меньше модель занимает места, тем больше обучающих данных мы можем использовать.

3. Опечатки можно разделить на две проблемные группы

- Низкочастотные опечатки. Рейтинг ошибочного варианта написания слова в модели в 2–3 раза ниже правильного написания слова.
- Частотные опечатки. Рейтинг ошибочного варианта близок или даже больше правильного написания слова. Например, слово «береме**нн**ая» в модели языка имеет рейтинг на 10% больше, чем у «береме**н**ая».

Для низкочастотных и частотных ошибок мы используем разные подходы, которые опишем далее.

4. Удаление низкочастотных опечаток

Согласно закону Ципфа в модели языка значительную часть составляют малочастотные слова, и среди них чаще всего встречаются опечатки. Самый простой способ обработки малочастотных слов — удалить их, но при этом теряется много ценной информации в виде редких слов и словосочетаний.

Данный этап ориентирован на удаление низкочастотных ошибок, их влияние на качество сказывается например в случаях двойных ошибок. Так как слово с двойной опечаткой более вероятно исправится на слово с одинарной опечаткой.

4.1. Простое удаление «невероятных» слов и словосочетаний

Языковая модель представляет собой словари униграм, биграмм и триграмм с рейтингами, характеризующими частотность этих слов. Суть очистки от «невероятных» слов состоит в том, чтобы найти в словаре слова, которые не могут быть результатом работы опечаточника. В общем случае рейтинг исправления слова (ска**чч**ать) «c» на вариант «w» (скачать) определяется по формуле Байеса:

$$Ratio(c \to w) = P_L(w) * P_E(w \mid c) \quad (1)$$

В которой:

- «c» — ска**чч**ать» — невероятное слово, которое мы хотим удалить из словаря, частотность 2.
- «w» — «скачать» — вариант исправления, частотность 45 354.
- $P_L(w)$ — языковая модель. Вероятность варианта «w», характеризуемая его частотностью.
- $P_E(w \mid c)$ — Модель ошибок. Вероятность изменения строки «c» до строки «w»
- $q \in L(c, 1)$— множество исправлений «c» с расстоянием 1 по Левеншнейну[2]

В качестве примера для «q» мы будем использовать «ска**ччч**ать». Чтобы «c» никогда не появлялось на выходе опечаточника, нам для любых «q» нужно потребовать условие:

$$P_L(c) * P_E(c \mid w) < P_L(w) P_E(w \mid q) \quad (2)$$

То есть «q» = «ска**ччч**ать» не будет исправляться на «c» = «ска**чч**ать», так как есть лучший вариант «w» = «скачать»

$$P_L(c) max P_E(c \mid q) < P_L(w) min P_E(w \mid q) \quad (3)$$

Так как вероятность $max P_E(c \mid q)$ не может быть больше 1, поэтому

[2] http://ru.wikipedia.org/wiki/Расстояние_Левенштейна

$$P_L < P_W minP(w \mid q)$$
$$\frac{1}{minP_E(w \mid q)} < \frac{P_L(w)}{P_L(c)} \quad (4)$$

У модели ошибок $P_E(w \mid q)$ всегда есть некоторая минимально возможная оценка $minP_E(w \mid q)$ (например 1/1000) и если частотность «w» = «скачать» в (45 354/2) раз больше частотности «c» = «ска**чч**ать», то мы можем смело удалять «c»= «ска**чч**ать»

Результаты очистки представлены ниже в таблице 1.

Таблица 1. Изменение параметров после удаления низкочастотных опечаток

Тип модели языка	F мера	Скорость работы (запросов/сек)	Объем модели. (млн. элементов)
Изначальный вариант	0.430	11.2	7.24
Очищенный вариант	0.430	19.3	3.91

Такой подход позволяет удалить 46% модели языка без потери качества, что приводит к повышению скорости на 72%.

4.2. Самоочистка

В нашем случае статистическая языковая модель строится на основе запросов пользователей к поисковой системе Rambler. Из этой статистики запросов формируется языковая модель в виде словарей униграмм, биграмм и триграмм. Рейтинг слова в словаре является частотностью слова или словосочетания в статистике запросов. Среди этих запросов встречаются запросы с опечаткой. Предположим, что есть запрос «w», в котором две опечатки (ск**оча**ть) и правильный вариант исправления «c» (скачать), при этом в словаре есть похожий на запрос вариант «c'» (скач**аа**ть). В случае использования, описанной выше формулы (1), если:

$$P_L(c')P_E(c' \mid w) > P_L(c)P_E(c \mid w) > P_L(w)P_E(w \mid w) \quad (5)$$

То результатом исправления будет неправильный вариант «c'», при этом настроить модель ошибок, для корректной обработки подобных опечаток очень проблематично. Доля случаев подобных опечаток составляет 2–3%, и чтобы их правильно исправлять, необходимо очистить словарь от опечаток.

Разбиваем данные для обучения модели языка на несколько частей (например 4), после чего строим для каждой части языковой модели и свою систему проверки, которую используем для проверки изначальных данных (списка

запросов). Если запрос определился системой как опечатка в 1 из 4-х случаев, то его рейтинг при обучении будет 1/4. Если запрос определился системой как опечатка в 4-х из 4-х случаев, то мы удаляем его из обучающих данных. Операцию самоочистки можно проводить несколько раз, при этом ее эффективность с каждым разом снижается, так как данных, которые будут определяться как опечатки, становится меньше. На первом этапе из общей модели выделяется 15–20 % запросов, которые считаются неправильными и требуют коррекции.

Таблица 2. Изменение характеристик после самоочистки

Этапы очистки	F мера	Скорость работы (запросов/сек)	Объем модели языка (млн. элементов)
Изначальный вариант	0.430	19.3	3.91
Первый этап	0.430	20.3	3.84
Второй этап	0.431	21.0	3.81

Такой подход позволяет лучше исправлять опечатки с расстоянием 2, а так же позволяет удалить низкочастотные опечатки.

4.3. Использование результатов поиска

Мы так же можем использовать количество результатов поиска по запросу. На рис. 1 представлена статистика результатов поиска по словарю униграмм модели языка.

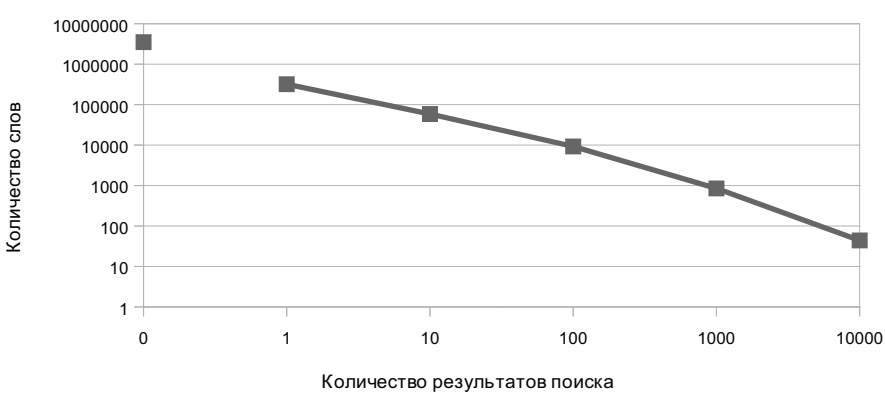


Рис. 1. График количества результатов поиска у слов модели языка

Как видно из графика, у значительной части словаря нет результатов поиска, поэтому мы можем удалить их из словаря. Этот вариант может удалить часть правдоподобной информации, но ее наличие или отсутствие никак не сказывается на качестве опечаточника в контексте работы поисковой системы.

5. Удаление высокочастотных опечаток

Это наиболее проблемные опечатки, так как их сложней обнаружить и исправить в автоматическом режиме. При этом, если опечатка популярна, ее чаще вводят пользователи. Основная проблема в таких случаях состоит в том, что мы не можем исправить запрос. Например, в модели языка есть два слова:

- «бессоница» имеет рейтинг 33 735
- «бессонница» его рейтинг 34 471

Согласно формуле (1), если:

$$P_L(c)P_E(c \mid w) < P_L(w)P_E(w \mid w) \quad (6)$$

То запрос с опечаткой не исправится. Модель ошибок, предполагает, что вероятность замены «н» на «нн» велика, но не перекрывает высокий рейтинг слова с опечаткой.

Для удаления этих опечаток используется два шага:

- Выявление опечатки с помощью метода общего контекста [6][12] и анализа поведения пользователей [4]
- Исправление опечатки, используя дополнительные материалы, в которых подобные опечатки маловероятны.

5.1. Выявление высокочастотных опечаток

Для обнаружения высокочастотных ошибок используются ниже описанные методики:

5.1.1. Соотношение вариантов

Для наиболее частотных слов модели языка строятся варианты их ошибок, среди которых ищутся варианты с большим рейтингом. Если вариант опечатки имеет достаточно большой рейтинг и при этом модель ошибок показывает, что данная замена очень вероятна, то считаем данное слово кандидатом в «высокочастотные опечатки», таких кандидатов набирается 14% от объема языковой модели.

5.1.2. Использование общего контекста

Подобный подход, описывается в работе Кейси Уайтлоу [16]. Суть идеи состоит в том, что если слово и вариант опечатки близки с точки зрения модели ошибок и при этом у них общий контекст, то с достаточной степенью

уверенности можно утверждать, что данный вариант опечатки действительно является опечаткой данного слова, а не отдельным абсолютно другим словом. Под общим контекстом понимается одинаковые слова и словосочетания, используемые совместно со словом и его вариантом опечатки.

Например:

- программы ***для android***
- програ**м**ы ***для android***
- пр**а**граммы ***для android***

В данном случае у слова «программы» и вариантов ошибок: «програмы», «праграммы» есть общий контекст «для android», мы можем утверждать, что данные варианты действительно являются опечатками слова «программы».

К контексту накладывается некоторые полезные ограничения:

- контекстом не являются слова короче 3–4 букв, но в сочетании слов их можно использовать («для android», «для машины», «тур в»)
- контекстом не являются слова общеупотребительной лексики (который, около …)

Рис. 2. Статистика общих контекстов

На Рис. 2. Представлена статистика, которая показывает, что достаточно часто в (60 %) случаев, у слова и варианта исправления есть как минимум одно слово употребляемое совместно с данным словом и вариантом исправления.

Например у слов «автомобилей» и «автамобилей» есть несколько слов контекстов:

- ***покупка*** автомобилей/авт**а**мобилей
- ***продажа*** автомобилей/авт**а**мобилей

- ***настройка*** автомобилей/авт**а**мобилей
- ***ремонт*** автомобилей/авт**а**мобилей
- ***запчасти для*** автомобилей/авт**а**мобилей

У пары слов «автомобилей» и «авт**а**мобилей»представлено 5 контекстов, по этому мы можем предположить, что одно из слов является ошибкой. Слово «автомобилей» будет считаться правильным, так как у него будет больше:

- Рейтинг языковой модели.
- Количество слов контекстов.

При этом похожие слова как «рамблер» и «трамблер» означают абсолютно разные вещи, но в написании отличаются на одну букву. Благодаря контексту мы можем определить их в разные группы и не исправлять одно на другое.

Слово «рамблер» употребляется в следующих словосочетаниях:

- ***почта*** рамблер
- ***поиск*** рамблер
- рамблер ***фото***
- рамблер ***тв программа***
- рамблер ***знакомства***
- рамблер ***игры***
- …

А слово «трамблер» больше употребляется в случаях:

- ***купить*** трамблер
- ***цена*** трамблер
- трамблер ***на ваз 21***
- ***ремонтировать*** трамблер
- …

Как видим из примера, пересечений контекстов в данном случае нет.

5.2. Исправление высокочастотных опечаток

Так как опечатки высокочастотные и прошли несколько стадий анализа, то их количество уменьшается в разы и при желании их даже можно проверить вручную. Например:

- девчонка/девч**е**нка
- лестница/ле**сн**ица

Но есть более эффективный способ. Полученный список кандидатов анализируется с помощью дополнительных источников данных:

- Материалы, в которых вероятность ошибок наименьшая (новостные, литературные тексты)

- Списки названий:
  - Списки фамилий, имен
  - Списки географических объектов, адресов
  - Списки лекарственных средств
  - Списки названий компаний

Анализируется наличие данного и исправленного слова в данных списках, если слово есть в списках, то мы его не правим. А так же анализируется общий контекст у исправления и слова кандидата в высокочастотные опечатки. Считать данный вариант опечаткой или нет определяется порогом для соотношения объемов контекста. Вариант, у которого больше контекста, считается правильным [16].

Например, как было показано выше, у слов «автомобилей» и «автамобилей» есть 6 контекстов. Таким образом мы можем предположить, что эти слова связаны и одно из них является ошибкой другого. Так как:

1. $\frac{P_L(\text{автомобилей})}{P_L(\text{автамобилей})} = \frac{0{,}0000546}{0{,}0000017} = 321 > 20$

2. Количество слов контекста у «автомобилей» — 114, а количество слов у «автамобилей» — 23 и 114/23=4,96 > порогового значения 3.

Доработка модели языка материалами, в которых вероятность ошибок наименьшая заключается в следующем:

- Ищем пересечение моделей языка
- Повышаем рейтинг словам, общим для моделей.

Вместо добавления новых слов мы используем дополнительные данные как критерий, повышающий рейтинг правильных слов над ошибочными.

Результаты исправления высокочастотных опечаток на корпусе 20 000 запросов представлены ниже.

Таблица 3. Изменение параметров после удаления высокочастотных опечаток

Вариант словаря	F мера	Скорость работы (запросов/сек)	Объем модели (млн. элементов)
Изначальный вариант	0.431	21.0	3.81
Доработанный вариант	0.442	21.3	3.77

6. Выводы

Описанные методы позволяют значительно сократить объем модели языка, при этом повысив ее качество.

В нашем случае:

- Очистка позволила удалить 48 % модели языка
- Улучшить качество работы на 2.7 %
- Увеличить скорость работы на 87 %

Так как качество опечаточника увеличивается при росте модели языка, то при фиксированном объеме быстрой оперативной памяти сервера очистка модели позволяет повысить качество за счет большего объема исходных данных.

References

1. *Berget A. L., Della Piertra S. A., Della Pietra. V. J.* 1996. Maximum Entropy Approach to Natural Language Processing. Computational Linguistics, 22 (1).
2. *Boldi P., Bonchi F., Castillo C., Donato D., Vigna S.* 2009. Query Suggestions Using Query-Flow Graphs. Workshop on Web Search Click Data.
3. *Brill E., Moore R. C.* 2000. An Improved Error Model for Noisy Channel Spelling Correction. Association for Computational Linguistics Stroudsburg.
4. *Cherkasskii V., Vassilas N., Brodt G. L., Wagner R. A., Fisher M. J.* 1974. The String to String Correction Problem.
5. *Cucerzan S., Brill E.* 2004. Spelling Correction as an Iterative Process that Exploits the Collective Knowledge of Web Users. Conference on Empirical Methods in Natural Language Processing.
6. *Damerau F. J.* 1964. A Technique for Computer Detection and Correction of Spelling Errors. Journal of the Royal Statistical Society, 39 (1): 1–21.
7. *De Mori R.* 2001. Spoken dialogues with computers.
8. *Dempster A., Laird N.,* et al. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39 (1): 1–38.
9. *Golding A. R., Roth D.* 1996. Applying Winnow to Context-Sensitive Spelling. Correction.International Conference on Machine Learning (ICML) : 182–190.
10. *Levenshtein V.* 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Doklady Soveta Fizikov, 10 (8) : 707–710
11. *Maning S.* 1999. Foundations of statistical natural language processing.
12. *Mayes E., Damerau F.,* et al. 1991. Context Based Spelling Correction. Information Processing and Management. International Journal, 27 (5).
13. *Mu Li, Muhua Zhu, Yang Zhang, Ming Zhou.* 2006 Exploring Distributional Similarity Based Models for Query Spelling Correction. ACL-44 Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics.
14. *Norvig P.* How to Write a Spelling Corrector, available at: http://norvig.com/
15. *Sampson G.* 2001. Empirical linguistics. Continuum International.
16. *Toutanova K., Moore R. C.* 2002. Pronunciation Modeling for Improved Spelling Correction. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6–12 : 144–151

17. *Whitelaw C., Hutchinson B., Y Chung G., Ellis G.* 2009. Using the Web for Language Independent Spellchecking and Autocorrection. EMNLP '09 Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2.
18. *Zobel J., Dart P. W.* 1996. Phonetic String Matching: Lessons from Information Retrieval. SIGIR '96 Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.
19. *Zobel J., Dart P. W.* 1995. Finding Approximate Matches in Large Lexicons. Software—Practice & Experience, 25 (3).

ОЦЕНОЧНЫЕ ЗНАЧЕНИЯ РЕБРЕНДИНГОВОГО ТИПА В ПРИЗНАКОВОЙ ЛЕКСИКЕ* (ПО МАТЕРИАЛАМ БАЗЫ ДАННЫХ СЕМАНТИЧЕСКИХ ПЕРЕХОДОВ В КАЧЕСТВЕННЫХ ПРИЛАГАТЕЛЬНЫХ И НАРЕЧИЯХ)

О. С. Карпова (o_k@inbox.ru)
РГГУ, Москва, Россия

Е. В. Рахилина (rakhilina@gmail.com)
Институт русского языка РАН, Москва, Россия

Т. И. Резникова (tanja.reznikova@gmail.com)
ВИНИТИ РАН, Москва, Россия

Д. А. Рыжова (daska1990R@yandex.ru)
МГУ им. М.В.Ломоносова, Москва, Россия

В статье отражены результаты исследования прилагательных с семантикой положительной и отрицательной оценки, оценочное значение которых образовано посредством семантического сдвига особого типа, называемого нами ребрендингом. Исследование выполнено на материале Базы данных семантических переходов в русских качественных прилагательных и наречиях. В работе обсуждаются различные аспекты функционирования оценочных значений: механизм их образования, лексическая сочетаемость, взаимодействие с другими значениями ребрендингового типа.

Ключевые слова: признаковая лексика, качественные прилагательные, качественные наречия, семантический переход, ребрендинг.

* Настоящее исследование выполнено при поддержке РФФИ, грант № 11-06-00385-а

MEANING OF ESTIMATION IN SEMANTIC SHIFTS OF REBRANDING TYPE IN ADJECTIVES AND ADVERBS (ON THE MATERIAL OF THE DATABASE OF SEMANTIC SHIFTS IN RUSSIAN ADJECTIVES AND ADVERBS)

O. S. Karpova (o_k@inbox.ru)

Russian State University for the Humanities, Moscow, Russian Federation

E. V. Rakhilina (rakhilina@gmail.com)

Russian Language Institute, Russian Academy of Sciences, Moscow, Russian Federation

T. I. Reznikova (tanja.reznikova@gmail.com)

VINITI, Russian Academy of Sciences, Moscow, Russian Federation

D. A. Ryzhova (daska1990R@yandex.ru)

Lomonosov Moscow State University, Moscow, Russian Federation

The article is focused on the description of meaning of positive and negative estimation of rebranding type in qualitative adjectives and adverbs (for example, bezumnyj ‘mad’: chelovek ‘man’ / plat’e ‘dress’; blestyashchij ‘shining’: pugovica ‘button’ / obrazovanie ‘education’; dikij ‘wild’ zver’ ‘animal’ / pricheska ‘hairdo’; zolotoj ‘golden’: slitok ‘bar’ / detstvo ‘childhood’; uzhasnij ‘terrible’: zver’ ‘animal’ / vkus ‘taste’ etc.). The investigation is fulfi lled on the material of the Database of semantic shifts in Russian adjectives and adverbs. The work contains analysis of different aspects of functioning of estimation meanings derived by the semantic shift “re-branding”: semantic zones as sources of estimation meanings, mechanism of their generation, lexical combinatory. We also discuss the interaction of estimation meanings with other meanings of re-branding type: combinations of estimation meaning with meanings of intensity, quantity, size, and variety.

Key words: estimation, adjective, adverb, semantic shift, rebrending.

1. Ребрендинг как теоретическая проблема

Настоящая статья продолжает серию наших публикаций в «Диалоге», посвященных исследованию моделей полисемии в русской признаковой лексике (Рахилина и др. 2009, Карпова и др. 2010). Специфика нашего подхода к этой проблематике, в сравнении с предшествующими работами, обсуждавшими регулярные модели многозначности на различном языковом материале (ср., например, Апресян 1971, 1974, Падучева 1988, 2004, Бирих 1995, Кустова 2004, Lakoff, Johnson 1980, Ostler, Atkins 1991, Pustejovsky 1995, Radden, Kövecses 1999, Peirsman, Geeraerts 2006 и мн. др.), заключается в «сплошном» изучении семантики лексических единиц. Иными словами, мы не выбираем отдельные примеры реализации той или иной полисемической модели, а рассматриваем языковой материал в его совокупности (в нашем случае — все частотные многозначные качественные прилагательные и соответствующие им наречия) и классифицируем его с точки зрения засвидетельствованных моделей семантических переходов.

Выбранный нами подход, будучи крайне трудоемким, оказался вместе с тем чрезвычайно продуктивным с теоретической точки зрения. Выяснилось, что при анализе и классификации всех переходов (пусть и на ограниченном участке лексики) невозможно обойтись двумя типами семантических сдвигов, которыми традиционно исчерпывается описание механизмов семантической деривации, а именно, метафорой и метонимией. Действительно, в целом ряде случаев в нашем материале производное значение признакового слова нельзя трактовать ни как метафорический, ни как метонимический сдвиг от исходного значения. Если учесть, что доля таких «неклассифицируемых» случаев составляет более 15 %, то очевидно, что их роль в системе семантической деривации достаточно велика.

Существенно, что эти переходы не носят, так сказать, случайного характера: большинство из них устроено сходным образом, подчиняется одинаковым правилам семантической эволюции. Тем самым они не выбиваются из системы, а образуют особый фрагмент общей системы семантических сдвигов. К основным чертам интересующего нас типа сдвигов относятся утрата («выветривание») части исходной семантики и возникновение нового значения на базе конвенциализованной импликатуры, выводимой из исходного значения.

Так, прилагательное *здоровый* в первом значении указывает на идею физического здоровья одушевленного существа (ср. *здоровый мальчик*). Однако в сочетаниях типа *здоровый кусок*, *здоровая палка* идея физического здоровья полностью пропадает (= выветривается), а ей на смену приходит идея большого размера, которая развивается из представления о здоровом теле как мощном и — следовательно — большом физическом объекте. Прилагательное *кривой* обозначает форму объекта (ср. *кривой нож*), но для конструкций типа *кривое объяснение* идея формы абсолютно нерелевантна: здесь *кривой* выражает отрицательную оценку. Значение оценки очевидным образом развивается как следствие из исходной семантики формы: действительно, по отношению

к некоторым объектам кривизна является отрицательной характеристикой (ср. *кривые ноги*), тем самым производное значение здесь, как и в случае прилагательного *здоровый*, является импликатурой из исходного.

Подчеркнем, что ни один из разобранных выше примеров не укладывается в стандартные представления о метафоре или метонимии. В самом деле, под метонимией обычно понимается сдвиг, при котором новое значение остается в той же семантической зоне (домене, фрейме), что и исходное (ср. Croft 1993, 2006, Radden, Kövecses 1999, Blank 1999, Peirsman, Geeraerts 2006 и мн.др.). В нашем же случае имеет место явная мена семантического домена (*здоровый*: физиологическое свойство → размер; *кривой*: форма → отрицательная оценка).

В то же время эти переходы нельзя классифицировать и как метафоры: в основе метафорического переноса лежит уподобление конечной зоны зоне-источнику, ср. понятие проекции (mapping) элементов структуры одной области на структуру другой в Lakoff, Johnson 1980. Так, примером стандартной метафоры в нашем материале может служить переход *пустая коробка — пустое обещание*. В данном случае исходное значение подразумевает признак физического объекта-контейнера. Применяя этот признак к *обещанию*, мы концептуализуем последнее как контейнер, наполнением которого выступает реализация обещанного. Тем самым *пустое обещание* понимается как отсутствие этого наполнения. Иначе говоря, и в исходном, и в производном значении *пустой* обозначает отсутствие чего-л., только исходно речь идет об отсутствии физических объектов, а в результате перехода — об абстрактных сущностях.

В прилагательных *здоровый* и *кривой* отношения уподобления между областью-источником и областью-целью установить невозможно: в сочетаниях *здоровый кусок / здоровая палка* физические объекты не наделяются признаками одушевленного существа, и сам признак существенно изменяется, а не сохраняет — с незначительной переинтерпретацией — свое основное содержание, как в случае *пустого*. Так же и в случае прилагательного *кривой*: *объяснение* не осмысляется как обладающее признаком формы. Все это не позволяет нам считать эти переходы метафорами.

Многочисленность примеров такого рода — с импликативным механизмом образования значения, отличным от метафоры и метонимии, — послужила основанием для постулирования отдельного — третьего — типа семантических переходов, который мы назвали «ребрендингом»[1]. Все выявленные в ходе семантического анализа ребрендинги, как и переходы стандартного типа — метафоры и метонимии — мы заносили в Базу данных, которая тем самым представляет собой полный каталог семантических сдвигов в частотной (со встречаемостью свыше 2000 употреблений в Национальном корпусе русского языка) признаковой лексике русского языка (в общей сложности База включает классификацию переходов для 300 признаковых слов, подробнее см. Карпова и др. 2010). Построенная таким образом База открывает возможности для самых различных исследований в сфере многозначности и закономерностей

[1] Подробнее о механизме переходов ребрендингового типа и его соотношении с метафорой и метонимией см. Рахилина и др. 2010.

деривации адъективных и адвербиальных значений (ср. в этой связи статью, посвященную метонимическим переходам Карпова и др. в печ.).

Особый интерес для нас представляют, конечно, случаи ребрендинга как не исследованного ранее класса переходов. Все встретившиеся в нашем материале сдвиги ребрендингового типа мы расклассифицировали с точки зрения семантики образующегося значения. В итоге мы получили следующие категории: оценка (положительная/отрицательная), степень (высокая/низкая), количество (большое/малое), разнообразие/неразнообразие, большой размер, одновременность, иммедиатность, апроксиматив, безальтернативность. Каждая из этих категорий заслуживает отдельного исследования в отношении природы сдвига, источников ребрендингового значения, системных связей с другими категориями. В настоящей работе мы обратимся к анализу оценочных значений.

2. Оценочный ребрендинг: источники и механизм семантических переходов

Признаковые слова, развивающие оценочную семантику посредством ребрендинга, естественным образом разбиваются на два класса: со значением положительной и отрицательной оценки. Проследим, какие исходные значения могут приводить к возникновению каждой из них.

2.1. Ребрендинг с семантикой положительной оценки

В качестве источника положительной оценки выступают следующие классы значений:

а) относящийся к сфере фантастического, ср. прилагательные *фантастический*, *нереальный, чудесный, сказочный, волшебный, невероятный*. Исходно эти лексемы характеризуют объекты и явления воображаемого мира, не существующие или едва ли возможные в реальности, ср. *фантастическое животное*, *нереальные существа*, *чудесное превращение, сказочный герой, волшебная палочка, невероятная история*. По-видимому, идея воображаемого импликативно связана с фантазией, мечтой, что, в свою очередь, порождает положительную оценку. Конвенционализация этой импликатуры приводит к стиранию исходной семантики и развитию чисто оценочного значения, ср.:

(1) *Владислав Игнатьевич любил, чтобы нигде не было ни пылинки, и Людмила была* ***фантастической*** *хозяйкой: стёкла блестели так, как будто их нет, зеркала сияли, в них можно было войти.* [Сати Спивакова. Не всё (2002)]

(2) *Очень большая актриса, она совершенно* ***невероятно*** *играла последний акт.* [Игорь Изгаршев. Мужчина и женщины (2001) // «Аргументы и факты», 2001.03.07]

(3) *Просто он **нереально** поет, по крайней мере сколько раз я его видела (много) никогда не разочаровывалась* [http://www.diary.ru/~russian-musicals/p42423314.htm]

b) существующий «здесь и сейчас», определенный, хорошо видимый, ср. *реальный*, *конкретный*, *четкий*. Интересно, что эта группа в некотором смысле противоположна предыдущей (особенно наглядно это проявляется в паре *реальный — нереальный*), тем не менее в сфере оценки их семантика сходится. И это вполне закономерно: достижимость здесь противопоставляется неосуществимости, недоступности и тем самым оценивается положительно. Любопытно, что, в отличие от зоны фантастического, для прилагательных зоны реального оценочное значение — это тенденция самого последнего времени, ср. следующие контексты:

(4) *Bugatti Veyron — вот это **реальная** тачка! До ста за 2.5 с., 406 км/ч максималка, 1001 л. с.! А выглядит как, ух*! [http://forum.igromania.ru/printthread.php?t=4822&pp=200&page=3]

(5) *Рёмо Симэй — вот это **конкретная** девчонка!!! Красотка!!! Выше всяких похвал!!! Я с самого начала болел только за неё.* [http://remake.net.ru/forums/index.php?showtopic=41546]

Остановимся подробнее на развитии ребрендингового значения в прилагательном *четкий*. В исходном значении *четкий* содержит указание на хорошую видимость объекта, возможность его различить во всех подробностях (ср. *четкая картинка, четкий шрифт*). Зрительное восприятие стандартным образом метафорически переносится на ментальное, ср. *четкое изложение*. Однако если существительное из класса ментально воспринимаемых сущностей заменить на имя лица, ср. *четкий пацан / девушка* как в примере (6), то получившуюся семантику уже невозможно будет связать метафорическим отношением ни с исходным, ни с производным метафорическим значением. Действительно, *четкий пацан / девушка* — это не характеристика лица с точки зрения легкости его восприятия — зрительного или ментального, это просто положительная характеристика этого лица. Таким образом, мы здесь наблюдаем те же процессы, что и в предыдущих примерах ребрендинга: имеет место, во-первых, стирание исходного значения (заметим, что в метафоре исходное значение — пусть и в несколько видоизмененном виде — но сохраняется) и, во-вторых, имплицитно заложенная и в исходном, и — особенно — в метафорическом значении идея положительной оценки эволюционирует в единственное семантическое содержание признакового слова.

(6) *очень **четкая** девушка, пример для подражания нашим гламурным персонам черпающим глупости из глянцевых журналов* [http://pyxc.ru/420-prezentaciya-knigi-kati-valievoj-mechtateli.html]

c) отвечающий некоторой норме, стандарту, ср. *приличный*, *правильный*. Прилагательное *приличный* (и наречие *прилично*) исходно относится

к человеку, соблюдающему установленные правила поведения, и — метонимически — к его манерам, внешнему виду и т. п., ср. *приличный человек*, *прилично одевается*. Лексема *правильный* описывает соответствие некоторым правилам или действительности, ср. *правильное произношение*, *правильный вывод*. Очевидно, что семантика и *приличного*, и *правильного* влечет за собой положительную оценку. Тем самым, при выветривании семантики соответствия правилам имеет место сдвиг значения ребрендингового типа:

(7) *<…> о встрече с Назымом Хикметом, который, будучи известным нам всем как турецкий писатель, вдруг обнаружил, помимо всего прочего,* ***приличное*** *знание русского языка.* [Алексей Козлов. Козел на саксе (1998)]

(8) *Гоблин* **правильный мужик***, и юмор у него остроумный*. [http://city.is74.ru/forum/showthread.php?t=406656&page=5]

d) лишенный рассудка, ср. *безумный*, *сумасшедший*. Исходно эти прилагательные характеризуют людей, страдающих душевной болезнью (ср. *безумный больной*). Казалось бы, эта семантика должна была бы способствовать развитию отрицательной оценки, однако в соответствующих контекстах эти лексемы выражают явно положительную оценку, ср.:

(9) *Он подарил мне очень красивое платье, просто* ***безумное*** *платье*. [http://36unise.ru/news/2/6298].

(10) *Стас Михайлов потрясающий певец:* ***безумно*** *поет, завораживает своим голосом… Очень сильно люблю его песни…* [http://www.youtube.com/all_comments?v=CTD-1RRF8n0]

(11) *«Остров проклятых» — очень понравился, Ди Каприо* ***безумно*** *играет …* [http://forum.plasticsur.ru/index.php?s=498c876b6e630ee410faba447e1894cc&showtopic=21803&st=140&p=889484&p=889484]

(12) *Да, Джовинко просто* ***сумасшедше*** *играет! Видно, что у парня есть будущее, не зря его так пресса восхваливала* [inter-fans.moy.su]

Несоответствие возникающей оценки нашим ожиданиям объясняется, по всей вероятности, тем, что оценка у этих прилагательных развивается не непосредственно из исходного значения, а через несколько промежуточных этапов. Значение душевной болезни метафорически переносится на характеристику здоровых людей, которые по своему поведению, поступкам напоминают людей с психическим расстройством (ср. *безумный исследователь / профессор*). От обоих значений — прямого и метафорического безумия — параллельно развиваются метонимии на проявления и поведение таких людей (ср. *безумный взгляд / хохот*, *безумные мысли* и под., ср. также в адвербиальном употреблении *безумно смотрел / хохотал / размахивал руками*). Заметим, что в некоторых сочетаниях такого рода

идея настоящего или метафорического безумия создает представление об интенсивности осуществляемого действия (так, *безумно хохотал* подразумевает громкий хохот, *безумно размахивал руками* — активные движения и т. д.). Эта импликатура создает предпосылки для развития ребрендингового значения интенсивности (ср. *безумная роскошь / усталость / ненависть; безумно устать / ненавидеть / дорогой*). В свою очередь, идея интенсивности регулярно связана с положительной оценкой, тем самым исходно отрицательная идея психической болезни развивает положительную оценку, которую мы наблюдали в примерах (9–12). Аналогичную эволюцию можно наблюдать и в случае прилагательного *сумасшедший*. Показательно, что значение оценки для обеих лексем является недавним (и встречается в основном в интернет-дискурсе), что является дополнительным подтверждением его вторичности по отношению к семантике интенсивности.

e) характеризующийся богатством и внешним великолепием, ср. *роскошный, шикарный*. Исходная семантика реализуется чаще всего в контекстах, описывающих дорогие дома / здания, интерьеры и — адвербиально — действия по их созданию, ср. *роскошное / шикарное убранство*, *роскошно / шикарно обставить*. Идея внешнего великолепия имплицитно уже содержит в себе положительную оценку, основное же значение выветривается при сочетании с лексемами, которые подразумевают не только и не столько зрительное восприятие. Тем самым в конструкциях такого типа реализуется значение чистой оценки, ср. *роскошно / шикарно отдохнули / провели время*.

(13) *Первоклассная русская баня.* ***Роскошно*** отдохнули *и душой, и телом*! [http://www.turpravda.com/ua/shodnica/?m=comment]

(14) *Берлинский кинофестиваль открылся двумя* ***шикарными*** *фильмами* … [Юрий Гладильщиков. Певцы Чикаго и Великой Китайской стены. Берлинский кинофестиваль открылся двумя шикарными фильмами (2003) // «Известия», 2003.02.07]

f) мало встречающийся, ср. *редкий*, *уникальный*. Прилагательное *редкий* проходит более длинный путь к ребрендинговому значению, *уникальный* — совсем короткий, однако механизм развития оценочной семантики у них сходный. *Редкий* исходно отсылает к физическому свойству объекта, заключающемуся в расположении его отдельных частей на относительно большом расстоянии друг от друга, ср. *редкий гребень / лес*. Метонимически свойство целого переносится на свойство его частей, ср. *редкие зубья / деревья*. На следующем этапе расстояние в пространстве метафорически интерпретируется как расстояние во времени (ср. *редкие свидания / события*). Далее в действие включается механизм импликатуры: большое расстояние во времени порождает идею того, что ситуаций / явлений, описываемых как *редкие*, очень мало, а это, в свою очередь, порождает идею их особой ценности. Заметим, что переход от идеи низкой частотности к малому количеству и оценке существенно расширяет сочетаемость прилагательного: в его контексте появляются теперь не только имена ситуаций, но и существительные других типов (ср. *редкие розы*, *редкий голос*). Последний переход, который мы наблюдали на примере

редкого — от малого количества к высокой ценности и — тем самым — положительной оценке — демонстрирует и прилагательное *уникальный*, ср. контекст, характерный для исходного значения (15) и производный оценочный (16):

(15) — *И он сердито положил на край кровати фотографию Кастанье —* ***уникальный*** *экземпляр, отысканный мной в старых архивных папках музея.* [Ю. О. Домбровский. Хранитель древностей, часть 1 (1964)]

(16) — *Поверьте, если приглашает Жозеф Надж, то отказаться невозможно. Он —* ***уникальный*** *человек, создает странные спектакли. В «Нет больше небесного свода» заняты акробаты, китайская танцовщица, японский актер и я.* [Елена Троицкая. Жан Бабиле: «Танец — не профессия, а тайна, которой делишься» // «Линия», 2005]

g) Единичным в отношении источника оценочного значения является ребрендинг в прилагательном *крутой*. По-видимому, основанием для развития оценки здесь послужило определенное свойство характера человека. Рассмотрим, однако, семантические переходы, приводящие к развитию значения положительной оценки, последовательно. Исходно *крутой* характеризует форму природных объектов, ср. *крутой склон / берег*. Метафорически резкость перепада высоты может переноситься на резкость в характере человека, ср. *крутой нрав / характер*:

(17) *Когда мамаша Мотрича, еще совсем не старая простая женщина с* ***крутым*** *характером, дома, Мотрич не решается приглашать к себе приятелей*. [Эдуард Лимонов. Молодой негодяй (1985)]

Метонимически этот признак может присваиваться самому человеку, ср:

(18) *И сам он, и продотрядовцы очень уж* ***круты****, революционно-беспощадно настроены, что и дает повод сбегать к белым посыльному и сообщить, что «Шелковку грабят».* [Виктор Астафьев. Зрячий посох (1978–1982)]

При этом, судя по данным НКРЯ, вплоть до конца 80-х гг. XX в. этим признаком в основном характеризуются люди, имеющие определенную власть, влиятельность. В этой ситуации обладание ‘крутым нравом’ создает преимущества для эффективного ведения дел. Соответственно, во многих контекстах к описанию самого свойства характера добавляется положительная оценка, ср.:

(19) *Дядя Володи Горячева был начальником Вейского отделения железной дороги,* ***крутой****, видный местный руководитель и общественный деятель, много полезного делавший для транспорта, города и народа.* [Виктор Астафьев. Печальный детектив (1982–1985)]

В ходе дальнейшего семантического развития положительная оценка прочно утверждается за прилагательным *крутой*, происходит существенное

расширение его сочетаемости, и семантика конкретного типа человеческого характера выветривается, ср. употребление *крутого* в конструкции с неодушевленными предметными или абстрактными именами, а также в адвербиальных конструкциях:

(20) На самом деле очень ***крутая книга***, одна из лучших по математике абитуриентам! Всем поступающим советую! [http://darudar.org/gift/513889/]

(21) Это очень ***крутая выставка*** и присутствие на ней уже почетно! [http://nkp-bulldogru.ru/forum/index.php?topic=899.30]

(22) ***Круто*** *посидели вчера, всегда бы так!* [*http://www.absent.ru/restoran/all/pyanyj-dyatel/reviews/1117/*]

Таковы выявленные нами классы значений, на базе которых посредством ребрендинга может возникать семантика положительной оценки. Обратимся теперь к источникам отрицательного оценочного значения.

2.2. Ребрендинг с семантикой отрицательной оценки

На материале частотной признаковой лексики зон-источников для отрицательного ребрендинга оказалось существенно меньше, чем для положительного, причем «полноценный» переход ребрендингового типа можно постулировать лишь для одной семантической группы слов. В остальных случаях речь идет, скорее, о частичном стирании исходного значения, то есть о промежуточной стадии процесса ребрендинга (о градуальной природе ребрендинга см. Рахилина и др. 2010). Кратко охарактеризуем основные значения-источники:

а) вызывающий страх, ужас, ср. *ужасный*, *жуткий*, *чудовищный*, *кошмарный*. Это главная и наиболее продуктивная зона для развития семантики отрицательной оценки. Исходно эти прилагательные характеризовали объекты, которые способны вселять страх, ср. *ужасный / чудовищный зверь*. Однако в современных употреблениях это значение встречается крайне редко — его в значительной степени вытеснила оценочная семантика. Основания для ее появления совершенно прозрачны: очевидно, что то, что внушает страх, оценивается отрицательно. Оценка содержится тем самым уже в исходных употреблениях этой группы лексики. Когда же идея страха, ужаса полностью выветривается, можно говорить об осуществлении перехода ребрендингового типа, ср. *ужасный / чудовищный: матч / поездка / новость / запах / платье / характер* и т. д. или адвербиальные употребления *ужасно / чудовищно: отдохнуть / пахнуть / сыграть на турнире и под.*:

(23) ***Чудовищный*** *матч! — говорил российский теннисист, уже выиграв.* [Роман Средиземский. Мертвее всех мертвых. Марат Сафин: «Я больше не мог. Это было ужасно» (2002) // «Известия», 2002.08.28]

(24) *Мерзко поют,* ***ужасно*** *поют, но с большим удовольствием и очень охотно!* [А. А. Яблоновский. Египет (1920–1921). Гости английского короля (1920–1921)]

b) неяркий, плохо различимый, ср. *бледный*, *тусклый*. Обратим внимание, что эта группа дает зрительный образ, противоположный прилагательному *четкий* (см. выше), так что вполне закономерно здесь развитие обратной оценки. Исходное значение *бледного* представлено сочетаниями типа *бледное лицо, бледные щеки* и описывает физиологический признак, соотносящийся с конкретной цветовой характеристикой. На основе цветового сходства развивается метафорический переход на природные объекты (*бледный небосклон*) или артефакты (*бледная фотография*). Это же значение может выражаться и адвербиально (*бледно напечатал*). Артефакты особенно наглядно выявляют в *бледном* отрицательную коннотацию, которая, однако, в этих контекстах еще остается следствием основной — цветовой — семантики. Но последняя может практически полностью стираться, как в следующих примерах:

(25) *Почему сборная России в последние годы столь* ***бледно*** *представляет страну?* [Труд-7, 2005.10.20]

(26) *Имеющиеся персоны крайне* ***бледно*** *вели себя на телеэкране.* [Новый регион 2, 2007.12.03]

В аналогичных контекстах может употребляться и наречие *тускло*, ср.:

(27) *Защитник «Интера» Майкон вспомнил слова журналистов, которые часто упоминали о том, что миланский клуб* ***тускло*** *выступает на евроарене.* [http://www.rufoot.ru/node?page=59&cat=forg]

Заметим, что пока что признаковые слова этой группы выражают отрицательную оценку в отношении зрительно воспринимаемых ситуаций, что является своего рода «наследием» их исходного цветового значения. Соответственно, переход ребрендингового типа для этих слов осуществился еще не полностью, но если их контекстная сочетаемость будет продолжать расширяться, то процесс перехода можно будет считать завершенным.

c) имеющий неправильную форму, ср. *кривой*. Пока что в нашем материале эта зона представлена одним признаковым словом, но когнитивный механизм развития отрицательной оценки здесь настолько прозрачен, что можно ожидать, что эта область окажется релевантной в типологической перспективе. Выше (см. Раздел 1) мы уже обсуждали основания для развития в *кривом* семантики отрицательной оценки, так что ограничимся только примерами ребрендинговых употреблений:

(28) ***Кривое*** *объяснение, немного, я просто не знаю как сказать иначе.* [www.diary.ru/~justforus/?userid=595485]

(29) ***Криво*** *объяснил, но смысл думаю понятен*.
[http://forum.windowsfaq.ru/archive/index.php/t-53170.html]

Итак, мы обсудили основные зоны-источники ребрендинговых оценочных значений. В следующем разделе мы рассмотрим, как оценочная семантика видоизменяется в зависимости от контекста и как возникающие в ходе этого взаимодействия частные значения соотносятся друг с другом.

3. Оценочный ребрендинг: взаимодействие с контекстом

До сих пор мы анализировали случаи ребрендинга, прослеживая за развитием преимущественно оценочных значений. Между тем очевидно, что семантика, возникающая у исследованных слов на базе ребрендинга, только оценкой не ограничивается. Легко заметить, что оценочная семантика, накладываясь на значение определяемого существительного или глагола, может реализовываться в виде различных частных ребрендинговых значений.

Так, лексема *фантастический* наряду с оценкой (ср. *фантастическая хозяйка / фантастически готовит* = ‘хорошо / хорошая’) может выражать, во-первых, степень, если выступает при существительных или глаголах с градуируемым значением (ср. *фантастическое везение, фантастически повезло*), а также при прилагательных (ср. *фантастически красивый* = ‘очень’), а во-вторых, количество, если определяемое слово задает счетное множество (ср. *фантастический выигрыш* = ‘много’). Тем самым понятно, что число ребрендинговых значений, или, что то же самое, число реализаций общего ребрендингового значения, напрямую зависит от широты сочетаемости данного слова. Среди нашего материала обнаруживаются как слова, не выходящие за пределы собственно оценки (ср. *бледный / тусклый, чудесный*), так и комбинации из двух (оценка + степень, ср. *редкий голос* <оценка> */ редкий негодяй* <степень>) или трех (оценка + степень + количество, ср. выше *фантастический*, а также, например, *чудовищный*: *чудовищный случай* <оценка> */ чудовищный лентяй* <степень> */ чудовищная зарплата* <количество>).

Интересно было бы выявить факторы, которые обуславливают различия в сочетаемости оценочных слов. По-видимому, в некоторой мере сочетаемость коррелирует с уровнем «освоенности» данного ребрендингового значения языком, т. е. его новизной. Действительно, как можно заметить при сопоставлении текстов разных временных периодов в НКРЯ, ребрендинговое значение, появляясь в языке, употребляется сначала лишь в ограниченном числе контекстов, но их число со временем имеет тенденцию к расширению.

Кроме того, на сочетаемость влияет исходная — доребрендинговая — семантика лексемы. Так, хотя группа *чудовищный / жуткий* и под., равно как и группа *бледный / тусклый*, имеют значение отрицательной оценки, от *бледного / тусклого*, в отличие от *чудовищного*, странно ожидать появления значения высокой степени, поскольку в их семантике исходно заложена как раз обратная идея низкой интенсивности. Однако чтобы системно проследить все

закономерности такого рода и описать общие принципы организации системных связей между отдельными ребрендинговыми значениями, данных одного языка явно недостаточно: надежные обобщения можно будет строить только с привлечением типологического материала.

Мы обсуждали реализации ребрендинговых значений в терминах типов значений (оценка, степень, количество). Между тем существенно, каким образом в языке реализуются комбинации отдельных подзначений этих типов (положительная/отрицательная оценка, высокая/низкая степень, большое/малое количество). Признаковые слова со значением положительной оценки ведут себя вполне предсказуемо: если они допускают семантику степени и количества, то это всегда высокая степень и большое количество, ср. *сказочное путешествие* <положительная оценка> */ сказочная удача* <высокая степень> */ сказочная сумма* <большое количество>.

Отрицательная оценка образует более сложную систему. Она может сочетаться со значением высокой степени, ср. *чудовищно рад / замерз* (отметим, что тем самым значение низкой интенсивности в комбинации с каким-либо оценочным значением нам не встретилось) и большого или малого количества. При этом идея большого количества возникает в контексте имен / глаголов, называющих неприятную для говорящего сущность / ситуацию, ср. *чудовищные налоги* (= ‘много’). Напротив, малое количество соотносится с чем-то, что хорошо для говорящего, ср. *чудовищная зарплата* (= ‘мало’). Тем самым предсказуемым образом добавление отрицательной оценки к идее плохого дает представление о большом количестве этого плохого, а ее комбинация с хорошим — о малом количестве хорошего. Интересно в этой связи, что положительная оценка такой дифференциации не проводит: так, и *приличная зарплата*, и *приличный долг* соотносятся с идеей большого количества.

Таким образом, изучение семантических переходов особого типа — полученных в результате стирания основного семантического содержания признакового слова — открывает целый ряд увлекательных исследовательских задач. В настоящей статье нам было интересно проследить, из каких семантических зон язык заимствует единицы для выражения чистого оценочного значения, какие механизмы переходов при этом задействуются, как оценка, вступая во взаимодействие с контекстом, переходит в другие типы ребрендинга. Детальный анализ каждого из этих типов, а также сопоставление полученных результатов с данными других языков, как мы надеемся, существенно расширит наши представления о природе и закономерностях семантических изменений в языке.

References

1. *Apresian Iu. D.* 1971. On Regular Polysemanticism [O Reguliarnoi Mnogoznachnosti]. Izvestiia Akademii Nauk SSSR. Otdelenie Literatury I Iazyka.
2. *Apresian Iu. D.* 1974. Lexical Semantics (Synonimical Language Means) [Leksicheskaia Semantika (Sinonimicheskie Sredstva Iazyka)].
3. *Birikh A.* 1995. Metonymy in the Modern Russian Language: Semantic and Grammar Aspects.

4. *Blank A.* 1999. Co-presence and Succession: A Cognitive Typology of Metonymy. Metonymy in Language and Thought : 169–191.
5. *Croft W.* 1993. The Role of Domains in the Interpretation of Metaphors and Metonymy. Cognitive Linguistics, 4 : 335–370.
6. *Croft W.* 2006. On Explaining Metonymy: Comment on Geeraerts and Piersman, "Metonymy as a prototypical category". Cognitive Linguistics, 17(3) : 317–26.
7. *Karpova O. S., Reznikova T. I., Arkhangel'skii T.A., Kiuseva M. V., Rakhilina E. V., Ryzhova D. A., Tagabileva M. G.* 2010. Database on Russian Polysemantic Qualitative Adjectives and Adverbs [Baza Dannykh po Mnogoznachnym Kachestvennym Prilagatel'nym I Narechiiam Russkogo Iazyka]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2010" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2010"), 9(16) : 163–168.
8. *Karpova O. S., Kiuseva M. V., Ryzhova D. A., Tagabileva M. G.* Menomymical Transition in Russian Indication Vocabulary [Metonimicheskie Perekhody v Russkoi Priznakovoi Leksike]. m*OST: Österreichische Studierendenkonferenz für junge SlawistInnen.
9. *Kustova G. I.* 2004. Types of Derrivative Meanings and Mechanisms of Language Extension [Tipy Proizvodnykh Znachenii I Mekhanizmy Iazykovogo Rasshireniia].
10. *Lakoff G., Johnson M.* 1980. Metaphors We Live By.
11. *Ostler N., Atkins B. T. S.* 1991.Predictable Meaning Shift: Some Linguistic Properties of Lexical Implication Rules. Proceedings of the First SIGLEX Workshop on Lexical Semantics and Knowledge Representation (Springer-Verlag), 627 : 87–100.
12. *Paducheva E. V.* 1988. On the Paradigm of Regular Polysemanticism (On the Material of Sound Verbs) [O Reguliarnoi Mnogoznachnosti (na Primere Glagolov Zvuka)]. NTI, 2 (432).
13. *Paducheva E. V.* 2004. Dynamic Models in Lexical Semantics [Dinamicheskie Modeli v Semantike Leksiki].
14. *Pustejovsky, J.* 1995. The Generative Lexicon.
15. *Peirsman Y., Geeraerts D.* 2006. Metonymy as a Prototypical category. Cognitive Linguistics, 17(3) : 269–316.
16. *Radden G., Kövecses Z.* 1999. Towards a Theory of Metonymy. Metonymy in Language and Thought : 17–59
17. *Rakhilina E. V., Karpova O. S., Reznikova T. I.* 2009. Models of Polysemantic Qualitative Adjectives Semantic Derication: Metaphor, Metonymy, and its Interaction [Modeli Semanticheskoi Derivatsii Mnogoznachnykh Kachestvennykh Prilagatel'nykh: Metafora, Metonimiia I ikh Vzaimodeistvie]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2009" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2009"), 8 (15) : 2009.
18. *Rakhilina E. V., Karpova O. S., Reznikova T. I.* 2010. Semantic Transitions in Attributive Constructions: Metaphor, Metonymy and Rebrending [Semanticheskie Perekhody v Atributivnykh Konstruktsiiakh: Metafora, Metonimiia I Rebrending].Lingvistika Konstruktsii :396–455.

АНТОНИМИЯ ВО ФРАЗЕОЛОГИИ: ФОРМАЛЬНОЕ СХОДСТВО КАК УСЛОВИЕ ПРОТИВОПОЛОЖНОСТИ ЗНАЧЕНИЙ

К. Л. Киселева (xenkis@mail.ru)

Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Показаны условия возникновения отношения антонимии между двумя идиомами и многоуровневый характер этого явления во фразеологии; обсуждаются формальные признаки и свойства антонимичных пар идиом, в которых участвуют лексические антонимы, контекстные антонимы, отрицательная частица *не*.

Ключевые слова: антонимия, фразеология, антонимичные идиомы, противоположные значения.

ANTONYMS IN PHRASEOLOGY: FORMAL SIMILARITY AS A CONDITION OF THE SEMANTIC OPPOSITENESS

K. L. Kiseleva (xenkis@mail.ru)

Vinogradov Institute of Russian Language of the Russian, Academy of Sciences, Moscow, Russian Federation

The paper deals with semantic oppositeness on various levels of the phraseological system. The data come from an attempt to look at the Russian phraseology in one particular perspective, i. e. to investigate the role that the oppositeness plays between and within idioms. We propose that the semantic oppositeness between idioms can be formed by lexical, contextual and grammatical means. The paper argues that strict antonymy emerges when two idioms have similar structures and are based on the same image. The paper focuses, in particular, on the cases when this oppositeness is formed by negation. Some ways to represent different degrees of negative polarity in the phraseological dictionary are discussed. Different semantic effects related to the negative particle ne in idioms, as in blizhnij svet — neblizhnij svet, k licu — ne k licu, are examined. Finally, we account for the oppositeness as a regular model of the inner form that manifests itself in series of idioms like k mestu i ne k mestu, star i mlad, ni sest' ni vstat' etc.

Key words: antonyms, phraseology, antonymical idioms, semantic oppositeness.

0. Данное исследование представляет собой фрагмент анализа антонимических отношений на разных уровнях фразеологической системы. Работа не претендует на теоретическую новизну, однако обладает, как представляется, практической ценностью как попытка рассмотреть фразеологический материал в узком ракурсе — с точки зрения того, как проявляет себя на этом материале такая лексическая «корреляция широкого охвата» (Кобозева 2000), как антонимия. Этот ракурс предполагает обзор средств выражения отношения антонимии, связывающего разные идиомы, разные значения одной идиомы, а также компоненты внутренней формы идиомы.

1. Применительно к обычной лексике «антонимами могут быть признаны слова, которые противопоставлены по самому общему и существенному для их значения семантическому признаку, причем находятся на крайних точках соответствующей лексико-семантической парадигмы» (Шмелев 1977). Вполне естественно, что антонимия как семантическая противопоставленность двух языковых единиц наблюдается и в сфере фразеологии. Это явление неоднократно описывалось, в частности, в работах А. И. Молоткова, Н. Ф. Алефиренко и Л. Г. Золотых, диссертациях Г. И. Волкотруб, Э. Р. Мардиевой и других. В частности, в Волкотруб 1991 утверждается, что «антонимичность ФЕ <фразеологических единиц> не зависит от совпадения их структуры и наличия в составе антонимичных компонентов». С этой точкой зрения согласна Мардиева 2003, для которой «пары с тождественным компонентным составом не являются единичным типом фразеологических антонимов», однако мы скорее склонны вслед за А. И. Молотковым считать тождество компонентного состава (за исключением одного компонента) признаком чистых фразеологических антонимов, рассматривая иные случаи противопоставленности фразеологических единиц как квазиантонимию.

В целом точная антонимия между двумя идиомами затруднена, на наш взгляд, следующими факторами:

- разная частеречная принадлежность и валентная структура идиом с противоположной семантикой (*собаку съесть* на чем-л. — *ни в зуб ногой* в чем-л.);
- разная внутренняя форма идиом с противоположной семантикой, которая неизбежно накладывает отпечаток на актуальное значение (см. об этом Баранов, Добровольский (в печати));
- сложность/нетривиальность актуального значения, которое невыводимо из внутренней формы (см. 2.4. ниже).
- В Алефиренко, Золотых 2008 фразеологизмы, приводимые в словарных статьях как антонимы, противопоставлены лишь на основе некоторого фрагмента толкования, но это противопоставление никак не подтверждается употреблениями в контекстах. Так, для выражения *вернемся к нашим баранам* предлагается считать антонимом, в частности, *размениваться на мелочи:* речь идет явно о противопоставлении по признаку «главное — неглавное», который является лишь одним из компонентов значения каждой из этих идиом. Мы считаем необходимыми (но недостаточными)

условиями антонимии двух идиом (1) наличие **сильного формального сходства** между ними: в общем случае их внутренняя форма должна совпадать с точностью до отрицания или с точностью до одного элемента, причем несовпадающие элементы должны сами быть антонимами — лексическими (*сильный — слабый, легкий — тяжелый*) или контекстными (*ваш — наш, отцы — сыновья*). (2) их **семантическую членимость**, т. е. относительную семантическую автономность отдельных компонентов (см. об этом Добровольский 2007), позволяющую производить «замену» одного из компонентов на противоположный по смыслу.

2. Пары идиом с одинаковой структурой, содержащие противопоставленные лексические элементы:

2.1. лексические антонимы
со знаком плюс — со знаком минус; этот свет — тот свет; Старый Свет — Новый Свет; левый уклон — правый уклон; сильный пол — слабый пол; белая кость — черная кость; с тяжелым сердцем — с легким сердцем; черти унесли — черти принесли; развязывать руки — связать руки; говорить на одном языке — говорить на разных языках.

На первый взгляд по крайней мере к некоторым из этих пар применима типология лексической антонимии, изложенная в Апресян 1995: *черти унесли — черти принесли* можно отнести к типу НАЧИНАТЬ — ПЕРЕСТАВАТЬ (Anti1), *связать руки — развязать руки* — к типу КАУЗИРОВАТЬ — ЛИКВИДИРОВАТЬ (тоже Anti1); *говорить на одном языке — говорить на разных языках* — к типу P — НЕ P (Anti2); наконец, *сильный пол — слабый пол* попадает в группу антонимов типа БОЛЬШЕ — МЕНЬШЕ (Anti3). Однако последний пример заставляет задуматься о том, что мы на самом деле описываем: антонимию внутренней формы (*сильный* vs *слабый* противопоставлены по типу Anti3) или антонимию актуальных значений (пара *мужчина — женщина* в Апресян 1995 отнесена к особому типу противопоставления по признаку пола). Еще более интересная ситуация возникает в случае пары *Старый Свет — Новый Свет*: при синонимической замене (*Европа — Америка*) антонимия утрачивается, поскольку она «держится» целиком на внутренней форме. Тем самым можно предположить, что взаимодействие актуального значения идиомы и образа, лежащего в его основе, не позволяет напрямую использовать во фразеологии типологию антонимии, разработанную для лексики.

2.2. контекстные антонимы
ваш брат — наш брат; попасть в яблочко — попасть в молоко; белые воротнички — синие воротнички; красная суббота — черная суббота; важная птица — невелика птица.

Существует ряд идиом, чей состав делает их удачными кандидатами в этот класс, однако парная идиома во фразеологической системе отсутствует:

средней руки, страны третьего мира, высшей пробы, с большой буквы. Наличие лексического или контекстного антонима у одного из элементов пары ни в коей мере не предсказывает наличие соответствующего антонимичного фразеологизма: можно *копить на черный день*, но нельзя «копить на белый день»; можно *играть первую скрипку*, но нельзя «играть последнюю скрипку», и т. д. При наличии контекстного противопоставления выражений *зеленый свет* и *красный свет* применительно к светофору в сфере фразеологии в этой паре возникает лакуна: выражение *зеленый свет* и *зеленая улица* существуют, а *красный свет* и *красная улица* в аналогичном смысле — нет. Идиомы, в состав которых входят цветообозначения, вообще часто оказываются непарными: *голубые береты, белая смерть, красный угол, синий чулок, черный континент.* Исходное выражение, по отношению к которому идиома возникает как антоним, может подразумеваться, но реально не встречаться, как в случае *холодной войны* (в противоположность войне с применением оружия, которая мыслится как горячая), *Третьего Рима* (первые два — собственно Рим и Константинополь) и *белой вороны* (по умолчанию все вороны — черные).

2.3. попарно-противопоставленные элементы

Первые компоненты сравнительных оборотов, входящих в этот класс, являются лексическими антонимами, а противопоставление вторых компонентов порождается данной конструкцией и ограничивается ею: *трусливый как* ***заяц*** *— смелый как* ***лев;*** *красив как* ***бог*** *— страшна как* ***смертный грех;*** *богат как* ***крез*** *— беден как* ***церковная мышь.***

2.4. Завершая этот раздел, следует заметить, что идиомы с антонимичной внутренней формой не обязаны быть антонимами. Они могут быть **квазиантонимами**: *спустя рукава* означает работать в первую очередь «плохо», «лениво» (и необязательно мало), а *засучив рукава* — в первую очередь «много», «энергично» (и только во вторую — хорошо). К квазиантонимам можно отнести также пары *выйти из себя — прийти в себя, распустить хвост — поджать хвост, отправиться на тот свет — вернуться с того света.* Далее, они могут иметь никак **не соотносимые актуальные значения** (*черное золото — белое золото; наставить рога — обломать рога*). Наконец, они могут иметь **сходные значения** (*Большая медведица — Малая медведица*), быть **вариантами одной идиомы** (*с головы до ног — с ног до головы; сверху донизу — снизу доверху*) или даже в отдельных экзотических случаях быть **квазисинонимами**: помимо классического примера про Венеру, которую можно назвать и *вечерней звездой,* и *утренней звездой,* существуют и другие, например: *полоскать мозги — сушить мозги.*

3. Пары идиом, противопоставленные грамматически

3.1. содержащие предлоги с противоположным значением

в глаза — за глаза; с оглядкой — без оглядки, с ведома — без ведома, с учетом — без учета.

3.2. различающиеся на отрицание

3.2.1. омонимы

к лицу (об одежде) — *не к лицу* (несоответствие статусу); *каши не просит (что-л.)* (много ресурса) — *просить каши* (про обувь).

3.2.2. квазисинонимы

ближний свет — неближний свет (далеко), *бог весть какой — не бог весть какой* (средненький, плохонький); *черт поймет/разберет — сам черт не поймет/ не разберет; мокрое место осталось (от кого-л.) — мокрого места не останется (от кого-л.).*

Подобное явление, когда введение отрицания не приводит к изменению значения на противоположное, наблюдается и в обычной лексике: *истовый=неистовый.*

3.2.3. антонимы

по душе — не по душе, персона грата — персона нон грата, игра стоит свеч — игра не стоит свеч, прийтись ко двору — прийтись не ко двору.

В связи с описанием и словарным представлением этого (самого многочисленного) класса идиом возникают несколько проблем, которые будут рассмотрены в следующем разделе.

4. Отрицание в идиомах

4.1. антонимы как формальные варианты

Мы исходим из того, что появление отрицания при идиоме в некоторых случаях является композициональным (*Он не **лезет в бутылку***), в других приводит к появлению варианта исходной идиомы (*Он пришелся не **ко двору***), а в третьих вообще затруднено в нейтральном контексте (*?Ему учеба не **по боку**. ?Я не **умываю руки**.*) Мы сейчас оставляем в стороне вопрос о сфере действия отрицания, грамматических и семантических свойствах идиом, влияющих на взаимодействие с отрицанием, и лишь констатируем, что в некоторых случаях у идиомы существует два варианта с противоположной семантикой, которые так или иначе фиксируются в словарях: *(не) для/ради красного словца; (не) отдавать себе отчет(а);(не) бросаться в глаза; (не) в курсе; (не) по душе.* Желательно две идиоматичных единицы, совпадающие по форме и по значению с точностью до отрицания, считать вариантами одной идиомы и описывать в одной словарной статье.

4.2. шкала приемлемости отрицания и способы ее отражения в словаре

Отрицание в идиоме может быть обязательным, оно может опускаться в некоторых строго ограниченных случаях, может опускаться достаточно легко при сохранении формы с отрицанием как исходной и, наконец, варианты

с отрицанием и без него могут быть (почти) равноправны. Желательно не только определить степень допустимости/обязательности отрицания в составе идиомы, но и последовательно отразить эту информацию в словарной статье, как это делается, например, для порядка слов во Фразеологическом объяснительном словаре. Для тех идиом, для которых ведение/снятие отрицания вообще релевантно, мы предлагаем выделить несколько разных классов и, соответственно, представлять их в словаре разными способами.

Эти классы идиом располагаются на шкале приемлемости отрицания так:

идиомы, всегда содержащие отрицание — сильные эксплицитно-негативные — слабые эксплицитно-негативные — идиомы, у которых формы с отрицанием и без отрицания равнозначны — слабые эксплицитно-позитивные — сильные эксплицитно-позитивные — не допускающие отрицания.

4.2.1. идиома всегда содержит отрицание

не то слово, не без того, не баран чихнул, не за горами, не долго думая, не отходя от кассы, не занимать, не ахти, не находить себе места, не с руки; не первой молодости; не по-детски, не по годам, не к спеху.

Вариант без «не» в некоторых контекстах возможен, но является однословным, имеет другой, неидиоматичный смысл *(долго думая, отходя от кассы)* или употребляется как цитата в диалоге (- *Вам не к спеху? — Очень даже к спеху*). Очевидно, что в этом случае вопрос о словарной форме идиомы не возникает — она (как и все стандартные примеры употребления) обязательно содержит отрицание.

4.2.2. сильные эксплицитно-негативные

не бери в голову, не лезть за словом в карман, палец в рот не клади, не (чьего-л.) ума дело

В Баранов, Юшманова 2000 сильными эксплицитно-негативными названы глагольные идиомы «с неустранимым отрицанием»: мы этот ярлык расширительно и полагаем, что отрицание здесь действительно неустранимо, но может выражаться не только в виде отрицательной частицы «не» (хотя ее опущение сильно затруднено). В Апресян 1995 такие идиомы описаны как тяготеющие к закреплению в отрицательной форме, но способные «употребляться в вопросительных, «сомнительных» и модальных предложениях, содержащих в лучшем случае лишь имплицитное отрицание».

Для данного класса мы предлагаем в лемме ограничиться вариантом с отрицанием, а в специальной зоне примеров, расположенной между стандартными и нестандартными употреблениями, приводить регулярные трансформации, при которых возможно синтаксическое перемещение отрицания или его выражение иными средствами. Например, для идиом *палец в рот не клади (кому-л.)* и *не бери в голову* можно привести такие трансформации:

(1) *Во главе каждого «вида» журналистики <…> стоят не мальчики-паиньки и не девочки-модницы, а матерые львы и львицы, которым* ***палец в рот***

класть *не рекомендовано: откусят. (В. Аграновский. Вторая древнейшая. Беседы о журналистике)*

(2) *Ну-ну, не шалите… Вам только* ***палец в рот положи****… Да я знаю, все я знаю — и как вам трудно живется, и что многие недовольны ролями и моим характером; ничего не поделаешь — <…> терпите, какой есть… (В. Смехов. Театр моей памяти)*

(3) *Маша, меньше* ***в голову бери*** *— больше в желудок. (Н. Горланова. История озера Веселого)*

(4) *— А! Брось, баба Леля! — отмахнулась Римма. — Кто это сегодня* ***берет в голову*** *«амур налево»? Ну, даст жена по морде… И сама сходит в этом же направлении… (Г. Щербакова. Митина любовь)*

Анализируя регулярные трансформации, допустимые для сильных эксплицитно-негативных идиом, важно отделять их от нестандартных употреблений. Так, пример (5) можно считать регулярной трансформацией идиомы *не (чьего-л.) ума дело (что-л.),* а пример (6), по-видимому, находится за пределами стандарта:

(5) ***Нашего*** *ли* ***ума дело*** *хлопотать о справедливости, если огненными письменами начертано: «Мне отмщение, и Аз воздам». (Т. Набатникова. День рождения кошки)*

(6) *Повеселели ребята. И Иван Денисычу тоже тихо говорят: бригадир процентовку хорошо закрыл. Веселый пришел. Уж где он там работу нашел, какую — это* ***его****, бригадирова,* ***ума дело****. Сегодня вот за полдня что сделали? Ничего. (А. Солженицын. Один день Ивана Денисовича)*

4.2.3. слабые эксплицитно-негативные
чаша сия ***не*** *миновала/минует;* ***не*** *болит голова (у кого-л. о чем-л.); конца* ***не*** *видно; путь* ***не*** *усыпан розами*

Для данной группы оба варианта идиомы возможны, но при этом исходной является форма с отрицанием, а форма без отрицания производна. В лемме приводится только исходная форма с отрицанием, а в зоне примеров помещаются употребления идиомы в обеих формах.

(7) *Вот — радовалась на большие окна, а сколько с ними возни — клеишь и клеишь, а* ***конца не видно****. (И. Грекова. Первый налет)*

(8) *Учусь. Зверски устаю. Особенно мои несчастные глазки. Но* ***конец виден****, да. Осталось всего-то ничего: четыре экзамена, четыре редактирования и Фаулз. (katesina.diary.ru)*

(9) *Могу предположить, что* ***путь*** *«Нестле» на российский рынок* ***не был усыпан розами****. (Дипломатический вестник)*

(10) *Бывают <…> великие артисты, творческий* ***путь*** *которых* ***усыпан розами****. (Ю. Елагин. Укрощение искусств)*

4.2.4. негативно-нейтральные: варианты равноправны

Самый многочисленный подкласс данного типа идиом — сочетания имени с предлогом: *(не) на высоте;(не) по пути; (не) по карману; (не) по плечу; (не) в ударе.* Следует отметить, однако, что оба члена этих пар являются именно идиомами, которые могут употребляться самостоятельно, вне контраста: *Он был сегодня (не) на высоте. Ему это (не) по плечу.* Далеко не все идиомы такой структуры имеют «полноценные» пары. Например, идиомы *в пику, по боку, в годах, до гроба* не имеют парных. Если такие употребления и встречаются, то только цитатно или в контрастивном контексте. *?У них любовь не до гроба. *Он мужчина не в годах.* Есть и другие негативно-нейтральные идиомы: *(не) дай бог; в (не)урочный час понюхать пороха — не нюхать пороха.* Помещая такие идиомы в словарь, мы сталкиваемся с проблемой выбора формы леммы. Одно из возможных решений: в соответствии с лексикографической традицией выбрать в качестве основной формы леммы позитивную, а в качестве варианта указать негативную. Такое решение принято в Лубенская 2004 применительно к идиомам *по душе, бросаться в глаза, играть в бирюльки*: вариант с отрицанием приводится в конце леммы в зоне Neg. Однако для некоторых идиом (например, *не по карману*) в качестве основной предлагается форма с отрицанием, а форма *по карману* приводится в конце леммы. В любом случае в зоне примеров должны быть представлены употребления в обоих вариантах.

На этом мы заканчиваем описание возможных способов отражения антонимичных форм идиомы во фразеологическом словаре. По другую сторону «нуля», где располагаются слабые эксплицитно-позитивные, сильные эксплицитно-позитивные и не допускающие отрицания идиомы, можно действовать аналогично.

5. Антонимия внутри идиомы

К данному виду фразеологической антонимии мы относим несколько разных явлений.

Во-первых, имеется в виду **энантиосемия**, т. е. противопоставление разных значений одной идиомы (Кравцова 2006). Об этом явлении дают представление такие идиомы, как

- *выйти из окопов* (прекратить воевать vs пойти в атаку)
- *на днях* (прошлое vs будущее)
- *проливать кровь* (свою/чужую)
- *иди с богом* (выгнать или пожелать счастливого пути).

Во-вторых, идиома, представляющая собой глагольную группу или именную группу с глагольными словами-сопроводителями (Жуков, Сидоренко, Шкляров 1987), может иметь разного рода **противопоставленные варианты.** В основном в рамках одной идиомы встречаются конверсивы (*брать*

на лапу — дать на лапу) и каузативы (*снять голову — полетели головы*), но возможны и антонимические противопоставления, например *держать бразды правления — выпустить бразды правления*.

В-третьих, антонимия является очень распространенным **приемом построения внутренней формы**. «Строительным материалом» могут оказаться, в частности:

- единицы, одна из которой является отрицанием другой: *знать не знаю, видимо-невидимо, была не была, к месту и не к месту, волей-неволей, хочешь — не хочешь*.
- векторные антонимы по Новикову 1973: *вдоль и поперек, направо и налево, взад-вперед, ни сесть ни встать, бог дал, бог взял, туда-сюда*.
- противоположные «концы» некоторой сущности (*альфа и омега, с утра до вечера, ни дна ни покрышки, с головы до пят, оставить рожки да ножки, и в хвост, и в гриву*)
- лексические и контекстные антонимы в конструкции с семантикой охвата всех элементов некоторого множества: *стар и млад, города и веси, кнут и пряник*.
- лексические и контекстные антонимы в конструкции с семантикой тотального отсутствия качеств, поддающихся позитивному определению: *ни то ни сё, ни горячо ни холодно, ни рыба ни мясо, ни два ни полтора, ни к селу ни к городу*.
- лексические антонимы в конструкции с семантикой неадекватности: *вливать новое вино в старые мехи, валить с больной головы на здоровую*.
- антонимы, указывающие на контрастные состояния: *ни жив ни мертв, и смех и слезы*.

Нельзя не отметить, что в качестве такого строительного материала выступают не только антонимы, но также **синонимы** (*целиком и полностью, в пух и прах, в целости и сохранности, худо-бедно, серединка на половинку, всем и каждому)*, **однокоренные элементы** (*день-деньской, чин-чинарем, мало-помалу, рад-радешенек, валом валить, перво-наперво*), **фонетические дубли** (*фигли-мигли, шуры-муры, тары-бары, йоксель-моксель*) и **два вхождения одной единицы** (*от корки до корки, капля за каплей, со дня на день, нос к носу, бок о бок, вот-вот, еле-еле, тайное тайных, с рук на руки, в конце концов*). Данное явление можно считать структурообразующим во фразеологии как системе и поэтому оно заслуживает отдельного изучения (Киселева (в печати)).

6. Заключение

Предложенный фрагмент описания антонимии во фразеологии может быть полезен (1) для описания фразеологии как многомерной системы, в которой такое явление, как антонимия, присутствует в разных формах на разных уровнях; (2) для лучшего понимания нетривиального соотношения внутренней формы и актуального значения идиом; (3) для формирования критериев разграничения антонимов и квазиантонимов во фразеологии; (4) для анализа особого вклада отрицательной частицы «не» в формирование семантики фразеологических единиц.

References

1. *Alefirenko N. F., Zolotykh L. G.* 2008. Phraseology Dictionary. Cultural and Cognitive Space of Russian Idiomatic Expresiions [Frazeologicheskii Slovar'. Kul'turno-Poznavatel'noe Prostranstvo Russkoi Idiomatiki].
2. *Apresian Iu. D.* 1995. Proceedings. I–II.
3. *Baranov A. N., Dobrovol'skii D. O.* 2008. Aspects of the Theory of Phraseology [Aspekty Teorii Freazeologii].
4. *Baranov A. N., Dobrovol'skii D. O.* The Fundaments of Phraseology [Osnovy Frazeologii].
5. *Baranov A. N., Iushmanova S. I.* 2000. Denial in Idiomatic Expressions: Semantic-Syntactic Limitations [Otritsanie v Idiomakh: Semantiko-Sintaksicheskie Ogranicheniia]. Voprosy Iazykoznaniia,1.
6. *Dobrovol'skii D. O.* 2007. Semantic Parcelling as a Factor of Idiom's Variability [Semanticheskaia Chlenimost' kak Faktor Variativnosti Idiomy]. Iazyk kak Materiia Smysla. Sbornik Statei v Chest Akademika N.Iu. Shvedovoi.
7. *Kiseleva K. L.* (In press). On the Types of Formal and Semantic Duplication in Idiomatic Field: Synonyms, Tautology, Replications, Fonetic Duplets [O Tipakh Formal'nogo I Semanticheskogo Dublirovavniia v Idiomatike: Sinonimy, Tavtologiia, Povtory, Foneticheskie Duplety].
8. *Kobozeva I. M.* 2000. Linguistic Semantics [Lingvisticheskaia Semantika].
9. *Kravtsova V. Iu.* 2006. Enantiosemy of Lexical and Phraseological Unities: Language and Speech [Enantiosemiia Leksicheskikh I Frazeologicheskikh Edinits: Iazyk I Rrech'].
10. *Mardieva E. R.* 2003. Principles of Formation of the Russian Phraseological Antonyms Dictionary [Printsipy Sostavleniia Slovaria Frazeologicheskikh Antonimov Russkogo iazyka].
11. *Lubenskaia S. I.* 2004. Large Russian-English Phraseological Dictionary [Bol'shoi Russko-Angliiskii Frazeologicheskii Slovar'].
12. *Novikov L. A.* 1973. Antonymy in Russian Language [Antonimiia v Russkom Iazyke].
13. *Russian* Phraseology Explicative Dictionary [Frazeologicheskii Ob"iasnitel'nyi Slovar' Russkogo Iazyka]. 2007.
14. *Shmelev D. N.* 1977. Modern Russian Language [Sovremennyi Russkii Iazyk].
15. *Volkotrub G. I.* 1991. Phraseological Antonymous-Synonymous Paradigms of Modern Russian Language [Frazeologicheskie Antonimiko-Sinonimicheskkie Paradigmy Sovremennogog Russkogo Iazyka].
16. *Zhukov V. P., Sidorenko M. I., Shkliarov V. T.* 1987. Russian Phraseological Synonyms Dictionary [Slovar' Frazeologicheskikh Sinonimov Russkogo Iazyka].

ВИДЫ ИМИТИРУЕМЫХ ЭМОЦИОНАЛЬНЫХ ЭКСПРЕССИВНЫХ СОСТОЯНИЙ В РУССКОЯЗЫЧНОМ ЭМОЦИОНАЛЬНОМ КОРПУСЕ

А. А. Котов (kotov@harpia.ru)

Национальный Исследовательский Центр
Курчатовский Институт, Москва, Россия

Люди часто имитируют выражение эмоций в коммуникации. Мы предполагаем, что такая имитация управляется другими скрытыми реакциями, и предлагаем первичную классификацию. Мы также дополняем архитектуру компьютерного агента, чтобы он смог использовать имитируемые эмоции.

Ключевые слова: эмоции, имитация эмоций, имитируемые эмоции, классификация, компьютерный агент.

TYPES OF SIMULATED EMOTIONAL EXPRESSIVE STATES IN THE RUSSIAN EMOTIONAL CORPUS

A. A. Kotov (kotov@harpia.ru)

National Research Centre “Kurchatov Institute”,
Moscow, Russian Federation

People often simulate expression of emotions in communication without actual emotional arousal. We suggest that such simulation is forced by other hidden reactions and propose an initial classifi cation. We also extend the architecture of a computer agent to make it able to produce simulated emotions.

Key words: emotions, simulation, simulated emotions, classification, computer agent.

1. Introduction

Imagine that we ask a person if he likes bananas, and he replies *No!,* wrinkling his nose and showing great disgust. Although the expression can be quite evident, the stimulus for the emotion — "imagined banana" — is too weak to force such a strong feeling. We rather say that the person simulates the emotion in order to communicate his appraisal and make the communication more effective or enjoyable. Sometimes such simulated or intentionally performed emotions are described as "pull-emotions" (in contrast to "push-emotions" where expression is forced by an incoming stimulus and internal arousal) [1].

Combination of different types of emotions is important for the development of emotional computer agents, which should understand human emotions or produce emotional cues in communication [2]. Sophisticated computer agents should not only balance between several emotional states, but also activate several "emotions" at the same time, simulate internal emotional conflicts and produce compound emotional patterns. In particular, Greta agent simulates "blended emotions" where the initial arousal (despair) is masked by the superficial pattern (anger), both emotions simultaneously control the facial expression, creating a complicated expressive pattern [3].

The studies of emotional communication are carried out on the basis of emotional corpora, and further allow modeling of the agent's behavior [4]. We collect and annotate the Russian Emotional Corpus[1] (REC) [5]. It contains 295 audiovisual recordings or oral university exams (total length 29.5 hours) and 510 recordings of interactions with clients at a municipal office on the questions of payment for municipal utilities (total length 32 hours). The material of REC shows that people quite often simulate different emotions in order to color up their speech, to make the communication more effective or to manipulate the addressee. Our goal is to create a typology of simulated emotions (based on REC) and extend the architecture of a computer agent to make it simulate not only "original" (or "push"), but also "simulated" ("pull") emotions.

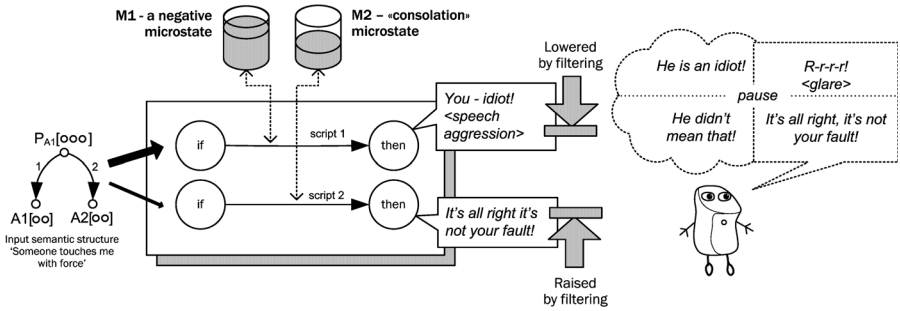
2. Mechanism of reaction substitution

We develop software agents, which react to incoming phrases, encoded as semantic trees, and provide rich emotional reactions.

The agent is controlled by a set of scripts (*rule-based* or *productive* model), where premises and implications are represented by semantic trees with sets of semantic markers in tree nodes [6, 7]. Scripts are activated by incoming semantic trees: events or phrases after parsing. The trees may contain "emotional" markers, but the agent may ignore or even challenge incoming emotional appraisal, depending on its internal state. Activated scripts may generate a set of reactions: (a) output semantic trees for speech synthesis, (b) ready-made phrases and (c) gestures in BML format [8, 9]. This output protocol is used to animate the behavior of 2D and 3D computer figures.

---

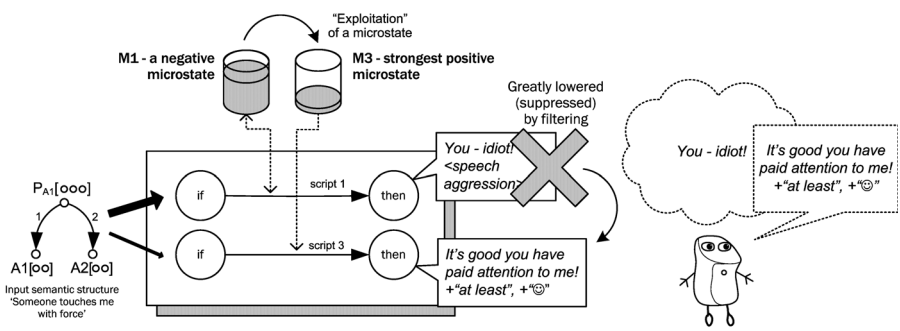
[1] Information page for REC is available at http://www.harpia.ru/rec/
Corpus is not available for copying or distribution, but is available for research purposes and verification, please, contact authors.



Scheme 1. Reactive scheme for the computer agent simulating emotional oscillation

Scripts are controlled by *microstates* — short emotional or communicative states. An input event may match several scripts, so the agent may activate several microstates, which further discharge, compete and control the agent's behavior for several minutes, simulating the "emotional oscillation" [10]. Microstates can be used, in particular, to simulate the case of *sarcasm*, where a negative emotional appraisal in speech is replaced by a positive utterance with irony markers [11].

As represented on Scheme 2, if the agent activates and suppresses Script 1, corresponding to a negative microstate, and looks for the best script, matching the input and corresponding to the opposite microstate — Script 3. Initially Script 3 has received too low activation to output its phrases and gestures (other — negative — reactions were more appropriate), but here it is exploited by Script 1 for sarcastic output with irony markers. A similar substitution framework may apply to make the agent generate simulated emotions.



Scheme 2. Reactive scheme for the computer agent who has been 'hit' and uses sarcasm in his reply

We suggest the following definition for the phenomena. Simulated emotional expressive states appear when some suppressed emotional or rational reaction (script) exploits another emotional reaction (script) to output its expressive patterns, because this output is considered more effective to achieve the communication goals.

As the superficial reaction is required to be emotional, scratching and fiddling (which may be resulted by nervousness) are not considered in our classification.

Of course, the superficial reaction may simply cover an inappropriate internal reaction, but we also need a selector to suggest which of the numerous superficial reactions to choose, and why this superficial reaction should be an emotional pattern: we cannot choose a random emotional reaction in case of confusion. So we involve the notion of "goal" in our definition, suggesting that the superficial reaction is chosen for some emotional interaction with the addressee in the difficult situation, where we have to hide our primary reaction.

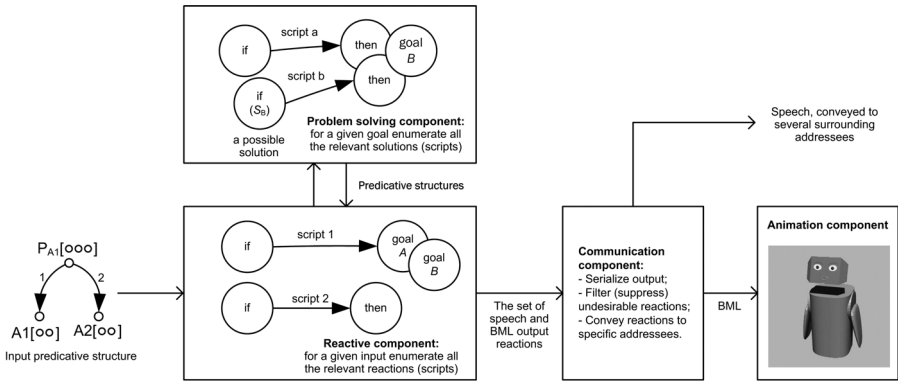
3. Goal processing by the computer agent

The central processor of the agent contains two main components: *reactive* and *problem solving*. In the reactive component the agent receives an input event, enumerates all the scripts with the corresponding premises, and generates a number of microstates with expressive cues (phrases and gestures) for the output. In the problem solving component the agent for a given goal enumerates all the scripts with the corresponding implication in order to find a script which leads to (implies) the given goal.

Goals cannot be directly loaded at the input: all the instructions and commands should meet the agen't internal motivation and thus should be processed by the reactive component (some commands may be ignored, some may even "irritate" the agent). To set up a new goal the agent should receive an input predicative structure, activate a reactive script (script 1), which constructs a single goal predicative structure or a composition of such goals. For the problem-solving task we use a combination of two goals (goals *A* and *B*). A student in an oral exam has to find a solution S_B for a theoretical question (goal *B*) and satisfy the examiner with his answer (goal *A*). We distinguish theoretical goals, which only require a theoretical solution, and performance goals, which should be reached through action. For an exam task goal *B* is theoretical (we only need to find a theoretical solution), while goal *A* is a performance goal (we have to influence the examiner). A robot may be asked to clean water on the floor, so it has to find a way to do it (theoretical goal *B*), perform and clean the water (performance goal *B*), and check if the human is satisfied (performance goal *A*).

Goals may be linked with actions (to be executed to achieve the goal), but if the agent doesn't know a specific action to be performed for the goal, the goal is transferred to the problem solving component (like goal *B*). In this component the agent enumerates all the possible scripts (scripts *a, b*), which may lead to the successful conclusion (goal *B*), and for the best of these scripts the problem solving component sends back the initial model S_B. S_B may suggest an action to be performed or an answer to be communicated, in both cases the agent finally should satisfy the main goal *A*.

If the agent experiences difficulties with achieving goal *A* or goal *B*, it may use simulated emotional cues to bypass the task and achieve a successful conclusion through emotional performance. In this architecture, the simulated expressive states are divided into two groups: states forced by the problem solver (while searching for a solution for goal *B*), and states aimed at influencing the addressee (goal *A*).



Scheme 3. Interaction between reactive and problem solving components

4. Types of simulated emotional expressive states

Following the definition of the phenomena, in our classification we have to answer the following main questions. What is the master reaction to provoke the simulated expression? Why are usual expressive means for the reaction not available? What subordinate reaction is used for expression? How are the expressive cues modified?

Classes of simulated emotional expressive reactions in REC are represented in Table 1.

Table 1. Classes of the simulated emotional expressive states

	Function of the simulated emotional expressive state	**Motive for the simulated emotion**	**Simulated emotional patterns**	**Imaginary statement of the speaker**
1.	Good performance during problem solving (PS)	Social situation during PS, failure during PS	Interest, inspiration	"You should appreciate my efforts"
2.	Bad performance during PS	Failure during PS, request	Frustration, caprice, nervousness, pain	"I feel bad, help me to find the solution"
3.	Shared appraisal	Hedge, <pride>	Disregard, <admiration>	"We all understand, that it's not so important", <You should admire my efforts!>

	Function of the simulated emotional expressive state	**Motive for the simulated emotion**	**Simulated emotional patterns**	**Imaginary statement of the speaker**
4.	Negative influence, pressure	Undesirable action of the addressee	Blame, indignation (sarcasm)	"You are doing something bad, this causes my negative emotion, you should take it into account and stop doing that!"
5.	Positive influence, manipulation	Failure, request	Provoke pity or look appealing, tiredness	"I'm nice and tired, I deserve some mercy"

In all the cases the original underlying reaction affects the social face of the speaker (when students fail to answer the question or clients have to acknowledge mistakes in their payments) or the social face of the listener (when students try to receive a positive mark or clients ask the officer to solve their question). This "face threatening act" forces the speaker to conceal the initial reaction and replace it with a simulated expression in order to influence the addressee through a symmetric communication (shared appraisal) or through a positive or negative emotional influence.

4.1. Good performance during the PS

Students may simulate "good" performance and try to satisfy the addressee (achieve goal *A*) through superficial cues and simulated emotions, even if the solution S_B is not found. They may show great concern and fussiness, perform different iconic gestures and manipulative cues (as if operating with physical objects) when explaining their answer or when looking for a solution. The intensity of these actions should convince the examiner of the student's competence. Student also show intense emotional cooperation with the examiner: they nod and repeat answers, accepted by the examiner. Students may also simulate emotions, linked with the problem solving task: exaggerated interest or inspiration.

Interest — Students change their gaze direction and look aside, as if trying to discern the answer in the air. They intensely squint and frequently change the gaze direction (for example, in 20081225-fipp-a02, 01:08).

Inspiration — Students show inspiration when starting to look for the solution: they sit up, frown, cough, say "*Ok!*" (or a similar interjection), look around or start to organize objects on the table (for

example, in 20080717-c01, 00:42).
These actions should "exteriorize" the process of problem solving for the examiner and through the intensity of actions show a great motivation of the student to solve the problem. For the emotional agent it means, that if the problem solving task includes social aspect and the solution has to be communicated and evaluated by the addressee (goal *A*), the agent may exploit and exaggerate different expressive means, normally evoked by the problem solving component.

4.2. Bad performance during the PS

People may show failures and negative emotional appraisal of the situation in order to provoke the opponent to correct the situation. During an oral exam student may start to answer and then express sudden failures in order to force the examiner to suggest the answer. For example in (20080717-f-psy, 00:41) during an answer a student starts to count types of psychological scales (the exam question) on the fingers of her left hand, and then hesitates on "the last finger", tapping it with a right hand finger and forcing the examiner to suggest the name of the last type.

Tiredness	In (20081225-fipp-b3, 01:07) a student simulates tiredness and exhales deeply while trying to find an answer. In (090623-a17) a client (female) expresses exhaustion and says *I'm tired because I keep having to come to you!* (implying: 'You make too many mistakes each time'). Through this superficial emotional pattern she manages to conceal her face threatening act (accuse the officer) and provoke assistance.
Caprice	In (20081230-a13, 01:45 and 01:51) a student (female) simulates child-like caprice when being asked to suggest an example; immediately after she asks the examiner to suggest the example himself, so that she analyses it.
Frustration	In (20081225-fipp-a02, 04:04) a student reports with regret and concern: *I don't seem to be able to remember it today* (also reducing the significance as in 4.3).
Emotional appraisal	A person may show negative emotions about the current situation (which should be fixed) or about a potential situation (which should be avoided) to conduct the dialogue in a desirable direction. For example in (091005-b11, 01:04) when a client (female) asks how to fill in the form, she adds *Because I was getting so nervous about that!* Here she refers to own negative emotions to motivate the opponent and to justify her request.[2] In (091005-b17, 00:45) a client worries that he has received

[2] This also corresponds to strategy #6 for negative politeness "give overwhelming reasons" [12: 189].

the same bill twice and paid it only once. When the officer confirms, that he paid correctly, he adds: *Or, I thought, they could snatch* [the money] *once again!* Here he conveys his emotional appraisal by using the high intensity verb *snatch* [13], representing to the officer a negative situation to be avoided.
As shown by the examples, if the agent fails to find a solution S_B for a local goal B (answer to the current question) he may choose and express moderate negative emotions in order to force a cooperative addressee to help him with the task.

Students can show expressive patterns for strong negative emotions — like pain and aggression — modifying these patterns with markers of irony: they smile and turn their head aside from the addressee to dissimilate the patterns with real emotions.

Pain — In (20081231-a2, 01:06) a student (female) performs an expressive pattern for pain (squints, bares teeth) when answering a question.

Aggression — In (20081230-a24, 01:58) a student (female) shows pattern of aggression (growls, squints) modified by smile, when she receives a question, which she failed to answer last time.
To perform this expression with a computer agent, we have to execute the following activation pattern. If a computer agent activates and suppresses a negative emotional reaction, it may choose a more "prototypic" reaction for that emotion (e. g. choose 'pain' for being displeased, choose 'aggression' for resentment) and express this reaction accompanied with markers of irony (look aside, smile).

4.3. Shared appraisal

In negative situations where the speaker wants to reconcile with the addressee, he may try to reduce the significance of his own negative action (or the action of the addressee) by saying *Oh! It's not a big deal!* He expects that the addressee shares appraisal and shall not be upset or angry as a result of the situation. The same can be achieved by demonstrating disregard for the result of these negative actions, e. g. by squinting, wrinkling nose, waving a hand. This emotional mechanism extends the notion of "hedge" — "a particle, word or phrase, that modifies the degree of membership of a predicate or noun phrase in a set", e. g. *You are quite right* [12: 145]. A pattern of disregard makes a phrase less definitive and secures the speaker from possible mistakes, as the mistakes will look less significant.

Disregard — In (20081209-zhurn6, 01:22) a student shows disregard (squints, wrinkles nose, turns the head slightly aside), when hesitating and reporting a rejected answer immediately followed by a suggested answer: *It's not with different meanings,*

it's in different situations.
When forced to report the answer (goal *A*) and not being confident about the answer (S_B) the agent may exploit the expression of disregard to reduce the definitiveness of the answer, and bypass the expected criticism from the addressee.

We expect, that on a wider corpus the shared appraisal substitution may also apply to positive emotions, for example, when a speaker wants to confirm his pride (which is not very modest) and communicates admiration for the situation (and for his own actions) to be shared by the addressee.

4.4. Negative influence, pressure

If a speaker is not satisfied by the action of the addressee, he may show blame, indignation or even aggression to change the addressee's behavior.

Blame, indignation

In (20081717-c15, 00:14) the examiner is surprised to meet a student (female) who is on the list, but whom he doesn't remember (surprise here may serve as a simulated pattern, concealing a reproach). The student replies with exaggerated indignation: *How can it be? We talked so nicely during the consultation? How could you forget?* Here the emotional simulation allows the bypassing of unpleasant explanations. Exaggerated indignation is expressed by clients of the municipal office, for example, in (20090623-a35).
These emotional reactions are normally activated by undesirable actions of the opponent (for example, his question). The agent may simulate indignation or laughter (2009sp02, 07:25), based on the incoming request, where this request applies to him undesirable obligations or otherwise threatens his face.

4.5. Positive influence, manipulation

If we have a higher level goal *A* 'to satisfy the listener' and all the subordinate means to achieve this goal fail, the agent may look for an "influence" script to satisfy the goal through emotional manipulation.

Provocation of pity / tender emotions

In (20081230-b4, 08:35) a student (female) whimpers and holds her hands up to her chest like a rabbit or puppy. In (20081225-fipp-a14, 03:17) a student (female) intentionally deeply breathes and watches the addressee (this also corresponds to 4.2 "Tiredness").

The agent can use the problem solver to find the best influence script for the given circumstances: goal *A* (not goal *B*). It may look for the best action, which influences the speaker and leads to goal *A* not through a successful performance, but through emotional manipulation. This strategy should not be always condemned: a mobile home robot should maintain a good impression even if it fails to execute the users instructions.

5. Conclusion

The suggested classification has many limitations. All the situations in REC include problem solving in a situation of a long communicative distance. There is no friendly chat, no romantic or tender relations, where we can expect numerous types of emotional simulation and games. Due to the same limitation we didn't observe emotional games, when both parties simulate similar (or the opposite) emotions and take into account the mutual role play.

We expect that emotional roles can be explored on a wider corpus and should be simulated by a more complicated architecture, which can apply an emotional role and play it through numerous dialogue turns, not only to trigger several emotional reactions, based on 1–2 incoming events, as it is presently the case in our architecture.

At the same time, classifications of emotional expressions in natural corpora are useful for several reasons: (a) they help to collect and organize expressive patterns for naturally occurring emotional states (not only for basic emotions), (b) they help to design the "alphabet" of expressive means/states to be used by emotional agents and mobile robots, (c) they help to understand emotional dynamics and the structure of cognitive mechanisms, involving rational and emotional processing during communication. All this may advance our design of mobile robots and computer agents, supporting believable emotional interaction with humans.

References

1. *Brown P., Levinson S. C.* 1987. Politeness: Some Universals in Language Usage (Studies in Interactional Sociolinguistics).
2. *Cassell J., Sullivan J., Prevost S.,* et al. 2000. Embodied Conversational Agents.
3. *Glovinskaia M. Ia.* 2004. Hidden Hyperbola as a Manifestation and Excuse of Speech Aggression [Skrytaia Giperbola kak Proiavlenie I Opravdanie Rechevoi Agressii]. Sokrovennye Smysly : 69–76.
4. *Kotov A. A.* 2009. Patterns of Emotional Communicative Reactions: the Problems of Creating of a Corpus and Computational Agents Transfer [Patterny Emotsional'nykh Kommunikativnykh Reaktsii: Problemy Sozdaniia Korpusa I Perenos na Komp'iutenrykh Agentov]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2009" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2009"), 8 (15) : 211–218.

5. *Kotov A.* 2005. Application of Psychological Characteristics to D-Script Model for Emotional Speech Processing. ACII 2005, LNCS 3784 : 294–302.
6. *Kotov A.* 2007. Simulating Dynamic Speech Behaviour for Virtual Agents in Emotional Situations. Affective Computing and Intelligent Interaction : 714–715.
7. *Kopp S., Krenn B., Marsella S.,* et al. 2006. Towards a Common Framework for Multimodal Generation: The Behavior Markup Language. Intelligent Virtual Agents : 205–217.
8. *Kotov A. A.* 2009. Management of the Speech Behavior Dynamics of Virtual Computational Agents [Upravlenie Dinamikoi Rechevogo Povedeniia Virtual'nykh Komp'iuternykh Agentov]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2009" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2009"), 7 (14) : 241–247.
9. *Kotov A.* 2009. Accounting for Irony and Emotional Oscillation in Computer Architectures. Proceedings of International Conference on Affective Computing and Intelligent Interaction ACII 2009 : 506–511.
10. *Ochs M., Niewiadomski R., Pelachaud C.,* et al. 2005. Intelligent Expressions of Emotions. ACII 2005, LNCS 3784 : 707–714.
11. *Rehm M., André, E.* 2008. From Annotated Multimodal Corpora to Simulated Human-Like Behaviors. Modeling Communication with Robots and Virtual Humans : 1–17.
12. *Scherer U., Helfrich H., Scherer K. R.* 1980. Paralinguistic Behaviour: Internal Push or External Pull? Language: Social psychological perspectives : 279–282.
13. *Vilhjálmsson H., Cantelmo N., Cassell J.,* et al. 2007. The Behavior Markup Language: Recent Developments and Challenges. Intelligent Virtual Agents : 99–111.

ЖЕСТОВЫЕ ИДИОМЫ И ЖЕСТЫ: ТИПЫ СООТВЕТСТВИЙ

А. Д. Козеренко (akozerenko@mail.ru)

Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

В работе производится семантический анализ русских идиом, внутренняя форма которых содержит жест. Изучается связь между семантикой идиомы и соответствующего жеста. На примере толкований конкретных идиом проиллюстрированы разнообразные типы соответствий жеста и жестовой идиомы.

Ключевые слова: идиома, жест, внутренняя форма, семантика, жестовая идиома.

GESTURE IDIOMS AND GESTURES: TYPES OF CORRESPONDENCE

A. D. Kozerenko (akozerenko@mail.ru)

Russian Language Institute Russian Academy of Sciences, Moscow, Russian Federation

The paper considers semantic analysis of Russian idioms, depicting gestures in their inner form. The relationship between the meaning of a gesture and that of the corresponding idiom is examined, as well as polysemy and synonymy relations between idioms, corresponding to the same or different gestures. Defi nitions of some idioms of the semantic fi eld SADNESS, REGRET, DESPONDENCY are demonstrated. Statements made on the semantics of idioms are illustrated with examples of idiom usage in contemporary texts.

Key words: idiom, gesture, inner form, semantics, gesture idioms.

0. В данной работе мы хотим затронуть тему соотношения жестовых идиом и жестов. Фактически эта область находится на стыке фразеологии и невербальной семиотики, однако нас в первую очередь будет интересовать языковая сторона вопроса: какие идиомы мы называем жестовыми, как они семантически соотносятся с соответствующими жестами (в особенности это интересно в случаях многозначности идиом или многозначности жестов),

и как это учитывается в толковании идиомы. Исследование семантики идиом проводится в рамках проекта Отдела экспериментальной лексикографии ИРЯ РАН «Теоретические основы описания русской идиоматики», осуществляемого под руководством А. Н. Баранова и Д. О. Добровольского. Участниками проекта уже выпущены книги «Словарь-тезаурус современной русской идиоматики» (2007) и «Фразеологический объяснительный словарь русского языка» (2009), а основные положения используемого теоретического подхода отражены в монографии «Аспекты теории фразеологии» (2008). Поскольку речь далее будет идти и о жестах, мы в определенной степени будем также опираться на ряд работ в области кинесики, прежде всего на «Словарь языка русских жестов», выпущенный в 2001г. коллективом исследователей под руководством Г. Е. Крейдлина и монографию Г. Е. Крейдлина «Невербальная семиотика: Язык тела и естественный язык» (1-е изд. — 2002г.).

1. Жестовые идиомы

1.1. Свободное сочетание vs. номинация жеста vs. идиома, производная от жеста

Для начала определим, какие именно идиомы находятся в сфере нашего рассмотрения.

Жестовой идиомой в узком смысле мы будем называть идиому, в основе которой лежит жест, т. е. знак языка тела, несущий определенный смысл. Поясним это на конкретном примере. Выражение *махнуть рукой* может быть свободным сочетанием, номинацией (т. е. устойчивым наименованием) нескольких жестов, или производной от жеста идиомой, не обязательно сопровождающейся жестом.

В данном примере *махнуть рукой* выступает как **свободное сочетание**:

(1) Если неожиданно и резко *махнуть рукой* перед лицом человека, то, пребывая в определенном состоянии духа, он не успеет среагировать и останется неподвижным. В. В. Шлахтер. Человек — оружие. Курс профессиональной психофизической подготовки бойца.[1]

Выражение *махнуть рукой* не несет здесь никакого дополнительного смысла.

В примерах (2)–(5) это же выражение является **номинацией** четырех различных **жестов**:

1. указательный жест «махнуть рукой 1. <в сторону чего-л.>»:

(2) А где Боговизна, знаешь? — Боговизна? Да там, — *махнул рукой* Костя в сторону леса. Огрызков нахмурился, подумав: не туда ли этот подросток показывает, откуда они шли ночь? Но что же тогда получается? Скверно тогда получается… [В. Быков. Болото]

[1] Поиск примеров производился в Национальном корпусе русского языка, Базе данных по русской фразеологии (ИРЯ им. В. В. Виноградова РАН, Отдел экспериментальной лексикографии), а также в русскоязычном сегменте Интернета.

2. жест «махнуть рукой 2.», призывающий следовать за кем-л.:

(3) Он встал, оказался коротконог, на полголовы ниже Тани, но двигался быстро и резко, как теннисный мяч. И *махнул рукой*, чтобы шли за ним… [Л. Улицкая. Путешествие в седьмую сторону света]

3. жест «махнуть рукой 3.», употребляющийся при прощании:

(4) Что касается хозяина, почтенного Георгия Романовича, то он остался дома для беседы с управляющим <…> и в данный момент стоял на крыльце веранды, грузный и краснолицый, собираясь *махнуть нам рукой* на прощанье. [Б. Хазанов. Далекое зрелище лесов]

4. жест «махнуть рукой 4.», выражающий безразличие:

(5) <…> Анна Федоровна спросила его в совершеннейшем изумлении: — Как ты можешь с ней так разговаривать? Он небрежно *махнул рукой*: — Опыт. У меня в клинике восемьдесят процентов пациентов старше восьмидесяти, все богатые и капризные. Пять лет учился с ними ладить. [Л. Улицкая. Пиковая дама]

Отметим, что с нашей точки зрения номинация жеста сама по себе является идиомой. Действительно, это словосочетание обладает высокой степенью устойчивости и идиоматичности. Его значение является результатом переинтерпретации исходного свободного сочетания, и это значение невозможно восстановить без обращения к языку жестов, т. е. без обращения к значению соответствующего жеста. Таким образом, принципиальное отличие свободного сочетания от идиомы-номинации жеста заключается в семантической переинтерпретации выражения.

Во всех вышеперечисленных примерах употребление номинации жеста сопровождается выполнением самого жеста, что хорошо видно из контекстов. Однако то же языковое выражение *махнуть рукой* со значением безразличия может употребляться в тексте и не сопровождаясь соответствующим жестом. Ср. следующие примеры на употребление **идиомы, производной от жеста**:

(6) Поначалу я сердился, возражал, сопротивлялся, если искажали мою фамилию, но когда получил красноармейскую книжку перед отправкой на сталинградскую мясорубку, *махнул рукой*: не всё ли равно, убьют меня Слюсаревым или Слесаревым — какое это будет иметь значение перед историей? [В. Астафьев. Обертон]

(7) Видя, что прокурор настолько ошалел и зазнался от высочайшей благодарности, что никого знать не хочет, все в конце концов *махнули* на него *рукой*, выпустили его из виду, и потом никто не мог вспомнить в точности, когда и как он ушел. [В. Войнович. Жизнь и необычайные приключения солдата Ивана Чонкина]

(8) <…> он стал еще невнимательнее, еще небрежнее, пропускал важные свидания, не являлся туда, куда было нужно, и вообще, казалось, *махнул рукой* на все. [Г. Газданов. Пробуждение]

Так, в примере (6) Слесареву стало безразлично, что искажают его фамилию, однако он мог вовсе не совершать соответствующий жест. В примере (7) на прокурора махнули рукой, т. е. он или его поступки стали всем безразличны, однако это не означает, что все выполнили в его адрес жест «махнуть рукой». И, наконец, в (8) лирическому герою все стало безразлично, однако он не совершал жест «махнуть рукой» в адрес всего. На это с определенностью указывает вводное слово *казалось*.

Таким образом, мы видим два принципиальных отличия идиомы, производной от жеста от идиомы-номинации жеста: во-первых, она может употребляться самостоятельно, вне контекста употребления соответствующего жеста, и во-вторых, происходят определенные сдвиги в значении и сочетаемости идиомы. Так, если субъектом и адресатом идиомы-номинации жеста должны быть конкретные объекты (субъектом — человек), то в примерах (7) и (8) на употребление производной от жеста идиомы мы наблюдаем генерализацию субъекта и объекта, ср. *все в конце концов махнули на него рукой, махнул рукой на все*.

Заметим, что в терминологии Г. Е. Крейдлина именно такие идиомы называются жестовыми фразеологизмами или жестовыми фраземами. При такой перспективе при изучении жестовых идиом в центре внимания оказываются именно механизмы семантической деривации, позволяющие перейти от значения жеста к значению производной от жеста идиомы, ср. сравнительное описание жеста *хлопнуть дверью* и идиомы *хлопнуть дверью* в статье А. Козеренко и Г. Крейдлина «Русские жесты и русские фразеологизмы II (тело как объект природы и тело как объект культуры)» (1999 г.).

Часто бывает, что жест, лежащий в основе идиомы, давно вышел из употребления, ср. такие идиомы, как *снимать шляпу* (перед кем-л./чем-л.), *бросить перчатку* (кому-л.), *преклонить колена* (перед кем-л.), *падать/простираться ниц* (перед кем-л.) и т. п. В этом случае номинация жеста реально не употребляется или же употребляется крайне редко (например, при описании фильмов и театральных постановок, в которых этот жест еще можно встретить), в то время как идиома, производная от жеста, остается в широком употреблении.

1.2. Жестовые идиомы в широком смысле

Помимо идиом, в основе которых лежит жест, нас также будут интересовать идиомы, так или иначе семантически связанные с каким-либо значимым, т. е. несущим определенный смысл телесным положением или движением (не обязательно жестом). Такие идиомы можно назвать **жестовыми в широком смысле слова.**

Поясним это на примере идиомы *как в воду опущенный*. Ее значение можно описать квазисинонимичными выражениями *огорченный, расстроенный, в унынии*. Толкование этой идиомы состоит из части, описывающей

психологическое состояние субъекта и части, указывающей на внешние, телесные и/или поведенческие, проявления, по которым заметно это состояние[2]:

КАК В ВОДУ ОПУЩЕННЫЙ (после какой-л. неприятности) находясь в негативном психологическом состоянии, характеризующемся подавленностью и общим восприятием окружающего как плохого, *что заметно по внешним проявлениям, часто сопровождающим это состояние — опущенным плечам, голове, сгорбленности и т. п., замедленной реакции на вопросы, нежеланию общаться и т. п.*

Внутренняя форма данной идиомы не является полностью прозрачной, однако те следствия из внутренней формы, которые, как нам кажется, имеются, мы учитываем в толковании (они выделены курсивом).

В основе этой идиомы, как мы видим, находится некоторое характерное положение тела, косвенным способом переданное во внутренней форме идиомы и несущее определенный смысл — ему соответствует ощущение подавленности. Это знание о соответствии определенного положения тела определенному психологическому состоянию отражается в толковании идиомы.

2. ПЕЧАЛЬ, СОЖАЛЕНИЕ, УНЫНИЕ: идиомы — номинации жестов

Далее мы хотели бы рассмотреть несколько жестовых идиом семантического поля ПЕЧАЛЬ, СОЖАЛЕНИЕ, УНЫНИЕ, являющихся номинациями жестов, и на их примере продемонстрировать, как могут соотноситься множества жестов и их номинаций, и как это фиксируется в словарном описании идиом.

2.1. Рассмотрим идиому *качать головой*. По контекстам употребления у этой идиомы выделяется три значения, соответствующие разным психологическим и ментальным состояниям: 1) ‘горе, печаль’, 2) ‘неодобрение’ и 3) ‘жалость, сочувствие’:

КАЧАТЬ ГОЛОВОЙ 1. *нейтр.* из-за того, что, случилось нечто плохое (неудача, неприятности и т. п.), находиться в негативном психологическом состоянии горя, *что передается описанием характерного непроизвольного жеста, сопровождающего такое состояние — человек опускает голову и несильно качает ей из стороны в сторону* ❖ печалиться, горевать 🗐 Я вспомнил, как она взяла меня за руку, *качая головой* и стараясь не плакать. Почему я ничего не сказал ей? Она очень ждала хоть одного слова. [В. Каверин. Два капитана]

КАЧАТЬ ГОЛОВОЙ 2. *нейтр.* отрицательно оценивать слова, действия собеседника или ситуацию, *что передается описанием характерного непроизвольного жеста, сопровождающего такую оценку — человек несильно качает головой из стороны в сторону* ❖ не одобрять 🗐 Поправив пенсне, Галина Николаевна внимательно посмотрела на дочь и с неодобрением *покачала головой*: — Просто не понимаю, что с тобой делается… Ты меня просто огорчаешь! И в кого только ты могла пойти? [Ю. Слепухин Перекресток]

---

[2] Здесь и далее толкования приводятся в формате, принятом во Фразеологическом объяснительном словаре русского языка (2009).

Если обратить внимание на часть толкования, которая описывает телесную оставляющую, становится видно, что в основе этих двух значений лежат разные жесты. В первом случае (‘горе, печаль’) голова опускается вниз, во втором (‘неодобрение’) — не опускается, а взгляд говорящего направлен на адресата неодобрения. Номинации этих двух жестов совпадают, следовательно, при описании соответствующей идиомы мы учитываем два значения идиомы, восходящие к разным жестам, что и отражается в приводимых толкованиях.

Третье значение идиомы восходит к тому же жесту, что и второе:

КАЧАТЬ ГОЛОВОЙ 3. *нейтр.* с сочувствием отнестись к кому-л., оказавшемуся в неприятной ситуации, *что передается описанием характерного непроизвольного жеста, сопровождающего такое выражение сочувствия — человек несильно качает головой из стороны в сторону* ❖ сочувствовать 🗐 По дороге Юрий разговорился с шофёром, глядя ему в спину; рассказал о своей неприятности. Тот выслушал его с сочувствием, *качая головой*, пожелал ему успеха в деле с билетами и предложил: «Если негде будет ночевать — езжай ко мне». [И. Грекова. Знакомые люди]

Как мы видим, в первом жесте голова опускается вниз, и жест выражает ‘горе, печаль’, во втором жесте голова не опускается вниз и он употребляется в двух значениях: ‘неодобрение’ и ‘жалость, сочувствие’. Поскольку номинации этих двух жестов совпадают, в языке этим трем жестовым значениям соответствуют три значения одной идиомы *качать головой*.

2.2. Рассмотрим идиомы того же семантического поля *поникнуть головой, повесить голову, понурить голову***.**

Это три номинации одного и того же жеста, имеющего 3 значения: ‘печаль’, ‘смущение’ и ‘вина’. Одно из значений этого жеста — ‘печаль’ — есть у всех трех идиом, однако далее наблюдается несимметричная картина: контексты употребления этих идиом показывают, что в значении ‘смущения’ употребляется только идиома *поникнуть головой*, в значении ‘вины’ — только *понурить голову*, а *повесить голову* употребляется только в одном общем для трех идиом значении ‘печали’.

Рассмотрим подробнее, как это происходит, и почему разные номинации оказываются закреплены за разными значениями одного жеста.

Идиома *поникнуть головой* получает следующее толкование:

ПОНИКНУТЬ ГОЛОВОЙ 1. из-за того, что, случилось нечто плохое (неудача, неприятности и т. п.), находиться в негативном психологическом состоянии, характеризующемся подавленностью, *что передается описанием характерного непроизвольного жеста, сопровождающего такое состояние — голова человека опускается вниз* ❖ опечалиться, приуныть 🗐 — Сейчас, как видите, я один. Молодой человек *поник головой* и сидел совершенно убитый — вцепившись тонкими бледными пальцами в худые колени. Виктору стало жалко его. [С. Бабаян. Ротмистр Неженцев]

ПОНИКНУТЬ ГОЛОВОЙ 2. из-за того, что стало известно что-то плохое о субъекте и он понимает, что это осуждается, прийти в негативное психологическое состояние, характеризующееся желанием спрятаться от окружающих, *что передается описанием характерного непроизвольного жеста, сопровождающего такое состояние — голова человека опускается вниз, как бы для того, чтобы*

не было видно проявления эмоций на лице ❖ смутиться, устыдиться 🗐 — Ладно, не переживай, ошибку твою я исправил. Кстати, ты женат? Поняв, что разговор перешел на другую тему, Ягафаров облегченно вздохнул. Вместо ответа смущенно *поник головой*. — До сих пор выбираешь? Заруби себе на носу, Ягафаров. <…> — У врача, особенно у хирурга, должна быть крепкая семья. Только тогда он сможет чувствовать себя спокойно и уверенно. [Д. Буляков. Жизнь дается однажды]

Как видно из толкования идиомы, опускание головы в случае смущения интерпретируется как желание спрятаться или спрятать лицо, скрыть проявление своих эмоций, выдающих понимание человеком существования угрозы его общественному лицу.

Идиома *повесить голову* имеет следующее толкование:

ПОВЕСИТЬ ГОЛОВУ из-за того, что, случилось нечто плохое (неудача, неприятности и т. п.), находиться в негативном психологическом состоянии, характеризующемся подавленностью, *что передается описанием характерного непроизвольного жеста, сопровождающего такое состояние — человек опускает голову вниз, как бы перестав ее контролировать* ❖ опечалиться, приуныть 🗐 — Жена вчера ушла, собрала вещи и ушла. — Студент замолчал, *повесив голову*. Потом, встряхнувшись, налил бокал до краев и залпом выпил. [Б. Акунин. Азазель]

Для номинации жеста в языке в данном случае был выбран глагол *повесить*, обозначающий действие, которое, как мы считаем, соответствует идее потеря контроля, отраженной в толковании. Однако эта идея не может сочетаться с желанием спрятать проявление своих эмоций, когда субъект как раз контролирует свои действия, что и объясняет отсутствие у данной идиомы второго значения — ‘смущения’. Иными словами, соответствующий жест в значении смущения не может быть назван выражением *повесить голову*.

Идиома *понурить голову* имеет следующее толкование:

ПОНУРИТЬ ГОЛОВУ *книжн.* **1.** из-за того, что, случилось нечто плохое (неудача, неприятности и т. п.), находиться в негативном психологическом состоянии, характеризующемся подавленностью, *что передается описанием характерного непроизвольного жеста, сопровождающего такое состояние — человек опускает голову вниз* ❖ опечалиться, приуныть 🗐 Новиков трясет прибор, дует на него, щелкает по лампам. Ничего не помогает. Проходит час, другой, последние попытки кончились, все сидят, *понурив голову*, пришибленные, не в силах уже ничего понять. [Д. Гранин. Искатели]

ПОНУРИТЬ ГОЛОВУ *книжн.* **2.** сделав что-л. неправильное и испытывая вину, прийти в негативное психологическое состояние, характеризующееся подавленностью, *что передается описанием характерного непроизвольного жеста, сопровождающего такое состояние — человек опускает голову вниз, как бы почувствовав ее тяжесть* ❖ испытывать вину 🗐 Он встал на колени, пустил слезу и ударил себя кулаком в грудь, виновато *понурив голову*. [А. Иванов. Лилипут — сын Великана]

Чувство вины метафорически осмысляется как тяжесть (ср. также выражения типа *он сгорбился под тяжестью вины*), а опускание головы в данном случае интерпретируется как следствие этого ощущения тяжести.

Таким образом, мы наблюдаем следующую ситуацию: жест, который можно кратко описать выражением *опустить голову вниз,* имеет три значения: в первом значении он имеет три устойчивых номинации, во втором и в третьем значении — по одной. Значения жеста следующим образом соответствуют значениям трех идиом — номинаций жестов:

1. ‘печаль’
ПОНИКНУТЬ ГОЛОВОЙ 1.
ПОВЕСИТЬ ГОЛОВУ
ПОНУРИТЬ ГОЛОВУ 1.
2. ‘смущение’
ПОНИКНУТЬ ГОЛОВОЙ 2.
3. ‘вина’
ПОНУРИТЬ ГОЛОВУ 2.

Как мы видим на примере разобранных идиом, номинация жеста является своего рода языковой интерпретацией жеста, а выбор номинации коррелирует со значением идиомы. В случае, когда опускание головы интерпретируется как желание спрятать проявления своих эмоций, идиома может употребляться в значении ‘смущения’. В случае, когда опускание головы интерпретируется как следствие ощущения тяжести, идиома может употребляться в значении ‘вины’. И, наконец, в случае, когда из-за выбора глагола (*повесить голову*) в языковом выражении появляется идея потери контроля, употребление идиомы в значении ‘смущения’ или ‘вины’ становится невозможным, и идиома употребляется только в одном значении.

Итак, в настоящей статье мы ввели понятия «жестовой идиомы в узком понимании» и «жестовой идиомы в широком понимании». На примере выражения *махнуть рукой* как свободного сочетания и как идиомы (в четырех значениях) было продемонстрировано, что жестовая идиома в узком понимании может быть номинацией жеста или производной от жеста идиомой, которая употребляется самостоятельно, не сопровождаясь никаким жестом.

Были рассмотрены идиомы-номинации жестов, принадлежащие семантическому полю ПЕЧАЛЬ, СОЖАЛЕНИЕ, УНЫНИЕ. Их анализ показывает, что разным жестам может соответствовать одна идиома, а одному жесту — несколько идиом, причем в последнем случае выбор номинации жеста неслучаен и коррелирует с семантическими компонентами значения идиомы.

Мы видим несколько возможных направлений исследования затронутой здесь темы. Для идиом — номинаций жестов хотелось бы составить по возможности полную типологию соответствий жестов и их номинаций, особое внимание при этом уделяя мотивации выбора номинации. Для идиом, производных от жеста, в фокус внимания уместно поместить разные виды семантической деривации от значения жеста к значению производной идиомы.

В обоих случаях на передний план выступает изучение семантики идиом и связь значения идиомы с ее внутренней формой (в данном случае, жестом) и с другими — метафорическими, символическими и т. п. — компонентами значения.

References

1. *Baranov A. N., Dobrovol'skii D.O., Kiseleva K. L., Kozerenko A. D.* 2007. Modern Russian Idiomatic Expressions Dictionary [Slovar'-tezaurus Sovremennoi Russkoi Idiomatiki].
2. *Baranov A. N., Dobrovol'skii D. O.* 2008. Aspects of the Theory of Phraseology [Aspekty Teorii Freazeologii].
3. *Russian* Phraseology Explicative Dictionary [Frazeologicheskii Ob"iasnitel'nyi Slovar' Russkogo Iazyka]. 2009.
4. *Grigor'eva S. A., Grigor'ev N. V., Kreidlin G. E.* 2001. Russian Gestures Dictionary [Slovar' Iazyka Russkikh Zhestov].
5. *Kozerenko A., Kreidlin G.* 1999. Russian Gestures and Russian Phraseologisms II (The Body as Natural Object and the Body as Cultural Object) [Russkie Zhesty I Russkie Frazeologizmy II (Telo kak Ob"ekt Prirody I Telo kak Ob"ekt Kul'tury)] : 269–277
6. *Kreidlin G. E.* 2002. Non-verbal Semiotics: Body Language and Natural Language [Neverbal'naia Semiotika: Iazyk Tela I Estestvennyi Iazyk].

ЛИНГВИСТИЧЕСКАЯ МОТИВИРОВКА ДЛЯ СТАТИСТИЧЕСКИХ МОДЕЛЕЙ ПЕРЕВОДА

Е. Б. Козеренко (kozerenko@mail.ru)

ИПИ РАН Москва, Россия

В данной статье рассматриваются проблемы выравнивания параллельных текстов для повышения достоверности перевода. Представлены статистическая и лингвистически-мотивированная модели выравнивания параллельных текстов и перевода методом трансфера. Предлагаемые решения основаны на гибридной грамматике, которая включает лингвистические правила и вероятностные характеристики структур языка. Поскольку сходные значения могут быть представлены различными способами, особенно важны описания синонимии языковых структур. Цель наших исследований — установление соответствий между структурами различных языков на уровне смысла.

Ключевые слова: перевод, параллельные тексты, выравнивание, модели выравнивания.

LINGUISTIC MOTIVATION FOR STATISTICAL TRANSLATION MODELS

E. B. Kozerenko (kozerenko@mail.ru)

Institute for Informatics Problems of the Russian Academy of Sciences, Moscow, Russian Federation

The paper deals with the problems of parallel texts alignment for enhancing the accuracy and adequacy of translation. Statistical and heuristic models of alignment and transfer are given. The solutions are proposed on the basis of a hybrid grammar, which includes linguistic rules and probabilities of language structures. The goal of the current development is the establishment of matches at the level of meaning, i. e. semantic matches. The meaning can be “packed” in different language structures, so the establishment of cross-language matches and inter-structural synonymy is of prime importance.

Key words: translation, parallel texts, alignment, alignment models.

1. Introduction

The paper is focused on discovering the ways of the two research paradigms combination, namely, introducing statistical methods into the rule-based systems of machine translation and employment of the methods and presentations capturing human language intuition in statistical translation models with the view of enhancing the existing language processing technologies.

In statistical machine translation (SMT) the task of translating from one natural language into another is treated as a machine learning problem. This means that via training on a very large number of hand-made translation samples the SMT algorithms master the rules of translation automatically. The first SMT developments were presented in [1,2].

The application of statistical models has considerably advanced the area of machine translation since the last decade of the previous century, however now new ideas and methods appear aimed at creating systems that efficiently combine symbolic and statistical approaches comprising different models. Both the paradigms move towards each other: more and more linguistics is being introduced into stochastic models of machine translation, and the rule-based systems include statistics into their linguistic rule systems. The procedures of analysis and translation are enhanced by the statistical data, which are taken into consideration by the "translation engine" for disambiguation of language structures. The stochastic approach to natural language processing originates from the projects in speech and characters recognition and spellcheckers. The main method for solving numerous problems, including the part of speech establishment and tagging, is the Bayesian approach. The architecture of stochastic systems is based on the dynamic programming algorithm.

Machine learning is rooted in the stochastic research paradigm. The training algorithms can be of the two types: supervised and unsupervised. An unsupervised algorithm should infer a model capable for generalization of the new data, and this inference should be based on the data alone. A supervised algorithm is trained on a set of correct responses to the data from the training set, so that the inferred model would provide more accurate decisions. The object of machine learning is the automatic inference of the model for some subject area basing on the data from this area. Thus a system learning, for example, syntactic rules should be supplied with a basic set of phrase structure rules. The widely used methods lately have been the N-grams which capture many intricacies of syntactic and semantic structures [3, 4], N-grams of variable length in particular [5], introduction of semantic information into N-grams. In [6] a detailed description is given of the approach to creating a statistical machine translation based on N-grams of bilingual units called "tuples" and the four special attribute functions.

The statistical models are built on the data obtained from the parallel corpora in different languages. Usually the texts are compared within language pairs. The text in the language from which the translation should be done is called the source text, and the text which is its translation is called the target text. Correspondently the languages are also called the source language and the target language (i. e. the language of translation).

The main method of extracting the data about the matches between the source and target languages and texts is the alignment of parallel texts. The result of this procedure is also called alignment and it is designated by A. The probability characteristics of alignments are employed in the algorithms of statistical machine translation. Hence, the alignment and the probability distribution are the key notions in these models description.

The following notations are employed in this paper: the symbol P denotes the probability distributions in the most general sense, and the symbol p denotes the probability distribution based on some particular model. The main attention in this paper is given to the description of various methods employed for parallel texts alignment, as the results of the alignment procedure determine the accuracy and adequacy of translation. We focus on the linguistic filters that are being introduced in the form of data structures and rules into the statistical translation models. The models under consideration are illustrated basing on the bilingual model for the Russian and English language pair. However, the similar methods are applicable for the alignments and translations of the Russian texts into the French and German languages, as well as other European languages.

2. Methods of parallel texts alignment

The statistical approaches to parallel texts alignment are aimed at establishing the most probable alignment A for the two given parallel texts S and T:

$$\arg\max_A P(A \mid S,T) = \arg\max_A P(A,S,T) \tag{1}$$

For estimation of the probability values indicated in this expression the most frequently used methods present the parallel texts in the form of aligned sentence sequences $(B_1,...,B_K)$. The probability of each sequence is independent from the probabilities of other sequences, and it depends on the sentences in the given sequence only [7]. Then

$$P(A,S,T) \approx \prod_{k=1}^{K} P(B_k) \tag{2}$$

This method takes into account the length of sentences in the source language and in the target language measured in symbols. The longer sentence in one language will correspond to the longer sentence in the other language. This approach gives stable results for similar languages and literal translation. The more finely tuned mechanisms of matching are provided by the methods of lexical alignment. Thus in [8] the method of alignment by means of creating the model for consecutive word-by-word translation is presented. The best alignment result will be the one which maximizes the probability of a corpus generation with the given translation model. For the alignment of the two texts S and T they should be split into the sequences of sentence chains. A chain contains zero or more sentences in each of the two languages, and the sequence of chains covers the whole corpus

$$B_k = (S_{a_k},...,S_{b_k}; t_{c_k},...,t_{d_k}) \tag{3}$$

Then the most probable alignment $A = B_1, ..., B_{m_A}$ of the given corpus is determined by the following expression, and the chains of sentences do not depend on each other:

$$\arg\max_A P(S,T,A) = \arg\max_A P(L)\prod_{k=1}^{m_A} P(B_k), \quad (4)$$

where *P(L)* denotes the probability of the *L* chains being generated. The translation model employed in this approach is extremely simplified and does not take into account the factor of the word order in a sentence and the possibility of the fact that a word in the source text can correspond to more than one word in the text of translation. In this model the word chains are used, and they are limited to the 1:1, 0:1 и 1:0 matches. The essence of the model consists in the idea that if one word is usually translated by the word of another language, then the probability of the word chains matches 1:1 will be very high, and much higher than the product of probabilities of the 1:0 and 0:1 word chains matches where the given word occurs. And the program chooses the most probable alignment variant.

The translation model based on the word-by-word alignment (we employ this model for the Russian and English parallel texts) will be as follows:

$$P(r \mid e) = \frac{1}{Z}\sum_{a_1=0}^{l}\cdots\sum_{a_m=0}^{l}\prod_{j=1}^{m} P(r_j \mid e_{a_j}), \quad (5)$$

where *e* is a sentence in English; *l* is the length of *e* expressed in words; *r* is a sentence in Russian; *m* is the length of *r*; r_j is the j-th word in *r*; a_j is the position in *e*, with which the r_j is aligned; $P(w_r \mid w_e)$ is the probability of translation, i. e. the probability of the w_r appearing in the Russian sentence if the corresponding w_e occurs in the English sentence, and *Z* is the normalization constant.

However, the above stated approach based on the word-by-word comparison and in no way accounting for the links between words and phrases does not give optimal results for the alignment of the Russian language and the English language texts, for there are certain structural differencies between these languages, and in translation there can be considerable transformations. If the languages under consideration are structurally different, the methods are used oriented at the introduction of grammar knowledge, for example, the alignment methods based on the words that belong to particular parts of speech [9] are employed. In this case the auxiliary words are not taken into account. For the employment of these methods the part of speech tagging of the parallel texts should be performed. The most general definition of the word-based alignment is given in [10]. Suppose the two word chains are given, one in the source text (for example, in Russian — r) $r_1^J = r_1, ..., r_j, ..., r_J$, and the other one is in the target language (English — e) $e_1^I = e_1, ..., e_i, ..., e_I$, and for these chains it is necessary to establish the alignment. The alignment between the two chains of words is a subset of a Cartesian product of the positions of words, i. e. the alignment A is defined as follows:

$$A \subseteq \{(j,i) : j = 1, ..., J; i = 1, ..., I\}. \quad (6)$$

In machine translation based on statistical methods an attempt is made to construct a model of the translation probability $P(r_1^J \mid e_1^I)$, which describes the correlation between some chain r_1^J in the source language and the chain e_1^I in the target language. In statistical texts alignment model $P(r_1^J, a_1^J \mid e_1^I)$ a "hidden" alignment a_1^J is introduced which describes the mapping from the source position j into the target position a_j. The correlation between the translation model and the alignment model is given in the following way:

$$P(r_1^J \mid e_1^I) = \sum_{a_1^J} P(r_1^J, a_1^J \mid e_1^I). \tag{7}$$

The alignment a_1^J can contain the alignments $a_j = 0$ with the empty word e_0 for the words of the source language which had not been aligned with any word in the source language. On the whole the statistical model depends on the set of unknown parameters θ which are extracted from the training data set in the course of learning. The following presentation is used to express the dependence of the model on the set of parameters:

$$P(r_1^J, a_1^J \mid e_1^I) = p_\theta(r_1^J, a_1^J \mid e_1^I) \tag{8}$$

The technique of statistical modeling consists in the development of specific statistical models which would capture the most relevant features of the subject area under consideration. Thus a statistical model of alignment should adequately describe the correlation between the chain in the source language and the chain in the target language.

For detection of the unknown parameters θ a training corpus of parallel texts is given containing S sentence pairs $\{(r_s, e_s) : s = 1, ..., S\}$. For each pair (r_s, e_s) the alignment variable is designated by $a = a_1^J$. The unknown parameters are established by means of maximization of the parallel texts similarity in the corpus:

$$\hat{\theta} = \arg\max_{\theta} \prod_{s=1}^{S} \sum_{\mathrm{a}} p_\theta(\mathrm{r}_s, \mathrm{a} \mid \mathrm{e}_s). \tag{9}$$

As a rule the maximization for such models is performed on the basis of the expectation maximization algorithm [11] or the similar ones. Such algorithm is useful for the solution of the parameters estimation problem, but it is not indispensable for the statistical approach.

Hence despite the fact that there exist a large number of alignments for a given pair of sentences, it is always possible to find the best alignment:

$$\hat{a}_1^J = \arg\max_{a_1^J} p_{\hat{\theta}}(r_1^J, a_1^J \mid e_1^I). \tag{10}$$

The alignment $\hat{a}_1^J$ is also called the Viterbi alignment for the pair of sentences (r_1^J, e_1^I). The estimation of the Viterbi alignment quality is performed by means of comparison with some reference alignment carried out manually. The parameters of statistical alignment models are optimized with the consideration of the maximal likelihood criterion which does not always reflect the quality of alignment.

The most frequently used statistical model which is used for parallel texts alignment is the hidden Markov model [13]. The alignment model $P(r_1^J, a_1^J \mid e_1^I)$ can be structured without the loss of generality in the following way:

$$\begin{aligned} P(r_1^J, a_1^J \mid e_1^I) &= P(J \mid e_1^I) \cdot \prod_{j=1}^{J} P(r_j, a_j \mid r_1^{j-1}, a_1^{j-1}, e_1^I) = \\ &= P(J \mid e_1^I) \cdot \prod_{j=1}^{J} P(a_j \mid r_1^{j-1}, a_1^{j-1}, e_1^I) \cdot P(r_j \mid r_1^{j-1}, a_1^j, e_1^I) \end{aligned} \tag{11}$$

When using this alignment the three probabilities are obtained: a length probability $P(J \mid e_1^I)$, an alignment probability $P(a_j \mid r_1^{j-1}, a_1^{j-1}, e_1^I)$ and a lexicon probability $P(r_j \mid r_1^{j-1}, a_1^j, e_1^I)$. In the hidden Markov alignment model the first order dependence for the alignments a_j is assumed, and it is assumed that the lexicon probability depends only on the word at position a_j:

$$P(a_j \mid r_1^{j-1}, a_1^{j-1}, e_1^I) = p(a_j \mid a_{j-1}, I), \tag{12}$$

$$P(r_j \mid r_1^{j-1}, a_1^j, e_1^I) = p(r_j \mid e_{a_j}). \tag{13}$$

If a simple length model is assumed $P(J \mid e_1^I) = p(J \mid I)$, then for $p(r_1^J \mid e_1^I)$ the following decomposition based on the hidden Markov model is obtained:

$$p(r_1^J \mid e_1^I) = p(J \mid I) \cdot \sum_{a_1^J} \prod_{j=1}^{J} [p(a_{j-1}, I) \cdot p(r_j \mid e_{a_j})] \tag{14}$$

with the alignment probability $p(i \mid i', I)$ and the translation probability $p(r \mid e)$. In order to make the alignment parameters independent from the absolute values of word positions, it is assumed that the alignment probabilities $p(i \mid i', I)$ depend only on the jump width $(i - i')$. Using a set of non-negative parameters $\{c(i - i')\}$, it is possible to present the alignment probabilities in the following way:

$$p(i \mid i', I) = \frac{c(i - i')}{\sum_{i''=1}^{I} c(i'' - i')}. \tag{15}$$

This form ensures that the alignment probabilities satisfy the normalization constraint for each conditioning word position $i', i' = 1, \ldots, I$. This model is also called the homogeneous hidden Markov model [12]. The original formulation of the hidden Markov alignment model did not comprise the empty word generating source words which have no directly aligned word in the target text. In [13] the empty word is introduced and the hidden Markov model network is extended by means of I empty words e_{1+1}^{2I}.

The existing methods basically employ either sentence alignment or word alignment some experiments are made with phrase alignment and recently a mixed sentence-word approach has been developed to explore the paraphrases in the aligned parallel corpora. These attempts to consider linguistic information mark a step forward to acknowledging the intricate character of natural language if compared with other types of data. The mixed approach employs both sentence and word alignments [14, 15]. However, all these methods deal with

the structural elements without considering the semantic aspects of the aligned language units.

The phrase-based translation model, or the alignment template model [16] and other similar approaches have greatly advanced [17] the development of machine translation technology due to the extension of the basic translation units from words to phrases, i. e. the substrings of arbitrary size. However, the phrases of this statistical machine translation model are not the phrases in the meaning of any existing syntax theory or grammar formalism, thus, for example, a phrase can be like "alignments the", etc. A real challenge is the cross-level (e. g. morphology-to-syntax) matching of language structures in parallel texts [18]. New research and development results demonstrate the growing awareness of the demand for enhancing linguistic motivation in statistical translation models and machine learning techniques [21,22].

3. Intertext development: establishment of semantic matches

The above stated methods are being employed for design and development of a linguistic knowledge base Intertext. It is a linguistic resource with semantic grouping of phrase structure patterns provided with the links to synonymous structures at all language levels for the languages included into the linguistic base.

Our focus on configurations provides high portability to the language processing software designed under these principles: we can operate with a lexicon which has only standard linguistic information including morphological characteristics, part of speech information and the indication of transitivity for verbs.

The Intertext linguistic knowledge base comprises the following components:

- parallel texts database: the texts are segmented into the functionally relevant structures that are semantically aligned;
- a bilingual Treebank (under development at present);
- structural parse editor (under development at present) which displays the parse and transfer schemes for indicated text segments;
- the inventory of structural configurations arranged on Cognitive Semantic principle.

4. Establishment of cross-language matches and inter-structural synonymy

Translation activity involves the search for equivalence between structures of different languages. However, to establish whether the structures and units are equal or not, we need some general equivalent against which the language phenomena would be matched. Our approach based on the principle "from the meaning to the form" focusing on Functional Syntax would yield the necessary basis for equivalence search.

4.1. Types of matches

The following types of structural semantic matches have been observed:

word → word, phrase structure → phrase structure, word → phrase structure, morpheme → word, morpheme → phrase structure.

Syntactically languages are most different in the basic word order of verbs, subjects, and objects in declarative clauses. English is an SVO language, while Russian has a comparatively flexible word order. The syntactic distinction is connected with a semantic distinction in the way languages map underlying cognitive structures onto language patterns, which should be envisaged in MT implementations [20].

The basis of Cognitive Transfer Grammar (CTG) is composed of the proto-typical structures of the languages (in the initial model Russian and English) being investigated, their most probable positions in a sentence, statistical data about the distributive characteristics of structures (the information about the contextual conditions of the use of the investigated objects, i. e. the information about the structural contexts), the schemes of the complete parse of sentences.

The creation and development of the CTG assumes: — the semantic approach to the analysis of language meaning and language form (forms); — the construction of formal grammar presentations taking into account the structures of components and mechanisms of linearization, and also the relations of dependence between the units of a syntactic tree (the approach, which has the features of similarity to HPSG: the inheritance of the features via the head elements of phrase structures); — the inclusion of the probability characteristics of language objects; — the creation of Cognitive Transfer Spaces (CTS), represented in the form of expert linguistic rules, which can be extended by means of the establishment of synonymous language structures of parallel texts in different languages. The notion of Cognitive Transfer Spaces is the elaboration of the Functional Transfer Fields idea (see Section 5) for the multivariant translations of language structures.

In contrast to the approaches on the basis of "translation memory" that provide the increase of a machine translation system language competence by accumulating the previously translated text fragments and mainly based on regular expressions, Cognitive Transfer Grammar is intended for the realization of the mechanism of structural memory, which simulates language competence of an adult learner ("Adult Learning Memory"). Thus, structural memory comprises the following components:

1) The initial basic collection of grammar rules represented in the formalized form (CTG);
2) The mechanisms of expansion and refinement of the system of rules, implemented by means of the methods of machine learning on parallel texts.

Our studies are based on the concepts of the functional approach, which we have used for the multilingual situation. With the development of the linguistic processor, which ensures English — Russian and Russian — English transfer, we introduced the concept of functional transfer fields (FTF) [19] that served the basis for the segmentation of language structures for the solution of machine translation problems. The basic

idea of FTF consists in the adoption of the hypothesis about the fact that at the basis of grammatical structures there lie the cognitive structures (mental frames); a functional transfer field reflects the interaction of elements from different language levels.

The basic design unit of the spaces of cognitive transfer is a *transfeme*.

Definition. *Transfeme* is a unit of cognitive transfer the, i. e. a semantic element embodied in a translatable semantically relevant language segment taken in the unity of its categorial and functional characteristics, that establishes the semantic correspondence between the language structures, which belong to different language levels and systems. The types of transfemes are determined by the rank of transfemes.

We distinguish the following ranks of transfemes:

- rank 1: lexemes as structural signs, i. e., a word, considered as a categorical — functional unit without taking into account the specific lexical value of this word;
- rank 2: a word combination, i. e., the syntactic structure, which consists of two and more syntactically connected words, but never a complete sentence (clause);
- rank 3: a clausal unit, i. e., dependent (subordinate) clause;
- rank 4: a sentence (either a simple sentence or the main clause of a complex sentence);
- rank 5: a scattered structure, i. e., a word group, which is characterized by a syntactic and semantic unity, but is discontinuous, i. e., between the members of the group there appear other language objects, which are not the members of this group;
- rank 0: the morphological units, which are not independent words, but which form a part of a lexeme of a source language, and in the language of transfer can be expressed by a clause and the units of other ranks, for example: the suffixes — *ible, — able* which are synonymous to the construction *"which can be"*, e. g. *extensible — which can be extended.*

4.2. Cross-level focus

Our studies focus on particular situations when the semantic match goes across language levels. The segmentation of phrase patterns used for the input language parse was carried out with the consideration of semantics to be reproduced via the target language means. Both the most important universals such as enumeration, comparison, modality patterns, etc., and less general structures were singled out and assigned corresponding target language equivalents.

Consider an example of a phrase structure conveying the modal meaning of obligation: "*…the task to be carried out…*". In other words, the meaning of this phrase can be rendered as "*…the task that should be carried out…*". The Infinitive phrase in the Engish language gives the regular way of expressive means compression without the loss of semantic value. A literary translation in Russian requires the second way of presenting the same idea of obligation. However in this specific case a "reduced" translation variant is also possible which consists in the introduction of the subordinate conjunction "*chtoby*" — "*so that*", between the noun and the modifying Infinitive. The parse rule would look like: *NP(to)* → *NP VPto; a*nd the generation rule would be presented as: *NP(to)* → *NP Punct.{comma} Conj.(chtoby) VPto*.

Special attention is required for the problem of passive constructions transfer. As in the phrase "*was considered*". The rules for simultaneous translation (which in many cases is similar to the real time machine translation performance and can be a source of compromise decisions for phrase structure design) requires the transformation of the English Subject into the Direct Object (Russian, Accusative Case) standing in the first position in a sentence and the passive verbal form would produce an impersonal verbal form in Russian.

Actually the process of transfer goes across the functional — categorial values of language units. A language structure which can be subjected to transfer has to be semantically complete from the point of view of its function. The cases of categorial shifts, in particular, when the technique of conversion is employed, require special treatment: the categorial shift of a syntax unit is determined by the functional role of this unit in a sentence (e. g. noun as a modifier → adjective).

Sometimes, a word may be translated by a word of another part-of-speech in the target language, a word combination, or even a clause, as the English *implementable* is best translated into Russian as *kotoryi vozmozhno realizovat* (*which can be implemented*). To overcome these differences the categorial and functional features of the two languages were considered, and the structures of the input were made conformed to the rules of the target language by applying contrastive linguistic knowledge for implementation of the transfer model. A suitable formalism is indispensable for an algorithmic presentation of the established language transfer rules, and the language of Cognitive Transfer Structures (CTS) was developed based on rational mechanisms for language structures generation and feature unification.

We apply multivariant CTG constraints to our parse and transfer algoritm to choose the optimal variants for translations from English into Russian (and from Russian into English). Each phrase (transfeme) has a set of different CTG labels, and we need a way of choosing which label to use when applying the constraint. At present we choose the best label for the phrase in a parse tree and the best transfer variant in the language of translation:

$$e = \arg\max_{e} \arg\max_{s \in CTG-labels(e,P)} p(e \mid r, s) \tag{16}$$

where *e* is an Enlish sentence, *r* is a Russian sentence, *P* is an English parse tree, *s* is a syntactic type of e belonging to the Cognitive Transfer Grammar.

Our linguistic simulation efforts are aimed at capturing the cross-level synonymy of language means and cross-linguistic semantic configurational matches for the English and Russian languages. The emphasis on the practical human translation experience gives the reliable foundation for statistical studies of parallel text corpora and automated rule extraction in further studies.

5. Rule set for training data: cognitive semantic approach

The establishment of structures equivalence on the basis of functional semantics proved to be useful for developing the syntactic parse and transfer rules module

for the English — Russian machine translation. This rule module was implemented in the first release of the Cognitive Translator system [19,20]. Generally, major efforts connected with natural language modeling lay emphasis at lexical semantics presentations and less attention is paid to the semantics of structures and establishment of functional similarity of language patterns as a core problem in multilingual systems design.

The set of functional meanings together with their categorial embodiments serves the source of constraints for the unification mechanism in the formal presentation of our grammar. The formalism developed employs feature-based parse, and head-feature inheritance for phrase structures which are singled out on the basis of functional identity in the source and target languages. The transferability of phrase structures is conditioned by the choice of language units in the source and target languages belonging to the same functional transfer fields (FTF), notwithstanding the difference or coincidence of their traditional categorial values. A set of basic FTF was singled out and language patterns employed for conveying the functional meanings of interest were examined:

- Primary Predication FTF (non-inverted) bearing the Tense — Aspect — Voice features; this field mainly includes all possible complexes of finite verbal forms and tensed verbal phrase structures.
- Secondary Predication FTF bearing the features of verbal modifiers for the Primary Predication FTF. Included here are the non-finite verbal forms and constructions, and subordinate clauses comprising the finite verbal forms. All these are united by the functional meanings they convey, e. g. qualification, circumstance, taxis (ordering of actions), etc.
- Nomination and Relativity FTF: language structures performing the nominative functions (including the sentential units) comprise this field.
- Modality and Mood FTF: language means expressing modality, subjunctivity and conditionality are included here. Here the transfer goes across the regular grammatical forms and lexical means (modal verbs and word combinations) including phrasal units.
- Connectivity FTF: included here are lexical — syntactic means employed for concatenation of similar syntactic groups and subordination of syntactic structures.
- Attributiveness FTF: adjectives and adjectival phrases in all possible forms and degrees comprise the semantic backbone of this field; included here are also other nominal modifiers, such as nominative language units and structures (*stone wall* constructions, prepositional genitives — *of* –phrases), and other dispersed language means which are isofunctional to the backbone units.
- Metrics and Parameters FTF: this field comprises language means for presenting entities in terms of parameters and values, measures, numerical information.
- Partition FTF: included in this field are language units and phrase structures conveying partition and quantification (e. g. *some of, part of*, *each of*, etc.).
- Orientation FTF: this field comprises language means for rendering the meaning of space orientation (both static, and dynamic).
- Determination FTF: a very specific field which comprises the units and structures that perform the function of determiner (e. g. the Article, which is a good

example for grammar — lexical transfer from English into Russian, since in Russian there exist no such grammatical category; demonstrative pronouns, etc.).
- Existentiality FTF: language means based on *be*-group constructions and synonymous structures (e. g. sentential units with existential *there* and *it* as a subject: *there is…*; *there exists…*; etc.).
- Negation FTF: lexical — syntactic structures conveying negation (e. g. *nowhere to be seen*, etc.).
- Reflexivity FTF: this field is of specific character since the transfer of reflexivity meaning goes across lexical — syntactic — morphological levels.
- Emphasis — Interrogation FTF: language means comprising this field are grouped together since they employ grammar inversion in English.
- Dispersion FTF: individual language structures specific for a given language are included here; these are presented as phrasal templates which include constant and variable elements. We single out 3 major types of the Dispersion FTF.

Interpretation techniques employ the segmentation of structures carried out on the basis of the functional transfer principle. The principal criterion for including a language structure into a field is the possibility to convey the same functional meaning by another structure of the field, i. e. the interchangeability of language structures. A constraint-based formalism which is called the Multivariant Cognitive Transfer Grammar has been developed and. It comprises about 350 transferable phrase structures together with the multiple transfer rules combined within the same pattern. Such patterns, or Cognitive Transfer Structures (CTS), serve constitutional components of the declarative syntactical processor module and encode both linear precedence and dependency relations within phrase structures. Consider, for example, the functional meaning of *Possessiveness*, which belongs to the Functional Transfer Field of *Attributiveness* in the following phrases: *Peter's house; the house of Peter.*

However, we see our main objective not in creation of an abstract semantic meta language, but in a careful research of all possible kinds of configurations of language patterns used by natural languages for expression of functional meanings.

5.1. Linguistic filters on the basis of the Cognitive Transfer Grammar

The key idea of our linguistic framework is cognitive cross-linguistic study of what can be called *configurational* semantics, i. e. the systemic study of the language mechanisms of patterns production, and what meanings are conveyed by the established types of configurations. We explore the sets of meanings fixed in grammar systems of the languages under study. Our studies are focused on the types of meanings outside the scope of lexical semantics, and we consider the lexical semantics when the meanings which we denote as configurational, have expression at the lexical level. The importance of this aspect is connected with the fact that natural languages are selective as to the specific structures they employ to represent the referential situation. However, it is always possible to establish

configurations which perform the same function across different languages (i. e. isofunctional structures). The parse aimed at transfer procedures requires a semantic grammar and cannot be efficiently implemented through a combination of monolingual grammars.

In the previously formulated Cognitive Transfer Grammar (CTG) [19, 20] the functional meanings of language structures are determined by the categorial values of head elements. The probability characteristics are introduced into the rules of the unification grammar as weights assigned to the parse trees.

In the Cognitive Transfer Grammar the basic structures are the *transfemes*. A *transfeme* is a unit of cognitive transfer establishing the functional semantic correspondence between the structures of the source language L_S and the structures of the target language L_T. For the alignment of parallel texts the transfemes are given as the rewrite rules in which the left part is a nonterminal symbol, and the right part are the aligned pairs of chains of terminal and nonterminal symbols which belong to the source and target languages :

$$T \rightarrow \langle \rho, \alpha, \sim \rangle, \tag{17}$$

where T is a nonterminal symbol, ρ and α are chains on terminal and nonterminal symbols which belong to the Russian and English languages, and ~ is a symbol of correspondence between the nonterminal symbols occuring in ρ and the nonterminal symbols occuring in α. In the course of parallel texts alignment on the basis of the CTG the derivation process begins with a pair of the linked starting symbols S_r and S_α, then at each step the linked nonterminal symbols are rewritten pairwise with the use of the two components of a single rule.

5.2. CTG-alignment

For automatic extraction of the rules on the basis of CTG from parallel texts these texts should be previously aligned by sentences and words. The extracted rules base on the wordwise alignments in such a way that at first the the starting phrase pairs are identified with the use of the same criterion as the majority of statistical models of translation employing the phrase-based approach [16], which means that there should be at least one word inside a phrase in one language aligned with some word inside a phrase in another language, but no word inside a phrase in one language can be aligned with any word outside its pair phrase in another language.

Definition 1. Assume that a pair of sentences $\langle r, e, \sim \rangle$ aligned wordwise is given, assume that r_1^j denotes a substring r from the position i to the position j inclusive, and correspondingly, $e_{i'}^{j'}$ denotes a substring e from the position i' to the position j' inclusive. Then the rule $\langle r_1^j, e_{i'}^{j'}, \sim \rangle$ is a starting phrase pair.

In order to continue the extraction of rules from the singled out phrases we find the phrases which contain other phrases and substitute them by nonterminal symbols.

Thus the mechanism of rules embedding is implemented which reflects the hierarchical structure of the natural language.

The next step is the formation of the rule system in the CTG notation. Cognitive Transfer Grammar is a generative unification grammar having a hierarchical structure and reflecting a major part of language transformations employed in the process of translation from one language into another. Besides, basing on the experimental data obtained from the corpora study the CTG rules are supplied with the weights of possible derivation variants.

Definition 2. Cognitive Transfer Grammar G_{CT} is a set

$$G_{CT} = \{T_{L_1}, T_{L_2}, N_{L_1}, N_{L_2}, P_{CA}, P_{CT}, S_{L_1}, S_{L_2}, M, D\}, \quad (18)$$

Where T_{L_1}, T_{L_2} are the sets of terminal symbols of the languages L_1 and L_2; N_{L_1}, N_{L_2} are the sets of non-terminal symbols of the languages L_1 and L_2; P_{CA}, P_{CT} are the rules of analysis and synthesis on the basis of the cognitive transfer ; S_{L_1}, S_{L_2} are a pair of the starting symbols of the languages L_1 и L_2 with which the process of analysis and alignment of sentences is initiated; M is the function of establishing the correlations between the structures of the languages L_1 and L_2; D is the function assigning the probability values to each rule from the sets P_{CA}, P_{CT}.

Ambiguity is an immanent feature of the natural language and it is a cause of major difficulties in machine translation implementation. Ambiguous and polysemous syntactic structures are taken into account in the further development of the CTG mechanisms, which is the multivariant CTG, and the implementations of the multivariant CTG data structures are designed as linguistic filters for statistical translation models. These data structures are called multivariant cognitive transfer structures (MCTS). The general presentation of the MCTS syntax is as follows :

MCTS {MCTS <identifier> MCTS <weight> MCTS <tag>}→
<Input phrase structure and the set of its features and values > →
<Head-driven transfer scheme> →
<Generated phrase structure and its set of features and values — variant 1>
<weight 1>
<Generated phrase structure and its set of features and values — variant 2>
<weight 2>
<Generated phrase structure and its set of features and values — variant N>
<weight N> .

The new multivariant CTG captures the polysemy of syntactic structures, the mechanisms of disambiguation basing on statistical data are introduced into the systems of parse and transfer rules, possible contexts of language structures are taken into account.

The multivariant CTG provides an extensible platform for the development of machine translation and knowledge extraction systems. At present the CTG principles are employed for development of the rule systems for the Russian-French and

Russian-German language pairs. A new hybrid approach to construction of the models for machine translation and other natural language processing systems bridges the gap between symbolic and stochastic paradigms. The new training data sets are introduced into the linguistic knowledge base for upgrading the rule systems. The linguistic filters employed for reduction of the noise rules generated in the process of learning are based on the cognitive transfer spaces which comprise major groups of cross-lingual functional synonyms.

Conclusions The urgency of the new hybrid methods of language objects presentation is caused by the demand for the optimal combination of advantages of the two research paradigms: logical linguistic modelling employing the designed rules and stochastic approach based on machine learning. This development is of special importance for the tasks of structural analysis and computer modelling of the full text scientific and patent documents. The work with patent documents requires the introduction of specific features of patent texts: such as employment of certain language constructions, the syntax of patent formulae, the extensive use of templates, domain-oriented lexicons. The Intertext base comprises a collection of scientific and patent texts in the Russian and English languages from the areas of Computer Science, Social Monitoring, Chemical Technology and other areas. One of the latest developments is connected with implementing the natural language web service for the multilingual search and analysis of financial information.

The objectives of the prospective research and development efforts consist in the inclusion of parallel texts and language processing features for the French, German and Italian languages, and evolving the Intertext into a multilingual knowledge base. Our focus on configurations provides high portability to the language processing software designed under these principles: we can operate with a lexicon which has only standard linguistic information including morphological characteristics, part of speech information and the indication of transitivity for verbs. The approach taken would be important in further development of educational programs for computer science and computational linguistics courses. Educational relevance of the methods discussed in the paper lies in deeper understanding of uniform cognitive mechanisms employed in particular language embodiments of semantic structures.

References

1. *Brown P. F., Cocke J., Della Pietra S. A., Della Pietra V. J., Jelinek F., Lafferty J. D., Mercer R. L., Roossin P. S.* 1990. A Statistical Approach to Machine Translation. Computational Linguistics,16 : 79–85.
2. *Brown P. F., Della Pietra S. A., Della Pietra V. J., Mercer R. L.* 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, 19 (2) : 263–311.
3. *Callison-Burch C.* 2008. Syntactic Constraints on Paraphrases Extracted from Parallel Corpora. Proceedings of EMNLP-2008.

4. *Chen S. F.* 1993. Aligning Sentences in Bilingual Corpora Using Lexical Information. Proceedings of the 31st Annual Conference of the Association for Computational Linguistics : 9–16.
5. *Dempster A. P., Laird N. M., Rubin D. B.* 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Ser. B, 39 (1) : 1–22.
6. *Gale W. A., Church K. W.* 1993. A Program for Aligning Sentences in Bilingual Corpora. Computational Linguistics, 19 : 75–102.
7. *Niesler T. R., Woodland P. C.* 1999. Modelling Word-Pair Relations in a Category-based Language Model. IEEE ICASSP-99, IEEE : 795–798.
8. *Ney H., Essen U., Kneser R.* 1994. On Structuring Probabilistic Dependencies in Stochastic Language Modeling. Computer Speech and Language, 8 : 1–38.
9. *Marino J. B., Banchs R. E., Crego J. M., de Gispert A., Lambert P., Fonollosa J. A. R., Costa-Jussa M. R.* 2006. N-gram-based Machine Translation. Computational Linguistics, 32 (4) : 527–549.
10. *Masahiko H., Yamazaki T.* 1996. High-Performance Bilingual Text Alignment Using Statistical and Dictionary Information. ACL 34 : 131–138.
11. *Och F. J., Ney H.* 2000. A Comparison of Alignment Models for Statistical Machine TransLation. COLING'00: The 18th International Conference on Computational Linguistics : 1086–1090.
12. *Och F. J., Ney H.* 2003. A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, 29 (1) : 19–51.
13. *Rosenfeld R.* 1996. A Maximum Entropy Approach to Adaptive Statistical Language Modeling. Computer Speech and Language, 10 : 187–228.
14. *Vogel S., Ney H., Tillmann Ch.* 1996. HMM-based Word Alignment in Statistical Translation. COLING'96: The 16th International Conference on Computational Linguistics : 836–841.
15. *Callison-Burch C., Koehn P., Monz C., Schroeder J.* 2009. Findings of the 2009 Workshop on Statistical Machine Translation. Proceedings of Workshop on Statistical Machine Translation (WMT09).
16. *Och F. J., Ney H.* 2004. The Alignment Template Approach to Statistical Machine Translation. Computational Linguistics, 30 : 417–449.
17. *Koehn P., Hoang H.* 2007. Factored Translation Models. Proceedings of the 2007Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) : 868–876.
18. *Yeniterzi R., Oflazer K.* 2010. Syntax-to-Morphology Mapping in Factored PhraseBased Statistical Machine Translation from English to Turkish. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics : 454–464.
19. *Kozerenko E. B.* 2003. Cognitive Approach to Language Structure Segmentation for Machine Translation Algorithms. Proceedings of the International Conference on Machine Learning, Models, Technologies and Applications : 49–55.
20. *Kozerenko E.* 2008. Features and Categories Design for the English-Russian Transfer Model. Advances in Natural Language Processing and Applications Research in Computing Science, 33 : 123–138.

21. *Wang W., May J., Knight K., Marcu D.* 2010. Re-Structuring, Re-Labeling, and ReAligning for Syntax-Based Statistical Machine Translation. Computational Linguistics, 36(2).
22. *Zhang H., Gildea D., Chiang D.* 2008. Extracting Synchronous Grammar Rules from Word-Level Alignments in Linear Time. Proceedings of the 22nd International Conference on Computational Linguistics (COLING-08).

НЕВЕРБАЛЬНЫЙ ДИАЛОГ В ИСТОРИИ КИНЕСИКИ

Г. Е. Крейдлин (gekr@iitp.ru)

Российский государственный гуманитарный университет, Москва, Россия

В докладе анализируются некоторые телесные знаки и невербальные формы диалога на материале книг Джонатана Свифта «Путешествия Гулливера» и «Эротические приключения Гулливера». Реконструируются жесты, наиболее распространенные в XVII и XVIII вв., а также представления Свифта об общественной морали, о социальной и личной жизни человека и о возможности использования в актах коммуникации естественного языка и языка тела в качестве «лингва франка».

Ключевые слова: кинесика, невербальный диалог, жесты, язык тела.

NONVERBAL DIALOG IN THE HISTORY OF KINESICS

G. E. Kreidlin (gekr@iitp.ru)

Russian State University for the Humanities, Moscow, Russian Federation

The traditional distinction between synchronic and diachronic gesture studies, which has been the cornerstone of nonverbal semiotics and kinesics, is being partly erased if one regards the refl ection of gestures in fi ction. This paper analyses the descriptions of several somatic signs and nonverbal forms of dialog in the two books of J. Swift — "Gulliver's Travels" and "Gulliver's Erotic Adventures". It argues that these remarkable works of art implement both some of the most common 17th and 18th century gestures and Swift's philosophical and scientifi c ideas concerning public morals, social and personal activities, communication and notions of language and gestures as lingua franca. I mean to discuss nonverbal acts of that time, purposely performed or uncontrollably leaked, that enhance, improve or disguise verbal messages in the texts. Nonverbal behavior can replace, multiply, or complement language, and Swift's books demonstrate these primary functions of nonverbal sign units vividly and convincingly. In Swift's time the gestural, or body language, as opposed to the natural languages has been considered common, plain, comprehensible, pure, forthright, and therefore the most effective in human communication. Face-to-face dialogs of the

author's characters and their corporeal activity incorporate many nonverbal signs that most of his contemporaries regarded as universal. However, Swift mocks and even jeers sometimes at these prevalent viewpoints because he would not believe in the uniqueness and in the universality of the body language.

Key words: kinesics, nonverbal dialogue, gestures, body language.

I

История кинесики — науки о жестах, жестовых процессах и жестовых системах — область для изучения не менее интересная, чем её современное состояние[1]. Ведь по жестам и жестовому поведению реконструируется отношение людей к внешней и внутренней жизни, к предметам и событиям, к телу и телесным характеристикам и еще очень многое другое. Раннему средневековью в Европе, например, было свойственно презрение к телу и стремление обуздать его различные проявления, усмирить, или, как говорили тогда, «смирить» тело. Позже, в XII–XIII веках, отношение к телу европейцев резко меняется: тело начинает рассматриваться ими как правильная и оптимальная оболочка для души, а телесная красота — как прямо соответствующая красоте человеческой души. В последующие исторические периоды внимание ученых и просто путешественников и бытописателей начинают постепенно привлекать не только телесные очертания и пропорции тела и его частей, но также знаковые движения и действия с телом и над телом.

Первые известные нам исследовательские работы, посвященные жестам (которые тут понимаются широко, а именно как лексические единицы языка тела разной природы[2]), появились в XVII веке. Эти работы были связаны с такими сферами человеческой деятельности, как риторика, медицина, психология, педагогика, искусство и физиогномика (физиогномика представляет собой учение о том, как отражаются в чертах, формах и выражениях лица человека его внутренние — психологические и ментальные — качества). Сюда же добавлю и некоторые другие важные в культурном и социальном отношении области, такие, как хирология (язык линий ладоней и бугорков рук), хирономия (мануальная риторика) и хиромантия (искусство гадания по руке). Все они составили предмет едва ли не самой первой книги, относящейся к кинесике. Я имею в виду руководство, написанное Дж. Балвером и вышедшее в свет в 1644 году[3].

По Балверу именно язык рук является естественным языком, будучи подлинно природным образованием, в отличие от искусственного, придуманного,

---

[1] Данная работа выполнена в рамках исследовательского проекта «Тело и его части в разных языках и культурах: типологическое описание» (2010–2012 гг.), получившего поддержку Российского гуманитарного научного фонда (грант РГНФ 10-04-00125а).

[2] О языке жестов, или, иначе, языке тела (body language) — главным образом, о русском языке тела — см. подробно в книге Крейдлин 2002.

[3] См. Bulwer 1974 / 1644.

языка слов, и уже по одному только этому язык рук заслуживает самого серьезного исследования. А человеком, с которого началось систематическое научное изучение другого рода телесных знаков — выражений лица и мимических жестов, — был Джон Каспер Лафатер, который в 1792 году опубликовал «Эссе по физиогномике». Он был первым из европейцев, кто провел подробные наблюдения и описал различные соотношения между выражениями лица и конфигурациями тела, с одной стороны, и типами внутренних, личностных свойств человека, с другой.

Если исследования Лафатера оказали огромное влияние на русскую культуру и науку, о чем мне уже доводилось как-то писать[4], то труд Балвера в значительной степени повлиял на культурную и общественную жизнь тогдашней Западной Европы. Дж. Балвер выделил и описал большое число кодифицированных и понятных нам жестов и даже классов жестов — это молитвенные жесты, жесты мольбы, аплодисменты, некоторые экспрессивные жесты, жесты защиты и др. Но вместе с тем в его руководстве содержится и описание телесных знаков, крайне неопределённых по смыслу и форме. Вот, например, как представлен в нём диалогический жест, который, по словам Балвера, выражает намерение человека открыто высказать перед собеседником свой взгляд на некоторый вопрос или своё отношение к некоторому событию: «Пальцы руки сомкнуты, смотрят вниз, затем рука разворачивается ладонью вверх и раскрывается» (Bulwer 1974 / 1644: 171). Отмечу, впрочем, что жестовая составляющая «открытая и направленная вверх ладонь» входит в состав форм многих современных иллюстративных жестов, и не только русского языка тела[5]. Она обозначает открытие некой новой темы или начало высказывания мнения по какому-то вопросу, что вполне согласуется с описанием Балвера.

Мысли Балвера и Лафатера нашли отражение в целом ряде культурно значимых текстов разных эпох, прежде всего, в литературных текстах. И потому, казалось бы, проблемы представления в художественных произведениях невербальных аспектов устного диалога должны были сразу стать предметом пристального внимания и изучения. Однако жесты рук, ног и головы, выражения лица, касания и другие единицы языка тела лишь относительно недавно стали интересовать лингвистов, филологов, психологов и других гуманитариев. В чем же причина такого к ним отношения со стороны учёных?

Этому можно дать разные объяснения, вполне совместимые одно с другим. О некоторых я уже писал, а потому упомяну их здесь кратко просто для того, чтобы составилась более полная картина.

Вернусь к истории кинесики. С XVIII века до середины XIX века проблема порождения и понимания жестовых и смешанных, жестово-речевых текстов начинает занимать многих. Среди них, однако, едва ли можно насчитать больше десятка известных нам лингвистов, литературоведов и психологов. Преобладали биологические (физиологические, медицинские) и философские идеи. Теоретические изыскания проводились преимущественно в трёх странах — Германии, Франции

---

[4] См. Крейдлин 2008.

[5] Об иллюстративных жестах и об их отличии от других семиотических типов жестов — эмблематических и регулятивных жестов — см., например, в книге Крейдлин 2002.

и (весьма незначительно) Англии. Психологи в то время осознанно пренебрегали жестами, по всей видимости, по той причине, что жесты казались слишком тесно связанными с намеренными действиями людей. Считалось, что жесты имели сугубо инструментальную, техническую природу, что мешало им участвовать в понимании интуитивного и иррационального моментов, особенно интересовавших психологов. Забыли о жестах и лингвисты, поскольку лингвисты думали о них тогда (да и сейчас часто тоже) как об исключительно индивидуальных выражениях, не поддающихся сколько-нибудь лингвистически интересным укрупнениям, группировкам и классификациям. Считалось, что очень трудно, если возможно вообще, объединить жесты в достаточно стройные системы типа фонологической или грамматической, которые главным образом занимали тогда лингвистов.

По мере появления новых способов аналитического изучения телесных знаков и осознания системного характера языков тела наблюдается постепенный переход от изучения индивидуального жестового выражения к анализу жестовой системы, и устный диалог начинает пониматься как область сложного взаимодействия естественного языка и невербальных знаковых кодов. Примерно со второй половины XIX века на жесты начинают смотреть как на знаки, способствующие проникновению в естественную историю, культуру, общественную жизнь, в области мышления, чувствования и понимания.

Тем не менее, изучение единиц и моделей невербального поведения и того, как оно управляет взаимодействием людей, не слишком увеличили интерес учёных к проблемам отражения телесных знаков в художественных текстах. Жесты, которые попадали в сферу внимания лингвистов, казались столь тесно связанными с языковыми и речевыми единицами, что отдельного интереса не вызывали: хорошо изучив язык и речь, можно узнать всё, что нужно, и о жестах — вот как примерно рассуждали тогда учёные. Не изменилась кардинально ситуация и тогда, когда к анализу языка тела приступили специалисты по семиотике — науке о знаках и знаковом поведении. Между тем, очевидно, что всякая символическая знаковая деятельность подлежит внимательному смысловому прочтению и разгадке. Так, много раз было показано, например, что жестикуляция может не только дополнять и усиливать речь, но и противоречить ей, а тогда человеку приходится выбирать, чему верить — речи или жестам.

Еще одна видимая причина почти полного пренебрежения в то время невербальными аспектами коммуникации — это существовавшая в европейской культуре мода на проблемы и темы, далекие от анализа устного диалога. Соотношение слов и называемых ими объектов, языковые средства выражения мыслей, механизмы и способы овладения языком детьми и иностранцами, взаимодействие языка, культуры и общества, многообразные проблемы, относящиеся к языкам мира, например, новое открытие или реконструкция языков, которые считались первоначальными или совершенными, создание искусственных языков, в частности, языков естественных наук, философских и логических языков, построение типологии языков, фонетические и грамматические вопросы, совершенствование владения иностранными языками в практическом (переводческом, дидактическом, редакторском, риторическом) плане — вот далеко не полный перечень вопросов, волновавших в то время европейских лингвистов.

Природный язык жестов (или язык тела, как его иногда называют) в те времена вслед за Балвером считали языком универсальным, избежавшим Вавилонского смешения языков. Считали, что он прост в использовании, нагляден и удобен в общении, что его все знают и понимают. А сложившееся разноязычие, согласно библейскому учению, было «наказание Божие, наложенное на людей с целью затруднить сношения их между собою, так как, в силу греховной наклонности сердца человеческого, подобными сношениями люди по преимуществу пользуются ко злу» (цит. по «Хронос», статья «Вавилонское столпотворение»[6]). Между тем жесты, как тогда полагали, могли вполне успешно применяться для общения с людьми не только своей культуры, но других культур. Например, Блаженный Августин в своей «Исповеди» (VIII, 13) пишет: «Я схватывал памятью, когда взрослые называли какую-нибудь вещь и по этому слову оборачивались к ней; я видел это и запоминал: прозвучавшим словом называлась именно эта вещь. Что взрослые хотели ее назвать, это было видно по их жестам, **по этому естественному языку всех народов** (выделено мной — Г. К.), слагающемуся из выражения лица, подмигиванья, разных телодвижений и звуков, выражающих состояние души, которая просит, получает, отбрасывает, избегает».

Не удивительно, что единицы универсального языка жестов и отношение к нему общества можно пытаться реконструировать по художественным текстам разных стран и культур, относящимся к тому времени — особенно по тем, которые были тогда популярны. В частности, мы можем узнать («прочесть», но также «вычислить» или просто «догадаться»), какие употреблялись тогда жесты (какую имели форму, что значили и в каком контексте преимущественно использовались) и что думали люди вообще о языке тела, если обратиться замечательному литературному памятнику Европы XVII–XVIII веков, роману Джонатана Свифта «Путешествия Гулливера»[7].

Кроме того, в 2006 году в Санкт-Петербурге вышла в переводе с английского книга Дж. Свифта «Эротические приключения <в некоторых отдаленных частях света> Лемюэля Гулливера, сначала хирурга, а потом капитана нескольких кораблей». Во вступительной статье, озаглавленной «От издательства», отмечается, что данная книга является переводом ранее не издававшейся подлинной рукописи Свифта 1727 года, представляющей собой главы и фрагменты эротического содержания, изъятые издателем «Путешествий Гулливера». Эта книга тоже содержит (прямо или косвенно) информацию о невербальном поведении людей, а потому выводы и соображения, которые мы далее делаем и приводим, опираются на материал обеих книг.

---

[6] Вавилонское столпотворение // Хронос. Всемирная история в Интеренете. http://www.hrono.ru/religia/spravka/vavilon_stolp.php.

[7] Полное название книги таково: "Путешествия в некоторые удалённые страны мира в четырёх частях: сочинение Лемюэля Гулливера, сначала хирурга, а затем капитана нескольких кораблей" (Travels into Several Remote Nations of the World, in Four Parts. By Lemuel Gulliver, First a Surgeon, and then a Captain of several Ships). Первое, лондонское, издание книги вышло в 1726–1727 годах.

Эти взаимодополняющие произведения порождают разного рода реминисценции и аллюзии к другим текстам, прежде всего, к художественным произведениям, которые появились раньше и которые в сильной степени повлияли на отношение людей к языку тела. В качестве примера назову лишь одну книгу, а именно «Гаргантюа и Пантагрюэль» замечательного французского писателя XV века Франсуа Рабле, и многие произведения живописи мастеров прошлого, в которых жесты, позы, выражения лица и взгляды представлены в исключительно яркой, экспрессивной манере[8].

Дальнейшую часть работы мы посвятим реконструкции и обсуждению отдельных существовавших тогда жестов и представлений людей о языке жестов. Такая реконструкция, как кажется, может в каких-то моментах не только дополнить историю кинесики, но и глубже понять отдельные места и идеи романа.

II

Жесты в книгах Свифта составляют существенный элемент текста, потому что они служат основным, естественным и эффективным средством общения главного героя с другими персонажами. Именно роман-путешествие, к жанру которого принадлежат обе книги, позволяет продемонстрировать универсальность языка жестов и некоторые его характеристики, как они тогда виделись людям. Попадание Гулливера сначала в страну маленьких человечков Лилипутию, затем в страну великанов Бробдингнег, потом попадание в ещё одну вымышленную страну Бальнибари со столицей, летающим островом, Лапута и др. — все путешествия демонстрируют сравнительную лёгкость межкультурной коммуникации при помощи жестов. По сути дела эти знаки являются единственным способом общения Гулливера с экзотическими народами, которые не владеют «хоть кому-то понятными» естественными языками. Жесты же понятны всем его собеседникам. Ср.:

(1) *Я не мог сдержать своего нетерпения и <…> несколько раз поднес палец ко рту, желая показать, что хочу есть. Гурго <…> отлично понял меня* (Путешествия Гулливера, ч. 1);

(2) *О том, что это была именно особа, я, невзирая на её малые размеры, безошибочно догадался по походке: лилипутские женщины, как и наши, ходят, чуть покачивая бёдрами* (Эротические приключения Гулливера, ч. 1);

(3) *Я протянул вперёд ладонь, положив её тыльной стороной на пол, приглашая тем самым даму ступить на неё* (там же, ч. 1);

[8] См., например, великолепные живописные и скульптурные изображения, содержащиеся на страницах энциклопедического издания Пасквинелли Б. Жест и экспрессия. // М.: Омега, 2009.

(4) *Я знаками дал понять, что они могут делать со мной всё, что им угодно* (Путешествия Гулливера, ч. 2).

Некоторые невербальные знаки, встречающиеся в романах, носят, однако, символический, а не иконический или инструментальный характер. Поэтому они сегодня кажутся нам и трудно воспроизводимыми и плохо понимаемыми. Тезис об универсальности и общепонятности жестов писатель фактически не только ставит под сомнение, но и иногда высмеивает его. Когда коммуникация Гулливера с лилипутами оказывается или может оказаться провальной, он прибегает к словам, как в случае (5), где он, видимо, использует слова «иностранного» лилипутского, языка:

(5) *Я просил его величество быть спокойным на этот счет, заявив, что готов раздеться и вывернуть карманы в его присутствии. Все это я объяснил частью словами, частью знаками* (Путешествия Гулливера, ч. 2).

Впрочем, слова далеко не всегда помогают пониманию смысла сообщения.

Тем не менее, именно жесты позволяют Гулливеру общаться с другими. Не случайно Дж. Свифту — через Гулливера — удается на их основе или с их помощью сравнительно простым способом донести до читателя собственные философские и социальные воззрения. Так, он выразил в романах своё отношение к Академии, которую называет *Академией Прожектёров*, и её «научным» изысканиям в области механики и математики, обработки земли и астрономии, высмеял существовавшую тогда практическую медицину, ср. фрагменты (6) и (7) из «Путешествий Гулливера»:

(6) *Я посетил также математическую школу, где учитель преподает по такомуметоду, какой едва ли возможно представить себе у нас в Европе. Каждаятеорема с доказательством тщательно переписывается на тоненькой облаткечернилами, составленными из микстуры против головной боли. Ученик глотает облатку натощак и в течение трех следующих дней не ест ничего, кроме хлеба и воды. Когда облатка переваривается, микстура поднимается в его мозг, принося с собой туда же теорему <…>.*

(7) *Я пожаловался <…> на легкие колики, и мой спутник привел меня в комнату знаменитого медика, особенно прославившегося лечением этой болезни путем двух противоположных операций, производимых одним и тем же инструментом. У него был большой раздувальный мех с длинным и тонким наконечником из слоновой кости. Доктор утверждал, что, вводя трубку на восемь дюймов в задний проход и втягивая ветры, он может привести кишки в такое состояние, что они станут похожими на высохший пузырь. Но если болезнь более упорна и жестока, доктор вводит трубку, когда мехи наполнены воздухом, и вгоняет этот воздух в тело больного; затем он вынимает трубку, чтобы вновь наполнить мехи, плотно закрывая на это время большим пальцем заднепроходное отверстие. Эту операцию*

он повторяет три или четыре раза, после чего введенный в желудок воздух быстро устремляется наружу, увлекая с собой все вредные вещества (как вода из насоса), и больной выздоравливает.

В своих книгах Свифт сатирически высмеивает самомнение и тщеславие людей. В стране Glubbdubdrib ‘Глаббдобдриб’ Гулливер знакомится с чародеями, способными вызывать тени умерших, и беседует со специалистами по древней истории, обнаруживая, что, в сущности, всё было не так, как пишут в исторических сочинениях. Там же он узнает про struldbrugs ‘струльдбругов’ — людей, рожденных вроде бы нормальными, но фактически бессмертных. Бессмертие, однако, не приносит струльдбругам счастья, поскольку они обречены на бесконечную старость, страдания, и болезни, которые приходят к ним сразу после тридцати лет (в частности, в это время они теряют зрение и волосы).

Высказывает Свифт также оригинальные суждения о свободной и подлинной любви, человеческой преданности и предательстве, нравственности и морали и отношения к ним людей, ср.:

(8) *Мне еще предстояло узнать, что нравы в Лилипутии довольно свободные, и хотя начальство и пытается насаждать нравственность<…>, население да и сама власть имущие таковой не следуют; одобряя мораль только на словах, они в реальной жизни действуют, согласуясь более со своими подспудными и явными желаниями <.,.>* (Эротические приключения Гулливера), но для нас интереснее всего его отношение к языку как средству общения.

Писатель считал и устную, и письменную формы естественных языков весьма несовершенными. Большинство лилипутов у него говорят резкими и визгливыми (*shrill*) голосами, их разговоры большей частью весьма эмоциональные и ведутся на повышенных тонах, многие слова лилипутского языка и языков других стран, в которых побывал Гулливер, содержат трудно произносимые сочетания звуков. Ср., например, английские имена *Brobdingnagians* (название страны), *ihnuwnh* (слово языка гуигнгнмов, а вот к нему авторский комментарий: *the word is strongly expressive in their language, but not easily rendered into English; it signifies ‘to retire to his first mother’* (англ. версия Путешествий Гулливера, кн. 4, с. 176), *gnnayh* ‘хищная птица’, *Houyhnhnms* ‘гуигнгнмы’. Фонетика слов их языка «очень неприятная» — в словах много назальных и задненёбных согласных: *In speaking they pronounce through the Nose and Throat* (там же, с. 219). Ср. также русские соответствия английским именам *Лаггнегг, Глаббдобдриб, Глюмдальклич* или такие фразы, как (9):

(9) *Она дала мне имя Грильдриг, которое утвердилось за мной сперва в семье, а потом и во всем королевстве. Это слово означает то же, что латинское «homunculus», итальянское «homynceletino» и английское «mannikin».*

А вот еще одна показательная цитата из «Путешествий Гулливера» (гл. V):

(10) <…> *Мы пошли в школу языкознания, где заседали три профессора на совещании, посвященном вопросам усовершенствования родного языка. Обсуждались разные проекты. Первый проект предлагал сократить разговорную речь путем сведения многосложных слов к односложным и упразднения глаголов и причастий, так как в действительности все мыслимые вещи суть только имена. Второй проект требовал полного упразднения всех слов; автор этого проекта ссылался главным образом на его пользу для здоровья и сбережение времени. Ведь очевидно, что каждое произносимое нами слово сопряжено с некоторым изнашиванием легких и, следовательно, приводит к сокращению нашей жизни. А так как слова суть только названия вещей* (эта мысль была доминирующей во времена Свифта — Г. К.), *то автор проекта высказывает предположение, что для нас будет гораздо удобнее носить при себе вещи, необходимые для выражения наших мыслей и желаний.* <…> *Многие весьма ученые и мудрые люди пользуются этим новым способом выражения своих мыслей при помощи вещей. Единственным его неудобством является то обстоятельство, что* <…> *собеседникам приходится таскать на плечах большие узлы с вещами, если средства не позволяют нанять для этого одного или двух дюжих парней. Мне часто случалось видеть двух таких мудрецов, изнемогавших под тяжестью ноши, подобно нашим торговцам вразнос. При встрече на улице они снимали с плеч мешки, открывали их и, достав оттуда необходимые вещи, вели таким образом беседу в продолжение часа; затем складывали свою утварь, помогали друг другу взваливать груз на плечи, прощались и расходились.*

Мудрецы, над которыми издевается Свифт, желающие разговаривать с другими людьми, вынуждены взваливать на плечи и повсюду таскать с собой тяжёлые мешки, что, по его мнению, и есть «достоинство» языка.

Естественные языки подверглись критике Свифта также и совершенно в другой связи. Они, по мнению Свифта, меняются столь неупорядоченно, незакономерно и настолько быстро, что даже если вы отсутствуете в родной стране и не имеете возможности говорить на родном вам языке всего лишь 16 лет, то по возвращению домой не сможете разговаривать с соотечественниками. Такую мысль Свифт высказывает в письме к своему родственнику Ричарду Симпсону на самых первых страницах романа. Язык не может, таким образом, служить надёжным и удобным средством общения.

Письменные формы языков тоже не следуют строгим нормам и не придерживаются орфографических традиций и принятых когда-то моделей. Лилипуты, например, вместо того, чтобы писать слева направо, справа налево или вверх и вниз, пишут наискось от одного угла страницы до другого, и в этом Свифт видит «их единственное сходство с порочной практикой английских леди».

Наконец, естественный язык не может воспрепятствовать человеческой лжи, обману, политической интриге и дипломатической игре. Он не эффективен и не обладает нужной образностью для передачи идей. Язык жестов,

напротив, непосредственен, понятен и легко дешифруется. Когда в книге 1 «Путешествий Гулливера» герой дважды не сумел понять, какую информацию лилипуты передали ему жестами, весьма напоминающими по форме жест из руководства Дж. Балвера, который я приводил выше, Свифт объясняет, что это произошло скорее из-за недостаточной сметливости Гулливера или плохого знания им языка жестов, чем из-за национальных или культурных границ, — ведь во всех странах люди пользуются тем же языком жестов, что и Гулливер. И этот язык в хорошем смысле консервативный и устойчивый; смысл и форма телесных знаков не меняются так быстро, как у языковых знаков.

Однако, чтобы быть в стране и общаться с народом, необходимо знать не только знаки и способы их комбинирования, нужно знать нормы общения, модели коммуникативного поведения, обычаи и любимые стереотипные жестовые формы. Например, надо знать, что в Лилипутии существуют невербальные ритуалы, такие как ритуал вызова соперника на дуэль, когда оскорблённый должен явиться к оскорбителю и притоптыванием и испусканием ветров известить его о желании драться, знать, что идеи непонимания и недоумения в проблемных ситуациях могут передаваться трясущейся головой или что в целом ряде эмоционально насыщенных ситуаций общения лилипуты «жестикулируют больше лицом, чем руками».

Подведём итог.

Из обеих книг мы узнаём, что основные жесты и классы жестов прошлого в Англии того времени были хорошо известны. Это поклоны, рукопожатия, поцелуи, например, в щёку или поцелуй кончиков пальцев, прикладывание руки к груди в качестве жеста клятвы и благодарности, поднятие головы и закатывание глаз как выражение размышления, жест «перст указующий», поднятие вверх рук и глаз в знак восхищения. Речь при этом как бы не обязательна: лилипуты переходят на исключительно речь, как правило, желая скрыть от собеседника или от наблюдателя информацию, а не сообщить ему что-то.

Проведенная реконструкция, таким образом, показывает, что во времена, когда жил и работал Дж. Свифт, отношение к жестам и их роли в устной коммуникации было прямо противоположным тому, которое повсеместно принято сегодня, а именно жесты играли тогда доминирующую, а не подчинённую роль.

References

1. *Bulwer J.* 1974/1644. Chirologia: Or the Natural Language of the Hand and Chironomia: Or the Art of Manual Rhetoric.
2. *Kreidlin G. E.* 2002. Non-verbal Semiotics: Body Language and Natural Language [Neverbal'naia Semiotika: Iazyk Tela I Estestvennyi Iazyk].
3. *Kreidlin G. E.* 2008. The History of Kinesics and the Problems of the Description of one Gesture.

4. *Paskvinelli B.* 2009. Gesture and Expression [Zhest I Ekspressiia].
5. *Semantic* Class [Istoriia Kinesiki I Problemy Opisaniia Odnogo Semanticheskogo Klassa Zhestov]. Lingvistika dla Vsekh. Letnie Lingvisticheskie Shkoly 2005 I 2006 : 100–117, available at: http://llsh.ru/books/llsh0506/llsch_2005_2006.pdf#section.1.10).
6. *Swift J.* 1987. Gulliver's Travels.
7. *Swift J.* 2006. Erotic Travels <into Several Remote Nations of the World, in Four Parts>, by Lemuel Gulliver, first a surgeon, and then a captain of several ships.

КОРПУС РУССКОЙ ДИАЛЕКТНОЙ РЕЧИ: КОНЦЕПЦИЯ И ПАРАМЕТРЫ ОЦЕНКИ[1]

О. Ю. Крючкова (vpks@rambler.ru)

В. Е. Гольдин (goldinve@yandex.ru)

Саратовский государственный университет
им. Н. Г. Чернышевского, Саратов, Россия

На основе сравнительной оценки двух диалектологических корпусов — диалектного корпуса в составе Национального корпуса русского языка (ДК НКРЯ) и Саратовского диалектологического корпуса (СарДК) — обсуждаются концепция и параметры оценки корпуса русской диалектной речи.

Ключевые слова: диалект, диалектный корпус, диалектная речь, параметры.

A CORPUS OF RUSSIAN DIALECTAL SPEECH: THE CONCEPT AND PARAMETERS OF EVALUATION

O. Iu. Kriuchkova (vpks@rambler.ru)

V. E. Gol'din (vpks@rambler.ru)

Saratov State University, Saratov, Russian Federation

The concept and parameters of evaluation of a corpus of Russian dialectal speech are discussed based on the comparative assessment of two dialect corpora — dialect corpus within the National Corpus of the Russian Language (DC NCRL) and the Saratov Dialect Corpus (SarDC): the principles of selection of the dialect materials and the criteria of the dialect corpus representativeness; the principles of the speech continuum partition in the corpus; the parameters of textual fragments return; the forms of representation of the dialect texts in the corpus; the types and rules of annotation of the corpus textual basis; the parameters of the dialect texts meta-marking; the representation of nonlinguistic information in the corpus; the possibilities of retrieval queries, optimal for dialect research. The paper proves that the dialect corpus cannot be based on the same model as the corpus

[1] Работа выполнена при поддержке Фонда «Русский мир» (грант № 354 Гр/1232-10).

of standard language because of the specifi c character of the dialect material. The dialect corpus must be modeled as a system of corpora of different dialects, representing the main dialect types of the Russian speech. According to the proportionality principle, the textual basis of the corpus of a separate dialect must be aimed at the modeling of communication in this specifi c dialect, refl ecting the main types and forms of the dialect speech, as well as social differentiation of the dialect native speakers and genre and theme structure of the dialect communication.

Key words: dialect, dialectal corpus, dialectal speech, parameters.

1. Введение

Сделанная в Национальном корпусе русского языка (НКРЯ) попытка создать на его лингвистической и программной платформе диалектный корпус русского языка, с одной стороны, продемонстрировала несомненную ценность применения к диалектному материалу общих для НКРЯ классификационных и структурных принципов; с другой стороны, результаты этой попытки и анализ параллельного развития других диалектных корпусов (например, Саратовского диалектологического корпуса) высветили целый ряд проблем, без решения которых корпусная диалектология и общерусский диалектный корпус не могут эффективно развиваться. Сейчас, когда диалектологами и разработчиками НКРЯ осознано неудовлетворительное состояние ДК НКРЯ, когда «пришло понимание, что этот проект полезно было бы перестроить так, чтобы он служил самим диалектологам — и как удобно организованный ресурс для учебного процесса, и как инструмент для исследовательской деятельности» [Рахилина 2009: 15], существует острая потребность в новых концептуальных решениях.

Необходимы выработка четкой концепции корпуса, базирующейся на выделении системы важнейших его параметров, и принятие в соответствии с каждым из них определенных решений. К числу таких параметров относятся, по нашему мнению, следующие:

1. принципы отбора диалектного материала и критерии репрезентативности диалектного корпуса;
2. принципы членения речевого континуума в корпусе;
3. параметры выдачи текстовых фрагментов;
4. формы представления диалектных текстов в корпусе;
5. виды и правила аннотирования текстовой базы корпуса;
6. параметры метаразметки диалектных текстов;
7. представление в диалектном корпусе нелингвистической информации;
8. оптимальные для диалектологических исследований возможности пользовательских запросов.

Дальнейшее изложение посвящено обсуждению названных параметров, которое проводится на основе сравнительной оценки двух диалектологических

корпусов — диалектного корпуса в составе Национального корпуса русского языка (ДК НКРЯ) и Саратовского диалектологического корпуса (СарДК[2]).

2. Параметры оценки диалектного корпуса

1. Принципы отбора диалектного материала и критерии репрезентативности диалектного корпуса

Необходимым условием решения данного вопроса является определение ***цели*** диалектного корпуса. Он может создаваться либо как корпус иллюстративного типа, единственная цель которого — демонстрация территориальной неоднородности национального языка, либо как научный источник нового типа, соответствующий общей идеологии корпусной лингвистики.

В первом случае репрезентативность диалектного корпуса определяется прежде всего максимальным охватом территорий распространения национального языка; при этом допустимы фрагментарность материала и отсутствие системности его представления. Такой корпус носит ознакомительный, популяризаторский характер, но не имеет ценности в качестве источника лингвистических или лингвокультурологических исследований[3]. Подобной модели диалектного корпуса соответствует сегодня ДК НКРЯ: «Корпус проектировался и создавался, — пишет Е.В. Рахилина, — с ориентацией на… рядовых пользователей…, большинство из которых никогда в жизни не видело ни одного диалектного текста. В то время задачей было сделать своего рода «научную игрушку», которая наглядно демонстрировала бы разнообразие русского языка в его региональных вариантах» [Рахилина 2009: 15]. ДК НКРЯ собирает сегодня по-разному записанные и по-разному же обработанные какие угодно текстовые фрагменты любых русских говоров и территориально соотносит их с областными центрами или с еще более крупными географическими ориентирами (*Архангельск, Вологда, Курск, Саратов, Забайкалье, Карелия*). Такой материал при любом увеличении его объема не может стать репрезентативным представлением ни отдельных русских говоров как особых полносистемных образований, ни русской диалектной речи в целом.

Диалектный текстовый корпус не может строиться по той же модели, что и корпус «стандартного» (литературного) языка ввиду специфичности диалектного материала, учет которой необходим при создании адекватных природе говоров и по-настоящему репрезентативных диалектных корпусов. Представим часть из этих различий в табличной форме (см. Табл. 1)

---

[2] Саратовский диалектологический корпус разрабатывается в Центре изучения народно-речевой культуры им. профессора Л.И. Баранниковой Института филологии и журналистики Саратовского государственного университета им. Н.Г. Чернышевского.

[3] Именно так строятся диалектологические корпуса многих других языков.

Таблица 1. Различия между текстами «стандартного языка» и текстами диалектными

Тексты «стандартного» языка	**Диалектные тексты**
1. «Стандартные тексты» (прежде всего письменные) в основном закреплены в составе каких-то изданий, собраний, библиотек, фонотек и т. п., в этом виде они естественным образом функционируют в преимущественно городской коммуникации и доступны наблюдателям.	1. Диалектные тексты не закреплены в их естественной (традиционной сельской) устной коммуникации, они получают закрепление и становятся доступными наблюдателям лишь в составе внеположенных диалектной коммуникации диалектных корпусов, хрестоматий и др. научных источников.
2. Достижение пропорциональности и репрезентативности корпуса русской литературной речи становится всё более простым и быстрым благодаря распространению электронных форм воплощения текстов.	2. Вследствие исключительно устного воплощения диалектной речи и множества относительно самостоятельных русских говоров при специфичности их внутренней организации репрезентативность национального корпуса по отношению ко всему «русскому диалектному языку» в составе его «микросистем» реально не может быть в обозримое время достигнута, тогда как репрезентативность материалов по отдельным давно и подробно изучаемым говорам, хорошо отражающим его главные структурные части (наречия, группы, зоны), вполне достижима уже сегодня при условии трансформации собранных диалектологами материалов в корпусную форму и целенаправленном их пополнении.
3. «Стандартные тексты» подготовлены самими авторами или публикаторами к использованию в публичном общении, рассчитаны на это, и, следовательно, открытие их в корпусе не является неразрешенным вторжением в чью-либо личную сферу.	3.Диалектные тексты в основном имеют более личный, часто даже глубоко интимный и просто наивно-открытый, незащищенный характер и были бы совершенно другими (или вовсе не состоялись бы), если бы говорящие предполагали, что их речь будет вынесена на всеобщее обозрение.

Тексты «стандартного» языка	Диалектные тексты
4. «Стандартные тексты» являются частью той культуры (в том числе языковой), к представителям которой относятся создатели корпуса и предполагаемые пользователи, поэтому тексты относительно легко могут подготавливаться к включению в корпус любыми филологически грамотными людьми и/или специальными компьютерными программами. По той же причине содержание текстов (упоминаемые в «стандартных текстах» события, лица, природные объекты, артефакты, идеи и т. п.) в большинстве случаев не требуют специального комментирования.	Диалектные тексты представляют совершенно особую, так называемую «традиционную культуру», особые самодостаточные языковые системы и автономные коммуникативные образования. Они воплощают специфическое содержание и специфические формы коммуникации, поэтому без специального лингвистического и культурологического сопровождения эти тексты могут лишь казаться понятными предполагаемым пользователям корпуса.

Следствиями отмеченных различий между «стандартными» текстами и текстами диалектными являются следующие общие принципы отбора и корпусного представления материала, без соблюдения которых репрезентативность диалектного корпуса как научного источника не может быть достигнута:

1.1. Диалектный корпус должен делать доступными по запросам не только фрагменты текстов, но и целые тексты, то есть диалектный корпус, в отличие от корпуса литературного языка, должен быть одновременно и диалектной библиотекой или архивным собранием материалов.

1.2. Диалектный корпус, как и обычные архивы, должен предусматривать различные степени допуска к различным его материалам (одну степень — для составителей, исследователей, другую для любых желающих) и постепенно открывать текстовые фонды, когда это становится возможным.

1.3.1. Диалектные тексты не могут адекватно пониматься, будучи вырванными из общего контекста родной для них традиционной культуры, поэтому они требуют воссоздания в корпусе соответствующего лингвистического и культурного (в самом широком смысле) фона в целом и в связи с содержанием каждого конкретного текста в отдельности. Эта проблема решается отсылками к размещенным в корпусе историческим, этнографическим, географическим и др. энциклопедическим данным мультимедийного характера, а также специальными комментариями к упоминаемым в конкретных текстах событиям, лицам, природным объектам, артефактам, идеям и т. п. Подобные комментарии целесообразно ориентировать не на традиционную культуру в целом и не на диалекты вообще, а на комплексы текстов конкретных говоров.

1.3.2. Подготовка диалектных текстов к введению в корпус — процесс более сложный и трудоемкий, чем подготовка «стандартных» текстов.

Он включает установление контакта с диалектоносителями, организацию записи, расшифровку, значительную долю ручной разметки, необходимо дополняющей автоматическое аннотирование, семантический и грамматический анализ. Это не механическая, а исследовательская работа, серьезный и очень ответственный авторский труд. Он может выполняться только специалистом-диалектологом, профессионально изучающим конкретный говор.

1.4. Диалектный корпус должен строиться как система корпусов отдельных говоров, представляющих важнейшие диалектные типы (наречия, группы, зоны) русской речи. В соответствии с принципом пропорциональности текстовая база корпуса отдельного говора должна стремиться к моделированию коммуникации в конкретном говоре, отражая важнейшие типы и формы диалектной речи, социальную дифференциацию носителей говора, жанрово-тематическую структуру диалектного общения.

Следование этим принципам обеспечивает репрезентативность диалектологического корпуса как научно-исследовательского источника.

2. Членение речевого континуума в диалектном корпусе

В соответствии с различными целями диалектологических корпусов в них по-разному решается задача членения представляемой в корпусе речи.

Текстовая база ДК НКРЯ наполняется отрезками диалектной речи (как правило, небольшого объема — в среднем не более 1 тыс. словоупотреблений), предварительно (до включения в корпус) фрагментированными на тематической основе. Членение речевого потока в СарДК отвечает принципу максимального приближения модели к объекту — естественной коммуникации на диалекте. Текст в СарДК полностью соответствует зафиксированному аудио- или видеоаппаратурой участку непрерывного общения, поэтому границы текста не зависят от таких параметров, как смена темы, жанра, формы речи, частичное изменение коммуникативной ситуации и числа ее участников. Такое представление речи существенно расширяет возможности использования корпуса.

3. Параметры выдачи текстовых фрагментов

В ДК НКРЯ в настоящее время предусмотрены 2 возможности выдачи: минимальный контекст и его расширение (как правило, до 3–4 строк, хотя тип расширения пока не носит последовательного характера). Однако специфика диалектного материала требует контекстов большей протяженности и возможности получения целой записи, поэтому в СарДК минимальной выдачей является абзац, а максимальной — целый текст, обычно значительной протяженности.

4. Формы представления текстов в диалектном корпусе

В ДК НКРЯ диалектные тексты представлены только в виде полуорфографической записи. Такая фиксация диалектной речи не позволяет изучать ее фонетическую сторону, что вызывает обсуждение вопроса о необходимости параллельного представления в ДК НКРЯ фонетической транскрипции. Однако в значительных по объему текстовых корпусах, в наполнении которых принимают участие большие и часто разрозненные коллективы диалектологов, достичь единообразия при транскрибировании весьма различных

по фонетической структуре диалектных текстов практически невозможно. В этих условиях бóльшую актуальность приобретает вопрос о включении в корпус аудио- и видеозаписей диалектной коммуникации и формах их соотнесения с символьной расшифровкой. В НКРЯ уже имеется успешный опыт создания устного корпуса со звуковой составляющей [см.: Гришина 2009], который важно использовать и при разработке диалектного корпуса. Отрадно, что такая перспектива уже заявлена [см.: Рахилина 2009: 14].

В СарДК параллельное представление текстовых и аудио-/видеомодулей является одним из важнейших принципов его строения, обеспечивающим максимальную достоверность информации. Наличие в корпусе звукового компонента обусловливает достаточность полуорфографической расшифровки диалектных текстов. Использование этой формы символьного представления речи в свою очередь требует решения вопроса о необходимости единого формата или достаточной степени единообразия полуорфографического представления диалектного текста в корпусе. В настоящее время в ДК НКРЯ нет единообразия символьного представления записей: расшифровки выполнены в разных диалектологических центрах по разным правилам. В СарДК используется единый формат символьной записи, создана специальная инструкция, регламентирующая характер отражения в расшифровке диалектных особенностей, способ членения текста и использование знаков препинания, способы обозначения нераспознанных фрагментов речи и недоговоренных слов, способы дифференциации речевых отрезков, принадлежащих диалектологу и диалектоносителю, способы дифференциации речевых отрезков, принадлежащих разным диалектоносителям, способ подачи необходимых для понимания текста комментариев.

5. Виды и правила аннотирования текстовой базы корпуса

Цель создания корпуса и характер включаемых в него текстов определяют принятые в нем виды и правила аннотирования текстовой базы. ДК НКРЯ и СарДК различаются как применяемыми видами разметки, так и правилами аннотирования при использовании одного и того же вида разметки.

Основным видом разметки в ДК НКРЯ и в СарДК, как и в большинстве текстовых корпусов, является морфологическая разметка, при проведении которой между сопоставляемыми корпусами есть существенные различия. Иллюстративная по своей сути стратегия ДК НКРЯ обусловливает использование дифференциального подхода при морфологической разметке корпуса. Диалектная речь представляется в ДК НКРЯ через ее соотнесение с литературной, рассматривается как речевая среда, характеризуемая отклонениями от литературных форм. Диалектные формы характеризуются такими, например, параметрами, как «другая флексия», «нестандартная флексия». Так, в качестве примеров сущ. Им. п. мн. ч. с «другой флексией» выдаются «деулинские» *окунья* и *утяты* или «волгоградское» *соседы*. «Другими» или «нестандартными» приведенные формы являются бесспорно лишь по отношению к литературной норме, но могут быть вполне системными для соответствующих говоров. Однако вопрос о системности конкретных говоров вообще не может ставиться и решаться на материале диалектного корпуса, если он строится по модели, которую можно было бы охарактеризовать как иллюстративно-дифференциальную.

В ДК НКРЯ используется сложная система дифференциальных помет, описывающая типы «отклонений» от литературной нормы [см.: Летучий 2005, 2009]. Стремление к детальной характеризации диалектных особенностей не только противоречит общему принципу НКРЯ «не навязывать пользователю своих исследовательских решений», но и в большей мере, чем для других подкорпусов НКРЯ, создает опасность навязывания субъективных квалификаций. Диалект, в отличие от других нелитературных разновидностей национального языка, — полносистемное образование, грамматическая специфика которого, как и специфика любого самостоятельного языка, не может быть описана без специального научного изучения репрезентативного языкового материала. Такое изучение и такое описание возможны лишь после создания полноценного диалектного корпуса конкретного говора, но никак не до того, как репрезентативность корпуса будет достигнута. Неизбежные уточнения и изменения в системе дифференциальных помет, возникающие по мере накопления материала (ср., напр. [Летучий 2005] и [Летучий 2009]), увеличивая трудоемкость процесса создания корпуса, не решают проблемы достоверности разметки.

СарДК в отличие от ДК НКРЯ — корпус принципиально недифференциального и нелитературноцентрического типа. Этим обусловлен ряд его отличий СарДК от ДК НКРЯ в лексико-морфологической разметке текстов (см. [Крючкова 2007]).

В связи с морфологическим аннотированием диалектного корпуса своего решения ждут также вопросы о лемматизации диалектных словоформ, о целесообразности и способах выделения в разметке различных функционально маркированных элементов в диалектной речи (просторечных, архаических форм); о целесообразности специальной маркировки различных видов неоднословных и идиоматических единиц. В ДК НКРЯ позиция по этим вопросам строго не эксплицирована. В СарДК приняты соответствующие (хотя, возможно, и не окончательные) решения и действуют определенные правила разметки названных единиц.

ДК НКРЯ и СарДК различаются по характеру тематической и жанровой разметки корпусов. В ДК НКРЯ выделение текста по тематическому принципу делает избыточной тематическую разметку каждого отдельного текста. Незначительный объем тематически цельных текстов в ДК НКРЯ обусловливает также нецелесообразность их жанровой разметки: каждый текст обычно оказывается моножанровым («бытовая сфера»).

В СарДК значительный объем, политематичность и полижанровость текстов, напротив, обусловливают необходимость такой разметки. Тематическая и жанровая разметка значительного по объему корпуса диалектных текстов в перспективе даст возможность исследовать жанрово-тематическую структуру диалектной коммуникации, выявить соотношение различных тем и жанров в составе диалектной коммуникации.

6. Параметры метаразметки диалектных текстов

В ДК НКРЯ метаописание текста ограничивается указанием на место, время записи, эксплоратора, общую тему и объем текста. Ввиду указанной выше специфики диалектных текстов названных параметров явно недостаточно для понимания и анализа диалектной коммуникации. Важными оказываются сведения о конкретной ситуации записи текста (в доме информанта,

в поле, в огороде, в лесу и т. п.), об адресатах речи, об упоминаемых в тексте лицах, о времени описываемых в тексте событий, о жанрово-тематической структуре записанного фрагмента речи. Все эти сведения включены в метаразметку текстов в СарДК.

7. Представление в диалектном корпусе нелингвистической информации

В ДК НКРЯ нелингвистическая информация ограничена приведенными параметрами метаописания. В СарДК, который создается как модель традиционной сельской коммуникации на диалекте, отражающая речевое общение в конкретных условиях жизни конкретного речевого коллектива, нелингвистической информации отводится специальное место. База корпуса включает отдельные модули с нелингвистической информацией, часть из которых связана с конкретными текстами, другая часть такой привязки не имеет. Текстовую привязку имеют биографический (биографические сведения об информанте) и иллюстративный (фотографии информанта, фотоиллюстрации к данному тексту) модули. Самостоятельный блок информации (не связанный с конкретным текстом) образуют другие справочные материалы: сведения исторического, социокультурного характера, данные демографические, этнографические, географические.

8. Оптимальные для диалектологических исследований возможности пользовательских запросов

Оптимальной для диалектологических исследований будет система пользовательских запросов, удовлетворяющая как потребностям уровневых описаний диалектной речи, так и исследовательским потребностям в области коммуникативного, лингвокогнитивного и лингвокультурологического изучения диалектного общения.

Применение к диалектному материалу общей для НКРЯ детально разработанной и разветвленной системы поисковых запросов является, безусловно, сильной стороной ДК НКРЯ и теоретически открывает перед диалектологами новые перспективы в изучении диалектной коммуникации. Однако эти возможности существенно ограничены (или даже сводятся к нулю) принципиальной нерепрезентативностью корпуса, обусловленной отсутствием концепции корпуса, соответствующей диалектному материалу.

СарДК предоставляет пользователю комплексную информацию о каждом конкретном говоре. Система пользовательских запросов в СарДК позволяет составлять выборки по отдельным морфологическим и лексическим явлениям, по тематическому и жанровому критериям, по отдельному информанту, по отдельному подкорпусу (говору) или по всем включенным в корпус подкорпусам. От текстовых модулей возможен переход к звуковым модулям и наоборот, а также параллельное их воспроизведение.

3. Заключение

Разработка диалектологических корпусов в настоящее время находится на начальной стадии, постоянно уточняются общие принципы и частные

методики их построения, поэтому широкое обсуждение обозначенных в данной статье проблем является актуальной задачей корпусной диалектологии и залогом ее успешного развития.

При условии выработки общей концепции русского диалектного корпуса, отвечающей специфике диалектного материала и современным задачам диалектологии, создания четких инструкций по всем параметрам обработки диалектного материала для включения в корпус, организации постоянно действующих семинаров для диалектологов, участвующих в наполнении корпуса, можно безусловно надеяться на то, что стадия «мечтаний, споров, проб... ошибок», в которой сейчас находится разработка диалектного корпуса [Рахилина 2009: 15–16], сменится интенсивной коллективной работой по созданию общерусского корпуса диалектной речи в составе НКРЯ.

Диалектный корпус, организованный как ***совокупность корпусов отдельных говоров***, обеспеченный ясной концепцией и инструктивными материалами, имеет реальную перспективу достижения репрезентативности в короткие сроки и может уже в ближайшее время стать научным источником, обладающим значительным эвристическим потенциалом.

References

1. *Grishina E. A.* 2009. Russian Multimedia Corpus: The Problems of Annotation [Multimediinyi Russkii Korpus (MURKO): Problemy Annotatsii]. Natsional'nyi Korpus Russkogo Iazyka: 2006-2008. Novye Rezul'taty I Perspektivy.
2. *Kriuchkova O. Iu.* 2007. Electronic Corpus of Russian Dialectal Speech, and the Principles of its Tagging [Elektronnyi Korpus Russkoi Dialektnoi Rechi I Printsipy ego Razmetki]. Izvestiia Saratovskogo Universiteta. Novaia Seriia. Filologiia. Zhurnalistika, 7 (1).
3. *Letuchii A. B.* 2005. Corpus of Dialectal Texts: Goals and Problems [Korpus Dialektnykh Tekstov: Zadachi I Problemy]. Natsional'nyi Korpus Russkogo Iazyka: 2003-2005. Rezul'taty I Perspektivy.
4. *Letuchii A. B.* 2009. Corpus of Dialectal Texts: Contents and Tagging Characteristics [Dialektnyi Korpus: Sostav I Osobennosti Razmetki]. Natsional'nyi Korpus Russkogo Iazyka: 2006-2008. Novye Rezul'taty I Perspektivy.
5. *Rakhilina E. V.* 2009. Corpus as a Creative Project [Korpus kak Tvorcheskii Proekt]. Natsional'nyi Korpus Russkogo Iazyka: 2006-2008. Novye Rezul'taty I Perspektivy.

ВЫСОКОТОЧНЫЙ МЕТОД РАСПОЗНАВАНИЯ КОНЦОВ ПРЕДЛОЖЕНИЙ

А. С. Кудинов (al.kudinov@corp.mail.ru)

А. А. Воропаев (voropaev@corp.mail.ru)

А. Л. Калинин (kalinin@corp.mail.ru)

Проект Поиск@Mail.Ru, Москва, Россия

В статье описывается метод применения машинного обучения в задаче распознавания концов предложений. Предлагаемый способ успешно решает проблему идентификации знаков препинания, таких как точка и др., которые не являются знаками конца предложения. Несмотря на сравнительно небольшой объем обучающей выборки, подготовленной вручную, способ демонстрирует точность не менее 99 % на среднестатистическом web-документе.

Ключевые слова: машинное обучение, конец предложения, знаки препинания, идентификация знаков препинания.

A HIGH PRECISION METHOD FOR THE RECOGNITION OF SENTENCE BOUNDARIES

A. S. Kudinov (al.kudinov@corp.mail.ru)

A. A. Voropaev (voropaev@corp.mail.ru)

A. L. Kalinin (kalinin@corp.mail.ru)

Project Search@Mail.Ru, Moscow, Russian Federation

We present a machine-learning method of sentence boundary recognition. The approach successfully identifies punctuation marks, such as periods or question marks that are not sentence boundary markers. In spite of a relatively small initial learning set (which was prepared manually), the accuracy of this approach appears to be no less than 99 % when applied to an average web document. The method is based upon the decision tree technique combined with a tiny set of manually constructed rules that play the role of classifi cation features. The rules are built using a dedicated declarative language, which is briefl y described. A comparison of accuracy of the approach with two freely accessible software products is provided. According to our

estimates, the algorithm provides good enough performance to be used in real-time environment such as indexer component of a web search engine. It can also be used to produce large learning sets to train faster machine learning models such as the maximum entropy model.

Key words: machine learning, sentence boundary, punctuation marks, punctuation marks identification.

Введение

В процессе разработки систем автоматической обработки естественных текстов часто возникает задача о корректном разбиении текста на предложения. В отдельных случаях её решение имеет принципиальное значение:

1. При генерации контекстов словоупотребления, например, для задач снятия морфологической омонимии [3] или семантической омонимии [4].
2. В задачах автоматической классификации текстов (рубрицирования) [5], где структурными единицами документа являются предложения, заглавия, названия колонок в таблицах, элементы списка и др.
3. При реализации корректного графематического и синтаксического машинного анализа текста [6].
4. В частности, при построении сниппетов поисковых системах: найденные слова, употребленные в пределах одного предложения или его сложноподчиненной части могут иметь более высокий ранг по сравнению с теми же словами, употребленными в соседних предложениях.

и др.

Приведенные выше примеры требуют уточнения касательно того, что конкретно мы называем предложением. Речь идёт не о традиционном определении предложения, подразумевающим грамматически организованную единицу речи, обладающую смысловой и интонационной законченностью. Предложениями мы считаем более широкий класс единиц речи, включающий не столько «законченные» элементы, сколько сочетания слов, наиболее часто встречающиеся в реальных коллекциях документов, которые, в зависимости от задачи, удобно рассматривать как единое целое. Ясно, что подобные фрагменты далеко не всегда обладают смысловой полнотой, впрочем, иногда они могут обладать и некоторой избыточностью. Например, отдельный элемент библиографического списка удобнее рассматривать как целую единицу текста, не смотря на наличие в нём формально корректных концов предложения. Другим примером может быть заголовок таблицы, который также является целой единицей текста, хотя и заканчивается двоеточием.

На практике исходная задача разделяется на две подзадачи. Первая подзадача заключается в том, чтобы выяснить, является ли знак концом предложения. Вторая подзадача (см., например, [2]), гораздо более сложная как в реализации, так и вычислительно, подразумевает определение мест в тексте, где делитель

предложения был пропущен, например, по ошибке. К счастью, в решении второй подзадачи, как правило, нет острой необходимости, поскольку «подразумеваемые» концы предложений встречаются существенно реже, чем явные.

Как следует из данного выше определения, концы предложений могут обозначаться точками, знаками вопроса и восклицания, символами конца абзаца, многоточиями и, в отдельных случаях, двоеточиями. Основные проблемы создаёт символ точки, поскольку, он также используется в сокращениях и литеральных обозначениях — датах, шифрах, адресах электронной почты, web-адресах и т. п.

Для определения, является ли точка в данном контексте концом предложения, применяются различные методы, сложность которых в целом зависит от типа входных данных и требований к качеству результата. В силу своей простоты, наиболее распространены способы, предполагающие точное постулирование понятий начала и конца предложения (см., например, [6]), а также применение таблиц стандартных сокращений (в том числе генерируемых автоматически — см. [1]), или регулярных выражений — практика, получившая широкое распространение в среде разработчиков ПО. Подобные способы зачастую показывают хорошие результаты, но в достаточно специфичных случаях, на которые они рассчитаны. К примеру, для разметки газетных корпусов весьма эффективным оказывается метод автоматического распознавания аббревиатур [1]. Отметим, что для выполнения машинного обучения авторам метода потребовалось проанализировать весьма объемные текстовые корпусы (до 1 млн. слов для английского языка).

Недостатком непосредственного использования таблиц сокращений в качестве достаточного признака является неприменимость к случаям нестандартных (авторских) сокращений, кроме того, стандартные сокращения могут находиться в конце предложения, а некоторые сокращения, например, «и т. п.», «и др.» часто обозначают конец предложения явно (что примечательно, данное предложение является исключением).

Очевидно, что информация о том, является ли знак концом предложения, хранится в его контексте. Также очевидно, что эта информация выводится из контекста по неким определенным правилам. Сформулировать эти правила, основываясь только на собственном опыте, оказывается не так уж просто, особенно с учётом того, что их требуется представить в форме алгоритма. Тем не менее, оказывается, что можно автоматически вывести правила, позволяющие различать знаки концов предложений более чем с 99 % точностью.

Метод

Предлагаемый нами метод заключается в первоначальном создании небольшого набора базовых правил (порядка 40 штук) и последующем автоматическом построении классификатора, опирающегося на результаты применения этих правил. Базовые правила делятся на два типа — подстановки и комбинации. Подстановки, в общем случае, задаются регулярными выражениями и проверяют конкретные простые признаки контекста такие как, например, «пробелы справа» или «одна прописная буква слева» и т. п. Комбинации являются

алгебраическими конструкциями, строящимися из подстановок, например, «прописная буква слева» + «титул справа» (титул — слово, начинающееся с прописной буквы). Каждое из базовых правил, будучи примененным к конкретному контексту, возвращает число начисленных очков, обычно это 0, -1 или +1. Конечный результат вычисляется из вектора очков посредством специального классификатора, обученного на заранее размеченном наборе текстов.

Обучение классификатора

Для построения классификатора нами был применен алгоритм машинного обучения на основе деревьев принятия решений. Начальным этапом обучения классификатора стала подготовка первичной выборки — сравнительно небольшого текста, порядка 1000 предложений, содержащего значительное количество «трудных» случаев: стандартных и авторских сокращений, как в середине, так и в конце предложения, специальных обозначений, дат, web-адресов, адресов электронной почты и т. п.

Выборка была подготовлена таким образом, чтобы простейший алгоритм (baseline), считающий делителями все точки, вопросительные и восклицательные знаки, ошибался по возможности в максимальном количестве случаев. Всего в подготовленном подготовленном тексте было 3220 потенциальных делителей, т. е. знаки являющиеся действительными концами предложений (согласно нашему определению) составляли около 30 %. Таким образом, на первичной выборке baseline ошибался в 70 % случаев.

Обученный на первичной выборке классификатор затем последовательно применялся к дополнительно отобранным документам различных типов — web-страницам, художественным и формальным текстам. Случаи, когда классификатор ошибался или «сомневался» поступали частично в обучающий и частично в проверочный наборы следующей итерации, после чего производилось переобучение. Целью этих итераций была оценка предельно возможной точности метода, которая по результатам тестирования составила 98,7 ± 0,4%. Окончательная проверка производилась вручную на 10 случайных web-документах, содержащих в сумме 2681 эпизод, из которых 28 оказались ошибочно не разделенными и 7 ошибочно разделенными. Всего потребовалось 6 итераций переобучения, за которые была накоплена финальная обучающая выборка порядка 10 тыс. эпизодов с отношением количества знаков-делителей предложений к знакам, не являющимся делителями, как 2 к 3. Попытки продолжать обучение точность не повышали, что связано, как выяснилось, с существованием случаев, когда знак препинания должен быть классифицируем как конец или не конец предложения, исходя из семантических признаков, т. е. требующих смысловой идентификации элементов предложения (см. пример 1).

***Пример 1**. Случай, когда точка является концом предложения по семантическим признакам.*
«Описание модели см. в А21. К-2301 т. IV, стр. 45. С 2003 г. изменена номенклатура.»

Принятие решения в примере 1 затруднено, поскольку в контексте каждой точки находятся последовательности, которые с точки зрения набора правил могут являться как стандартным или авторским сокращением, так и частью шифра. В подобных случаях разбиение предложения нашим классификатором, как правило, не происходит.

Сравнение с существующими методами

Для получения сравнительных оценок точности и производительности нашего метода мы воспользовались двумя открытыми разработками, предлагающими свои решения задачи определения концов предложений — модулем графематического анализа из проекта «АОТ» [6], основанном на эвристике, и компонентом Sentence Boundary Detector из проекта OpenNLP, использующим принцип максимальной энтропии [7]. Последний компонент мы обучили на той же финальной выборке, по которой обучался наш классификатор. В качестве тестовой коллекции была использована коллекция документов, полученных с новостных и аналитических сайтов, таких как news.mail.ru, lenta.ru, rian.ru, rbc.ru, wciom.ru и др., содержащая большое количество инициалов, сокращений, дат и других обозначений, включающих символы пунктуации. Для получения текстового содержания статей выкачивались в основном препринты (версии документов с минимальной разметкой, подходящей для печати). Очистка документов от включений тэгов и HTML-сущностей производилась посредством html-парсера, используемого в индексаторе нашей поисковой системы. Поскольку графематический модуль проекта «АОТ» все переносы строк однозначно считает концами предложений (абзацами), все переносы строк из документов тестовой коллекции были заменены пробелами. Результаты приведены в таб. 1.

Таблица 1. Сравнительные характеристики методов выделения концов предложений

	OpenNLP / Sentence Boundary Detector.	**Графематический модуль проекта «АОТ».**	**Наш классификатор.**
Алгоритм:	принцип максимальной энтропии	эвристический	набор правил и дерево принятия решений
Число знаков в обучающей выборке:	9820	Обучение не производилось.	9820 (та же выборка)
Число потенциальных делителей в тестовой выборке:	Всего: 3087 вхождений: « . » — 2980 вх., « ? » — 37 вх., « ! » — 70 вх.		

[1] Указанное время не должно расцениваться как показатель производительности графематического анализатора, входящего в состав АОТ, поскольку использованный нами компонент помимо графематического анализа также производил синтаксическую разметку входного текста.

	OpenNLP / Sentence Boundary Detector.		Графематический модуль проекта «АОТ».		Наш классификатор.	
Общее число знаков, признанных делителями:	2760		2522		2069	
Размер случайной выборки для ручной проверки:	500 элементов					
Время выполнения:	0,42 с		12,1 с[1]		0,72 с	
	Верные:	**Неверные:**	**Верные:**	**Неверные:**	**Верные:**	**Неверные:**
Число разбиений:	339	161	386	114	499	1
Число слияний:	455	45	462	38	497	3
Общий процент ошибок (*Error rate*):	41,2%		30,4%		0,8%	
Точность (*Precision*):	0,678		0,772		0,998	
Полнота (*Recall)*:	0,883		0,910		0,994	
F-мера Ван Ризбергена:	0,767		0,835		0,996	

Отметим, что невысокая точность метода, использующего принцип максимальной энтропии, в целом обусловлена небольшим объемом обучающей выборки и может быть значительно улучшена.

О способе описания правил классификатора

Подбор правил производился вручную, для чего использовался специально разработанный декларативный язык. Каждое декларируемое правило снабжается уникальным именем, посредством которого на него могут ссылаться другие правила. При объявлении указывается класс правила, задающий способ вычисления, и, собственно, вычисляемая формула. Классом правила может являться регулярное выражение, алгебраическая конструкция и сам классификатор — объект, опрашивающий ассессора или получающий информацию о графематической разметке из файла.

Пример 2. Объявление правила, распознающего сокращения «т. к.», «т. е.», «т. н.».

```
ABBR3a_L := RegEx(L) <- $(open)т\z
ABBR3a_R := RegEx(R) <- \A$(s)*[енк]\.
ABBR3b_L := RegEx(L) <- $(open)т\.$(s)*[енк]\z
ABBR3 := Rule() <- (ABBR3a_L & ABBR3a_R ) | ABBR3b
```

В представленном выше примере объявлены четыре правила, три из которых являются регулярными выражениями, причем два применяются к левому контексту (класс RegEx(L)) и одно к правому (класс RegEx(R)). Последнее

правило выражает логическую связь между остальными правилами посредством операторов «&» (логическое И) и «|» (логическое ИЛИ).

Производительность и оптимизация

Исследование производительности алгоритма показало, что основная нагрузка приходится на обслуживание регулярных выражений, что не удивительно, поскольку большинство правил относятся к этому типу. В общем случае, если применять регулярные выражения ко всему левому и правому контексту, то сложность алгоритма получается квадратичной. В особенности это касается левого контекста, поскольку в этом случае затруднена привязка регулярного выражения к концу проверяемого диапазона посредством соответствующего нетерминала (\z).

Исследование возможности задать максимально допустимый диапазон для проверки показало (см. Рис. 1), что для русскоязычных текстов достаточно 28 символов слева и 16 справа (при этих ограничениях на нашем наборе правил не происходила потеря точности).

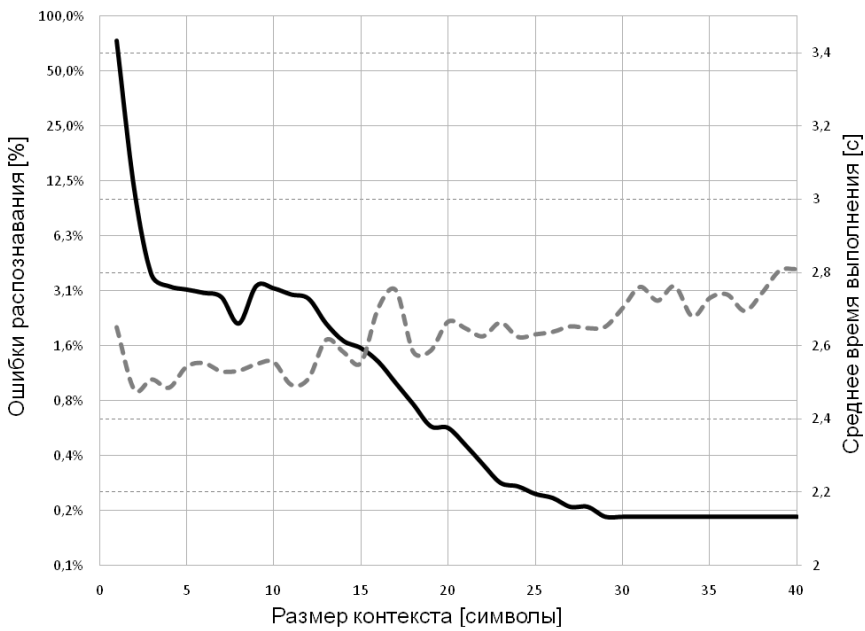


Рис. 1. Зависимость точности и скорости от размера контекста (для 50 тыс. эпизодов)

Для повышения производительности также были применены дополнительные приёмы:

1. По набору правил автоматически сгенерирован эквивалентный программный код, снимающий вычислительную избыточность, связанную с обращениями к правилам по имени, а также виртуальными вычислениями. Эффект +30% производительности.
2. Вычисление не всех правил, а лишь тех, которые необходимы классификатору (по обратному запросу). Эффект +25% производительности.
3. Кэширование значений вычисленных правил на период разбора одного эпизода. Эффект +5% производительности.

В сумме применение всех перечисленных приёмов позволило получить приемлемую скорость для процесса индексирования в нашей поисковой системе.

Наиболее значимые правила

Значимость отдельных правил определяется автоматически в процессе обучения. Грубо говоря, чем выше степень корреляции оценки, получаемой некоторым правилом, с требуемым ответом, тем выше это правило будет в списке опроса, и тем чаще его выполнение будет проверяться. По итогам проведенных экспериментов, к наиболее значимым правилам относятся

1. «тип разделителя», 3 правила — определяют соответственно точку, знак вопроса и знак восклицания;
2. «пробелы справа/слева» — определяют наличие пробельных символов справа и слева от разделителя;
3. «символ пунктуации справа/слева»;
4. «цифра справа/слева»;
5. «прописная/строчная буква справа/слева»;
6. «открывающая/закрывающая скобка справа/слева»
7. «стандартные сокращения»;
8. «неизвестные сокращения общего вида ххх.-хх. хх.» и т. д.

Об алгоритме построения классификатора

В качестве основы классификатора мы использовали деревья принятия решений, широко применяемые в машинном обучении [8, 9, 10]. Пример реального дерева принятия решения из нашего классификатора показан на Рис. 2 (мы приводим фрагмент, так как всё дерево слишком велико). В узлах дерева находятся условия проверки выполнения правил. Запись вида «20 < 0.5» означает, что если правило №20 набрало менее 0.5 очков, то нужно перемещаться в левое поддерево, а в противном случае в правое поддерево. В листьях дерева хранятся оценки вероятности того, что данный знак препинания является концом предложения (чем больше значение, тем выше вероятность).

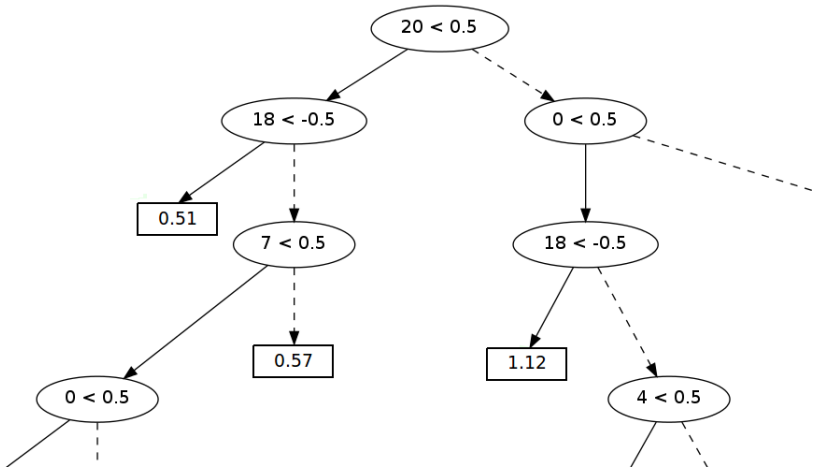


Рис. 2. Фрагмент дерева принятия решений

В приведенном примере (см. Рис. 2) классификатор принимает решение, обращаясь к правилам, перечисляемым ниже:

1. №0 — разделитель является точкой;
2. №4 — пробелы справа;
3. №7 — прописная буква слева;
4. №18 — многоточие (+2 для первой точки, +1 для средних, -1 для последней);
5. №20 — титул справа (титул — слово, записанное с прописной буквы).

К сожалению, одно дерево не может обеспечить приемлемое качество классификации. Для того, чтобы решить эту проблему, в качестве классификатора мы используем комбинацию нескольких десятков относительно простых деревьев. Каждое дерево строится так, чтобы максимально скомпенсировать ошибку работы остальных деревьев классификатора (такой подход известен в специальной литературе под названием boosting — см. [11, 12]). Дополнительно, для улучшения качества работы классификатора в целом мы использовали технику bootstrapping, которая заключается в том, что для построения каждого дерева используется случайное подмножество обучающего множества, меняющееся от дерева к дереву. В нашем случае было автоматически построено 540 деревьев, каждое из которых содержит до 8 уровней и от 300 до 350 листьев. Полученный таким образом композитный классификатор показывает значительно более высокую точность, чем отдельно взятое дерево.

Пример разбиения

В таб. 2 приведены оценки, сделанные классификатором для части web-документа. Порог отсечения (равный 0,64) выбирался методом наименьших

квадратов для оценок, полученных по обучающей выборке. Смещение порога объясняется несимметричностью обучающей выборки относительно числа случаев, когда знак является разделителем и когда не является.

Таблица 2. Пример вывода размечающей программы для «сложного» web-документа

Левый контекст:		**Правый контекст:**	**Оценка (0 — не конец предложения, 1 — конец).**	
/римером в круглых скобках	.	##Нумерация рисунков такж/	Да	1.000000
/я рисунков также сквозная	.	Подписи под рисунками до/	Да	1.000000
/лжны начинаться так: «Рис	.	<Номер рисунка> <Названи/	**Нет**	**1.004000 (ошибка)**
/унка>» и набраны курсивом	.	##Подзаголовки должны быт/	Да	1.000000
/браны полужирным курсивом	.	При делении на разделы р/	Да	1.000000
/разделов и подразделов (1	.	1., 1.2., 1.3, 2.1 и т.д./	Нет	0.001357
/зделов и подразделов (1.1	.	, 1.2., 1.3, 2.1 и т.д.)./	Нет	0.020330
/ов и подразделов (1.1., 1	.	2., 1.3, 2.1 и т.д.).##Сн/	Нет	-0.000760
/ и подразделов (1.1., 1.2	.	, 1.3, 2.1 и т.д.).##Снос/	Нет	0.006150
/одразделов (1.1., 1.2., 1	.	3, 2.1 и т.д.).##Сноски т/	Нет	-0.000760
/делов (1.1., 1.2., 1.3, 2	.	1 и т.д.).##Сноски также /	Нет	-0.000760
/(1.1., 1.2., 1.3, 2.1 и т	.	д.).##Сноски также должны/	Нет	0.051560
/.1., 1.2., 1.3, 2.1 и т.д	.	).##Сноски также должны б/	Нет	-0.006359
/., 1.2., 1.3, 2.1 и т.д.)	.	##Сноски также должны быт/	Да	1.000000
/и помещены внизу страницы	.	##Список литературы оформ/	Да	1.000000
/через запятую, год, точка	.	Затем указываются том (Т/	Да	1.000000
/ Затем указываются том (Т	.	), номер издания (№), стр/	Нет	-0.011740
/ издания (№), страницы (С	.	).##Пример оформления:##1/	Нет	-0.011740
/здания (№), страницы (С.)	.	##Пример оформления:##1. /	Да	1.000000
/).##Пример оформления:##1	.	Крейдлин Г. Е. Невербальн/	**Нет**	**0.498900**
/формления:##1. Крейдлин Г	.	Е. Невербальная семиотика/	Нет	-0.003087
/рмления:##1. Крейдлин Г.Е	.	Невербальная семиотика //	Да	1.000000
/вербальная семиотика // М	.	: Новое литературное обоз/	Нет	0.291300
/ературное обозрение, 2002	.	#2. Якобсон Р. О. О лингви/	Да	0.661100
/турное обозрение, 2002.#2	.	Якобсон Р. О. О лингвисти/	**Нет**	**0.498900**
/рение, 2002.#2. Якобсон Р	.	О. О лингвистических аспе/	Нет	0.001101
/ние, 2002.#2. Якобсон Р.О	.	О лингвистических аспект/	Да	1.000000
/ в зарубежной лингвистике	.	М.: 1978. С.16–24.#Остал/	Да	1.000000
/зарубежной лингвистике. М	.	: 1978. С.16–24.#Остальны/	Нет	-0.008412
/ной лингвистике. М.: 1978	.	С.16–24.#Остальные парам/	**Нет**	**0.481700**
/ лингвистике. М.: 1978. С	.	16–24.#Остальные параметр/	Нет	0.002148
/истике. М.: 1978. С.16–24	.	#Остальные параметры текс/	**Да**	**0.498900**
/, указанными верстальщику	.	#/	Да	1.000000

Заключение

Примененный нами подход показывает, что совмещение эффективного инструмента для ручного подбора правил с автоматическим классификатором (в частности, на основе деревьев принятия решений) позволяет использовать сравнительно небольшие тренировочные наборы, что, в свою очередь, не только экономит время работы асессоров, но и позволяет сделать их работу менее монотонной. После достижения необходимого уровня точности, метод, в зависимости от реализации, допускает существенную оптимизацию по скорости (в нашем случае до 50 %). В случаях, если скорость работы классификатора недостаточно высока для конкретной задачи, с его помощью можно обучать более быстрые модели, требующие, как правило, существенно бо́льших обучающих выборок.

References

1. *Berger A. L., Della Pietra V. J.* 1996. A Maximum Entropy Approach to Natural Language Processing. Computational Linguistics, 22.
2. *Bishop C.* 2006. Pattern Recognition and Machine Learning.
3. *Breiman L., Friedman J. H., Olshen R., Stone C. J.* 1984. Classification and Regression Tree.
4. *Friedman J. H.* 1999. Stochastic Gradient Boosting.
5. *Stevenson M., Gaizauskas R.* 2000. Experiments on Sentence Boundary Detection. Proceedings of the Sixth Conference on Applied Natural Language Processing and First Conference of the North American Chapter of the Association for Computational Linguistics.
6. *Hastie T., Tibshirani R., Friedman J. H.* The Elements of Statistical Learning. Second Edition.
7. *Kiss T., Strunk J.* 2006. Unsupervised Multilingual Sentence Boundary Detection. Computational Linguistics, 32 (4).
8. *Kobritsov B. P., Liashevskaia O. N., Shemanaeva O. Iu.* 2005. Superficial Filters for Semantic Omonymy Settlement in the Text Corpus [Poverkhnostnye Fil'try dla Razresheniia Semanticheskoi Omonimii v Tekstovom Korpuse]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2005" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2005").
9. *Petrovskii M. I., Glazkova V. V.* 2007. Machine Learning Algorythms for Electronic Documents Analysis and Rubrication Target [Algoritmy Mashinnogo Obucheniia dla Zadachi Analiza I Rubrukatsii Elektronnykh Dokumentov]. Vychislitel'nye Metody i Programmirovanie, 8.
10. *Sokirko A.* Description of the Graphematic Unit of the Project "AOT" [Opisanie Grafematicheskogo Modulia Proekta "AOT"], available at: http://www.aot.ru/docs/graphan.html

11. *Zelenkov Iu. G., Segalovich Iu. A., Titov V. A.* 2005. Probabilistic Model of Morphological Omonymy Elimination on the Base of Normalizing Substitutions and Adjacent Words Positions [Veroiatnostnaia Model' Sniatiia Morfologicheskoi Omonimii na Osnove Normalizuiushchikh Podstanovok I Pozitsii Sosednikh Slov]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2005" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2005").

КОНСТРУКЦИИ С АБСТРАКТНЫМИ СУЩЕСТВИТЕЛЬНЫМИ И ИХ ОТРАЖЕНИЕ В ЭЛЕКТРОННОМ СЛОВАРЕ*

Г. И. Кустова (galinak03@gmail.com)

Московский Педагогический Государственный Университет, Москва, Россия

В статье представлен проект создания электронного словаря конструкций с абстрактными существительными и обосновываются основные параметры их описания: структурные типы (предложный оборот, предложно-атрибутивный оборот, производный предлог), синтаксические функции, семантические типы, парадигматические связи.

Ключевые слова: абстрактные существительные, электронный словарь, параметры описания, словарь конструкций.

CONSTRUCTIONS WITH ABSTRACT NOUNS IN AN ELECTRONIC DATABASE

G. I. Kustova (galinak03@gmail.com)

Moscow State Pedagogical University, Moscow, Russian Federation

The paper discusses the types of abstract noun constructions and the types of information in an electronic dictionary (lexical database). The electronic dictionary includes “non-nominative” items which are used as predicates (e. g. X в плену, в обмороке, в отчаянии, на тренировке, под арестом), as sentence modifi ers (В заключении он научился шить рукавицы), as adverbial modifi ers (спрыгнул на ходу; ушел со службы под предлогом болезни), as parentheses (во всяком случае, он нам ничего не обещал). The electronic dictionary includes such types of information on abstract noun constructions as the formal structure, the syntactic function, and the semantic type.

Key words: abstract nouns, electronic database, parameters, constructions database.

* Работа выполнена при поддержке РГНФ, проект № 11-04-00223а.

В работе речь пойдет о типах и свойствах конструкций с непредметными существительными, построенных на базе свободно присоединяемых падежных форм (*в недоумении*; *при проверке*; *по соседству* и под.). Под свободно присоединяемыми понимаются формы, которые не предсказываются предикатным словом (в первую очередь — глаголом) и не входят в модель управления (ср. понятие синтаксических и наречных падежей у Е. Куриловича [8]).

В глагольной конструкции за понимание ситуации «отвечает» глагол. Свободные конструкции, именно потому, что их значение определяется за счет «наречной» семантики падежа или семантики предлога, представляют собой семантически самостоятельные единицы, «семантический» тип которых можно определить и без глагола, а иногда и вне предложения. Например, в выражении *прийти к выводу* предложная группа *к выводу* не считается конструкцией, т. к. обусловлена моделью управления глагола. Эта предложная группа не интерпретируется вне контекста управляющего глагола, ср. *пришел к выводу, привык к выводу, обратился к выводу, с недоверием отнесся к выводу* и т. д. А *к утру* или *к нашему удивлению* — конструкции (свободно присоединяемые формы), которые могут быть поняты вне контекста и могут употребляться в разных функциях и позициях. Свободно присоединяемыми могут быть и беспредложные падежи, которые превращаются в обстоятельственные обороты (*задним числом*), наречия (*боком*), отыменные предлоги (*путем чего*, *типа чего*). Далее речь пойдет, в основном, о предложных формах (под предложной формой понимается конструкция «предлог + существительное в косвенном падеже»).

Формы существительного входят в парадигму, где они считаются семантически тождественными. На самом деле это единство парадигмы и семантическое тождество ее членов достаточно иллюзорно. Точнее — это факт словаря. В реальных текстах семантические и синтаксические особенности разных падежных и особенно предложных форм меняются в зависимости от связей с другими единицами и позиции в предложении (наглядный пример — массовая «миграция» формы творительного падежа существительного и ряда предложных форм в наречия).

Если говорить об абстрактных существительных, то в предложных конструкциях они постепенно утрачивают субстантивные, именные свойства и превращаются во что-то другое. Результаты этого процесса могут существенно различаться (в силу разных причин — семантики предлога, семантического класса существительного, типа конструкции, позиции в предложении) — одни формы превращаются в настоящие наречия: *на боку*, *напоказ*, *враз*; вторые являются обстоятельственными характеристиками других ситуаций: ***В тишине*** *слышен каждый шорох*; ***В темноте*** *ничего нельзя найти*; третьи являются «свернутыми» ситуациями, редуцированными предикациями: ***В плену*** *он познакомился со своей будущей женой — Когда был в плену…*; ***При желании*** *можно подать заявку через Интернет — Если захочешь…*. При этом одна и та же конструкция может употребляться в разных позициях: *Весь день мы провели* ***в ожидании*** VS. ***В ожидании*** *автобуса пассажиры все время выбегали на проезжую часть*; обороты с одним и тем же предлогом могут иметь разное значение: *Лежал / стоял* ***в бесчувствии*** (= ‘состояние’) VS. *Упал* ***в бесчувствии*** (=

‘причина’); ***Со временем*** *у него не очень* (‘у него мало времени’) VS. ***Со временем*** *у него пропал интерес к шахматам* (‘постепенно’).

Поскольку семантические и грамматические изменения абстрактных существительных происходят по определенным направлениям и подчиняются определенным закономерностям, количество типов конструкций, типов «превращений» не столь уж велико. Так же, как бывает несколько агрегатных состояний вещества, причем одно и то же вещество может встречаться в разных состояниях (например, лед — вода — пар), так же бывает вполне обозримое количество «агрегатных состояний», языковых статусов форм существительных.

С точки зрения структуры выделяются: предложный оборот, ср. *в плену, в тоске, с удовольствием, по соседству, на слух* и т. д., предложно-атрибутивный оборот, ср. *в полной уверенности, по первому требованию, на всякий случай* (сразу отметим, что прилагательное может входить и в «обычный» предложный оборот, ср. *с большим удовольствием*, но при этом оно факультативно, а в атрибутивных оборотах прилагательное, по тем или иным причинам, необходимо: либо существительное неполнозначное, например, параметрическое, ср. *вещества* ***с большим содержанием*** *жиров* — *?с содержанием жиров*; *город* ***в праздничном убранстве*** — *?город в убранстве*; либо прилагательное (или местоимение) выражает валентность, ср. ***по вашему приказанию*** *прибыл* — *?по приказанию прибыл* (данная валентность, конечно, может быть выражена и иначе: *по приказанию начальника*); либо в силу полной фразеологизации, ср. *по большому счету, из первых рук*); производный предлог, ср. *по аналогии с чем, в форме чего, под предлогом чего*. Собственно фразеологизмы, т. е. идиомы, могут рассматриваться наряду с остальными конструкциями, если они построены по тем же грамматическим моделям, ср. идиому *из первых рук* и конструкцию *из достоверных источников*.

У предложных оборотов наиболее широкий набор синтаксических функций:

— предикат: *Наш начальник —* ***с принципами***; *И сама* ***в порядке****, и дети* ***под присмотром***;

— детерминант: ***С принципами*** *прожить нелегко*; ***Под присмотром дядюшки*** *Писанли мы разрабатывали его* [номер] *и репетировали*. [А. Н. Толстой. Рукопись, найденная под кроватью (1923–1924)][*];

— глагольный обстоятельственный распространитель: ***Из принципа*** *не пойду*; *Пока договорились* ***в принципе****, детали обсудим потом*; *Пошел в армию* ***по убеждению***; *Детям разрешается купаться только* ***под присмотром родителей***;

— приименный распространитель: *человек* ***с убеждениями***, *люди* ***с недостатками***;

— вводные конструкции: ***В принципе****, неплохо было бы заранее все выяснить*; ***По нашему убеждению****, этот план нуждается в пересмотре*.

[*] Все литературные примеры взяты из Национального корпус русского языка и далее могут приводиться в сокращении и без ссылки.

Предложно-адъективные обороты, ср. *по какому принципу*, *по этому принципу*, и производные предлоги (с зависимыми), ср. *по принципу чего* (*по принципу эквивалентности*), употребляются, в основном, в обстоятельственной функции.

Наибольшее внимание обсуждаемым единицам (для краткости далее будем называть их оборотами или конструкциями) уделялось в тех языковых теориях, в которых важное место отводится фразеологическому компоненту языка и процессам фразеологизации. В первую очередь необходимо упомянуть созданную в рамках модели «Смысл ⇔ Текст» (см. [11]) классификацию И. А. Мельчука, в которой фразеологические единицы (фраземы) делятся на идиомы, коллокации (полуфраземы) и квазифраземы, а кроме того, выделяются прагматемы и синтаксические фраземы (см. [20] и, особенно, [7]). В работах Московской семантической школы (Ю. Д. Апресян, И. М. Богуславский, Л. Л. Иомдин и др.), также тесно связанной исторически и идеологически с теорией «Смысл ⇔ Текст», большое место занимают исследования языковых свойств коллокаций и конструкций малого синтаксиса — синтаксических фразем (ср. *все равно*), фразеосхем (ср. *в Х-овую силу* — например, *в полную силу*), конструкций с производными предлогами и др. (см. [1], [4], [6]). С развитием разных направлений когнитивной лингвистики подобного рода единицы и конструкции стали одним из главных предметов описания в грамматике конструкций, получившей теоретическое обоснование в работах Ч. Филлмора и его последователей (см., в частности, [16], [17], [18], [19]). Впрочем, надо заметить, что отдельные идеи, касающиеся связи значения и конструкции, высказывались еще в работах классиков отечественной лингвистики (ср. идею конструктивно обусловленных значений у В. В. Виноградова или понятие фразеосхемы у Д. Н. Шмелева).

Обсуждаемые здесь конструкции образуются на пересечении процессов грамматикализации и фразеологизации и представляют интерес не только для грамматики, но и для лексикографии.

Превращение исходных «номинативных» единиц в предикатные и служебные, связанное с утратой первоначальных признаков и приобретением новых, может проходить разные ступени (фазы) и двигаться по разным линиям. Каждая предложная конструкция является особого рода единицей и обладает особыми грамматическими свойствами и особой семантикой по сравнению с исходной «номинативной» единицей. Множество этих свойств имеет индивидуальный («словарный») характер и требует словарной фиксации. Приведем несколько примеров в качестве иллюстрации этого тезиса.

Разные существительные образуют конструкции отнюдь не с любыми, а с совершенно определенными предлогами. Например, для эмоциональных существительных характерны конструкции «*В* + предл. п.» и «*С* + твор. п.»: *в беспокойстве* (*Страдает, **в беспокойстве** мечется по сцене моя Амнерис* [И. А. Архипова. Музыка жизни (1996)], *Нынче трикотажники **в беспокойстве**: у коз последнее время что-то не очень с подшерстком*. [«Домовой», 2002.11.04]) и *с беспокойством* (*Она заметила мрачную неподвижность Нади и поглядывала на неё **с беспокойством***. [А. Солженицын. В круге первом (1968)]); ср. также *в страхе* и *со страхом*, *в печали* и *с печалью*, *в ужасе* и *с ужасом* и т. д.; существительные со значением

физиологического состояния допускают *B*-конструкцию, но не *C*-конструкцию: *в обмороке, в бреду, в истощении, в сознании*, но не **с обмороком*, **с бредом*, **с сознанием* (**с сознанием подписывал бумаги*), — если же такие конструкции есть, то они не соотносительны с *B*-конструкциями или существительное в них имеет другое значение (*попал в больницу с истощением*; *с сознанием собственного превосходства разглядывал собравшихся*). Для существительного *память* возможны конструкции *по памяти*, *на память*, *для памяти*, *на* (*чьей*) *памяти*, *в* (*чьей*) *памяти* (а также, в другом значении, *ради памяти кого*, *в память о ком / чем*), но не *с памятью* или *при памяти*; ср. также: *польза — в пользу кого / чего* (или: *в чью пользу*), *на пользу*, *с пользой*, *без пользы* — но не *в пользе*; *порядок — в порядке*, *по порядку*, *для порядка*, *в порядке чего*, *на порядок* [*больше*, *лучше* и т. д.], *в* (*каком*) *порядке*, *без* (*какого*) *порядка* — но не *с порядком*, и т. д.

Одна и та же конструкция возможна не в любой синтаксической функции даже с близкими по семантике существительными, например, в позиции детерминанта обороты *в тревоге* и *в надежде* (*на что / инф*) возможны оба, ср.: ***В тревоге*** *вы срываетесь с работы, но, оказывается, все прошло…* [«Семейный доктор», 2002.05.15], ***В надежде обратить*** *всё в шутку он улыбнулся*. [А. Азольский. Лопушок // «Новый Мир», № 8, 1998], а в позиции предиката — только *в тревоге*, ср.: *Он в тревоге* VS. **Он в надежде* (ср., однако, другие ментальные состояния: *Он в раздумьях / в затруднении / в сомнении* и *в сомнениях*). То же относится и к существительным других семантических классов — далеко не все они могут выступать как сказуемые: *Он в плену*; *Мы в окружении*; *Он на занятии*; *Он давно уже на пенсии*, но не **Он в погоне*; **Он на преступлении*. Возможные для некоторого существительного конструкции и позиции — это словарная информация, однако в обычном словаре ее некуда поместить: там все формы «равны», и нет опции для пояснений и комментариев к каждой в отдельности.

У некоторых оборотов развиваются типичные глагольные значения, которые не могут фиксироваться в обычном («номинативном») словаре, т. к. существительное не обладает этим значением вне данной конструкции. Так, значение конструкции *в отъезде* основано на импликации: ‘X уехал’ → ‘X-а нет, X отсутствует’. Это значение состояния, которое наступило в результате события и занимает определенный интервал, т. е. значение перфекта, ср. *Начальник в отъезде, так что прием посетителей отменили*. В обычном «номинативном» словаре у слова *отъезд*, разумеется, нет такого значения и оно там не может быть отражено, поскольку *отъезд* — это само отправление в путь, и в этом смысле отъезд продолжается только до тех пор, пока человек еще не уехал, а не после того, как он уехал (ср.: *При отъезде каждому депутату вручили по ящику самого знаменитого тираспольского коньяка марки «Суворов»*). Аналогично в случаях *в отставке*, *под арестом*: эти существительные обозначают события, мероприятия, и значение состояния приобретают только в конструкции — точнее, значение состояния имеет сама конструкция (справедливости ради надо отметить, что в НКРЯ один раз все-таки встретилось выражение *двухсуточный арест*, т. е. у «номинативного» *ареста* тоже есть возможность развивать перфектное значение; у *отъезда*, однако, такой особенности нет: *Его отъезд длился сутки* значит ‘сутки пытался и не мог уехать’, а не ‘сутки отсутствовал’).

В-третьих, важно также отметить, что в наречных конструкциях имена с исходно предметной семантикой приобретают абстрактную семантику, ср. ***на глазах*** *у всех* (‘на виду’); *не говори / не попадайся мне* ***под руку***; *чтобы Х всегда был* ***под рукой*** (‘доступен’).

Подобные сведения должны получать лексикографическую фиксацию. Возникает вопрос: где и в какой форме?

Традиционные «номинативные» словари плохо приспособлены для отражения оборотов — не только потому, что принцип номинативных «входов» не предполагает отдельную фиксацию предложных конструкций (например, *ошибочно* как однословный элемент является отдельным входом, а синонимичное *по ошибке* нет, т. к. это и не наречие в полном смысле и не номинатив), но и потому, что сама идеология словаря «начальных форм» плохо совместима с «неноминативными» конструкциями (см. [9], [10]). В результате в существующих толковых словарях такие конструкции представлены неполно, не говоря уже о неудобствах поиска.

Какая-то часть оборотов находит отражение во фразеологических словарях, но и здесь те же проблемы — проблема неполноты и проблема поиска. Так, в словаре [15] при слове *страх* даны обороты *страх какой*, *страх сколько*, *в страхе* и *под страхом чего*, но нет *со страху*, *от страха* и *из страха* (а это коллокации, или полуфраземы, по И. А. Мельчуку, т. е. они должны фиксироваться в словаре). В словаре [3] есть обороты *на свой страх и риск*; *не за страх, а за совесть*; *без страха и упрека*; *страха ради*. В словаре [14] нет оборота *без страха* (хотя есть *без смеха*). Наконец, ни в одном из упомянутых словарей нет оборота *на страх кому* (ср. *на страх врагам*; *<…> первые попытки осознать новейшую историю России не как всплеск безумия* ***на страх остальному миру****, а как этап самопостижения*. [Лев Аннинский. Десять лет, которые растрясли мир (1999)]; *<…> заборы с гвоздями, длинными остриями торчащими* ***на страх ворам и разбойникам***. [И. Е. Репин. Далекое близкое (1912–1917)]). Между тем «*на* + сущ. с эмоциональным значением» — это фразеологизованный оборот (заметим в скобках, что, например, *назло* уже пишется слитно и считается наречием). Нельзя считать, что это обычная целевая конструкция с предлогом *на* типа *пойти на прогулку* или *отдать на воспитание*, т. к. здесь нарушается нормальное для целевой конструкции требование контролируемости: эмоциональные состояния страх, радость и под. — неконтролируемые ситуации. Кроме того: (а) многие эмоциональные существительные в такой конструкции не употребляются, хотя семантически вполне подходят: **на спокойствие*, **на ужас*, **на умиление* и др.; (б) конструкции с существительными, которые употребляются, могут быть устроены по-разному, например, *на радость* отличается от *на счастье*: *на радость* обычно употребляется в значении ‘радуя Х-а’, ср. *Кубок Президента Республики Башкортостан на удивление многим и* ***на радость*** *уфимским* ***болельщикам*** *выиграл «Салават Юлаев»*. [Хоккей-2 (форум) (2005)], *на счастье* обычно употребляется в значении ‘чтобы Р; для Р’, ср. *Во Франции один ландыш (буквально!) дарят 1 мая* ***на счастье***. [Сати Спивакова. Не всё (2002)], или в значении ‘к счастью для кого’, ср. *Но,* ***на счастье Дубова****, выстрела не произошло*. [«Ежедневные новости» (Владивосток),

2003.01.17] (а близкий к *на счастье* оборот *на удачу* чаще употребляется в значении ‘наобум’); (в) есть обороты с «одиночными» существительными, не входящими в классы (*борщ удался* ***на славу***; *поработали* ***на совесть***). Из всего перечисленного следует, что данная конструкция — объект словарной фиксации.

Наиболее полно предложные обороты представлены в словаре наречий и служебных слов (сост. В. В. Бурцева) [14], однако там они даны по алфавиту первого элемента, т. е. предлога, что неудобно для поиска по существительному и для сопоставления (например, *в страхе*, *от страха* и *со страху* находятся в разных местах словаря), и не в полном объеме (*под страхом чего* отсутствует; выражение *две большие разницы* есть, а *с разницей в* [*2 см*] — нет).

В синтаксическом словаре Г. А. Золотовой [5] приводятся именно конструкции (хотя только одного типа — синтаксемы, т. е. падежные и предложные формы), но нет списков существительных, которые в этих конструкциях участвуют. К тому же словарь Г. А. Золотовой в первую очередь дает синтаксемы с предметными и личными именами в качестве предицируемого компонента (*Над рекой туман*; *В семье — печаль*; *С ней обморок* и т. п.), а событийные имена даются во вторую очередь (и без разбиения на семантические классы).

По-видимому, самым подходящим местом для отражения обсуждаемых конструкций мог бы быть Толково-комбинаторный словарь [12] как он задумывался его разработчиками. Именно в этом словаре предусмотрены зоны, куда записываются идиомы и коллокации. Однако пока он не существует в таком объеме, чтобы отразить все подобные единицы.

Для того, чтобы иметь возможность системно анализировать процессы грамматикализации и фразеологизации и свойства участвующих в них единиц, та информация, о которой шла речь выше, — структурные типы конструкций, морфологические характеристики (например, ограничение на число), синтаксические функции, семантические типы, индивидуальные особенности, — должна быть собрана в одном месте и храниться в удобной для поиска форме. Оптимальной формой для словаря конструкций является электронная база данных, которая может отражать основные типы конструкций и основные линии, по которым происходит движение непредметных существительных в сторону грамматикализации, — независимо от того, завершился этот процесс или нет, а кроме того, включать разные виды информации (семантическую, морфологическую, синтаксическую) и обеспечивать поиск по разным признакам и параметрам. Создание такой базы предполагает решение целого ряда вопросов как технического, так и концептуального характера: какие бывают типы конструкций; есть ли корреляции между семантическим классом существительного и набором его конструкций; есть ли корреляция между типом конструкции и набором ее синтаксических функций; как на перечисленные характеристики влияет семантика существительного, семантика падежа и семантика предлога и т. п. Ниже будут затронуты некоторые из этих вопросов.

Рассмотрим один из основных классов конструкций, который в базе данных условно назван СИТУАЦИИ. К нему относятся предложные конструкции с предлогами *В* (+ предл. п.), *НА* (+ предл. п.) и *ПОД* (+ твор. п.), которые могут употребляться в функции сказуемого (а также, обычно, и в функции

детерминанта), т. е. являются редуцированным и неполноценным аналогом глагольного предиката: *в плену*, *в разведке*, *на занятии*, *под арестом* и т. п. Этот класс конструкций назван «ситуации», поскольку они могут выполнять функцию предиката в, так сказать, независимой клаузе (в отличие от них, конструкции, например, с предлогом *ПРИ* коррелируют с зависимой клаузой, ср.: *при пересадке / остановке / задержке / отправлении / регистрации* и т. д. *нажмите красную кнопку* — ‘Когда / если P, сделайте Q’). Кроме того, обсуждаемые конструкции — это особый способ представления ситуации, которая, вообще говоря, могла бы быть обозначена и глаголом. Они создают своего рода семантические модели ситуаций, образуют некоторый семантический шаблон (разумеется, в создании этого шаблона участвует и семантика предлога, и семантика существительного).

«Ситуации» — это некоторые событийные единицы, на которые членится жизненный поток, течение жизни. С помощью этих конструкций в жизненном потоке выделяются какие-то значимые события и состояния, периоды, содержательные (а не механические) части (см. [2]). Предикативные предложные конструкции обозначают не просто ситуации, а временный статус человека (ср. *под арестом*, иногда и длительный, ср. *в отставке*, *в рабстве*) или его временную занятость (*на тренировке*). Глаголы для этого приспособлены хуже, а иногда их просто нет, ср. *Он в отпуске* (выражения вроде *Его отпустили на месяц* этот смысл, конечно, не передают). Вместо *Он на занятии* мы могли бы говорить по-русски *Он занимается*. Но *на занятии* имеет специальный смысл — выделяет квант жизненного потока. На этом интервале человек имеет определенный статус и характеризуется определенным типом занятости — в обоих смыслах этого слова: с одной стороны, он занят каким-то видом деятельности, сопряженным с данным типом ситуации, подчиняется каким-то конвенциям (например, *X на работе* имплицирует (в идеале), что на определенном интервале X работает, а не ходит по магазинам), а с другой стороны, он занят в том смысле, что недоступен для других видов деятельности, ср. *Я не могу сейчас разговаривать, я на занятии* (или просто недоступен, как в случае *в плену*, *в командировке* и под.).

Рассматриваемые конструкции очень важны для русского языка. Не случайно в них участвует такое большое количество событийных существительных.

В зависимости от семантики существительного различаются ситуации «пассивные» и «активные». В «пассивных» ситуациях человек не свободен, не может располагать собой: *в плену, в осаде, в рабстве*, *под арестом* (по-видимому, это связано, в частности, с семантикой замкнутого пространства, выражаемой предлогом *в*, хотя в *в*-конструкциях не всегда описываются «пассивные» состояния, ср. *в разведке*, *в командировке*). «Активные» ситуации могут предполагать активную деятельность (*на тренировке*, *в разведке*), но не обязательно (*в дороге* — X-а везут), — поэтому их можно было бы называть нейтральными, но мы оставляем термин «активные», т. к. таковых большинство. Для «активных» существительных характерны конструкции с глаголами движения: целевая, ср. *Пошел на репетицию, на дежурство, на тренировку, на занятия, в разведку* — и «обратная», ср. *вернулся с дежурства, со службы, с тренировки*, *из разведки*. Для пассивных существительных характерно сочетание

с неконтролируемыми глаголами: *попал в плен, оказался в заключении* (или с пассивной формой глагола: *взят под арест*).

Основную массу класса ситуаций составляют конструкции с предлогами *в* и *на*. Они имеют множество особенностей, из которых мы сможем упомянуть лишь некоторые.

Существительные, которые имеют многозначность ‘здание / учреждение’, в этой конструкции обозначают мероприятия, действия или состояния (т. е. такое значение имеет конструкция в целом): *Он в больнице* значит ‘он на лечении’, *Он в школе* значит ‘он на занятиях’. Если за человеком гналась полиция и он забежал в ресторан, про него, конечно, можно сказать *Он в ресторане* — и это будет пониматься именно как ‘находится в помещении ресторана’, но конструкция *Он в ресторане* значит другое –‘проводит время и принимает пищу’.

В активной конструкции *Он на X-е* X имеет статус запланированного мероприятия, которое происходит «по расписанию», ср. *Он на занятии*, *на смене, на тренировке, на репетиции, на дежурстве* — или заранее намечено, готовилось самим субъектом и в каком-то смысле зависит от него: *Он на свидании, на рыбалке* (субъектом может контролироваться не само мероприятие, ср. *Он на концерте / на футболе*, а тот факт, что он принимает в нем участие).

У многих конструкций *Он на / в X-е* имеется соотносительная конструкция *У него X* (*У него занятие / смена / тренировка*), которую можно условно назвать «плановой». Характерным элементом плановой конструкции является указание на время: *У меня* ***сегодня*** *репетиция*; *У меня* ***завтра*** *дежурство*. Мероприятия-развлечения, которые не обязательны и могут быть отменены, плохо совместимы с этой конструкцией: *?Завтра у меня рыбалка*; *?В июне у нас путешествие*; для этой конструкции не подходят также «зависимые» состояния: нельзя сказать *У меня сегодня арест*, даже если человека заранее предупредили, что его арестуют (так может сказать только тот, кто сам арестовывает).

Не имеют соотносительной «плановой» конструкции сообщения о постоянном социальном статусе (*Он на пенсии, в отставке, в браке*, ср. *?У него отставка*) или о длительной занятости (*Он на фронте, на войне, в армии*, ср. *?У него война* — такое высказывание возможно, но в другом значении, бытийном: *У нас война с соседями* — ‘имеет место’). При этом конструкция *Он в институте* понимается как ‘Он на занятиях’, а не как ‘он студент; он учится’, — хотя учеба, как и служба, может занимать несколько лет, ср. *Он в армии* — ‘он служит’. И это конвенциональное понимание, т. е. тоже словарная информация.

Также не имеет соотносительной «плановой» конструкции военная лексика. *В бою, в дозоре, в разведке* устроены как *в плену* или *в рабстве*: нельзя сказать *?У него бой, дозор, разведка* — возможно, потому, что в военной иерархии человек сам мало что решает и, в основном, подчиняется приказам.

В- и *НА*-конструкции употребляются также в качестве детерминантов.

Детерминанты с событийными существительными (в отличие от личных — субъектных и объектных — детерминантов, ср. *У него гости*; *К вам гости*; *С ним одни проблемы*, и пространственных детерминантов, ср. *В лесу тихо*; *На улице шумно*) — это редуцированные предикации, и их, как правило, можно развернуть в полноценные предикации. Чаще всего ситуативные конструкции

можно развернуть в придаточные времени: ***В заключении*** *он научился шить рукавицы* — ‘когда был в заключении’; ***На тренировке*** *они отрабатывали новый прием* — ‘когда были на тренировке’.

Другую группу составляют конструкции со значением ОБСТАНОВКИ (внешние условия, состояние окружающей среды): *в тишине*, *в темноте*, *в холоде, в тепле, в тесноте, в сырости*. Они очень неохотно употребляются предикативно: **Город в тишине / в темноте* — хотя логически этому ничто не препятствует (в НКРЯ обнаружился всего один пример на предикативную конструкцию *в темноте*: *Играла музыка. Манеж* ***в темноте***. (*В это время униформисты ставили реквизит*). [Юрий Никулин. Как я стал клоуном (1979)]). Вместо этого употребляется настоящий предикатив или номинатив: *На улице темно*; *В городе тихо / тишина*; *На улице холодно / холод*. Зато конструкции обстановки свободно употребляются в функции детерминанта, часто — в причинном значении:

В тишине *было слышно, как шумит вода* — ‘благодаря тишине’;

В темноте *нельзя было разобрать дороги* — ‘из-за темноты’.

ЗАМЕЧАНИЕ

Есть еще разновидности *в*-конструкций, в которых употребляются большие классы существительных, — модальные и оценочные конструкции: *Мы в тупике, в ловушке, в опасности*; *Он в беде*; физиологические, эмоциональные, психологическике состояния и проявления человека: *Она в обмороке, в сознании, в шоке, в коме, в депрессии, в отчаянии, в раздумье, в слезах*. Но мы, за неимением места, их здесь не рассматриваем.

Кроме собственно предложных конструкций в базу должны быть включены производные предлоги. В современном языке образование предлогов — это активный процесс, и его результаты не всегда могут быть однозначно охарактеризованы: какие-то единицы уже зафиксированы словарями и грамматиками, а какие-то нет. Например, в Русской грамматике 1980 г. [13] список отыменных предлогов невелик, и туда входят, главным образом, предлоги, образованные на базе «грамматических» существительных (*по причине чего, с целью чего, в случае чего*; ср., однако, *без сопровождения кого / чего*), но не входят конструкции *в обмен на что*, *в ответ на что*, *на основе чего*, *в награду за что* и мн. др. В современных словарях служебной лексики список таких предлогов значительно больше. Так, в «Словаре наречий и служебных слов русского языка» 2005 г. [14] отсутствовавшие в грамматике 1980 г. конструкции (см. выше) уже зафиксированы как предлоги, но в нем нет, например, оборота *в режиме чего*. В базе данных имеет смысл отражать не только «обычные» предлоги, но и такие обороты, которые не зафиксированы в словарях, но по структуре и употреблению аналогичны отыменным предлогам (ср. *под градом чего*; *в припадке чего*; *с оглядкой на кого / что*).

Дело в том, что большинство подобных выражений — это не окончательно сформировавшиеся единицы. Черта, которая их сближает с предлогами, — обязательное заполнение валентности (обязательное заполнение валентности,

т. е. отсутствие абсолютивного употребления — как свидетельство лексической вырожденности, неполноценности, — характерно и для других классов абстрактной лексики, например, полуслужебных глаголов (лексических функций), ср. *принять участие*).

С другой стороны, в выражениях типа *под влиянием чего* или *под властью чего* существительные в большей или меньшей степени сохраняют именные свойства — в частности, способны присоединять прилагательные. Благодаря корпусу, мы можем обнаружить, что большинство из этих прилагательных (хотя и не все) — тоже «полуслужебные» по своему значению: степенные, модальные и оценочные — но тем не менее они допустимы, следовательно, предложная конструкция находится в промежуточной стадии грамматикализации:

под влиянием:

(а) со степенными и модальными прилагательными: *под большим / сильным / огромным / прямым / непосредственным / явным / заметным / несомненным влиянием чего*; (б) с оценочными прилагательными: *Целое поколение выросло под бесовским влиянием телевидения*; *под целительным влиянием времени боль стала немного успокаиваться*; (в) с «содержательными» характеристиками: *Все побережья Западной Европы и ключевые пункты Средиземного моря заштрихованы как «регионы» независимые, но под политическим влиянием Англии»*; *он заворожен ее* [смерти] *красотой, находясь под литературным влиянием поэтов-романтиков* (примеров (в) совсем немного);

ср. конструкции с другими существительными:

под реальной властью престола, под мудрой властью владык, под страшной властью оккупантов, под неограниченной властью помещиков, под загадочной властью смерти, под верховной властью Турции;

под страшным, тяжелым, постоянным, непосредственным, железным гнетом немцев, слухов, преданий, присяги; под нравственным гнетом субъекта;

под единоличным, личным, оперативным, особым, жестким, полным, постоянным, строгим контролем руководителя, министра, государства;

под заботливой, благодатной, нежной, бдительной, постоянной, вечной, тройной, полной, строжайшей опекой Х-а;

под неслыханным, яростным, жестоким, неотразимым, железным натиском кого / чего;

под двойным, сильным, неослабевающим, сокрушительным, жестоким, победным, постоянным напором Х-а (*стихий, половодья, пара, марксизма, противника, варварства, воображения*).

Фиксация всех вариантов конструкций с непредметными существительными в базе данных позволит составить общее представление об их типах, свойствах и связях и даст материал не только для решения уже известных проблем, но и для формулирования новых.

References

1. *Apresian Iu.D., Boguslavskii I. M., Iomdin L. L., Sannikov V. Z.* 2010. Theoretical Problems of Russian Syntax. Interaction of Grammar and Dictionary [Teoreticheskie Problemy Russkogo Sintaksisa. Vzaimodeistvie Grammatiki I Slovaria].
2. *Arutiunova N. D.* 1988. Types of Linguistic Meanings: Evaluation. Event. Fact. [Tipy Iazykovykh Znachenii: Otsenka. Sobytie. Fakt.].
3. *Baranov A. N., Dobrovol'skii D.O., Kiseleva K. L., Kozerenko A. D. 2007. Modern Russian Idiomatic Expressions Dictionary [Slovar'-tezaurus Sovremennoi Russkoi Idiomatiki].*
4. *Boguslavskii I. M., Iomdin L. L.* 1982. Unconditional Turns and Phrasems in the Explanatory-Combinatorial Dictionary [Bezuslovnye Oboroty I Frazemy v Tolkovo-Kombinatornom Slovare]. Aktual'nye Voprosy Prakticheskoi Realizatsii Sistem Avtomaticheskogo Perevoda, 2 : 210–222.
5. *Croft W. 2001. Radical Construction Grammar.*
6. *Fillmore Ch. J., Kay P., O'Connor C.* 1988. Regularity and Idiomaticity in Grammatical Constructions: The case of Let Alone. Language, 64 (3) : 501–38.
7. *Goldberg A., Ackerman F. 2001. The Pragmatics of Obligatory Adjuncts. Language, 77 (4) : 798–814.*
8. *Goldberg A.* 2006. Constructions at Work.
9. *Iomdin L. L.* 2006. Polysemantic Syntaxical Phrasems: Between Vocabulary and Syntaxis [Mnogoznachnye Sintaksicheskie Frazemy: Mezhdu Leksikoi I Sintaksisom]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2006" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2006"): 202–206.
10. *Iordanskaia L. N., Mel'chuk I. A.* 2007. Meaning and Combinatiry in a Dictionary [Smysl I Sochetaemost' v Slovare].
11. *Kurilovich E.* 1962. The Problem of Cases Classification [Problema Klassifikatsii Padezhei]. Ocherki po Lingvistike : 175–203.
12. *Kustova G. I.* 2008. Adverbial Modifier Groups of the kind of 'VO VSIAKOM SLUCHAE' in Modern Russian Language [Obstoiatel'stvennye Gruppy tipa 'VO VSIAKOM SLUCHAE' v Sovremennom Russkom Iazyke]. Instrumentarii Rusistiki: Korpusnye Podkhody. Slavica Helsingiensia, 34 : 126–139.
13. *Kustova G. I.* 2008. On "Non-Nominative" Electronic Dictionaries [O "Nenominativnykh" Elektronnykh Slovariakh]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2008" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2008"): 297–302.

14. *Mel'chuk I. A.* 1974. Experiment of the Theory of Linguistic Models "Meaning - Text" [Opyt Teorii Lingvisticheskikh Modelei "Smysl - Tekst"].
15. *Mel'chuk I. A., Zholkovskii A. K.* 1984. Explanatory-Combinatorial Dictionary of Modern Russian Language [Tolkovo-Kombinatornyi SLovar' Sovremennogo Russkogo Iazyka].
16. *Russian* Grammar [Russkaia Grammatika], I.1980.
17. *Russian* Adverb and Syntactic Words Dictionary [Slovar' Narechii I Sluzhebnykh Slov Russkogo Iazyka]. 2005.
18. *Russian* Phraseology Dictionary [Frazeologicheskii Slovar' Russkogo iazyka]. 1994.
19. *Mel'čuk I.* 1998. Collocations and Lexical Functions. Phraseology. Theory, Analysis, and Applications : 23–53.
20. *Zolotova G. A.* 1988. Syntaxical Dictionary. The Repertoire of Russian Syntax Elementary Unities [Sintaksicheskii Slovar'. Repertuar Elementarnykh Edinits Russkogo Sintaksisa].

ВЫЯВЛЕНИЕ РОЛЕВЫХ ФУНКЦИЙ ЛИЦ НА ОСНОВЕ СТРУКТУР ЗНАНИЙ

И. П. Кузнецов (igor-kuz@mtu-net.ru)

ИПИ РАН Вавилова, Москва, Россия

Рассматривается семантико-ориентированный лингвистический процессор, извлекающий из текстов естественного языка информационные объекты, их свойства и связи и формирующий на этой основе структуры знаний. Одно из направлений развития таких процессоров связано с выявлением имплицитной информации, которая рассматривается в узком плане — как выявление новых свойств объектов, заданных в неявном виде. Предлагается методика такого выявления, основанная на анализе структур знаний. В качестве примера рассматривается выявление ролевых функций фигурантов на базе их описаний в сводках происшествий.

Ключевые слова: ролевые функции, структуры знаний, лингвистический процессор, фигуранты, происшествия, сводки происшествий.

IDENTIFYING ROLE FUNCTIONS OF PEOPLE ON THE BASIS OF KNOWLEDGE STRUCTURES

I. P. Kuznetsov (igor-kuz@mtu-net.ru)

Institute for Informatics Problems of the Russian Academy of Sciences, Moscow, Russian Federation

The linguistic processor which extracts knowledge structures (information objects and their links) from natural language texts is considered.The development of the processor is connected with extracting implicit information, e. g. role functions of people. The proposed extraction methods are based on the analysis of knowledge structures. The methods are used for identifi cation of role functions of people involved in criminal cases reported in law texts.

Key words: linguistic processor, role functions, knowledge structures, figurants, criminal cases, reports.

Introduction

One of the primary tasks in the area of cognitive technologies is the automatic extraction of knowledge from natural language (NL) texts. It is a complex problem connected with developing linguistic processors which perform automatic formalization of texts, i. e. mapping the texts into formal models or knowledge structures. It should be noted that a lot of relevant information in NL texts is presented in a concealed form. This information is called implicit. An example is the task of assigning certain features to persons on the basis of actions performed by them. In the subject area of "Criminology" it is assigning such features as "victim", "suspect", etc. to persons. The methods of implicit information extraction should be considered in the context of the knowledge extraction task, and it is conditioned by the specific features of the linguistic processor. The given paper describes these methods within the framework of the object-oriented linguistic processor developed at the Institute for Informatics Problems of the Russian Academy of Sciences (IIPRAS).

1. The object-oriented linguistic processor

The research direction connected with unstructured NL texts processing has been developing for 20 years at the IIPRAS for particular application areas and specific user tasks [1,2]. One should consider that the large category of users have the specific official responsibilities, and respectively, constant interests. Completely concrete information is necessary for them. For example, a criminal inspector seeks to extract information on important figurants, their places of residence, telephones, criminal events, dates and other such facts [3]; a personnel manager is interested in the organizations, when and where a person worked and in what position [4]. Other people try to fish out from the media the information about the countries, important persons, catastrophes, places of interest and historical monuments [5]. We call this concrete information interesting for a user ***information objects***. Objects are distinguished by their types.

Let us note that the connections between the objects, which interest users, can have the high degree of variety. For example, not only the connection of the persons with their information from their questionnaires or the objects can present interest, but also the actions or the events, in which these persons participate. Such events are attached to the time and the place. Moreover, some events can be a component part of others. They can be connected with cause-effect and temporary relations. For the number of problems similar connections play an important role. They also must be revealed and processed. Therefore one should consider that events are also information objects, interconnected and connected with other information objects. Complex structures appear. For their representation within the framework of the projects of IIPRAS (Russian Academy of Science) the language of the extended semantic networks (ESN) has been developed, while for the processing the production rules the language DEKL [6] has been implemented.

The ESN networks are represented in the form of special graphs [6]. In the formal record they are the extension of the predicate logic language. ESN consist

of elementary fragments, each of which has its unique code (see Section 3), which can stand at the argument places of other fragments, and provide great possibilities for the representation of knowledge structures. The language DEKL is designed for the transformation of such structures. The problem of extracting knowledge from natural language texts is considered from the point of view of developing information objects and connections with the construction of the knowledge structures on the basis of which the solution of user problems is achieved. For this within the framework of the IIPRAS projects the ***object-oriented*** linguistic processor (LP) converting the natural language (NL) texts to the knowledge structures is developed and constantly updated. The processor LP achieves a deep NL text analysis with bringing of synonymous groups to one form, development of objects and their properties, identification of objects, elimination of ambiguities, development and unification of various forms, which present events or actions (including forms with the verbal noun, participial and verbal-adverbial constructions), which are connected with the time and the place [2–6]. As a result the structures of knowledge in the ESN formalism are created.

The linguistic processor (LP) is realized by means of the language DEKL and is controlled by the linguistic knowledge (LK) in the form of object dictionaries, means of parametric tuning, and also the rules of extracting objects and connections [2,4,5,6]. With the aid of LK the tuning of LP to the appropriate categories of users and text corpora is accomplished. Concrete realization appears as a result. Thus, the paper deals with the means of constructing a class of processors with powerful mechanisms for their tuning and updating. Further development of such processors (LP) is connected with the development of implicit information [7], which we will consider in a narrow plan, i. e. as the addition of the structures of knowledge by the new information, which is absent or assigned implicitly. In this article the procedure of this development is proposed, which consists in the use of LP for mapping NL texts onto the structures of knowledge (ESN) and the use of the means of logical-analytical processing (productions of the language DEKL) for the creation of new information.

Advantages and deficiencies of the proposed procedure will be examined on specific objectives from the area of "criminology", that is the role functions establishment for the persons (participants) on the basis of the acts performed by them or due to the participation in some specific events. We consider the problem of assignment of properties to the persons (basing on their participation in the acts of different kinds) — "the suffered", "the suspect" and others, if an explicit description of such properties is absent from the text. For example, if it is said in the text "suffered Ivanov I. I.", then another task appears, i. e. extraction of some property in the process of linguistic analysis and forming of the corresponding fragments in the knowledge structure. In this article the discussion will deal with LP, customized for the Russian language texts (NL), although the possibilities of LP are wider. There is a sufficient test of tuning LP to the English language texts [9,10].

2. The choice of method

The task of the role functions establishment for the information objects is a special case of the more general task, connected with the estimation of objects according

to their descriptions in the NL texts, for example, with the estimation of the stability of enterprise (according to the information from the Internet), by featuring political figures (positive or negative depending on the statements in the press), by the estimation of the role functions quality of product (basing on the statements of users) and so forth. Quite frequently, it is not said directly whether something is bad, or good. As a rule, in NL texts the events are described, the situations, in which one or other information object participated. On the basis of them the estimation is done, which is often represented in the form of a new (generated) property of object.

For the solution of this problem different methods are used [11–15]. The most common one is the method of the new properties of objects development by using the syntactical-semantic forms. For example:

<what-medicine> caused allergy in <who-human organism>…,
< what-medicine > has side effects …
<who-person> made scandal… and so forth.

The application of such forms to the NL texts consists in the search for "estimating" or "characterizing" words (of type "scandal") or for word combinations of the type "caused allergy" ("it can cause allergy"), it "has side effects" ("side-line actions"), "to make scandal" ("to brawl")… And then the environment is analyzed, i. e., the words, which stand to the left and to the right, their semantic classes (objects are recognized by them) and case forms. Estimations of information objects as a result are given. By the first two forms the "quality of medicines" is estimated, while by the latter it is recognized that a man performed "hooligan actions" or that he is "suspected". It is known that in NL many versions are possible for expressing the same idea — with the aid of different syntactic constructions, verbal groups, forms and so forth. Therefore the number of estimating word combinations will be sufficiently large. Moreover, the application of such forms requires different forms of analysis — morphological (in order to reduce different word forms to one form), syntactic (the trees are built of the selection of sentences in order to isolate the connected components and to find place for the estimated words) and semantic (in order to extract the objects, which are evaluated). The use of syntactical-semantic forms is connected with certain difficulties caused by special NL features: by the presence in texts of participial, verbal-adverbial constructions, different explanations, facultative components (time, place, purpose), anaphoric references and other language structures. As a result, information objects are frequently disconnected from the estimated words. Hence — the significant losses, which influence the quality of estimation.

Example 1 *(the text is taken from the summaries of incidents of the City Office of Home Affairs, Moscow):*
… Gorelov Peter Sergeyevich, 01.03.76 yr/bir, liv: c. Moscow, st. Young Leninists, h.71-6-12, does not work, 01.02.1998 yr. at 4.30 in his house out of hooligan motives in the state of alcoholic intoxication made scandal and broke window glass in the apartment of Litvinova Galina Ivanovna, 20.07.1961 yr/bir,…

Пример 1 (текст взят из сводок происшествий ГУВД г. Москвы):
… Горелов Петр Сергеевич,01.03.76 г/р, прож: г.Москва, ул.Юных Ленинцев, д.71-6-12,не работает, 01.02.1998 г. в 14.30 у своего дома из хулиганских побуждений в состоянии алкогольного опьянения учинил скандал и разбил оконное стекло в квартире Литвиновой Галины Ивановны,20.07.1961 г/р, …

In this example the estimating (characteristic) words are "made scandal" and "broke the window glass", they are located at a significant distance from the estimated person — "Gorelov Peter Sergeyevich". This limits the possibilities of applying the forms. It is required that the initial extraction of components, which must not be considered in the forms: the years of birth, addresses, specific properties ("he does not work", "in the state of alcoholic intoxication"), time, place and others, which requires sufficiently deep text analysis with the extraction of objects, their properties and attributes. In connection with the aforesaid, another more promising method is represented — when evaluation is accomplished at the level of knowledge structures. For their construction the objective-oriented LP is used producing the structures of knowledge in which the objects are directly connected with the events and the actions and excluding the above mentioned losses. For the development of implicit information (role functions of objects) the rules of the DEKL language are used which analyze the structures of knowledge (ESN) and form new properties of objects. In this case the structure of knowledge does not change, but it is only supplemented by new (useful) fragments.

3. Meaningful portraits of documents

Within the framework of the proposed procedure the development of the role functions of objects (implicit information) is achieved at the level of the structures of knowledge, called the meaningful portraits of documents (SS-documents). Let us examine how such structures appear in the ESN formalism [2,3,6].

Example 2 (translation of the Russian text given below). *Text N22 is taken from the summaries of incidents of the City Office of Home Affairs, Moscow*:
01.02.98 yr. 16–30 to the Home Office applied citizen Mitrofanov Victor Mikhaylovich, 1955 yr. bir., liv.: Bohr Highway 38–211, n/w. he stated that 01.02.98 yr. at 10-00 in house 3 at St. Fedosino the unknowns being found in the drunk state made scandal, they expressed themselves by unquotable swearing, they set dog. As a result of what Mitrofanov applied to trauma care center, where the diagnosis was set: the bite of foot.

Пример 2. Текст документа (с номером 22) взят из сводок ГУВД:
01.02.98 г. в 16-30 в ОВД обратился гр-н Митрофанов Виктор Михайлович, 1955 г.р., прож.: Боровское шоссе 38-211, н/р. Он заявил, что 01.02.98 г. в 10-00 у д.3 по ул. Федосьино неизвестные находясь в пьяном виде учинили скандал, выражались нецензурной бранью, натравили собаку. В результате

чего Митрофанов обратился в травмпункт, где был поставлен диагноз: укус ноги.

The objective-oriented LP performs the deep analysis of the text and automatically builds its meaningful portrait (SS- document, transliterated):

```
DOC_(22, "1-02-98", "SUMMARY; " /0+) 0 (RUS)
OVD_(OVD/1+)
FIO(MITROFANOV], VICTOR, MIKHAYLOVICH, 1955/2+) UNEMPLOYED (2-/3+) 3-
   (22, PROP)
ADR_(Borovskiy, Sh., 38,211/4+)
PROZH. (it is 2nd, 4)
ADR_(UL, FEDOSINO, HOUSE, 3/5+)
FIO (" " , " " , " " , NESKOLKO/6+)
UNKNOWN (6)
DRUNK (6-/7+) 7 (2, PROP_)
SCANDAL (6, PYANYY/8+)
IS EIGHTH (22, ACT_)
TO REPORT (IT IS 2ND, 8-/9+) 9 (22, ACT_)
DATA_(1998,02, ~01, " 10-00" /10+)
When (9, 10)
TO TURN (1, GR- N, 2-/11+) 11- (22, ACT_)
DATA_(1998,02, ~01, " 16-30" /12+)
When (11-, 12-)
EXPRESS (6, UNQUOTABLE, [BRAN]/13+) 13- (22, ACT_)
TO SET (6, [SOBAKA]/14+) 14 (0, ACT_)
TO TURN (IT IS 2ND, IN, [TRAVMPUNKT]/14+) 14 (0, ACT_)
TO PLACE (DIAGNOSIS, BITE, [NOGA]/16+) 16 (0, ACT_)
PREDL_(22,11-, 4, 3-, 9, 13-, 14-/17+) 17- (2,15,341)
PREDL_(22,15-, 16-/18+) 18- (6,342,448)
```

For the original Russian text the automatically generated SS- document looks as follows:

```
ДОК_(22,"1-02-98", "СВОДКА;"/0+) 0-(RUS)
ОВД_(ОВД/1+)
FIO(МИТРОФАНОВ,ВИКТОР,МИХАЙЛОВИЧ,1955/2+)
БЕЗРАБОТНЫЙ(2-/3+) 3-(22,PROP_)
АДР_(БОРОВСКИЙ,Ш.,38,211/4+)
ПРОЖ.(2-,4-)
АДР_(УЛ.,ФЕДОСЬИНО,ДОМ,3/5+)
FIO(" "," "," ",НЕСКОЛЬКО/6+)
НЕИЗВЕСТНЫЙ(6-)
ПЬЯНЫЙ(6-/7+) 7-(2,PROP_)
СКАНДАЛ(6-,ПЬЯНЫЙ/8+) 8-(22,ACT_)
СООБЩИТЬ(2-,8-/9+) 9-(22,ACT_)
ДАТА_(1998,02,~01,"10-00"/10+)
Когда(9-,10-)
ОБРАТИТЬСЯ(1-,ГР-Н,2-/11+) 11-(22,ACT_)
ДАТА_(1998,02,~01,"16-30"/12+)
Когда(11-,12-)
```

```
ВЫРАЖАТЬСЯ(6-,НЕЦЕНЗУРНЫЙ,БРАНЬ/13+) 13-(22,ACT_)
НАТРАВИТЬ(6-,СОБАКА/14+) 14-(0,ACT_)
ОБРАТИТЬСЯ(2-,В,ТРАВМПУНКТ/14+) 14-(0,ACT_) ПОСТАВИТЬ(ДИАГНОЗ,УКУС,НО
   ГА/16+) 16-(0,ACT_) ПРЕДЛ_(22,11-,4-,3-,9-,13-,14-/17+) 17-(2,15,341)
ПРЕДЛ_(22,15-,16-/18+) 18-(6,342,448)
```

A meaningful portrait consists of the elementary fragments, arguments of which are words in the normal form (necessarily for the search and processing). Each elementary fragment has its unique code, which is written in the form of the number with the sign + and is separated by a slash line. For example, in the fragment OVD_(OVD/1+) the sign 1+ is its code (but 1 is the reference to it). Fragments DOK_(22, "1-02-98.TXT", "SUMMARY; " /0+) 0 (RUS) indicate that the meaningful portrait is built on the basis of the Russian-language text of document with number 22 of the file of 1-02-98.TXT", which was processed as the summary of the incidents (linguistic knowledge depend on this). The following fragments present police department OVD_(… /1+), person's surname, name and patronymic FIO (… /2+), person's specific property UNEMPLOYED (2-/3+), address ADR_ (… /4+) and so forth; the signs 2+, it is 2nd, 3+, 3-,… are the codes of the fragments, with the aid of which their connections and relations are assigned. For example, the fragment PROZH (live) (it is 2nd, 4) represents the relation that the person (represented as FIO with code 2+) lives at the address (fragment [ADR]_ with code 4+). Actions are represented in the form of fragments of the type SCANDAL (6, PYANYY/8+) it is 8 (22, ACT_), where it is represented that "person (FIO with code 6+), being drunk, made scandal". With the aid of it is the fragment 8_(22, ACT_) indicates that the first fragment is SCANDAL (…./8+) presents the action and relates to the document with the number 22. A similar role is played by the fragments of the type 3-(22, PROP_), by which the properties are noted. The codes of fragments also serve for the idea of time, scene of action and cases, when one action is included in the composition of another. For example, the fragment TO REPORT (it is 2nd, 8-/9+) represents that the person (code 2+) "reported" (code 9+) about the action (code 8+), i.e., about "made scandal". The following fragments DATA]_(… /10+) when (9, 10) represent the time (DATA_), which relates (when) to the action "to report". Special role is played by the fragments PREDL_(...), which correspond to the sentences. They are filled up with the words, which did not enter the information objects (in this example they are absent), or with the codes of objects themselves. To these fragments the indicators of their position in the text are added. For example, the fragment PREDL_(22,11-, 3-, 9, 13-, 14-/17+) 17- (2,15,341) represents the fact that the objects with codes 11- (corresponding to the action "to turn"), 3- (corresponding to the property "unemployed") and others are located in the sentence, which begins from the 2nd line of the text of the document and they occupy the place from the 15-th to the 310-th byte. These means of positioning are necessary for the work of the reverse linguistic processor (LP).

Analyzing this example, it is possible to make the following conclusions: 1) In SS-document the estimating (characterizing) words occur either in one fragment with the object — SCANDAL (,…), or the next one, i.e., the codes of the actions, in which the object participates, are nearby in PREDL_(… 9, 13-, 14…). In this case the possibility of composite actions is considered. 2) On the actions, represented as SCANDAL (,…), it is possible to draw the conclusion that the discussion deals with "that suspected", and

TO REPORT (,) — that the person is "suffered" or "the applicant". Such conclusions are easily arrived at with the aid of the rules IF… THEN (productions) of the language DEKL, which are the basis for the extraction of role functions. 3) The particular difficulties of dividing the text into the sentences occur (in the old version). The reduction "of n/r" (with the point at the end) was not understood as the end of a sentence. 4) The linguistic processor (LP) correctly identified the pronoun "*he*", and also it knew how to reveal the participation of the subject ("*unknowns*") by the actions "*to be evinced by unquotable swearing*" and "*to set dog*", which also characterize subject. At the same time the LP could not connect the action "*diagnosis was set*" with the person — "*Mitrofanov…*" (the code is 2-nd). In this case an example proved to be successful. Also the processor LP (with its linguistic knowledge — LK) was developed for the tasks of the criminal police, connected with different forms of the objective searches: the search for similar pparticipants (addresses, and so forth), search according to the connections, precise search for objects, for the search by signs and other identifiers. In this case the analysis of some complex NL forms was not required, i. e. the cases of the enumeration of the objects participating in the uniform actions (they are described by one verb), the enumeration of the actions of one object and others in contrast to the aforesaid, with the extraction of role functions for each object the indication of its participation in each action is required. Hence it follows that with the use of the proposed procedure the more qualitative extraction of role functions is directly connected with the works on improvement of LP in the aspect of the development of objects and their actions. In many instances the numerous errors caused the inaccuracies in SS-document, e. g.: the absence of punctuation marks or their presence (where it was not required), the inappropriate reductions, gaps in the words and many others. The fact is that the documents, entering the summaries of incidents, are composed on the spot by people (militiamen) of different degree of literacy. Hence — the additional noise and loss. Thus, meaningful portraits are the collections of fragments of ESN which represent the sufficiently high level of formalization of NL texts and are convenient for the working — with the aid of the instrument means — DEKL [3]. Besides LP which analyzes texts and builds SS-documents, there is a reverse linguistic processor (LP) which on the basis of the fragments of the SS- document generates the NL texts presented to the user [6].

4. The means of the development of the role functions

As it has already been said, within the framework of the proposed procedure (instead of the application of syntactical-semantic forms to the documents) the rules are used for logical conclusion and transformation of the knowledge structures — the SS- documents, in which there are no morphological features (of type who, whom,…), and the subjects and the objects are distinguished by their arrangement in the fragments of ESN, which present actions. The names of fragments present the nature of actions. Syntactical-semantic forms are transformed into the fragments of ESN which determine conversions and logical conclusion achieved by productions of language DEKL. Such fragments play the role of the logical-semantic shell, which determines conversions and logical conclusion on the basis of SS-documents. After filling of the

shell by ontological-fragmental knowledge (OFK) which consist of the mentioned fragments (ESN), the program is formed, which accomplishes the development of role functions and completion of the SS-document by the appropriate fragments. With this approach it is possible to avoid many difficulties, connected with the design features of NL and the specific character of the use of syntactical-semantic forms. There are many versions of construction of the shells and representation of the corresponding knowledge which are distinguished according to the degree of their generality. Let us examine the version which is at present realized and verified.

Case 1. The role functions are determined by the names of actions. In this case for the extraction of objects (participants) which should be assigned properties (role functions), the fragments of the following form are used :

```
INTERPRET (MAN_2, FIO, "suffered") FORMA_CC (MAN_2, CLASS_D4, " ") CLASS_
   D4 (TO TURN, TO STATE, TO REPORT, TO PASS AWAY,…)
```

The first fragment INTERPRET (…) means that from the SS- document it is necessary to extract the fragments of the form FIO (…), that correspond to participants, and to analyze the possibility of assigning them the property "suffered". Such participants are conditionally designated as MAN_2. The second fragment FORMA_CC (…) specifies the conditions for assigning this property to MAN_2, determined by the constant CLASS_D4. In the third fragment CLASS_D3 (…) the words are given which present actions. It is represented that the words belong to the class CLASS_D3. If the participant occurs in one of the enumerated actions, then to this participant the property "suffered" is assigned. This participation is revealed via the analysis of the SS-document. If there is a fragment TO TURN (…, it is n-th,…) in it, the argument of which is the code FIO (… /N+), then the fragment N-("suffered") is added that represents the role function of the corresponding participant. Conformably for the SS-document represented in example 2 the analysis will occur as follows. Consecutively the extraction of fragments FIO (…) corresponding to the participants is performed. First FIO (MITROFANOV,… /2+) will be extracted . Its code is 2- is the argument of the fragment TO TURN (1, GR- N, 2-/11+), that presents the action. In connection with this to SS- document the fragment 11- ("suffered") will be added, which via the reverse LP will be transformed into the statement that "Mitrofanov Victor Mikhaylovich is a suffered person". These actions are realized within the framework of the logical-linguistic shell.

Case 2. Role functions are determined by the actions and elucidating words. For this the same fragments are used, as in the first case, but during the enumeration of the names of actions the additional fragments which present actions with the possible elucidating words, are introduced:

```
INTERPRET (MAN_1, FIO, "suspect")
FORMA_CC (MAN_1, CLASS_D3, " ")
FRAUD (USER, POKUPATEL/15+)
TO SET (SOBAKA/16+)
TO BE EXPRESSED (UNQUOTABLE, SWEARING, MATERNYY,… /17+)
CLASS_D3 (IS DELAYED, TO BE SOUGHT,…, 15,16-, 17-)
```

The given fragments determine actions of the extraction of persons (MAN_1), by which the property of "suspect" is assigned. For this at the level of the knowledge structures their participation is analyzed in the actions "*is delayed*", "*to be sought*", and also in the composite actions: "*to set dog*", "*to be expressed unquotable…*", "*to be expressed by swearing…"* and others. In example 2 the code of fragment FIO (" " , " " , " " , NESKOLKO]6+), that represents the unknown persons is the argument of the fragment TO SET (6, SOBAKA/14+), representing action *"to set"* with the elucidating word *"dog"* — *"sobaka"*. Therefore the fragment 6 is added ("*suspect*"), that represents that "*the unknown persons are suspected*", and through the reverse LP the explanation to this conclusion is offered, see below. A similar conclusion will be made on the basis of the fragment TO BE EXPRESSED (6, UNQUOTABLE, BRAN/13+), but with other explanations.

Case 3. The actions determine the role functions of several persons. For this (additionally to the fragments INTERPRET) the fragments are added: CLASS_D1 (TO STRIKE, TO BEAT UP,…) FORMA_CC (MAN_1, CLASS_D1, MAN_2), where FORMA_CC (…) indicates the need of the search of two persons — "*suspect*" and "*suffered*" (MAN_1 and MAN_2), that participate in one action, which are mentioned in the fragment CLASS_D1 (…). For example, "*certain person struck another…*". In the appropriate fragment TO STRIKE (…) the code FIO (…) that corresponds to the first person will stand in front of the second. The given fragments of ESN compose the knowledge OFK which are constantly supplemented — due to the filling of classes by the new words-actions and with the elucidating words. The process of filling is sufficiently simple. If role function is not revealed, then it is necessary to look in the SS-document in which the action of one or another participant (by the text its role is easily determined) occurs. Further, the corresponding constants are located, by which are supplemented the classes of knowledge OFK. Subsequently it is intended to automate the process of completing the knowledge OFK as follows. In the text the words, which determine role functions, are noted. Further, in the formed SS-document the corresponding constants which supplement knowledge OFK are located.

5. Explanation of the results

The explanation of results is accomplished through the reverse LP which on the basis of SS- document and additional fragments builds texts in natural language which are displayed to the user. The reverse LP through the codes of the fragments which correspond to object and actions, finds the sentence (PREDL]) and its location in the text of a document. Through the arguments of these fragments (the words in the normal form) the processor finds the components of the sentences in which the mentioned object and actions are described. These components are converted into the form suitable for the delivery to the user. The fact is that many components are transformed depending on the context. For example, "… *he threatened Petrov I. I.*…." — "… *ugrozhal Petrovu I. I.*…." , where during the delivery FIO *"Petrovu"* should be transformed into "*Petrov I. I.*" Further the description of the object is delivered, its generated property and actions which explain this property — role function.

Example 3. with the use of the above given knowledge OFK of the SS — document of example 1 the role functions will be generated which with the aid of the reverse LP will be given out to user in the following form:
unknowns — suspected,
since — unknowns, being in the drunk state made scandal,
they expressing themselves by unquotable swearing, they set dog
Mitrofanov Eugene Mikhaylovich, 1953 r. — suffered,
since — applied to OVD citizen Mitrofanov Eugene Mikhaylovich, 1953 yr. birth,
since Mitrofanov applied to trauma care centre.

It should be noted that all the actions considered in the paper connected with various cases of role functions establishment and explanations of the results are implemented within the framework of the logical semantic environment written in the logical programming language DEKL. Since the DEKL language is oriented at the processing of knowledge structures (represented in the form of the extended semantic networks — ESN) and since it features the generalized production rules [6], the program code in the DEKL language is very simple and concise: it comprises 16 productions and about 4 Kbyte of text.

Conclusions

The proposed procedure of the role functions extraction centered at the analysis of knowledge structures is sufficiently promising from the point of view of the knowledge bases technology development. The current task is to improve its performance for the documents comprising enumeration of the type: 1. *Ivanov I. I.*… 2. *Petrov A. A.*…. 3.… and further follows the continuation, which describes their acts, for example, "*were subjected to detention by…*" or "*who performed…*". The recognition of such cases requires further upgrade of the linguistic processor (LP) software. The quality of analysis is lowered by the breaks in significant words of the type "*Iva nov*" or "*Iva-nov*", which are typical for the summaries of incidents. The methods were tested on the basis of the summaries of incidents which contain about three thousand documents (each document consists of 10–80 lines). In the case of the summaries processing the documents with the mentioned enumerations (there were about 10% of them) were withdrawn and in the remained texts the gaps in the words were removed. At the current moment the program which realizes the proposed procedure gave about 80% of correct recognition of role functions, and about 65% of complete explanations with the indication of all acts. But these numbers rapidly change for the better due to the means (the LK and OFK knowledge) of tuning the LP to special features of the subject area texts. For this not much time is required. Let us note that tuning itself to the extraction of the role functions of persons from the mentioned summaries (with reaching the indicated percentages), required about two weeks of the work of one person. The development and fixing of the shell itself took about four days. The subsequent development is connected with the improvement and the tuning of LP to the work with complex NL forms. At present the extraction of actions is interfered with causal word combinations of the type "*out*

of the hooligan motives", "*owing to the hostile relations*" and so forth, which at present are introduced into the system. Difficulties appear with the transfer of the subject of action to other actions to which the subject is not assigned explicitly, but its presence is implied.

The second direction of research and development is connected with the extension of the shell features to the solution of other problems connected with the estimation of objects depending on the nature of statements about them in the texts of description. Within the framework of the studies conducted it is also intended to tune the shell to the work with the English language texts. Since the meaningful portraits of the English language and Russian language texts have the identical structure (SS-documents), this tuning cannot be labor-consuming.

References

1. *Banko M., Cafarella M., Soderland S., Broadhead M., Etzioni O.* 2007. Open Information Extraction from the Web. Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07) : 2670–2676.
2. *Clark P., Harrison P., Thompson J.* 2007. A Knowledge-Driven Approach to Text Meaning Processing. Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning : 1–6.
3. *Gildea D., M. Palmer.* 2002. The Necessity of Syntactic Parsing for Predicate Argument Recognition. Proceedings of the 40th Annual Conference of the Association for Computational Linguistics (ACL-02) : 239–246.
4. *Kuznetsov I. P.* 1986. Semantic Ideas [Semanticheskie Predstavleniia].
5. *Kuznetsov I., Kozerenko E.* 2003. The System for Extracting Semantic Information from Natural Language Texts. Proceeding of International Conference on Machine Learning. MLMTA-03 : 75–80.
6. *Kuznetsov I. P.* 1999. Methods of Reports Editing with Figurant and Incidents Characteristics Selection [Metody Obrabotki Svodok s Vydeleniem Osobennostei Figurantov I Prosshestvii]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 1999" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 1999").
7. *Kuznetsov I. P., Matskevich A. G.* 2006. Semantically Oriented Linguistic Precessor for Automatic Formalization of Autobiographical Data [Semantiko-Orientirovannyi Lingvisticheskii Protsessor dla Avtomaticheskoi Formalizatsii Avtobiograficheskikh Dannykh]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2006" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2006"): 317–322.
8. *Kuznetsov I. P., Efimov D. A.* 2008. The Peculiarities of Knowledge Extraction by Semantically Oriented Linguistic Processor Semantix [Osobennosti Izvlecheniia Znanii Semantiko-Orientirovannym Lingvisticheskim Protsessorom Semantix]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2008" (Computational Linguistics and

Intelligent Technologies: Proceedings of the International Conference "Dialog 2008"), 7 (14) : 281–291.
9. *Kuznetsov I. P., Matskevich A. G.* 2007. Semantically Oriented Systems with Knowledge Bases [Semantiko-Orientirovannye Sistemy na Osnovanii Baz Znanii].
10. *Asher N., Lascarides A.* 2003. Logics of Conversation.
11. *Kuznetsov I. P., Somin N. V.* 2008. The Facilites of Semantically Oriented Linguistic Processor Adjustment for Search and Finding Out the Objects [Sredstva Nastroiki Semantiko-Orientirovannogo Lingvisticheskogo Protsessora na Vydelenie I Poisk Ob"ektov]. Sbornik IPI RAN, 18 : 119–143.
12. *Kuznetsov I. P., Kozerenko E. B.* 2008. Linguistic Processor "Semantix" for Knowledge Extraction from Natural Texts in Russia and English. Proceeding of International Conference on Machine Learning, ISAT-2008 : 835–841.
13. *Kuznetsov I. P., Matskevich A. G.* 2005. English Version of the Automatic Information Detection System for Natural Language Texts [Angloiazychnaia Versiia Sistemy Avtomaticheskogo Vyiavleniia Znachimoi Informatsii iz Tekstov Estestvennogo Iazyka]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2005" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2005").
14. *Pasca M., Van Durme B.* 2007. What You Seek is What You Get: Extraction of Class Attributes from Query Logs. Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07) : 2832–2837.
15. *Punyakanok V., D. Roth, W. tau Yih.* 2008. The Importance of Syntactic Parsing and Inference in Semantic Role Labeling. Computational Linguistics, 34(2) : 257–287.

ПРОНОМИНАЛИЗАЦИЯ СЕНТЕНЦИАЛЬНОГО АРГУМЕНТА В РУССКОМ ЯЗЫКЕ

А. Б. Летучий (alexander.letuchiy@gmail.com)

Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Ключевые слова: сентенциальный аргумент, прономинализация, местоимение, модальный контекст.

PRONOMINALIZATION OF SENTENTIAL ARGUMENTS IN RUSSIAN[1]

A. B. Letuchii (alexander.letuchiy@gmail.com)

Russian Language Institute of Russian, Academy of Sciences, Moscow, Russian Federation

The article deals with the distribution of the three Russian pronouns referring to a sentential argument (e. g. — Vasja ne priedet. — Ja eto znaju '- Vasja will not come — I know it') — namely, eto, tak and takoe. Each of them has its particular distribution, including contexts where none of the other two pronouns can be used. I show that the pronoun takoe is usually used in the context of negation and modal operators, but only rarely occurs in affi rmative sentences with a verb in the indicative mood. The pronoun eto, contrary to tak and takoe, can be used with concrete descriptions of speech acts, including supplementary characteristics of speech, such as loudness, whereas tak is incompatible with these characteristics. The individual properties of the pronouns are refl ected in their distribution in the corpus data. For instance, the proportion of infi nitive clauses among the uses of the pronoun takoe is much greater than for the two other pronouns, which nicely agrees with the tendency of takoe to be used in modal contexts. Finally, I show the difference between the uses of takoe where the pronoun refers to an NP and those where it refers to a sentential argument. In the former case, takoe always denotes a class of entities, whereas in the latter case, takoe can denote one particular content of a speech act. This difference has to do with different referential properties of NPs (objects) vs. propositions.

Key words: sentential argument, pronominalization, modal contexts, pronoun.

[1] Исследование выполнено в рамках работ по гранту Президента РФ для поддержки молодых учёных — кандидатов наук МК-3522.2010.6 и гранту РГНФ 10-04-00256а.

Существует большое количество глаголов, у которых место подлежащего или одного из дополнений занимает так называемый сентенциальный аргумент (актант) — придаточное предложение или оборот, называющий ситуацию. Примером могут являться глаголы, присоединяющие придаточное с союзом *что*: ср. *Вася говорил, что не поедет в Киев.*

Как и стандартные аргументы, сентенциальные актанты способны к прономинализации, то есть могут заменяться на местоимения. Для именных аргументов главные способы прономинализации — местоимения *он*, *тот* и нулевое выражение. Вопрос об их соотношении до конца не решён, однако в целом прономинализация именных аргументов описана хорошо (см., например, [Крейдлин, Чехов 1988] о местоимении *тот*, [Kibrik, Prozorova 2007] о соотношении анафорических местоимений и синтаксического нуля).

Напротив, местоимения, относящиеся к сентенциальным актантам, в русистике анализировались мало. В типологии они тоже редко становились предметом исследования, преимущественно исследуются местоимения, замещающие именную группу.

Основным способом прономинализации сентенциальных актантов является местоимение *это*:

(1) *Но только он это подумал, как слышит, что Ак-Бозат заржала*. [Д. Н. Мамин-Сибиряк. Ак-Бозат (1895)]

(2) *На каком основании он это полагал? При наличии невестки и двух внучек, какие мог он питать надежды?* [Н. Н. Берберова. Железная женщина (1978–1980)]

Однако существуют и другие способы:

Местоимение-прилагательное *такой* (в форме *такое*):

(3) *— Он утверждал, что Солнце вертится вокруг Земли. — Как можно в девятнадцатом веке* ***такое*** *утверждать?*

Местоимение *так*:

(4) *Здесь нет автодорог —* ***так****, по крайней мере, утверждает атлас.*

Нашей задачей является проследить разницу между тремя способами: *это, так* и *такое.* Мы покажем, что различие здесь, как и в случае *он* vs. *тот*, не совсем тривиально.[2]

Отметим, что местоимение *это* часто упоминается в описаниях русского языка, в частности, в [Грамматика 1980], [Падучева 1985]. Местоимение *так*

[2] Отметим сразу, что в данной работе мы не анализируем распределение местоимений с точки зрения параметра активации референта в дискурсе. Данный параметр, как показано в [Poesio, Modjeska …], релевантен для распределения английских местоимений *it* и *this / that*.

не описано. Местоимению *такой* посвящена специальная небольшая работа [Куликов 1985]. Автор не рассматривает употребления *такое* как заместителя сентенциального актанта — однако. что важно, отмечает взаимозаменяемость в ряде контекстов местоимений *этот* и *такой* (*Некоторые люди страдают бессонницей. Эти / такие люди, как правило, раздражительны и вспыльчивы* [Куликов 1985: 73])

Некоторые примеры

Вначале приведём некоторые примеры, когда одно из местоимений неграмматично, а другие допустимы. Такие примеры существуют для всех трёх рассматриваемых слов, иначе говоря, ни одно из местоимений не совпадает по сфере употребления с другим и не включает другое.

(5) *Он сказал: «Вот скоро поедем на конгресс генетиков, там и решим вопрос о переезде Тимофеева-Ресовского». Но сказано это (*так, *такое) было как-то без обычного вавиловского оптимизма*. [Даниил Гранин. Зубр (1987)]

(6) *Лев Леонидович при этом знал, чем Артур занимается на самом-то деле, а Лариска — нет. Не положено женщине знать такое (*так, ?это) — все равно где-нибудь когда-нибудь проболтается!* [Андрей Грачев. Ярый-3. Ордер на смерть (2000)]

(7) *«Ты улыбайся, когда улыбаешься — не убьют», — так (*такое, *это) думал 10-летний киевский мальчик.* [Юлия Кантор. Вернувшиеся из бездны. ОРТ показало фильм к 60-летию трагедии Бабьего Яра (2001) // «Известия», 2001.10.05]

Нашей целью будет выявить признаки, которые различают три рассматриваемых местоимения. Для местоимений, замещающих именные группы (например, *он*) такая работа уже во многом проделана, см. хотя бы [Крейдлин, Чехов 1988] для *тот* или [Красавина 2004] для указательных групп *этот* + именная группа.

Речевые акты с конкретной характеристикой

Первым параметром, существенным для употребления местоимений, является степень конкретности характеристики речевого акта. Мы называем речевым актом с конкретной характеристикой такое описание речевого акта, в котором указываются какие-либо внешние признаки речи (*говорил это, смеясь; он бурчал это себе под нос*), а не только содержание высказывания.

Если перед нами описание речевого акта с конкретной характеристикой, то может употребляться только местоимение *это*:

(8) *Он, смеясь, говорил это (*так, *такое), вспомнив то, что теперь особенно мучило Бессонова, — причиненная когда-то сыну физическая боль.* [Юрий Бондарев. Горячий снег (1969)]

Поскольку с помощью деепричастия описывается то, как (*смеясь*) субъект произносил слова, ни местоимение *так*, ни местоимение *такое* в данном случае не допустимы. Напротив, если опустить деепричастие, местоимение *так* (но не *такое*) станет допустимым:

(9) *Он, говорил так / это, вспомнив то, что теперь особенно мучило Бессонова, — причиненная когда-то сыну физическая боль.* [Юрий Бондарев. Горячий снег (1969)]

С тем же самым связана неприемлемость всех местоимений, кроме *это*, в примере (9).

Набор глаголов

Набор глаголов, допускающих каждое из местоимений, тоже различается. Так, фактивные глаголы в понимании Ю. Д. Апресяна (2001), то есть глаголы, подразумевающие, что выражаемый зависимой клаузой факт существует в действительности (*знать, доказать*), как правило, не допускают *так*. Мы не нашли в Национальном корпусе русского языка ни одного примера на сочетание *так доказал*.

В то же время существуют глаголы, которые не допускают или с трудом допускают местоимение *это*. Так, глагол *подумать* редко встречается с этим местоимением, хотя примеры существуют:

(10) *Не успел Сашка это подумать, как услышал сквозь разрывы голос ротного: — Сашка! Где ты? Сашка!* [Вячеслав Кондратьев. Сашка (1979)]

Контексты снятой утвердительности

Интересной чертой местоимения *такое* является его частая встречаемость в контекстах снятой утвердительности, то есть, по [Падучева 1985], контекстах, где факт, излагаемый в предложении, отрицается или является частью модальной рамки. Так, словосочетание *такое говорил* встречается в 7 примерах из Корпуса.[3] Во всех примерах оно выступает либо в вопросе, либо с дополнительными модификаторами с модальным значением:

[3] Мы исключили из числа примеров те, где используются группы *что такое* или *что-то такое*, поскольку она имеет свойства, отличные от местоимения *такое* в независимом употреблении.

(11) — *Это где же он такое говорил? — А в «Онегине» своем: что без грамматической ошибки, мол, речи русской не люблю…* [Леонид Саксон. Принц Уэльский // «Октябрь», 2001]

Сочетание *такого не говорил* более частотно (11 примеров) — это показывает, что для местоимения *такое* контекст отрицания не менее характерен, чем модальный:

(12) *Вот ты нам, Иван Семёнович, про своего брата распелся, и товарищ Корнилов тебя поддержал, что он, мол, не виноват, а злодеи его погубили. — Я такого не говорил, — перебил бригадир.* [Ю. О. Домбровский. Хранитель древностей, часть 2 (1964)]

Пример (11) показывает, что просто понятия отрицания для описания местоимения *такой* недостаточно. Оно может встречаться также в контексте сомнения, который семантически подразумевает отрицания, но не содержит выраженного отрицательного маркера.

Местоимения *это* и *так* также встречаются в подобных контекстах, однако численное соотношение здесь другое, чем при *такое*. Мы встретили 24 примера на *так не говорил* и более 300 — на *так говорил* в рассматриваемом значении. Аналогичным образом, на *это / этого не говорил* было найдено 116 примеров, а на *это говорил* — 303. Тем самым, только местоимение *такое* встречается чаще с отрицанием, чем без отрицания.

Похожую картину мы получаем, сравнив употребления *такое, так* и *это* с финитными формами и с инфинитивом (см. Таблицу 1).

Отвлекаясь от того факта, что *такое* в целом гораздо менее частотно, чем другие два местоимения, можно заметить, что инфинитив при этом местоимении в процентном отношении встречается существенно чаще, чем при двух других.

Таблица 1. Местоимения в сочетании с формой прошедшего времени совершенного вида мужского и женского родов и с инфинитивом

	говорил(а)	*говорить*
это / этого	394	254
так	629	479
такое	27	38

При этом существуют контексты, где частотность *такое* выше всего. Примером является сочетание *можно + местоимение + инфинитив*:

(13) *И что это за телеграмму я получил: требуется ли приезд Олюни? Разве телеграммой можно такое спрашивать?* [Александр Морозов. Прежние слова (1985–2001) // «Знамя», 2002]

(14) *Что я вам сделала? Как это можно такое говорить? .. — Фенечка, — промолвил печальным голосом Павел Петрович, — ведь я видел…* [И. С. Тургенев. Отцы и дети (1862)]

Такое в сочетании с глаголами речи выступает в данном контексте в 18 примерах, и именно данная конструкция, вероятно, наиболее характерна для данного местоимения.

Такое *с глаголами речи и с другими предикатами: важные различия*

Казалось бы, местоимение *такое* нередко выступает и не при глаголах с сентенциальным актантом. Приведём несколько примеров:

(15) *Смотрел «Русский сувенир» Гр. Александрова. Как можно такое снимать?* [Василий Катанян. Лоскутное одеяло (1943–1999)]

(16) *— Обязательно попробуете! Но потом, когда мы уже… — последовала запинка и гримаска, означавшие поиск слова, — когда мы уже поладим. Если такое пить сразу, то это уже отчасти поддавки.* [Георгий Полонский. Роль в сказке для взрослых или «Таланты и Полковники» (1970–1980)]

Между двумя типами примеров — с глаголами, присоединяющими сентенциальный актант, и с другими предикатами, — есть два различия. Первое, так сказать, статистическое, состоит в том, что вне группы глаголов с сентенциальным актантом *такое* встречается гораздо реже. Среди примеров на *такое* + *инфинитив* большинство иллюстрируют глаголы с сентенциальным актантом.

Второе различие более существенно. Оно заключается в том, что для обозначения конкретного объекта *такое* встречается только при глаголах с сентенциальным актантом. Заметим, что в двух примерах выше речь идёт не о конкретной ситуации, а о целом классе ситуаций (хотя исходно суждение базируется на одной конкретной ситуации). Так, высказывание *Как можно такое снимать?* означает ‘Как можно снимать фильмы такого типа’, хотя поводом для высказывания послужил фильм «Русский сувенир». Точно так же в (16) речь идёт не о конкретном напитке, а о классе напитков, включающих данный напиток и похожие на него.

Напротив, в примере (12) высказывание бригадира *Я такого не говорил* вовсе не означает ‘Я не говорил, что мой брат не виноват, а также ничего похожего на это’. Бригадир имеет в виду, что не произносил ровно одного конкретного высказывания — *Мой брат не виноват, а злодеи его погубили*. Точно так же в (17):

(17) *— А вот насчет непригодности словаря для изучения частотности, употребимости и стилистической окраски позвольте с Вами не согласиться. — А я такого не утверждал.* [http://community.livejournal.com/

ru_ivrit/1889289.html#comments] — *такое* обозначает содержание конкретного высказывания (*Словарь непригоден для изучения частотности*), а не класса высказываний, описывающих словари и их использование.

Тем самым, при глаголах с сентенциальным актантом и при других типах предикатов различен референциальный статус актанта, который заменяется *такое*: только при глаголах с сентенциальным актантом он может быть конкретно-референтным. С чем же связано такое различие? По-видимому, нельзя сказать, что мы имеем дело с разными значениями *такой* (усмотреть разницу в лексических значениях *такой* в (13)–(14) и в (15)–(16) сложно). Скорее причина в том, что сентенциальные актанты (точнее, суждения, которые они выражают) и предметные имена в целом имеют различную природу.

Среди объектов (особенно среди материальных предметов) чётко выделяются классы, которые мы и называем некоторым словом в родовом употреблении (см. Падучева 1985), а в классах — их отдельные представители. Например, среди фруктов легко выделить подкласс яблок, а в этом подклассе — конкретное яблоко, которое мы при желании можем обозначить словом *яблоко* в конкретно-референтном употреблении в понимании Е. В. Падучевой. Тем самым, каждый конкретный объект всегда хорошо отделим от других предметов того же класса и от предметов других классов.

Напротив, ситуации — в частности, такие, которые занимают место дополнения при глаголах с сентенциальными актантами — имеют более сложные свойства. Само высказывание — например, *Мой брат не виноват!* — может быть отделено от других тождественных ему высказываний (в частности, по тому, кто является его автором, когда и в каких условиях оно сделано). Однако по форме и содержанию любое высказывание имеет множество «близнецов» — тождественных ему высказываний (ту же фразу *Мой брат не виноват!* с той же интонацией и семантикой в другое время мог произнести другой человек, помимо бригадира в примере (12)). Ещё больше набор похожих высказываний станет в том случае, если мы учтём синонимичные высказывания, не тождественные данному по форме (например, *Мой брат невиновен!*). Тем самым, употребление местоимения *такой* по отношению к одному конкретному высказыванию может быть связано с тем, что это высказывание всё равно является представителем класса синонимических высказываний, совпадающих или не совпадающих между собой по форме.

Теперь можно ответить на другой вопрос: почему *такое* чаще встречается с глаголами речи, чем с другими предикатами? Вероятно, потому, что ситуацию естественнее обозначить через класс, который она представляет (например, в *Как можно такое говорить учителю на уроке* — класс некоторых фамильярных и оскорбительных высказываний). Для объектов это менее характерно.

Впрочем, как уже говорилось, местоимение *такое* может обозначать и предметы.

(18) *В каждом буклете — статья об обстоятельствах, сопутствовавших выходу альбома: пишет их, к примеру, Александр Кушнир. Такое нельзя не купить.* [Алексей Мунипов. Про животных и людей. Обзор CD (2002) // «Известия», 2002.04.29]

Истолковать его можно примерно так: *такое* = ‘имеющее свойства, как у данного предмета’. Например, в (16) *такое* = ‘все напитки с такими свойствами’, в (18) — ‘все буклеты с такими свойствами (с дисками высокого качества и подробными описаниями)’ и т. п. Вероятно, основываясь на данном толковании, можно уточнить объяснение того, почему *такой* чаще обозначает ситуации: в отличие от предметов, ситуация всегда употребляется как представитель класса ситуаций с одними и теми же свойствами.

Местоимение такое *и местоимение* такой *в сочетании с существительным*

Заметим, что местоимение *такой* в функции определения к существительному (*такие люди, такой город, такие мысли*) отличается по свойствам от *такое* в функции существительного (без определения). *Такой + существительное* употребляется для обобщения утверждения, сделанного ранее об одном определённом предмете. Напротив, *такое* употребляется так только в контекстах снятой утвердительности:

(19) *Дом был серый и мрачный. Такие дома строили при Сталине.*

(20) *??? Дом был серый и мрачный. Такое строили при Сталине.*

Однако возможно:
(21) *Дом был серый и мрачный. При Хрущёве такое / такого не строили.*

Тем самым, предложенное выше толкование не покрывает всех свойств местоимения *такое*. Нам кажется, что два свойства местоимения *такое*:

1. неопределённость
2. способность выступать преимущественно при снятой утвердительности — несомненно, коррелируют между собой.

Сравним, например, употребление английского местоимения *any*. С одной стороны, оно способно употребляться при отрицании, а с другой, маркирует неопределённость объекта (‘любой, какой угодно’).

Местоимения это *и* так *и референция*

Вернёмся к местоимениям *так* и *это*. Местоимение *это* очевидным образом не может обозначать ни любое высказывание с данным содержанием, более того, для него крайне нехарактерно обозначать даже любое высказывание с данными формой и содержанием или любую ситуацию, идентичную данной. Оно относится именно к данному «экземпляру» высказывания или вообще ситуации. В этой связи естественно, что *это* допускает выражение таких параметров речевого акта, как голос, интонация говорящего, его сопутствующие жесты: определение конкретных параметров ситуации невозможно без подобных второстепенных параметров.

Наиболее ясно эта особенность местоимения *это* проявляется на примере глаголов *случиться* и *произойти* — они не являются глаголами речи, но их субъектом тоже является событие (предмет *произойти* не может). Приведём примеры на каждое из местоимений:

(22) *Да, однажды* <u>*такое*</u> *случилось: я забыла слова гениальной грузинской песни «Сулико»!* [Светлана Ткачева. Тамара Гвердцители: «Не умею учиться на чужих ошибках» (2003) // «100% здоровья», 2003.01.15]

(23) *Правда, однажды* <u>*такое*</u> *случилось, когда у Петьки Халютина ощенилась собака, и я ходил смотреть щенят.* [Виталий Губарев. Трое на острове (1950–1960)]

(24) *Если* <u>*такое*</u> *произошло, сразу уходи вбок от двери и вниз, ведь незваные «гости» могут открыть стрельбу.* [Место схватки — подъезд (2004) // «Солдат удачи», 2004.05.05]

(25) *Быть без вины виноватыми перед людьми, с которыми, казалось, контакт и взаимопонимание были полными, обидно.* <u>*Так*</u> *случилось с Темиркановым.* [Сати Спивакова. Не всё (2002)]

(26) *В случае же наших «застольных» песен ни текстовая, ни сюжетная память не действенны.* <u>*Так*</u> *произошло с песнями типа «Хаз-Булат удалой», «На Муромской дорожке», «По Дону гуляет» и др. Вследствие каких закономерностей происходят эти метаморфозы?* [Олег Николаев. Новый год: праздник или ожидание праздника? // «Отечественные записки», 2003]

(27) *Серьёзные занятия начались после моей киношной карьеры.* <u>*Это*</u> *случилось так. В 1937 году, в школу, где я учился, нагрянула съёмочная группа <…>.* [Марк Тайманов. Эмоции неисчерпаемы (2003) // «64 — Шахматное обозрение», 2003.10.15]

(28) *Однако при этом оказалось, что листочки отклонились ещё больше, вместо того чтобы приблизиться друг к другу. Почему* <u>*это*</u> *произошло?* [Владимир Лукашик, Елена Иванова. Сборник задач по физике. 7–9 кл. (2003)]

Местоимение *такое* в данных примерах подчиняется ранее отмеченным закономерностям: в примерах (22) и (23) *такое* обозначает целый класс ситуаций (слово *однажды* только подчёркивает, что, вообще говоря, ситуаций такого рода могло быть несколько). В (24) *такое* выступает в условном придаточном, а условие — это один из контекстов снятой утвердительности.

В примерах с *так* явного указания на класс ситуаций нет, но в обоих случаях (как и во множестве других примеров) в предтексте содержится общее утверждение, *так* обозначает любую ситуацию, отвечающую некоторым признакам (например, ‘случаи, когда текстовая и сюжетная память не действенны’), хотя и иллюстрирует её одним конкретным примером.

Напротив, *это* в (27) и (28) обозначает конкретный случай — так, в (27) *это* относится к ситуации, когда Марк Тайманов стал заниматься шахматами, снявшись в кино — но не к любой ситуации такого рода. В противном случае мы скорее использовали бы *так*:

(29) *Многие подростки приходят в шахматы из кино. Так случилось и со мной.*

Образуют ли местоимения шкалу определённости?

Из ранее сказанного, казалось бы, следует, что *это, так* и *такое* формируют своего рода шкалу определённости. *Это* обозначает конкретно-референтное событие или факт. В частности, при глаголах речи *это* относится к конкретному речевому акту со всеми его характеристики:

(30) *Сказано это было с таким апломбом, что я даже растерялся.*

Так, как было сказано выше, тоже может относиться к конкретному речевому акту, однако не фиксирует его характеристики:

(31) **Он так сказал с апломбом.*

(32) *Он так мне сказал вчера.*

Это связано с тем, что *так* относится не к речевому акту, а только к его содержанию.

Казалось бы, можно сказать, что *такое* обозначает ещё менее определённое высказывание, чем *так*: мы уже сказали, что для него характерны контексты типа (33), а не (34):

(33) *Такое было сказано вчера.*

(34) *Разве можно такое говорить*?

Однако в действительности *это*, *так* и *такое* противопоставлены не по одному признаку (определённость / неопределённость), а по нескольким. Отметим, что *такое* (как и *это*) способно выступать в случаях, если отмечены характеристики речевого акта:

(35) *Если бы такое (это, *так) было сказано с насмешкой, я бы обиделся.*

Тем самым, важны две характеристики каждого из местоимений:

1. *это* и *такое*, но не *так*, обозначают высказывание или другую ситуацию, вместе с условиями, в которых возникла ситуация или было произнесено высказывание;

2. *это* обозначает конкретный речевой акт, *такое* — нереферентный или не существовавший в реальности, *так* способно и к референтному, и к нереферетному употреблению.

Заключение

Итак, как мы выяснили, к имени ситуации в русском языке могут отсылать три местоимения: *так*, *это* и *такое*. Каждое из этих местоимений имеет свои особенности.

Местоимение *этот* может не только обозначать собственно ситуацию, но и фиксировать характеристики речевого акта, содержащего упоминание о ней. При этом местоимение *этот* отсылает именно к конкретному речевому акту, а не к совокупности речевых актов.

Местоимение *так* с глаголами речи обозначает либо конкретное высказывание, либо класс высказываний (при этом местоимение не конкретизирует условия, в которых высказывание было порождено. За пределами глаголов речи (например, с глаголами *случаться, происходить*), как правило, употребляется интенсионально, то есть обозначает ‘ситуацию, обладающую данными семантическими признаками, во всех её проявлениях’.

Наконец, местоимение *такое* интересным образом также имеет два разных употребления. Если оно относится к предметам, оно обозначает целый класс предметов, подобных данному. Напротив, по отношению к ситуации оно может выражать её конкретный «экземпляр», как в примерах (12) и (17). Данное различие связано с другим, более общим — а именно, с разной природой предметов и ситуаций. При этом по отношению к конкретной ситуации местоимение *такое* употребляется преимущественно в контекстах снятой утвердительности.

Ключевой вывод нашего доклада заключается в том, что система местоимений, обозначающих ситуацию, устроена не так, как система кодирования предметов. Именно в системе «ситуационных» местоимений наиболее важны признаки определённости и референтности[4]. Это связано с неоднозначной сущностью ситуаций и, в частности, высказываний. Высказывания могут быть конкретно-референтными как на уровне формы и содержания (в этом случае конкретно-референтным является класс высказываний с данным конкретным содержанием и, возможно, формой), так и на уровне собственно события, в частности, речевого акта (тогда конкретная референция относится к данному проявлению ситуации). Рассмотренные нами местоимения различают данные типы референции.

---

[4] Нет нужды говорить, что и при обозначении предметов сказываются признаки референтности и определённости (см. особенно [Падучева 1985], [Тестелец, Былинина 2005]), только проявляются они преимущественно в системе неопределённых, а не собственно анафорических местоимений.

References

1. *Apresian Iu. D.* 2001. Generating Meanings ´TO KNOW´ AND ´TO THINK´ in Russian Language [Sistemoobrazuiushchie Smysly 'ZNAT" I 'SCHITAT" v Russkom Iazyke].Russkii Iazyk v Nauchnom Osveshchenii, 1 : 5–26.
2. *Kibrik A., Prozorova E.* 2007. Referential Choice in Signed and Spoken Languages. DAARC 2007 (6th Discourse Anaphora and Anaphor Resolution Colloquium). Proceedings.
3. *Krasavina O. N.* 2004. The Use of Indication Group in Russian Narrative Discourse. Voprosy Iazykoznaniia, 3 : 51–68.
4. *Kreidlin G. E., Chekhov A. S.* 1988. Correlations of Semantics, Actual Segmentation and Pragmatics in the Lexicographical Description of Anaphoric Pronouns (On the Material of the Pronoun of the Group 'TOT') [Sootnoshenie Semantiki, Aktual'nogo Chleneniia I Pragmatiki v Leksikograficheskom Opisanii Anaforicheskikh Mestoimenii (Na Materiale Mestoimeniia Gruppy 'TOT')]. Institut Russkogo Iazyka AN SSSR. Problemnaia Gruppa po Eksperimental'noi I Prikladnoi Lingvistike. Predvaritel'nye Publikatsii,178.
5. *Shvedova N. Iu.* 1980. Russian Grammar [Russkaia Grammatika].
6. *Kulikov L. I.* 1985. On the Interchangeability of Anaphoric Pronouns 'ETOT' and 'TAKOI' [O Vzaimozameniaemosti Anaforicheskikh Mestoimenii 'ETOT' i 'TAKOI']. Vestnik Moskovskogo Universiteta, 9 (1) : 72–74.
7. *Paducheva E. V.* 1985. Statement and its Correlation with the Reality: Referential Aspects of Pronouns Semantics [Vyskazyvanie I ego Sootnesennost' s Deistvitel'nostiu: Referentsial'nye Aspekty Semantiki Mestoimenii].
8. *Poesio M., Modjeska N.* 2002. The THIS-NPs Hypothesis: A Corpus-Based Investigation. DAARC 2002 (1th Discourse Anaphora and Anaphor Resolution Colloquium). Proceedings.
9. *Testelets Ia. G., Bylinina E. G.* 2005. Some Constructions with the Meaning of Indefinite Pronouns in Russian Language [Nekotorye Konstruktsii so Znacheniem Neopredelennykh Mestoimenii v Russkom Iazyke].

О НЕКОТОРЫХ ГЛАГОЛАХ С НЕАССЕРТИВНЫМИ ЗНАЧЕНИЯМИ[1]

И. Б. Левонтина (irina.levontina@mail.ru)

Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Значение слова определяется не только тем, из каких компонентов оно состоит, но и тем, каков статус каждого компонента в логической структуре значения данного слова. Существует несколько семантических классов глаголов, у которых, во всяком случае в части употреблений, отсутствует полноценная ассерция, а все значение состоит из разного рода неассертивных компонентов (в частности, пресуппозиций и модальных рамок). Например, это глаголы, в которых содержится указание на некие моральные нормы или запреты, которые нарушает или не нарушает субъект: *(не) побрезговать, (не) погнушаться, (не) полениться, (не) постесняться* и т. п., глаголы, связанные с готовностью или неготовностью человека прилагать какие-либо усилия для совершения действия: как *удосужиться, потрудиться, позаботиться (Не позаботился оформить загранпаспорт)* и др. За последние два века у некоторых глаголов сформировались неассертивные значения, у других они усложнились, так что в целом в русском языке образовался мощный класс подобных сложных в семантическом и прагматическом отношении глаголов.

Ключевые слова: неассертивное значение, глагол, ассерция, семантический класс.

[1] Работа выполнена при финансовой поддержке Программы фундаментальных исследований отделения историко-филологических наук РАН «Генезис и взаимодействие социальных, культурных и языковых общностей», гранта НШ-4019.2010.6 для поддержки научных исследований, проводимых ведущими научными школами РФ, и гранта РГНФ № 10-04-00273а.

Данная работа едва ли была бы возможна без помощи Национального корпуса русского языка (ruscorpora.ru).

ON SOME NON-ASSERTIVE VERBS

I. B. Levontina (irina.levontina@mail.ru)

Russian Language Institute, Russian Academy of Sciences, Moscow, Russian Federation

The meaning of the word is determined not only by the components it consists of but also by the status of each component in the logical structure of this word's meaning. The paper deals with a group of Russian verbs with a very peculiar logical structure and unusual syntactic properties. Their meaning is confi ned to non-assertive components, while the assertion is conveyed by the subordinate verb. In their semantic structure they are therefore similar to some discourse markers (particles, etc.). The verbs in question are udat'sia'manage', ugorazdit', udosuzhit'sja, spodobit'sja, zablagorassudit'sja, soizvolit', soblagovolit', posmet' [≈dare], imet' smelost', vzjat' (vzjal i sdelal) etc., most of them hard for translation. Some of such phrases can be approximately translated into English with the verb to do [On spodobilsia prij ti ≈ He did come]. Thus the meaning of the sentence On soblagovolil prij ti is confi ned to the message 'He came' and a combination of speaker's attitudes and expectations. Partly these verbs are negative polarity items (e. g. udosuzhit'sja), while others have positive polarity (e. g. ugorazdit'). Special attention is given to the verbs udosuzhit'sja and potrudit'sja, which express the idea of being ready to make efforts. Interestingly, the meaning of these two verbs, including its logical structure, has be enchanging during the last 200 years. The paper demonstrates how their actual meaning has taken shape.

Key words: non-assertive verb, verb, assertion, semantic classe.

Вводные замечания

Хорошо известно, что значение слова определяется не только тем, из каких компонентов оно состоит, но и тем, каков статус каждого из компонентов в логической структуре значения данного слова. Так, по принятым в Московской семантической школе в настоящее время представлениям, толкование складывается из пяти частей, ни одна из которых не является обязательной: ассерции, пресуппозиции, модальной рамки, рамки наблюдения и мотивировки; см. [Апресян 2009]. При этом замечательно, что не является обязательной, в частности, и ассерция. Особенно неассертивные значения характерны для модальных частиц и других дискурсивных слов, однако не только для них. Ср., например, следующее толкование Ю. Д. Апресяна: *X-а угораздило сделать P* ≈ (X сделал P [ассерция]; говорящий не понимает, как получилось, что X сделал P, потому что он считает очевидным, что P плохо для самого X-а [модальная рамка]) [Апресян 2009: 514]. Как мы видим, толкуется выражение *X-а угораздило сделать P*, причем в ассерции здесь находится компонент 'X сделал P', то есть, в сущности, значение подчиненного глагола.

Заметим, что в работе [Зализняк Анна, Левонтина 1996] для *угораздило* предлагалось другое толкование, в котором была полноценная ассертивная часть: ‘Р плохо’[2]. Во многих подобных случаях локализация смысловых компонентов является спорной, поскольку слова с оценочными значениями, сложными пресуппозициями и т. п. зачастую плохо сочетаются с отрицанием, что делает невозможной стандартную логическую процедуру установления ассертивной части значения как части, которая попадает под отрицание.

Кроме того, хорошо известно, что у некоторых слов отдельные смысловые компоненты имеют неустойчивый логический статус и могут, в зависимости от контекста, перемещаться то в ассерцию, то в пресуппозицию. Такие случаи разбираются, в частности, в работах [Кустова 1996; Зализняк Анна, Левонтина 1996; Падучева 2005; Апресян 2006; Апресян В. 2010].

Обнаружилось, что эти разрозненные факты складываются в определенную систему: существует несколько семантических классов глаголов, для которых характерна такая структура значения, при которой, во всяком случае в части употреблений, отсутствует полноценная ассерция, а все значение состоит из разного рода неассертивных компонентов (в частности, пресуппозиций и модальных рамок). Ни в коей мере не претендуя на полноту, укажем некоторые из этих групп.

Уже упомянутый глагол *угораздило* является представителем целого класса слов, объединенных идеей неполного контроля человека над ситуацией: *удалось, посчастливилось*[3], *довелось, умудрился* и др.[4]; см. о них в [Зализняк Анна, Левонтина 1996].

Близкую группу составляют оценочные единицы типа *иметь счастье / несчастье / удовольствие*[5]. В данном случае особенно бросается в глаза сходство с дискурсивными словами; ср. *Я имел несчастье опоздать* и *К несчастью, я опоздал*.

Совершенно замечательна конструкция вида *взял и сказал* с трудноуловимым значением. О заполнении валентности при некоторых глаголах сочиненной группой см. [Богуславский 1996]. Ср. также сочетания типа *давай говори*, в которых, впрочем, глагол уже далеко продвинулся на пути превращения в частицу; см. [Левонтина 2005].

Интересную группу, в которой сложным образом взаимодействуют ассертивные и неассертивные компоненты значения, составляют некоторые глаголы намерения: *вздумать, вздуматься, собраться, заблагорассудиться, взбрести в голову*. Так, в работе [Зализняк Анна, Левонтина 1996] отмечалось, что фраза *Не собрался ей написать* не означает, что человек не имел такого намерения.

---

[2] При этом в указанной работе были другие глаголы, для которых предлагались толкования с выхолощенной ассерцией ‘Х сделал Р’: *удаться, успеть* и др., или ‘Р произошло’ — например, *умудриться*.

[3] В отличие от *повезло*, в котором оценка в ассерции; ср. [Зализняк Анна, Левонтина 1996; Апресян В. 2010].

[4] Сюда же относится и пушкинское *Догадал меня черт родиться в России с душою и талантом*.

[5] Ср. также несколько иное *иметь честь*.

Намерение он имел, но не сделал. Ср. также о таких словах, как *вспомнить, спохватиться* и др. в [Иомдин 2010][6].

Еще одна важная в рассматриваемом отношении группа — это глаголы *благоволить (Благоволите открыть!), соблаговолить, изволить, соизволить*. Так, смысл фразы *Он соблаговолил прийти* сводится к сообщению о его приходе (эта часть смысла локализуется в подчиненном предикате) и совокупности разного рода ожиданий и оценок говорящего.

Несколько особняком стоит яркое слово *сподобиться,* которое в разных употреблениях примыкает к разным группам. В прямом значении *сподобиться* имеет смысл, близкий к ‘удостоиться’, в ряде употреблений сближается с *довелось* (*На старости лет сподобилась увидеть море*), в других — с *удосужиться* (*Наконец сподобился принести дневник*).

Пожалуй, самую большую и разнообразную группу составляют глаголы, в которых содержится указание на некие моральные нормы или запреты, которые нарушает или не нарушает субъект: *(не) побрезговать, (не) погнушаться, (не) полениться, (не) постесняться, (не) сметь, (не) осмелиться, (не) позволить себе, иметь наглость* и т. п. Очень показательно, например, сравнение сочетания *иметь наглость (сделать что-то)* и глагола *обнаглеть*. Ассерция выражения *иметь наглость сделать что-то* целиком содержится в значении глагола, заполняющего вторую валентность при *иметь наглость*, значение же самого этого сочетания состоит только из модальной рамки, то есть неассертивного компонента. В глаголе *обнаглеть* аналогичный смысл составляет ассертивную часть его значения: *Он опять не помыл посуду? Совсем обнаглел*. Впрочем, при другой синтаксической организации высказывания *обнаглеть* сближается с *иметь наглость*: *Он обнаглел до того, что отказывается мыть за собой посуду* [ср. *Он имел наглость отказаться мыть за собой посуду*].

Интересно также сравнить глаголы *посметь* и *осмелиться*[7]. Оба они имеют неассертивные значения. Фразы *Он осмелился <посмел>* и *Он не осмелился <не посмел>* в равной степени указывают на то, что для совершения данного действия существовали определенные препятствия. Однако если у глагола *осмелиться* значение состоит из пресуппозиции ‘совершению данного действия субъектом препятствуют моральные запреты или страх’, то глагол *посметь* в большинстве употреблений имеет значение, состоящее скорее из модальной рамки: ‘говорящий считает, что данное действие опасно[8] или недопустимо с моральной точки зрения’. Естественно, что в первом лице это различие нейтрализуется. В императиве же в силу данного различия естественнее всего используется именно глагол *сметь* с отрицанием (*Не смей так говорить* — то есть, ‘Не говори так, потому что это недопустимо’). Особо стоит отметить

---

[6] К этой группе примыкают некоторые употребления глагола *спешить (поспешить)*: *Он не спешил возвращать долг*.

[7] Разумеется, можно упомянуть и близкие выражения *собраться с духом, решиться, отважиться.* Каждое из них имеет свои особенности.

[8] Этот компонент смысла обслуживает контексты типа *Он не посмел возражать отцу.*

слегка архаичный круг деонтических употреблений данного глагола в высказываниях типа *Ты не смеешь так говорить* — то есть, ты не должен себе этого позволять[9]. Следующие две фразы могут быть синонимичны:

А) Он смеет так обращаться с матерью! [Он обращается с матерью таким образом, и говорящий считает, что такое обращение недопустимо];

Б) Он не смеет так обращаться с матерью! [Недопустимо обращаться таким образом с матерью]. Фраза Б) имеет и другое понимание — он не позволяет себе так обращаться с матерью.

При этом формы совершенного вида *посметь* и *осмелиться* естественно понимать в том смысле, что человек совершил соответствующее действие. Вообще говоря, перераспределение смысла между пресуппозицией и ассерцией в разных видовых формах глагола совершенно естественно. Особенность рассматриваемых случаев в том, что в некоторых употреблениях у данных глаголов может не оставаться полноценной ассерции.

Глаголы *стыдиться (постыдиться), гнушаться (погнушаться), брезговать (побрезговать), стесняться (постесняться), лениться (полениться), бояться (побояться)* и некоторые другие устроены с точки зрения логической структуры еще более сложно. Они имеют разные понимания, причем отчасти это связано с видовыми формами как самих этих слов, так и подчиненных глаголов; ср. *Он постеснялся сказать / Он стеснялся сказать / Он постеснялся говорить (о себе)* — то есть, не сказал или не говорил о себе из-за чувства неловкости. Таким образом, в значение глагола *стесняться* здесь входит оператор ‘не’ и пресуппозиция ‘испытывая неловкость’. Фраза *Он стеснялся говорить о себе* двузначна: она значит либо ‘стеснялся, говоря о себе’ [ассертивное понимание глагола *стесняться*, само действие в пресуппозиции], либо ‘не говорил из-за чувства неловкости’ [в ассерции значение глагола, заполняющего вторую валентность *стесняться*, значение самого *стесняться* включает пресуппозицию и модальный оператор]. При этом близкие к стесняться глаголы *смущаться* и *конфузиться*, как отмечено в статье Ю. Д. Апресяна **СТЫДИТЬСЯ** в НОССе [Апресян 2004], вообще с трудом управляют инфинитивом. Ср., впрочем, приводимый в указанной работе пример: *А дама не двигается. И конфузиться докушивать* (М. Зощенко, Аристократка). Аналогичные и тоже не вполне стандартные для современного языка примеры есть и для глагола *смущаться*. Ср. *В 4 1/2 была Елка офицерам; бедная Аликс, понятно, смущалась разговаривать со всеми!* [Николай II. Дневники 1894–1896 (1894–1896)]; — *Я хотела вас увидеть, услышать ваш голос, — лживо пробормотала я, смущаясь объяснить ей,*

[9] Аналогично ведет себя слово *стыдно*, которое может указывать как на чувство, которое человек испытывает, так и на чувство, которое он, по мнению говорящего, должен испытывать; ср. *Вам не стыдно так поступать* [‘Вы не испытываете стыда?’] и *Стыдно вам так поступать!* [‘Вы должны испытывать стыд’]. Ср. выделение особого деонтического *стыдно* в [Булыгина, Шмелев 2000].

что я вовсе не стремлюсь послушать ее стихи, а одержима эгоистическим желанием почитать ей свои. [Н. Воронель. Без прикрас. Воспоминания (1975–2003)].

Ср. также обсуждение похожих свойств глаголов *хотеть* и *бояться* в [Зализняк Анна 1992] и в статьях Ю. Д. Апресяна **ХОТЕТЬ** и **БОЯТЬСЯ** в НОССе [Апресян 2004]. Например, возможны два понимания для фразы *Я боюсь летать на самолете* и одно — для фразы *Он побоялся лететь на самолете*, как и для фразы *Он не захотел помочь*. Немного другая ситуация с глаголом *мочь* обсуждается в [Булыгина, Шмелев 1999]. Фразы *Я рад, что мог вам помочь* и *Ты же мог ему помочь!* различаются импликатурами: в первом случае имплицируется, что помог, во втором — что не помог.

Далее будет чуть более подробно рассмотрены два представителя группы глаголов, связанной с готовностью или неготовностью человека прилагать какие-либо усилия для совершения действия. Это такие единицы, как *удосужиться, потрудиться, трудиться (не трудитесь), дать себе труд, почесаться (А он и не чешется возвращать долг), позаботиться (Не позаботился оформить загранпаспорт)* и др. При этом нас будет интересовать не только то, как в значении глаголов распределены ассертивные и неассертивные компоненты смысла, но и то, как это распределение изменялось на протяжении последних двух веков.

УДОСУЖИТЬСЯ

Примеры.

Во время прошлой беременности она даже не удосужилась завести медицинской карточки, не делала никаких там положенных анализов [Л. Улицкая. Путешествие в седьмую сторону света // Новый Мир, № 8–9, 2000]

Жаль только, что я не удосужился спросить у профессора, что такое шизофрения. [М. А. Булгаков. Мастер и Маргарита, часть 1 (1929–1940)]

Многие его мемуары уже долгое время существовали в форме вполне законченных устных «новелл», прежде чем он удосужился их записать. [К. И. Чуковский. Репин — писатель (1930–1950)]

Ф. Ф. Кузнецов не только не выполнил моей просьбы, но и ответить удосужился лишь через пять месяцев. [С. Резник. «Выбранные места из переписки с друзьями» (2003) // «Вестник США», 2003.10.15]

Если бы в то время кто-нибудь удосужился составлять рейтинги, то эти реликтовые программы завоевывали бы 98 % аудитории. [А. Беляков. Алка, Аллочка, Алла Борисовна (1998)]

— Дурак ты, — говорит старик. — Хоть бы на карту нашу удосужился взглянуть. Нет никакого Северного Архипелага… [А. и Б. Стругацкие. Жук в муравейнике (1979)]

Как видно из приведенных примеров, слово *удосужиться* характеризуется слабой отрицательной поляризованностью. Чаще всего оно употребляется в контексте эксплицитного отрицания, но также и в и в разного рода гипотетических, вопросительных, условных, модальных, уступительных и т. п. предложениях.

Вторая валентность *удосужиться* обычно заполняется инфинитивом. Однако возможно, что она может заполняться и сочиненной группой (*Хоть бы раз удосужился и позвонил бабушке!*). Удалось пока обнаружить только один реальный пример, притом несколько сомнительный[10]; ср. *Я хотел достать однотомник произведений Маяковского. Неоднократно говорил и Муле, и Вильмонту. Но никто не удосужился и не вспомнил.* [Г. С. Эфрон. Дневники. Т. 1. 1941 (1941)]

Толкование: *X удосужился сделать P ‘X сделал P [ассерция]; говорящий считает очевидным, что X-у давно следовало сделать P и что у X-а не было никаких причин не делать P [модальная рамка]’.*

Таким образом, *удосужиться* — глагол, не имеющий собственной ассерции. Так, однако, было не всегда.

Исторический экскурс.

В ранних примерах, конца XVIII — начала XIX в., глагол *удосужиться* встречается в значении ‘освободиться, завершить необходимые дела’; ср.

Не ставя в порок невинных удовольствий, я, удосужась от дел, езжал в клоб танцевать [И. М. Долгоруков. Повесть о рождении моем, происхождении и всей моей жизни… (1788–1822)]

В назначенное время, когда други наши все удосужились, князь Гаврило начал продолжение своего повествования следующим образом: [В. Т. Нарежный. Российский Жилблаз, или Похождения князя Гаврилы Симоновича Чистякова (1814)]

Мне посчастливилось удосужиться и вот я попал под вечер в семейный кружок, оторвавшись на несколько часов от лагерных занятий. [Константин Константинович (К. Р.). Письма И. А. Гончарову (1888)]

Данное значение к концу XIX в., видимо, устарело, возможно, еще какое-то время оставаясь в просторечии; ср. следующий пример, где *удосужиться* заключено в кавычки, что очень показательно:

[10] Сомнительный не в смысле допустимости, а в том смысле, что синтаксически можно считать, что имеется в виду, что никто *не удосужился прислать* и *не вспомнил* об этом. Однако сконструированный пример с бабушкой кажется вполне естественным.

Через Мишку Савелий узнал весь порядок генеральского дня, а главным образом, когда Мотька бывает свободной. Таких минут, когда Мотька могла «удосужиться», было, правда, немного, и их приходилось ловить. [Д. Н. Мамин-Сибиряк. Верный раб (1891)]

На базе этого значения развивается новое, более распространенное: ‘Найти свободное время для чего-л.’[11].

На вопрос императрицы, поедет ли государь в тот вечер в театр, он отвечал, что еще не знает, удосужится ли, потому что должен писать фельдмаршалу (князю Варшавскому). [М. А. Корф. Записки (1838–1852)]

Виноват я, что так давно не отвечал на твое письмо, привезенное мне Виктором. Но я очень занят и вот едва удосужился сказать тебе несколько слов. [И. А. Гончаров. Письма (1842–1859)]

Друг мой! удосужься, напиши к ней! [А. О. Корнилович. Михаилу Осиповичу Корниловичу (1833)]

Мне тоже, как и всегда, обо многом хочется говорить с вами. Когда удосужусь, напишу длиннее. [Л. Н. Толстой. Письма. 1894 (1894)]

Когда я, удосужившись, навестил свою благодетельницу, она даже попеняла мне, что долго не шел. [И. Е. Репин. Далекое близкое (1912–1917)]

Глагол *удосужиться* имел также безличное употребление, хотя оно встречалось и нечасто:

На Радуницу надо бы матушке Аркадии иную книгу в келарню внести, да за хлопотами ей не удосужилось. [П. И. Мельников-Печерский. В лесах. Книга первая (1871–1874)][12]

Как мы видим из приведенных примеров, долгое время значение глагола *удосужиться* не было оценочным. Однако уже в XIX в. у него стало развиваться ироническое употребление; ср.

Даже черногорский князь удосужился и съездил в Вену, где тоже был «сердечно» принят. Что все это означает, как не фабрикацию испугов в умах и без того взбудораженных простецов? [М. Е. Салтыков-Щедрин. Мелочи жизни (1886–1887)]

---

[11] Ср. современное выражение *выбрать время сделать что-л.*

[12] Ср. также следующий пример, по-видимому стилизованный: *Если б вам удосужилось бросить взгляд на то важное место, вы видели б только лак, только лоск* [Андрей Белый. Петербург (1913-1914)].

На базе этого иронического употребления и сложилось современное модальное значение глагола *удосужиться*[13].

ПОТРУДИТЬСЯ[14]

Примеры.

Характерно, что следствие даже не потрудилось установить источник поступления денег на счет Обозинцева. [А. Кучерена. Бал беззакония (2000)]

Галочка даже урну с теткой, сожженной семь лет назад в крематории, не потрудилась отнести на кладбище — как-то все было не до того. [Н. Катерли. Брызги шампанского (1998) // «Звезда», 2000]

Вы хотя бы потрудились подойти поближе к поврежденному вагону и осмотреть его, прежде чем начать дезинформировать людей. [Хроника происшествий (форум) (2007.03.04)]

— Голос спокойный, тембр приятный. Вежливый господин — потрудился выяснить отчество. — Я вас слушаю. [М. Баконина. Девять граммов пластита (2000)]

Как мы видим из приведенных примеров, *потрудиться* во многих случаях очень близко к *удосужиться*; ср. *даже не потрудились выяснить / даже не удосужились выяснить*. Для данного глагола также характерна слабая отрицательная поляризованность.

Толкование*: X потрудился сделать P 'X сделал P [ассерция]; говорящий считает очевидным, что X-у следовало сделать P и что от X-а не требовалось больших усилий для совершения P [модальная рамка]'.*

Как следует из приведенных толкований, если *удосужиться* намекает скорее на безалаберность или безответственность субъекта, то *потрудиться* — на его лень.

---

[13] Параллельно — и даже, видимо, раньше — развивалось оценочное значение и у прилагательного *досужий*; ср. старый пример: *Я приехал сюда не для масленой, а единственно провести несколько досужих дней с моим приятелем*. [В. Т. Нарежный. Российский Жилблаз, или Похождения князя Гаврилы Симоновича Чистякова (1814)].

[14] Здесь не будет рассматриваться прямое значение данного глагола — 'поработать'; ср. *Сегодня мы хорошо потрудились; Или не получалось, или было лень потрудиться, преодолеть себя…* [Михаил Козаков. Актерская книга (1978-1995)].

Удосужиться и *потрудиться* существенно различаются в следующем отношении. Для *потрудиться* очень типично использование в побудительных предложениях; ср.:

— Ваше присутствие на похоронах отменяется, — продолжал кот официальным голосом. — Потрудитесь уехать к месту жительства. [М. А. Булгаков. Мастер и Маргарита, часть 1 (1929–1940)]

Можно, конечно, подняться, зажечь свет и громко предложить своим гостям, как предлагает обычно Лев Семенович: «Потрудитесь выйти вон!» [В. Токарева. Ни сыну, ни жене, ни брату (1984)]

Это различие семантически объяснимо. *Удосужиться* включает идею временного отрезка, на протяжении которого действие может быть выполнено (ср. смысловой компонент ‘раньше’), который снижает эффективность побудительного высказывания. В *потрудиться* же этой идеи нет.

Исторический экскурс.

Значение данного глагола также менялось на протяжении XIX в. Ср. следующие примеры:

Я месяца четыре назад писал домой, просил Янковского, чтоб он потрудился съездить в Каменец, взять из тамошнего приказа моих 3000 рублей и из оных 1000 переслать ко мне. [А. О. Корнилович. Михаилу Осиповичу Корниловичу (1834)]

Мари, потрудитесь принести мой платок… Здесь что-то холодно. [И. И. Панаев. Прекрасный человек (1840)]

С этими же книгами потрудись прислать из моих книг сочинение Гнедича, книжечку стихов Языкова и Хомякова. [П. И. Бартенев. Письма (1852)]

Как видно из примеров, логическая структура здесь уже похожа на современную, но в модальной рамке представлены почти противоположные идеи: говорящий считает, что субъект не обязан совершать данное действие и вежливо преувеличивает усилия, которые требуются от субъекта.

Однако довольно рано появляются и иронические употребления; ср.

Я терпеть не могу лакейского круга: всегда развалится в передней, и хоть бы головою потрудился кивнуть. [Н. В. Гоголь. Записки сумасшедшего (1835)]

Далее происходит лексикализация иронического типа употреблений, подобно тому, как это произошло с такими глаголами, как *удосужиться, соблаговолить* и *соизволить*.

С глаголом *потрудиться* сближаются также такие единицы, как *трудиться* (*Не трудитесь меня провожать*), *утруждаться, утруждать себя*. Особенно интересно выражение *дать себе труд,* которое имеет некоторые тонкие отличия, с одной стороны, от *взять на себя труд*, а с другой — от *потрудиться*. К сожалению, объем не позволяет на них остановиться.

Если говорить не только о двух разобранных глаголах, но и о других упомянутых словах, можно отметить, что за последние два века у некоторых глаголов сформировались неассертивные значения, у других они усложнились, так что в целом в русском языке образовался мощный класс подобных сложных в семантическом и прагматическом отношении глаголов.

References

1. *Apresian Iu. D.* 1995. Proceedings. II : 348–386.
2. *Apresian Iu. D.* 2006. Fundamentals of System Lexicography [Osnovaniia Sistemnoi Leksikografii]. Iazykovaia Kartina Mira I Sistemnaia Leksikografiia : 145–160.
3. *Apresian Iu. D.* 2009. Researches on System Lexicography [Issledovaniia po Sistemnoi Leksikografii]. I.
4. *Apresian V. Iu*. 2010. Semantic Structure of Word and its Correlation with Negation [Semanticheskaia Struktura Slova I ego Vzaimodeistvie s Otritsaniem]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2010" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2010").
5. *Boguslavskii I. M.* 1996. The Sphere of Action of Lexical Unities [Sfera Deistviia Leksicheskikh Edinits].
6. *Bulygina T. V., Shmelev A. D.* 1999. The "Possibilities" of Natural Language and Modal Logic ["Vozmozhnosti" Estestvennogo Iazyka I Modal'naia Logika]. Iazykovaia Kontseptualizatsiia Mira (Na Materiale Russkoi Grammatiki).
7. *Bulygina T. V., Shmelev A. D.* The Grammar of Shame [Grammatika Pozora]. Logicheskii Analiz Iazyka: Iazyki Etiki : 216–234.
8. *Iomdin B. L.* 2010. Mental Vocabulary: Memory and its Functioning [Mental'naia Leksika: Pamiat' I ee Funktsionirovanie]. Prospekt Aktivnogo Slovaria Russkogo Iazyka.
9. *Kustova G. I.* 1996. On Communicative Structure of the Sentences with Event-trigger Causator [O Kommunikativnoi Strukture Predlozhenii s Sonytiinym Kauzatorom]. Moskovskii Lingvisticheskii Zhurnal : 240–261.
10. *Levontina I. B. 2005. 'DAVAI-DAVAI'. Language. Personality. Text. ['DAVAI-DAVAI'. Iazyk. Lichnost'. Tekst.]. Sbornik Statei k 70-letiiu T.M. Nikolaevoi.*
11. *NOSS.* New Explanatory Dictionary of Russian Synonyms [Novyi Ob"iasnitel'nyi Slovar' Sinonimov Russkogo Iazyka].2004.
12. *Paducheva E. V.* 2005. The Effects of Discarded Affirmation: Global Negation [Effekty Sniatoi Utverditel'nosti: Global'noe Otritsanie]. Russkii Iazyk v Nauchnom Osveshchenii, 10 (2) : 17–42.

13. *Zalizniak A. A.* 1992. Researches on Inner State Predicates Semantics [Issledovaniia po Semantike Vnutrennego Sostoiania]. Slavistische Beiträge, B. 298.
14. *Zalizniak A. A., Levontina I. B.* 1996. The Reflection of a "National Character" in Russian Vocabulary [Otrazhenie "Natsional'nogo Kharaktera" v Leksike Russkogo Iazyka]. Russian Linguistics, XX. Zalizniak A. A., Levontina I. B., Shmelev A. D.2005. Key Ideas of Russian Language Vision of the World [Kliuchevye Idei Russkoi Iazykovoi Kartiny Mira].

СТРАТЕГИИ ПЕРЕДАЧИ «ЧУЖОЙ РЕЧИ» В РАССКАЗАХ ПО КАРТИНКАМ (НА МАТЕРИАЛЕ РУССКОГО ЯЗЫКА)[1]

А. О. Литвиненко (allal1978@gmail.com)

Московский государственный университет им. М. В. Ломоносова, Москва, Россия

В докладе обсуждаются стратегии передачи так называемой «чужой речи» в устном русском дискурсе — рассказах по картинкам. Исследование, основанное на материале разрабатываемого русского устного корпуса «Истории о подарках и катании на лыжах», ставит целью выявить факторы, влияющие на выбор говорящим того или иного способа цитирования.

Ключевые слова: «чужая речь», устный дискурс, рассказ по картинкам, комиксы.

SPEECH REPORTING STRATEGIES IN RUSSIAN COMICS-BASED STORIES

A. O. Litvinenko (allal1978@gmail.com)

Lomonosov Moscow State University, Moscow, Russian Federation

The paper considers the factors that infl uence the choice of speech-reporting strategies in Russian spoken discourse. 10 speakers were asked to produce stories based on a series of pictures that included empty speech 'bubbles'. The experiment resulted in 2 sets of stories, the fi rst one being produced while looking at the pictures, and the second one — several hours later, without using the pictures. In order to be able to analyze the matching instances of reported speech from different speakers, we marked 10 positions in the pictures, where speech was possible. We will show that not all such positions are actually used by the speakers to produce reported speech; that direct speech seems not to be a prevailing type, at least in this case; that there is no signifi cant difference between telling and retelling a story as regards the choice of speech-reporting strategies. It is discussed

[1] Исследование выполнено при поддержке РФФИ (грант 10-06-00338а).

that the importance of an episode for the story, the need to portray the characters and personal preferences in style should be considered as signifi cant factors for a speaker choosing the most adequate form of speech reporting.

Key words: speech reporting, spoken discourse, picture story, comics.

1. Введение. Прямое, полупрямое и косвенное цитирование

Одно из распространенных явлений в устном дискурсе — передача говорящим так называемой «чужой речи», или цитирование. Одним из методов описания этого явления традиционно считается разделение чужой речи на прямую и косвенную, нередко с выделением промежуточных вариантов. Наряду с вопросами собственно классификации чужой речи и определения способа цитирования для каждого конкретного случая, открытым остается вопрос о том, является ли один из этих способов базовым, немаркированным и какие именно факторы влияют на выбор говорящим того или иного способа.

Коротко остановимся на проблеме определения способа цитирования. Традиционные грамматические описания и стилистические исследования, как правило, опираются на письменные, чаще художественные тексты, где цитирование подчиняется строгим формальным правилам (см., например, [Волошинов 1930], [Виноградов и др. 1954], [Есперсен 1958], [Гвоздев 1965], [Русская грамматика 1980], [Валгина и др. 2006] и многие другие). Между тем, в устном дискурсе цитирование встречается ничуть не реже, зато вариативность этого процесса намного выше, чем в кодифицированной письменной речи. В отличие от последней, где тип цитирования уже задан автором текста, а сама чужая речь оформлена соответствующим образом, при работе с устным дискурсом исследователь регулярно сталкивается с необходимостью сперва определить тип цитирования, опираясь на просодические, синтаксические, лексические и морфологические факторы, а уже затем выбрать адекватный способ транскрибирования. В силу того, что письменные тексты традиционного характера, как правило, являются нормативными и оформлены по определенным правилам, актуальные для них параметры классификации обычно не годятся для устного дискурса. Кроме того, опора на «оригинал» и степень его сохранения/искажения при передаче в принципе является до некоторой степени мифической. Как показано в [Nathan 1992] и [Aikhenvald 2008], нередко при образцовом по грамматическим признакам прямом цитировании порождаемый текст имеет очень мало сходства с оригиналом, а идеальное косвенное цитирование, за исключением необходимых грамматических преобразований сохраняет текст практически буквально. Наконец, в преобладающем числе случаев «оригинал» цитаты исследователю недоступен, а то и вовсе существует только в сознании говорящего. Все это приводит нас к необходимости научиться различать виды цитирования, не обращаясь к понятию «оригинала» или значительно расширив это понятие.

Подробно разработанная нами типология видов цитирования и методы классификации конкретных примеров в устном дискурсе изложены в [Литвиненко и др. 2009]; здесь повторим основные положения.

Прямым цитированием мы называем такое цитирование, когда говорящий подает цитируемую речь/мысли/письменный текст как не принадлежащие ему, приписывая все особенности интонации, лексики, грамматики и стиля автору оригинального дискурса. По умолчанию в такой ситуации и говорящий, и адресат «верят» в то, что цитация действительно идентична предполагаемому «оригиналу». Прямая цитация является в значительной степени иллокутивно независимой и «сохраняет» единство и целостность оригинала. **Косвенным цитированием** мы называем такое цитирование, когда цитация, становясь частью подчинительной конструкции, теряет иллокутивную независимость, утрачивает просодические и стилистические свойства «оригинала» и подвергается специальным грамматическим и лексическим преобразованиям. Это противопоставление не является бинарным; прямая и косвенная цитация представляют собой концы шкалы с множеством промежуточных значений. Наконец, **полупрямым цитированием** мы называем всякое цитирование промежуточного характера, например, когда цитация с точки зрения сегментного (лексического и грамматического) состава идентична прямой, а по просодическим характеристикам скорее совпадает с косвенной. К этой категории, в частности, следует отнести такой известный способ передачи чужой речи, как несобственная прямая (несобственно-прямая) речь.

Взаимодействие передаваемой и авторской речи — динамический процесс, совмещающий два разнонаправленных процесса: сохранение независимости, аутентичности и самостоятельности «оригинала» и творческая переработка, встраивание цитируемого в авторскую речь, подчинение этого чуждого материала целям и нуждам порождаемого здесь и сейчас дискурса.

Задача настоящего исследования — попытаться выявить, существует ли какой-либо приоритет при выборе говорящим способа цитирования и от каких факторов может зависеть этот выбор. Для этих целей нами была использована часть корпуса «Истории о подарках и катании на лыжах», собранного В. Г. Хуршудян, см. [Хуршудян 2006] и размечаемого в настоящее время исследовательским коллективом при участии автора данной статьи.

2. Корпус «Истории о подарках и катании на лыжах». Сценарий «рассказа о подарках»

Корпус «Истории о подарках и катании на лыжах» представляет собой 20 рассказов по картинкам (10 испытуемых в возрасте 20–30 лет, 2 набора картинок) и 20 пересказов по памяти, записанных 6–8 часов спустя.

Для изучения нами была отобрана половина этого корпуса, а именно, «Истории о подарках»: 10 рассказов по одному из наборов картинок и 10 пересказов по памяти от тех же испытуемых, всего около 2800 словоупотреблений.

Ниже на рис. 1 представлен набор картинок, послуживший основой для «Историй о подарках». Сюжет понимается всеми участниками более или менее

однозначно: герой не может найти подарок для своей жены, спрашивает совета у детей, обдумывает покупку машины, но решает, что это слишком дорого, и дарит жене игрушечную/сувенирную машинку.

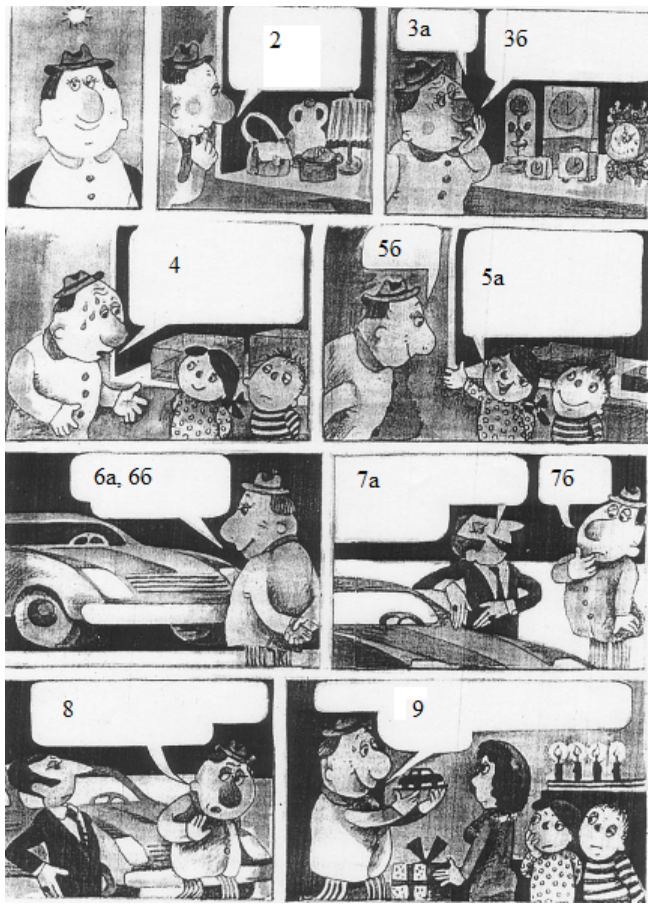


Рис. 1. Набор картинок «История о подарках»

Как мы видим, это 9 картинок, на 8 из которых представлены пустые области для передачи речи или мысли. Набор довольно легко делится на несколько эпизодов, которые в дальнейшем находят отражение в рассказах участников эксперимента.

- Картинка 1 — представление героя.
- Картинки 2–3 — затруднения героя при выборе подарка.
- Картинки 4–5 — беседа с детьми.
- Картинки 6–8 — беседа в автосалоне.
- Картинка 9 — поздравление жены и предподнесение подарка.

По отношению к тому, какой тип речи участнику эксперимента нужно вообразить и передать в рассказе, картинки неоднородны. Так, картинки 2–3 подразумевают монологический дискурс: герой либо говорит сам с собой, высказывая свои затруднения, либо просто размышляет. Картинки 4–5 подразумевают диалог с довольно четкой структурой: вопрос героя, ответ детей, реакция героя. Структура диалога героя с консультантом/продавцом в автосалоне несколько менее однозначна, подробнее об этом см. ниже. Наконец, завершается история преподнесением подарка жене.

Для того чтобы иметь возможность соотнести между собой интерпретацию этих эпизодов у разных рассказчиков, возможным позициям в монологе/диалоге нами были присвоены номера, которые указаны на рисунке. Самой большой вариативностью в интерпретациях обладают картинки, относящиеся к эпизоду в автосалоне. Так, картинка 6 частью рассказчиков воспринималась как монологический фрагмент рассказа — размышления/рассуждения героя о качестве машины. Другие рассказчики интерпретировали ее как начало беседы героя с продавцом. Некоторые совмещали эти два понимания, сначала описывая соображения героя, неважно, озвученные или нет, а затем — начало разговора с продавцом (например, вопрос о стоимости или свойствах машины). Чтобы как можно аккуратнее учесть все варианты, каждому случаю цитирования был приписан один номер из следующего возможного списка позиций:

- 2–3 — монологические рассуждения героя о затруднениях при выборе подарка;
- 4 — обращение героя к детям;
- 5а — ответ детей;
- 5б — реакция героя на предложение детей;
- 6а — размышления о машине (монологическая часть эпизода в салоне);
- 6б — обращение к продавцу
- 7а — речь/ответ продавца;
- 7б — реакция героя;
- 8 — решение героя об отказе от покупки;
- 9 — поздравление жены.

Рабочей гипотезой перед началом анализа являлось предположение, что прямое или полупрямое цитирование будет преобладать. Причин для такого предположения было две: во-первых, в предыдущих исследованиях (в корпусе «Рассказов о сновидениях», см. [Литвиненко 2006], [Литвиненко и др. 2009]) мы видели статистическое преобладание прямого и полупрямого цитирования над косвенным, а во-вторых, форма эксперимента и сам набор рисунков, как кажется, подталкивают к использованию прямого цитирования — по аналогии с тем, как устроены подобные тексты, например, в комиксах. Кроме того, предполагалось, что, возможно, в пересказе по памяти доля косвенного цитирования возрастет по отношению к непосредственно рассказу по картинкам, когда изображение находится перед глазами участника эксперимента.

3. Выявленные стратегии передачи «чужой речи»

Прежде всего, следует обратить внимание на то, что общее число цитаций оказалось значительно меньше того, чем предлагается визуальным материалом. С учетом приведенного в п. 2 списка эпизодов, где возможна передача речи или мысли, в 20 историях таких позиций не менее 200 (если объединять размышления героя на картинках 2 и 3), между тем цитаций в этой части корпуса всего 65, то есть около 30% от задуманных авторами рисунка.

Распределение случаев цитирования по типам указано в следующей таблице.

Таблица 1. Цитирование разных типов в «Историях о подарках»

	Прямое цитирование	**Полупрямое цитирование**	**Косвенное цитирование**	**Всего**
Рассказ	10	9	14	*33*
Пересказ	9	8	15	*32*
Всего	*19*	*17*	*29*	*65*

Характерные примеры основных типов цитирования приведены ниже.

(1) Прямое цитирование[2]
….(1.0) Пошел в \–автосало-он…
…(0.7) /выбирал-\выбирал,
«\Во!
Классная \машина!»
….(1.5) Решил \поинтересоваться:
«/Сколько же она \стоит?»
….(1.4) «Триста тысяч \долларов!»

(2) Полупрямое цитирование
..(0.3) Ну ему там /–↓реклами-руют…
типа «Вот там классный такой /\автомобиль…
вот этот вот /–возьми-те там¡»…

(3) Косвенное цитирование
…(0.5) Потом он встретил на улице своих /детей,
и \пожаловался,
что не может придумать маме \подарок.

[2] Здесь и ниже знаками /, \ и — обозначено направление тона в главном акценте, ударный слог в слове-акцентоносителе подчеркивается. Стрелка обозначает заударное движение тона, знак ¡ — директив. Подробнее о принципах транскрипции см. [Кибрик, Подлесская ред. 2009].

В примере (1) говорящий старательно «отыгрывает» интонации обоих участников; в примере (2), несмотря на сохранение грамматических форм, характерных для прямого цитирования, преобладает интонация сомнения-вспоминания (переданная многоточием) и встречается маркер «там», характерный для контекстов припоминания, мечтания или размышлений о будущем[3] и здесь отражающий отношение говорящего к пересказываемому (вряд ли такое «там» возможно в речи рекламирующего товар продавца). Пример (3) уже не содержит никаких указаний на вид и форму исходного высказывания/мысли, зато в авторской ремарке используется глагол *пожаловаться*, что характерно для косвенной речи, где вместо восклицаний, междометий и других подобных элементов применяются описательные средства передачи эмоций.

Часть позиций из потенциального списка проигнорирована говорящими целиком; в части случаев потенциальная речевая ситуация передана квазицитациями (пример 4) или другими описательными способами (пример 5).

(4) *…(0.5) ээ(0.2) /Дети ему /посоветовали-и …(0.1) купить \машину.*

(5) *Он \пошё-ёл в /магазин,*
…(0.6) долго /выбирал,
но-о так ничего и не \смог /придумать,
ему ничего не \понравилось.

Интересно, что позицию 9 проигнорировали все 10 участников эксперимента: ни в одном из рассказов в этом месте не используется цитация какого-либо вида; все говорящие воспринимают последнюю картинку набора как описание собственно акта дарения и делают акцент на поступках персонажей (что подарил герой, понравился ли подарок жене и детям). Характерным, например, является такой способ изложения этого эпизода, как в примере 6.

(6) *….(1.0) ээ(0.5) …(0.9) /Пошёл,*
\купил ….(1.0) –↑маленькую машинку,
….(1.0) подарил её \жене.

Следующий существенный момент — высокая общая доля косвенного цитирования: 29 из 65 случаев, то есть около 45%. С учетом того, что переходные случаи, как правило, нами интерпретируются как полупрямое цитирование, это очень высокий показатель. Уже на этой стадии анализа результатов можно заключить, что ни о каком априорном преобладании прямого цитирования речи не идет, несмотря на «подсказки», предлагаемые визуальным материалом. Для сравнения: в корпусе детских «Рассказов о сновидениях», где никакого общего материала для рассказчиков не было и опрашиваемых детей ничто не направляло и не сдерживало в выборе способа передачи чужой речи, доля косвенного цитирования не превышала 20% (подробнее см. [Литвиненко и др. 2009]).

[3] За указание на контексты, связанные с будущим, а не только с прошлым, спасибо анонимному рецензенту «Диалога».

Третий важный результат: соотношение долей разных видов цитирования практически идентичны в рассказе и пересказе. Гипотеза о возможной «косвеннизации» чужой речи в пересказе не подтверждается ни во всем корпусе целиком, ни для отдельных говорящих. Вообще, судя по количеству прямых буквальных повторов в рассказе и пересказе, несмотря на время между первой и второй частью эксперимента, говорящие очень хорошо запоминают свой рассказ и воспроизводят в значительной степени его, а не собственно историю в картинках.

Наконец, доли прямой, полупрямой и косвенной цитации в исследованном подкорпусе неодинаковы для разных позиций рассказа, причем монологичность или диалогичность эпизода не оказывают на распределение непосредственного влияния или, по крайней мере, не являются единственным фактором.

Значительное преобладание (60–80%) косвенного цитирования перед прямым и полупрямым свойственно позициям 2–3, 6б, 7а и 8. Характерные примеры заполнения этих позиций приведены ниже.

(7) позиции 2–3
....(1.0) /\'О-он ..(0.1) ходил по /\магази-ци-цнам...
..(0.2) /выбира-ал ==
..(0.4) "(0.1) думал
что бы ему /\купи-ить...

(8) позиция 6б
…(0.9) ээ(0.3) Он \спроси-л-л,
у другого \мужчи-ины-ы,
мм(0.4) сколько она \↑стоит,

(9) позиция 7а
…(0.8) {СМЕХ 0.4} /тот \сказа-ал ему,
что-о {СМЕХ 0.4} очень \/↑мно-ого,
(10) позиция 8
...(0.7) /мужи-ик \реши-и-ил,
что не \сто-о-оит покупать эту /машину,
/слишком уж \доrogow.

В остальных случаях преобладает (60–100%) прямое и полупрямое цитирование. Встает вопрос: в чем разница между одной группой ситуаций и другой? Диалогичность/монологичность визуального эпизода явным образом не играет существенной роли. Так, например, для монологического размышления о том, какой же выбрать подарок, действительно более характерно косвенное цитирование; однако для столь же монологического размышления о качествах машины в автосалоне характерно прямое/полупрямое цитирование.

Как представляется, выбор говорящего может объясняться следующим. Вне зависимости от наличия или отсутствия в предлагаемых рисунках позиций

для «чужой речи», говорящий воспринимает серию картинок прежде всего как сценарий, реализацию некоторого сюжета. Этот сценарий/сюжет состоит из последовательности действий, как речевых, так и неречевых, часть из которых является ключевой для продвижения сюжета, а часть — нет. Так, ситуации «затруднения с выбором подарка», «получение информации об автомобиле», «отказ от покупки» и «подарок жене» являются ключевыми *действиями* рассказа. По всей видимости, для рассказчика эти элементы сюжета представляются в первую очередь *поступками*, а не разговорами, вне зависимости от того, какой словесный материал, возможно, эти поступки сопровождал.

Поскольку косвенное цитирование значительно облегчает встраивание пересказываемого материала в нарративный дискурс, смещая акцент с передаваемых слов на сам *акт* их передачи, использование этого способа цитирования подчеркивает деятельный характер того или иного эпизода.

Напротив, в сюжетно менее важных эпизодах или эпизодах, где на первый план выходит коммуникация персонажей между собой, использование прямого или полупрямого цитирования позволяет говорящему посредством имитации чужой речи, интонаций и т. д. акцентировать внимание слушателя на характере героя и его переживаниях. Эта гипотеза подтверждается еще и тем, что в некоторых случаях рассказчики сначала вводили в рассказ речевую ситуацию в форме косвенного цитирования, а затем раскрывали ее характер уже в виде прямого.

(10) *..(0.1) мм(0.1) ...(0.7) /–Дети недолго /думаяh ..(0.1) /сказ<u>а</u>ли,*
ш= || чего бы хотела их \м<u>а</u>ма.
..(0.2) Так как /дети всегда \б<u>о</u>льше знают.
....(1.0) /Они ему \сказ<u>а</u>ли:
«/–Куп<u>и-ш</u>и ей \маш<u>и</u>ну¡!»

Наконец, следует признать, что, по-видимому, некоторые говорящие могут: а) предпочитать цитирование как таковое или избегать этого способа изложения; б) предпочитать косвенное или прямое/полупрямое цитирование как базовый способ цитирования. Так, например, один из участников эксперимента на 20 возможных позиций использовал 13 цитаций, из которых 9 были прямыми, 2 — полупрямыми и только 2 — косвенными. Другие два участника, напротив, почти не использовали цитирование в рассказах, а если использовали, то только косвенное. Таким образом, индивидуальные предпочтения также имеют значение при выборе стратегии передачи «чужой речи», хотя и в этом случае прослеживаются описанные выше общие тенденции.

4. Заключение

Безусловно, проведенное исследование следует считать пилотным в силу ограниченности материала и специфичности жанра исследуемых текстов — рассказ по картинкам. Однако, как кажется, оно представляет полезные данные хотя бы в силу однородности корпуса, трудно достижимой вне лабораторных

условий. Также оно позволяет выделить некоторый список факторов, влияющих на выбор способа цитирования, для дальнейшего изучения этого явления в нарративном дискурсе и данного, и других жанров. Это а) сюжетная значимость излагаемого эпизода (более важные моменты, «поступки», тяготеют к изложению в виде косвенной цитации, поскольку так легче встраиваются в контекст); б) возможность и/или необходимость акцентировать внимание на эмоциональных и психологических характеристиках персонажа (прямое цитирование позволяет это легче, чем косвенное); в) индивидуальные предпочтения говорящего, иногда явно предпочитающего ту или иную стратегию.

References

1. *Aikhenvald A. Y.* 2008. Semi-direct Speech: Manumbu and beyond. Language Sciences : 383–342.
2. *Espersen O.* 1958. Philosophy of Grammar [Filosofiia Grammatiki].
3. *Gvozdev A. N.* 1965. Sketches on Russian Stylistics [Ocherki po Stilistike Russkogo Iazyka].
4. *Khurshudian V. G.* 2006. Hesitation Expression Means in Oral Armenian Discourse in Typological Perspective [Sredstva Vyrazheniia Khezitatsii v Ustnom Armianskom Diskurse v Tipologicheskoi Perspektive].
5. *Litvinenko A. O.* 2006. The Strategies of Another's Speech Appearance in Childern'Oral Narration [Strategii Oformleniia Chuzhoi Rechi v Ustnom Detskom Narrative]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2006" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2006"): 353–356.
6. *Litvinenko A. O., Korotaev N. A., Kibrik A. A., Podlesskaia V. I.* 2009. Constructions with Citation, or with "Another's Speech"[Konstruktsii s Tsitatsiei, ili s "Chuzhoi Rech'iu"]. Rasskazy o Snovideniiakh. Korpusnoe Issledovanie Ustnogo Russkogo Diskursa : 288–308.
7. *Nathan D.* 1992. Can You Say 'That' Again? (The Status of dDrect and Indirect Speech as Grammatical Categories). La Trobe University Working Papers In Linguistics, 5.
8. *Russian* Grammar [Russkaia Grammatika], I, II. 1980.
9. *Valgina N. S., Rozental' D. E., Fomina M. I.* 2006. Modern Russian Language [Sovremennyi Russkii Iazyk].
10. *Vinogradov V. V.* 1954. Russian Grammar [Grammatika Russkogo Iazyka].
11. *Voloshinov V. N.* 1930. Marxism and the Phylosophy of Language [Marksism I Filosofiia Iazyka].

СТАТИСТИЧЕСКИЕ ХАРАКТЕРИСТИКИ СИНТАГМАТИЧЕСКОГО ЧЛЕНЕНИЯ ПРЕДЛОЖЕНИЙ В ПРИЛОЖЕНИИ К СИНТЕЗУ ВЫРАЗИТЕЛЬНОЙ РЕЧИ ПО ТЕКСТУ

Б. М. Лобанов (lobanov@newman.bas-net.by)

Ю. С. Гецевич (mix1122@gmail.com)

Объединенный институт проблем информатики
НАН Беларуси, Минск, Беларусь

В докладе описываются используемый экспериментальный материал, методика и результаты статистической обработки текстового и звукового файлов. Приводятся статистические характеристики особенностей членения предложений на синтагмы, реализуемого в процессе выразительного чтения художественного текста профессиональным диктором (актёром). Описываются пути использования полученных статистических данных в приложении к синтезу выразительной речи по тексту.

Ключевые слова: выразительная речь, синтез речи, синтез выразительной речи, синтагматическое членение.

STATISTICAL CHARACTERISTICS OF SYNTAGMATIC SEGMENTATION OF UTTERANCES FROM THE VIEWPOINT OF EXPRESSIVE TEXT-TO-SPEECH SYNTHESIS

B. M. Lobanov (lobanov@newman.bas-net.by)

Iu. S. Getsevich (mix1122@gmail.com)

Institute of Informatics Problems NAS Belarus, Minsk, Belarus

We describe the results of a statistical study of text segmentation into phrases that occurs during expressive reading of Russian fi ction by a professional speaker(actor),The pupose is to fi nd out whether part-of-speech tags could be used to predictbreaks between phrases in a sentence. The experimental material was Anton Chekhov's story. A Hunting Drama, presented in text (54 thousand words) and sound formats (an audio book with 7 hrs playing time). This material was divided into two parts: the initial segment of the tagged text of the story containing 420 sentences (ca. 6000 words) and the rest of the text (untagged). The untagged part was used for model evaluation. Prosodic phrases were manually tagged by a professional auditor — phonetician who listened to the text. The total number of tagged phrases in the initial 420 sentences was 1516 (of which 710 had pauses no longer than 100 msec and 380 had longer pauses). The average number of phrase breaks in a sentence was 3.6, while the average length of a phrase was 4 words. Pairs consisting of words belonging to 11 different parts of speech or POS-like morphological classes were investigated: adjective, adverb, conjunction, gerund, interjection, parenthetical word, noun, numeral, participle, pronoun, and fi nite verb. In addition to POS information, the statistical analysis takes account of punctuation marks appearing in the sentence (commas, hyphens, dashes, colons, semicolons and parentheses).. Quantitative distributions have als been obtained for phrase breaks occurring in the pairs: "punctuation mark — part of speech", "part of speech — punctuation mark", "space — part of speech", "part of speech — space". Potentials of using this data in expressive text-to-speech synthesis system are considered.

Key words: expressive text-to-speech, speech synthesis, expressive text-to-speech synthesis, syntagmatic segmentation.

Введение

К настоящему времени системы синтеза речи достигли определённого уровня развития и уже используются в ряде практических приложений. Однако комфортность восприятия синтезированной речи в реальных условиях систем массового обслуживания остаётся ещё не вполне удовлетворительной. Мировая тенденция развития речевых технологий указывает на актуальность создания систем синтеза выразительной речи (expressive text-to-speech) [1–3]. Понятие «выразительность речи» сформировалось как междисциплинарное понятие, характеризующее одну из функций устной речи человека [4]. Речевой опыт каждого из нас говорит о том, что, например, два доклада, прочтённые на одну и ту же тему, могут оказать на человека совершенно разный эффект, зависящий от степени выразительности речи диктора.

Одним из главных компонентов звуковой реализации выразительности устной речи является просодика речи и, в частности, при синтезе речи — правильность и качество просодической разметки. Просодическая разметка текста при синтезе речи заключается в членении предложений на синтагмы, в маркировке просодически выделенных слов в синтагме и в установке интонационного типа синтагмы [5]. В данной работе исследуется первый из указанных аспектов просодической разметки. Установка границ синтагм влияет на правильность передачи интонационных характеристик, а также на передачу смыслового содержания текста. Следует отметить, что при членении предложений на синтагмы особенно важно не поставить её границу там, где она может нарушить смысловое восприятие речи, например, между предметом и его признаком.

Подробный обзор и анализ проблем, связанных с локализацией синтагматических границ в естественной речи и возможных подходов к установлению границ синтагм при синтезе речи по тексту дан в работе О. Ф. Кривновой и И. С. Чардина [6]. При этом рассмотрены следующие возможные классы систем синтагматического членения предложений:

1. Системы, которые обходятся анализом структуры текста с помощью обнаруженных эвристик (экспертные системы).
2. Системы, в которых проводится синтаксический анализ с использованием формальных грамматик.
3. Системы, где используется вероятностный анализ текста, основанный на статистической модели, параметры которой получены через обучение по аннотированной тексто-речевой базе данных.

Для систем синтеза русской речи по тексту, относящихся к первому классу, были предложены достаточно простые правила синтагматического членения предложений, основанные на морфологической информации о словосочетаниях [7, 8]. Прототип системы синтеза русской речи, в которой используется глубокий синтаксический анализ, описан в [9]. Системы, где используется вероятностный анализ текста, разработаны для синтеза речи на ряде европейских языков (см., например, [10]). Первая попытка использования статистических особенностей для синтагматического членения русской речи на небольшом по объёму текстовом и аудио материале была описана в [11]. В данной работе продолжено развитие этого подхода с более детальным рассмотрением различных ситуаций

синтагматического членения. При этом использованы аудиозаписи выразительного чтения художественного текста профессиональным диктором (актёром) и существенно расширенный экспериментальный материал.

1. Экспериментальный материал

В качестве экспериментального материала использована повесть А. П. Чехова «Драма на охоте», представленная в текстовой форме (54 тыс. слов) и в звуковой (аудиокнига с временем звучания — 7 часов в исполнении профессионального диктора А. Балакирева). Экспериментальный материал состоит из двух частей: начального размеченного отрезка повести, состоящего из 420 предложений (около 6000 слов) и остальная неразмеченная её часть. Неразмеченная часть повести использовалась затем для контрольных экспериментов.

Текстовый и звуковой файлы начального отрезка текста предварительно были разбиты на отдельные предложения. Разметка на просодические синтагмы осуществлялась профессиональным аудитором-фонетистом в процессе прослушивания отдельных предложений. Затем эта разметка переносилась на звуковые файлы. Знаком [/] (короткая пауза) аудитором отмечались границы синтагм в тех случаях, когда отсутствовала заметная физическая пауза звука (менее 100 мс), а знаком [//] (долгая пауза) — когда физическая пауза присутствовала (более 100 мс). Наличие границы синтагмы при отсутствии физической звуковой паузы определялась аудитором на основе его представлений о достаточной в интонационном смысле самостоятельности синтагмы.

Общие количественные характеристики используемого размеченного экспериментального материала представлены таблице 1.

Таблица 1. Количественные характеристики используемого текста

Общ. кол. слов в тексте	Общ. кол. предл. в тексте	Общ. кол. внутр. зн.преп.	Общ. колич. синтагм в тексте	Кол. синтагм с короткой паузой	Кол. синтагм с длинной паузой	Средн. кол. синт. в предл.	Средн. кол. фонетич. слов в синт.
6234	424	485	1516	710	382	3,6	4,1

Ниже приведены 4 примера аудиторской разметки предложений на синтагмы.

(1) В один из апрельских полудней / тысяча восемьсот восьмидесятого года // в мой кабинет вошел сторож Андрей / и таинственно доложил мне, / что в редакцию явился какой-то господин // и убедительно просит свидания с редактором.

(2) К пишущим людям / не имею чести / принадлежать, // но, тем не менее, явился к вам / с чисто писательскими целями.

(3) Был, знаете ли, // судебным следователем в уезде, // прослужил пять с лишком лет, / но ни капитала не нажил, // ни невинности не сохранил...

(4) Лентяя вдруг ни с того ни с сего / осенила мысль, / что нос моего попугая / очень похож на нос / нашего деревенского лавочника / Ивана Демьяныча, // и с той поры / за попугаем навсегда осталось имя / и отчество // длинноносого лавочника.

Очевидно, что приведённые выше варианты разметки предложений на синтагмы не являются единственно возможными, однако, они объективно отражают те предпочтения, которым следовал данный диктор при их выразительном чтении.

2. Методика и результаты статистической обработки экспериментального материала

Для всех последующих этапов статистической обработки каждое слово анализируемого текста маркируется его грамматическим типом (в нашем исследовании названием части речи). Всего в нашем исследовании используется 11 различных частей речи: *Вводное слово, Глагол, Деепричастие, Междометие, Местоимение, Наречие, Прилагательное, Причастие, Союз, Существительное, Числительное*. В этом списке отсутствуют предлоги и частицы, которые считаются присоединёнными к соседним словам в соответствии с известными правилами образования фонетических слов, а также предикаты, которые мы условно присоединили в одну группу с наречиями. Количественное распределение встречаемости частей речи в анализируемом тексте представлено на рис. 1.

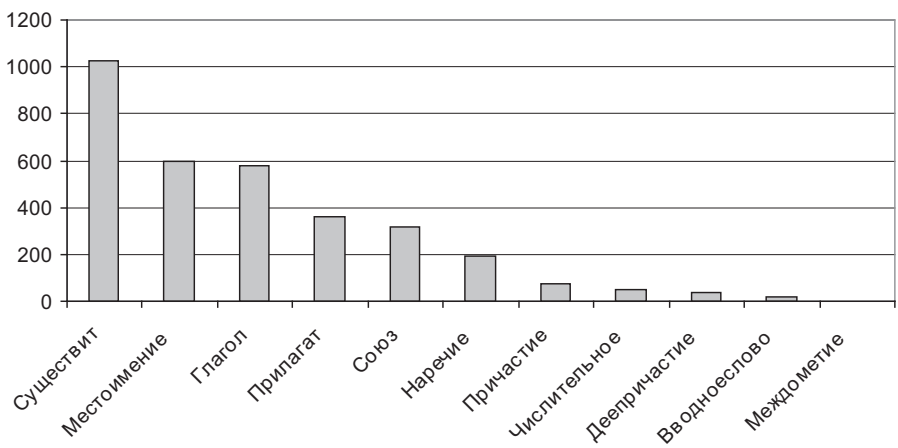


Рис. 1. Количественное распределение частей речи в анализируемом тексте

Кроме информации о частях речи при статистической обработке принимаются во внимание также присутствующие внутри предложений знаков препинания: (,) — запятая, (—) — тире, (, —) — запятая с тире, (:) — двоеточие, (;) — точка с запятой, (() — открывающая и ()) — закрывающая скобки. Количественное распределение встречаемости знаков препинания в анализируемом тексте представлено на рис. 2а,б.

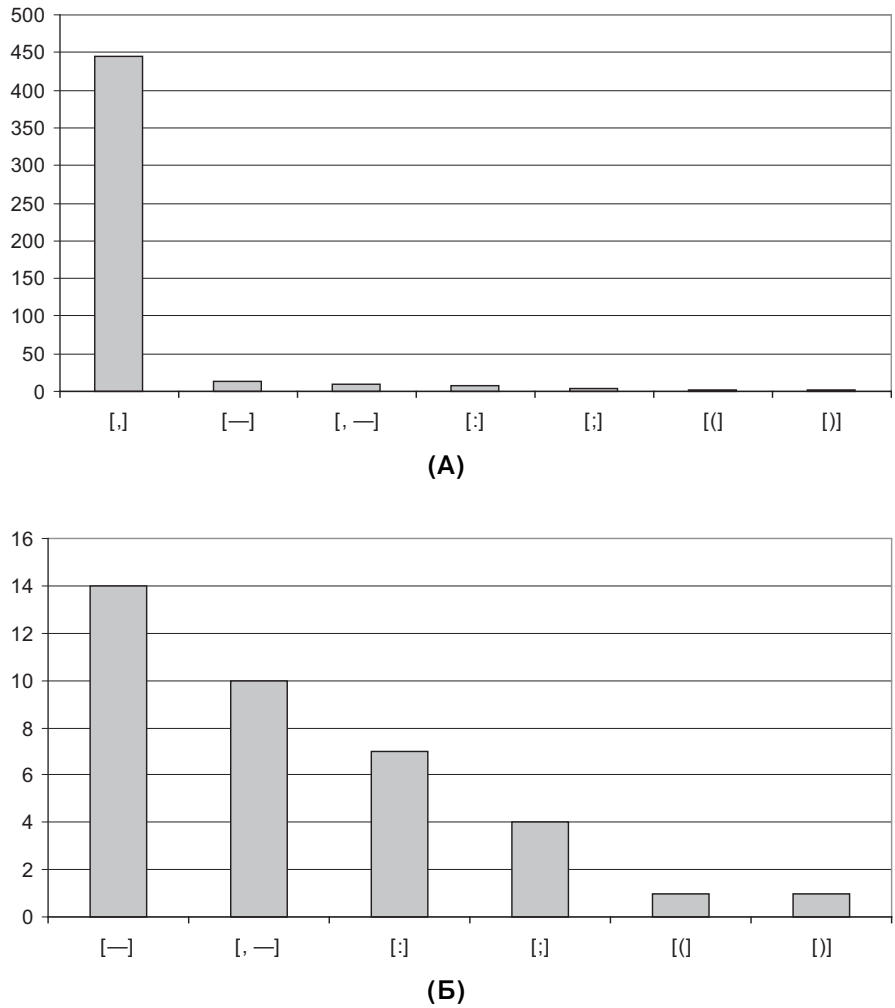


Рис. 2. Количественное распределение встречаемости всех знаков препинания (А) и малочастотных знаков (Б)

Дальнейшее исследование экспериментального материала посвящено подсчёту и анализу количественных распределений встречаемости синтагматической границы (с долгой или с короткой паузой) или её отсутствия для следующих пар:

- «знак препинания — часть речи»,
- «часть речи — знак препинания»,
- «пробел — часть речи»,
- «часть речи — пробел».

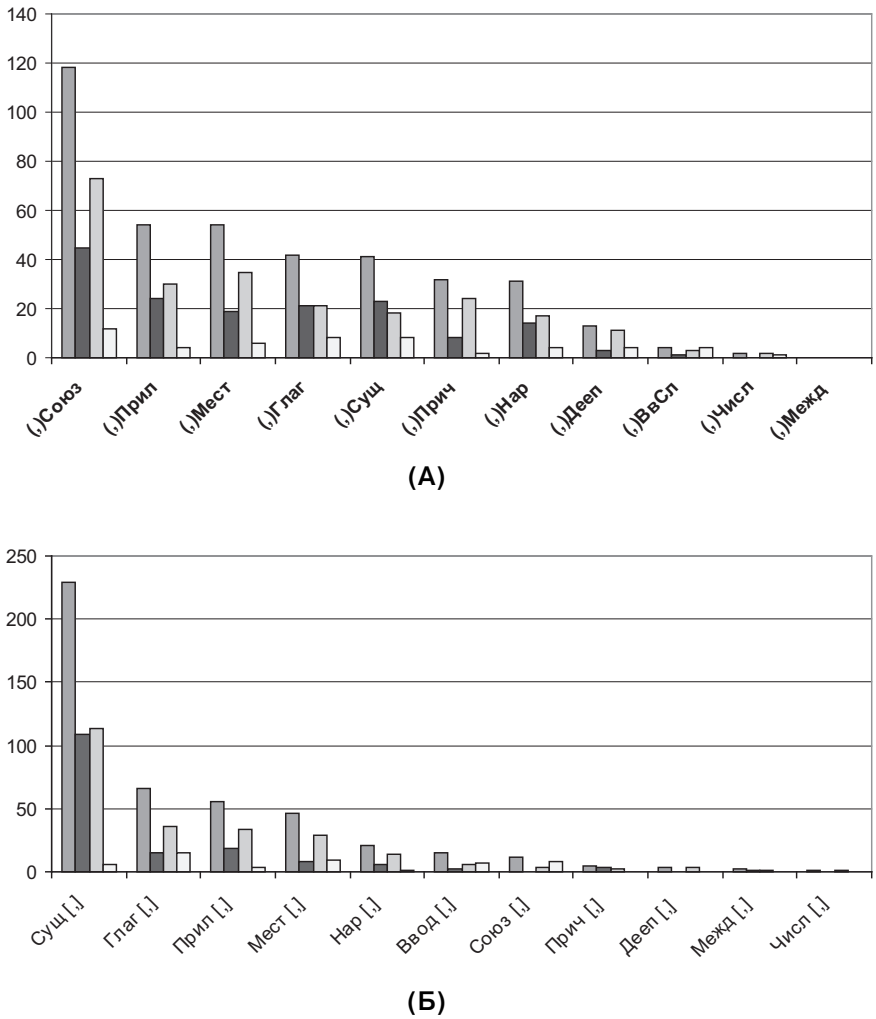


Рис. 3. Количественное распределение встречаемости синтагматической границы в присутствии знака препинания для пар: «запятая — часть речи» (3А), «часть речи — запятая» (3Б) (всего пар — ряд 1, долгая пауза — ряд 2, короткая пауза — ряд 3, нет границы — ряд 4)

В последних двух случаях под пробелом понимается пробел между фонетическими словами в отсутствие между ними знака препинания.

На рис.3а представлено количественное распределение встречаемости синтагматической границы для пар: «запятая — часть речи», а на рис. 3б — для пар «часть речи — запятая».

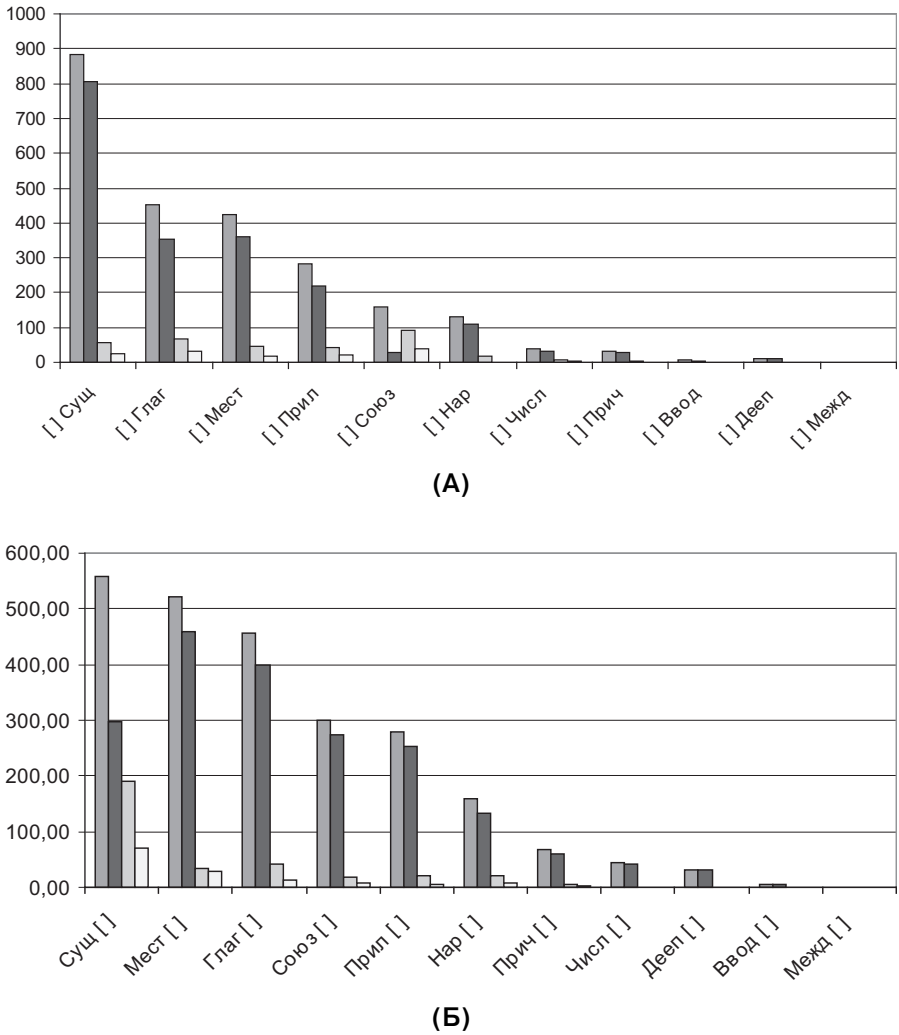


Рис. 4. Количественное распределение встречаемости синтагматической границы в отсутствие знака препинания для пар: «пробел — часть речи» (4а), «часть речи — пробел» (4б) (всего пар — ряд 1, нет границы — ряд 2, долгая пауза — ряд 3, короткая пауза — ряд 4)

Общий вид распределений на рисунках 3а,б позволяет сделать вывод о существенно большей частоте встречаемости короткой или долгой синтагматической границы в присутствии знака препинания в сравнении с частотой отсутствия границы. Из приведенных рисунков видно также, что при выразительном чтении границы синтагм реализуются предпочтительно в виде короткой паузы.

На рис. 4а представлено количественное распределение встречаемости синтагматической границы в отсутствие знака препинания для пар: «пробел — часть речи», а на рис. 4б — для пар «часть речи — пробел».

Общий вид распределений на рисунках 4а,б позволяет сделать вывод, что при отсутствии знака препинания, в отличие от случая рис. 3а,б, частота отсутствия синтагматической границы существенно выше в сравнении с частотой присутствия синтагматической паузы (короткой или долгой). Из приведённых рисунков видно также, что при выразительном чтении, в сравнении с рассмотренным ранее случаем, границы синтагм реализуются (за некоторыми исключениями) в виде долгой паузы.

3. Использование статистических данных в приложении к синтезу выразительной речи по тексту

Полученные статистические характеристики могут быть положены в основу вероятностного алгоритма синтагматического членения предложений в системе синтеза речи по тексту. Для этого приведенные на рисунках 3, 4 данные преобразуем в форму распределений статистической вероятности (в %) наступления полной группы событий: «отсутствие — наличие» синтагматической границы. На рис. 5 представлены распределения вероятностей встречаемости синтагматической границы при наличии знака препинания для пар: «запятая — часть речи» (рис. 5а) и «часть речи — запятая» (рис. 5б), а на рис. 6 — распределения вероятностей встречаемости синтагматической границы в отсутствие знака препинания между словами для пар: «пробел — часть речи» (рис. 6а) и «часть речи — пробел» (рис. 6б)

Исходя из данных, представленных на рис. 5, можно сделать следующие основные выводы. Вне зависимости от типа части речи, следующей за знаком препинания (рис. 5а), синтагматическая граница (с короткой либо долгой паузой) присутствует в подавляющем большинстве случаев. Исключение составляют пары с междометиями, вводными словами и числительными, где вероятность отсутствия синтагматической границы весьма высокая. Это хорошо согласуется с нашими априорными представлениями. Похожие выводы можно сделать и исходя из анализа данных рис. 5б. Вероятность отсутствия синтагматической границы весьма высокая для пар «союз — (,)» и «вводное слово — (,)». Отметим, что полученное 100% присутствие синтагматической границы для пар с причастиями, деепричастиями, вводными словами и междометиями хотя и правдоподобно, но объясняется, скорее всего, их низкой частотностью в исследуемом материале.

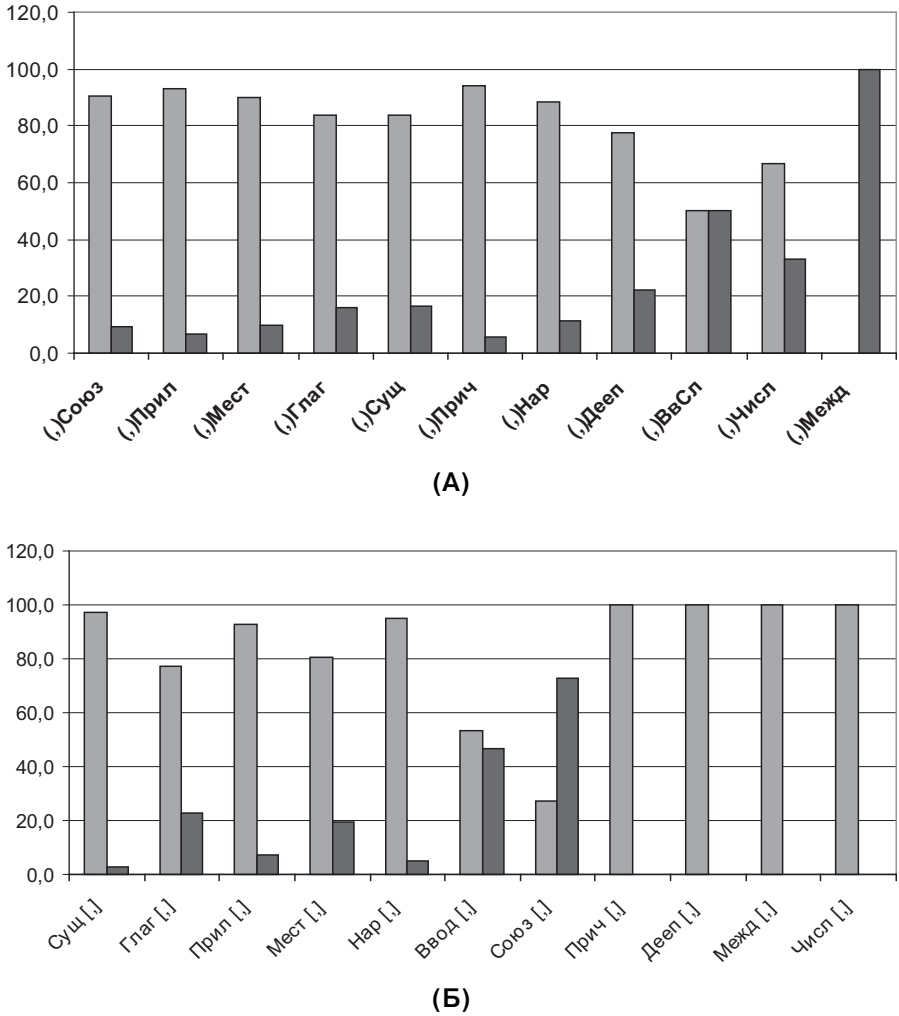


Рис. 5. Процентное соотношение встречаемости синтагматической границы при наличии знака препинания для пар: «запятая — часть речи» (а), «часть речи — запятая» (б) (есть граница с паузой короткой или долгой — ряд 1, граница отсутствует — ряд 2)

Исходя из данных, представленных на рис. 6, можно сделать следующие основные выводы. Вне зависимости от типа части речи, следующей за пробелом (рис. 6а), вероятность наличия синтагматической границы (с короткой либо долгой паузой) в большинстве случаев значительно более низкая, чем её присутствие. Исключение составляет пара с союзами (как правило, союзы «И», «ИЛИ»), для которой вероятность присутствия синтагматической границы весьма высокая. Это хорошо согласуется с нашими априорными представлениями. Похожие выводы можно сделать и исходя из анализа данных рис. 6б.

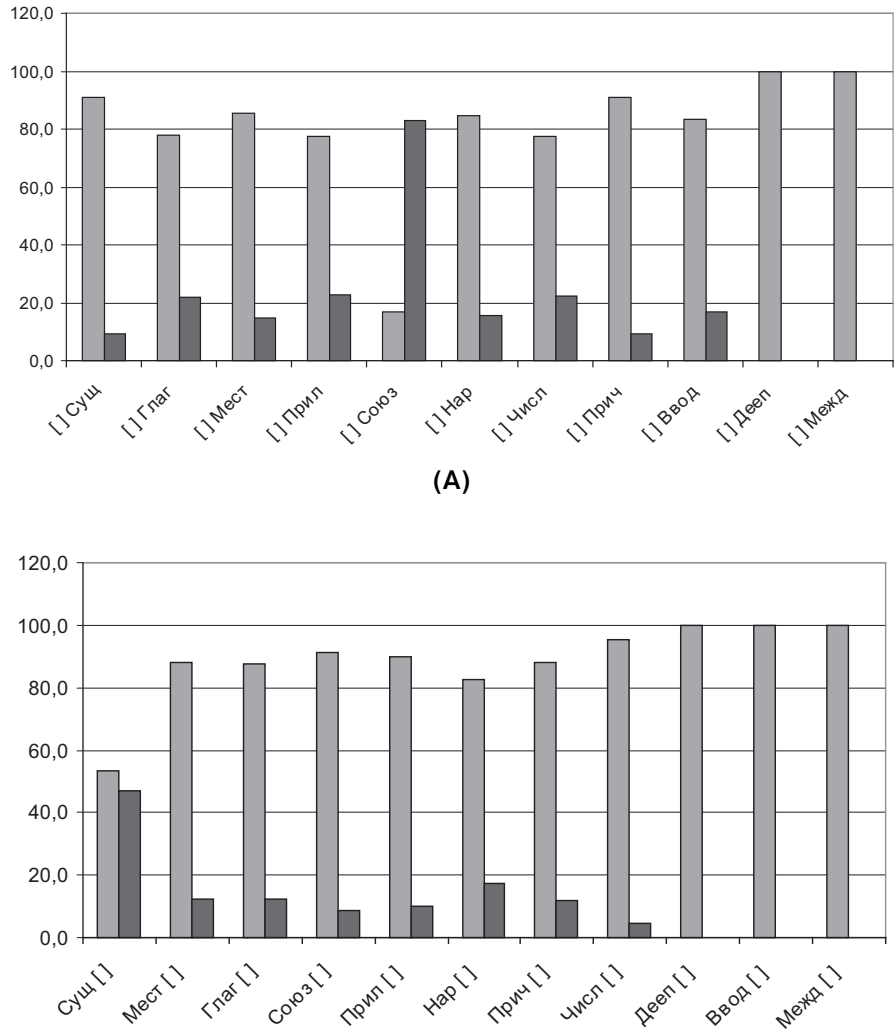


Рис. 6. Процентное соотношение встречаемости синтагматической границы в отсутствие знака препинания для пар: «пробел — часть речи» (а) и «часть речи — пробел» (б) (нет границы — ряд 1, есть граница с паузой короткой или долгой — ряд 2)

Опираясь на представленный выше комплекс статистических характеристик, решающие правила вероятностного синтагматического членения при синтезе речи в строгом математическом смысле должны основываться на формуле Байеса:

$$P(A \mid B_{ij}) = \frac{P(B_{ij} \mid A) \cdot P(A)}{P(B_{ij})}$$

Здесь:

A — интересующее нас событие: появление синтагматической границы между парами частей речи в тексте (паузы в речи)

B_{ij} — текущее событие в тексте: одна из множества возможных ij-пар частей речи в промежутке между словами;

P(*A*) — априорная вероятность гипотезы *A* (как часто в среднем появляется граница между любыми словами). Определяется исходя из подсчитанного среднего числа слов в синтагме (см. таблицу 1);

$P(B_{ij} \mid A) = P_b(ij)$ — вероятность наступления события B_{ij} при истинности гипотезы *A*. Найденное заранее распределение вероятностей пар частей речи B_{ij}, между которыми присутствует граница синтагмы (пауза);

$P(B_{ij})$ — вероятность наступления события B_{ij}. Найденное заранее распределение вероятностей пар частей речи B_{ij} вне зависимости от того присутствует либо отсутствует между ними граница синтагмы (пауза).

$P(A \mid Bij) = P_a(ij)$ — вероятность гипотезы *A* при наступлении события B_{ij} (апостериорная вероятность). Искомое распределение вероятностей присутствия границы синтагмы (паузы) при условии, что в данной позиции текста находится пара частей речи — B_{ij};

Рассмотрим далее в качестве иллюстрации упрощённый алгоритм. Процедура вероятностного синтагматического членения предложений с использованием данных, приведенных на рис. 5, 6, может быть построена следующим образом. На первом этапе определяются наиболее вероятные границы синтагм на стыке фонетических слов, разделённых знаком препинания (назовём их «пунктуационными» синтагмами). Для этого определяются среднее процентное соотношение встречаемости синтагматической границы при наличии знака препинания для пар: «запятая — часть речи» (а), «часть речи — запятая» (б), исходя из данных, приведенных на рисунках 5а и 5б. По некоторому заранее выбранному порогу принимаются решения и проставляются границы найденных таким образом пунктуационных синтагм. Далее все слова внутри пунктуационных синтагм считаются разделёнными только пробелами. К каждой из полученных таким образом последовательностей слов применяется описанная выше процедура, но только с использованием данных, приведенных на рисунках 6а и 6б. В результате каждая пунктуационная синтагма разбивается в свою очередь на последовательность внутренних синтагм (назовём их «синтаксическими» синтагмами).

Наибольший интерес, с нашей точки зрения, представляет проверка работоспособности изложенных правил при сегментации многословных предложений без единого знака препинания, которые часто встречаются в текстах. Рассмотрим пример такого рода предложения, не входившего в обучающую выборку (предлоги присоединены знаком «ъ»):

(1) «*Но молодая жена упорно продолжала отстирывать белую въкровавых пятнах рубаху мужа посиневшими отъхолода руками въжелезном тазике съледяной водой.*»

На рис. 7 для этого предложения графически представлены нормированные значения статистической вероятности наличия синтагматической границы между словами, рассчитанными в соответствии с данными, приведенными на рисунках 6а и 6б, а также возможные положения границ синтагм при двух порогах принятия решений — 0,45 и 0,25.

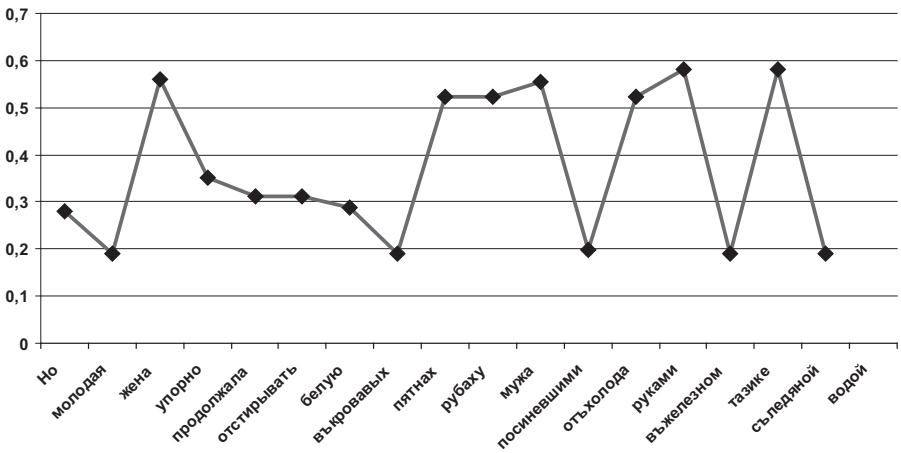


Рис. 7. Возможные положения границ синтагм при двух порогах принятия решений — 0,45 и 0,25 для предложения «Но молодая жена упорно продолжала отстирывать белую в кровавых пятнах рубаху мужа посиневшими от холода руками в железном тазике с ледяной водой»

Ниже на 3-х примерах показаны возможные границы синтагм для максимального объёма синтагм (1) — при пороге 0,45 и для минимального объёма: (2) — при пороге 0,45 и (3) — при пороге 0,25.

(1) *Но молодая жена // упорно продолжала отстирывать белую в кровавых пятнах рубаху мужа // посиневшими от холода руками // в железном тазике // с ледяной водой.*

(2) *Но молодая жена // упорно продолжала отстирывать белую в кровавых пятнах // рубаху // мужа // посиневшими от холода // руками // в железном тазике // с ледяной водой.*

(3) *Но // молодая жена // упорно // продолжала // отстирывать // белую // в кровавых пятнах // рубаху // мужа // посиневшими от холода // руками // в железном тазике // с ледяной водой.*

Приведенные варианты синтагматического членения, на наш взгляд, вполне допустимы. Как показано в [12], для перцептивной сегментации на синтагмы весьма характерна индивидуальная вариабельность стратегий, проявляющаяся в выделении синтагм существенно разного объёма — от одного фонетического слова до пяти-семи.

Заключение

В результате проведённого исследования получены объективные статистические данные особенностей синтагматического членения в процессе выразительного чтения текста профессиональным диктором. Намечены пути использования статистического подхода для реализации алгоритмов синтагматического членения в синтезаторе русской речи по тексту. Авторы ясно осознают, что полученных статистических данных во многих случаях недостаточно для безошибочного членения предложений на синтагмы. Однако, ввиду сравнительной простоты полученных правил, их использование, например, в приложениях с ограниченными вычислительнымими ресурсами, может оказаться целесообразным. Предполагается дальнейшее развитие данного подхода с использованием более детальной информации о морфологической структуре слов и синтаксической структуре предложений. Предварительные результаты синтеза речи по тексту с использованием статистического алгоритма синтагматического членения будут продемонстрированы во время доклада.

Авторы выражают свою признательность Белорусскому фонду фундаментальных исследований за поддержку данной работы.

References

1. *Golovin I. B.* 1983. Fundamentals of Speech Culture [Osnovy Kul'tury Rechi].
2. *Lobanov B. M.* 2008. Computational Synthesis and Speech Cloning [Komp'iuternyi Sintez I Klonirovanie Rechi].
3. *Hiromichi Kawanami* et al. Designing Speech Database with Prosodic Variety for Expressive TTS system, available at: http://gandalf.aksis.uib.no/lrec2002/pdf/337.pdf
4. *Hongwu Yang, Shuang Li, Lianhong Cai.* Toward Synthesizing Expressive Mandarin Speech, available at: http://www.w3.org/2005/08/SSML/Papers/Tsinghua.pdf
5. *Iomdin L. L., Lobanov B. M., Getsevich Iu. S.* 2004. The Talking ETAP. Using the ETAP Parser in Russian Speech Synthesis [Govoriashchii "ETAP": Opyt Ispol'zovaniia Sintaksicheskogo Analizatora Sistemy ETAP v Russkom Rechevom Sinteze]. This compendium : 874–877.
6. *Khomitsevich O. G., Solomennik M. V.* 2010. Automatic Pausing in the System of Russian Text-To-Speech Synthesis [Avtomaticheskaia Rasstanovka Pauz v Sisteme Sinteza Russkoi Rechi po Tekstu]. Komp'iuternaia Lingvistika

i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2010" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2010") : 531–537.

7. *Krivnova O. F., Chardin I. S.* 1999. Pausing in Natural and Synthesized Speech [Pauzirovanie v Estestvenooi I Sintezirovannoi Rechi]. Teoriia I Praktika Rechevykh Issledovanii (ARSO-99). Materialy Konferentsii, (Proc. of Conference "Theory and Practice of Linguistic Researches"), available at: http://www.russian.slavica.org/article9348.html
8. *Lobanov B. M.* 2008. The Algorithm of Text Segmentation into Syntactic Syntagmas for Speech Synthesis [Algoritm Segmentatsii Teksta na Sintaksicheskie Sintagmy dla Sinteza Rechi]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2008" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2008") : 323–329.
9. *Lobanov B., Tsirulnik L.* 2006. Statistical Study of Speaker's Peculiarities of Utterances into Phrases Segmentation. Speech Prosody: Proceedings of the 3-rd International Conference, 2 : 557–560.
10. *Pitrelli J. F.* et al. 2006. The IBM Expressive Text-to-speech Synthesis System for American English. Audio, Speech, and Language Processing, IEEE Transactions on V., 14 (4) : 1099 –1108.
11. *Ventsov A. V., Kasevich V. B., Slepokurova N. A.* 1993. Perceptive Segmentation of a Sounding Text [Pertseptivnaia Segmentatsiia Zvuchashchego Teksta]. Problemy Fonetiki : 242–273.

МЕТОД ПОРОЖДЕНИЯ ПРАВИЛ МЕЖЪЯЗЫКОВОЙ МАШИННОЙ ТРАСКРИПЦИИ

В. К. Логачева (logacheva_vk@mail.ru)

Е. С. Клышинский (klyshinsky@mail.ru)

Институт Прикладной Математики РАН, Москва, Россия

В статье рассматривается метод генерации правил машинной транскрипции. Метод пригоден для применения к языкам различных групп. Генерация правил проводится на основе анализа коллекции имен собственных, в которой представлено написание как на языке оригинала, так и на выходном языке. Работа поддержана грантом РФФИ № 10-01-00800.

Ключевые слова: машинная транскрипция, правила машинной траскрипции, генерация правил, имена собственные.

NON-STOCHASTIC LEARNING OF CROSS-LANGUAGE TRANSLITERATION RULES FROM A SMALL DATASET

V. K. Logacheva (logacheva_vk@mail.ru)

E. S. Klyshinskii (klyshinsky@mail.ru)

Keldysh IAM Russian Academy of Sciences, Moscow, Russia

We present a language-independent method of generating rules for machine transliteration. The generation of rules is based on the analysis of a test dataset, which contains names written in the source language and their transliterations into the target language.

Key words: transliteration rules, machine transliteration rules, rules generation, proper nouns.

1. Introduction

Proper names cross-language transcription is an essential problem in many spheres — from linguistic topics, like machine translation or information retrieval, to some purely practical ones — for example, when translating documents or maps.

There are several ways to reproduce names with the means of other language:

- Translation (for example, Easter Island — остров Пасхи). Interpreters rarely use this way. Translation is impossible in many cases as proper names usually don't have any lexical meaning.
- Transliteration:
  - Strict transliteration — every letter of alphabet of the source language is associated with a letter in target language. This way of transfer can misrepresent phonetic appearance of the word as almost every language has di- or trigraphs — that is set combinations of letters that should be read in a specific way. Even rules of extended transliteration (rules that allow to transform one letter of source alphabet to two or more letters of target alphabet) are not always sufficient to define all relations between phonetics and graphics.
  - Transliteration with regard for phonetic appearance of the word. This method is usually called practical transcription.

At different times different approaches to translation of names entities were popular among translators, but since the middle of 20th century most of translators agree that name should keep its sounding. Progress in computational linguistics has raised a question of automatic transcription of proper names.

Currently there already exist a lot of various methods of cross-language transliteration. They are based on different approaches and use different techniques: stochastic state-finite automata, Viterbi decoding algorithm, learning of statistic machine translation systems. Vast majority of existing methods are based on statistics. This approach is often effective because it doesn't need to involve specific linguistic information. But its simplicity is paid with necessity in huge amounts of learning data, which is often inaccessible. While other groups of researches develop methods of automatic data retrieval or generate cross-language phoneme or letter mappings using monolingual corpora, we tried to work out a "clever" method of rules generation.

Our work is based on the system "Transscriba" [11]. This is a rule-based system, that means that transliteration model is a set of rules, constructed manually by expert. Such approach provides very high accuracy of transliteration, but it takes from two weeks to six months of an expert's work to construct a system of rules for one pair of languages. Moreover, "Transscriba" has one more drawback. Its method of strings transformation is ineffective as speed of parsing depends on amount of rules in the system. In Section 4 we demonstrate effective method of strings processing with generated rules. Section 5 introduces the method of automatic learning of cross-language transcription rules from small training set.

2. Related works

First attempts to work out system of rules of transliteration relate to pre-computer epoch. There are plenty of works that should rather be treated as recommendations for translator than a code of laws, but these recommendations have later become basis of formal rules used in machine transliteration systems [9, 10].

As for machine transliteration itself, one of the first works that had determined direction of many researches in this area is work [3]. It describes transliteration and back-transliteration between English and Japanese languages. The model is trained with modified Viterbi algorithm. Transliteration is accomplished by a chain of statistical finite-state transducers. Output of every automaton is an input of the next one.

Later this method was adapted for the Arabic language [1]. However, fundamental principle of those works was recognition of separate characters and their groups. It didn't allow to raise quality of transliteration. So researches started working with chained substrings [5]. Such replacement allowed to improve transliteration accuracy from 30 % to 90 %.

The above mentioned chain of finite-state automata served as a basis for many other methods. Jonathan Grael has developed an instrument that constructs a chain of automata that can be trained on user's data. This tool was widely used in many works on machine transliteration [2].

Another popular method is learning of cross-language mappings using phrase-based machine translation systems. While during translation minimal unit is a word and sentences are regarded as word successions, during transliteration minimal unit is a letter and main analyzed unit is a word.

There are a lot of techniques of machine transliteration. As we can't specify all of them in this paper, we will name main parameters in which different methods vary:

- **Letter / phoneme substitution** — some methods work with letters and substrings [5] and others transform letters in phonetic notation and look for phoneme cross-language mappings [3]
- **Statistical / rule-based models** — cross-language mappings can be acquired with statistic analysis of test data or using some heuristic model
- **Manual / automatic generation of learning data** — for statistic-based models size and quality of test data is very important. Some researches are satisfied with manually-constructed sets, others use multilingual dictionaries of names and terms [3, 5, 7], acquire parallel examples from bilingual corpora [6], or even learn on unilingual data [4].

3. Preliminaries

We will speak about transduction of word from one language to another. So our method deals with pairs of languages, one of which is a source language (language of original) and another is a target language (language of translation). Let we denote alphabet of source language as $\mathbf{V}_I$, and alphabet of target language as $\mathbf{V}_O$.

Let we denote letters of $\mathbf{V_O}$ and $\mathbf{V_I}$ with letters of Latin alphabet, strings of letters from $\mathbf{V_I}$ and $\mathbf{V_O}$ are denoted with letters of Greek alphabet. Letters *i*, *j*, *m* and *n* are reserved for enumerations.

The aim of present research is to work out a method that allows to generate cross-language letter mappings: in other words, to determine substitutions for letters or substrings of source language among letters or substrings of target language. Let we define rules of transliteration from source language into target language as these letter mappings.

In our implementation rule is a pair $\mathbf{r} = <\mathbf{p}, \beta>$, where:

$\mathbf{p} = <\mathbf{p_l}, \alpha, \mathbf{p_r}>$, where $\mathbf{p_l}$ and $\mathbf{p_r}$ are pre-condition and post-condition, α is a transduced string, $\mathbf{p_l} = \{\gamma_1, \gamma_2, \ldots, \gamma_n\}$, $\gamma_i \in V_I^+$, $\mathbf{p_r} = \{\delta_1, \delta_2, \ldots, \delta_m\}$, $\delta_i \in V_I^+$,

β — output string. String that substitutes for string α in target language.

Consequently the rule is applicable to the current position means that symbols from α are following from the current position, α is preceded by a substring that has symbols from $\mathbf{p_l}$ on appropriate positions and followed by a substring that has symbols from $\mathbf{p_r}$ on appropriate positions.

We sequentially look for rules that can be applied to the input string. If we find one, we move the current position by $|\alpha|$ symbols from current position and return β.

We learn rules of transliteration from a manually-constructed learning set, which consists of proper names in source language and their transliterations into target language.

During the process of rules generation we need some more information about the rule: how many times and in which situations it was applied. So when learning we use full rule format, that is represented as a triplet $\mathbf{r} = <\mathbf{p}, \beta, \mathbf{w}>$, where **p** and β are the same as in format listed above and $\mathbf{w} = \{<w_1, pos_1>, \ldots, <w_n, pos_n>\}$, where w_i is a word from the learning set that satisfies the rule **r**, pos_i is a number that indicates position of the first symbol from α in w_i.

4. Method of string transformation

We needed a method of strings processing which is linear with respect to length of string and doesn't depend on amount of rules in the system.

We have decided to transform strings with state finite machine as it provides linear speed of parsing. We use an extended finite state transducer of the following structure: $g = <\mathbf{V_I}, \mathbf{V_O}, \mathbf{Q}, q_0, \mathbf{F}, \delta>$

$\mathbf{V_I}$ — input alphabet (source alphabet);

$\mathbf{V_O}$ — output alphabet (target alphabet) ;

Q — automaton's states set;

q_0 — initial state;

F — final states set;

δ — next-state function $Q \times VI \rightarrow <Q, a>$. It defines state we should move from the current state if we receive a certain input symbol, $a \in A$ is a set of actions committed during transition to the new state. A = {Out(), Shift()}:

Out(ω) — function returning a substring ω (can be empty) from the target alphabet.

Shift(*n*) — procedure that skips *n* symbols of the processed string. The default value of *n* is 1, that means that during a transition automaton moves to the next symbol.

Every final state has a transition to the initial state by the empty symbol.

The automaton is constructed from a set of rules (see section 3). Rules are sequentially added to the automaton. We start adding every rule from the initial state of the automaton.

If rule doesn't have context, that is $r(\mathbf{p}_l) = r(\mathbf{p}_r) = \varnothing$, we use the following algorithm of transformation.

Algorithm 1. Transformation of rule without context.

1. Add new state and move from the initial state to the new state by the first symbol from α
2. Move from current state n_i to state n_{i+1} by every following symbol of α (state n_{i+1} is newly constructed)
3. State n_z which we have come by the last symbol of α becomes a final state
4. Add action Out(β) to the transition to the final state n_z. In other words, if we meet substring α in a processed string, we add its transliteration (β) to the output string.

Note that value of function Shift() for every constructed transition is 1.

If rule has contexts, the procedure of its transformation is a little different.

Algorithm 2. Transformation of rule with context.

1. Add new state and move from the initial state to the new state by the first symbol from α and set $\text{Shift}() = |\gamma| \times (-1)$, $\gamma \in \mathbf{r}(\mathbf{p}_l)$ — after checking the first symbol we return back to check context;
2. Commit step 2 of Algorithm1 for every $\gamma_i \in \mathbf{r}(\mathbf{p}_l)$,
3. Commit step 2 of Algorithm1 for α
4. Commit steps 2–4 of Algorithm1 for every $\delta_i \in \mathbf{r}(\mathbf{p}_r)$ BUT set meaning of Shift() for the last transition to $|\delta\text{-}1| \times (-1)$, $\delta \in \mathbf{r}(\mathbf{p}_r)$ — after checking the rule and its post-context we return back to parse context separately.

Note that transformation demands that all $\gamma \in \mathbf{r}(\mathbf{p}_l)$ have the same length and all $\delta \in \mathbf{r}(\mathbf{p}_r)$ have the same length, yet there is no request to $|\gamma|$ and $|\delta|$ to be equal. Moreover, presence of one of contexts doesn't mean presence of the other.

Automaton constructed from system of rules using Algorithms 1 and 2 can turn out to be nondeterministic if there is a pair of rules r_1 and r_2 such that $r_1(\alpha) = <u_1, u_2, \ldots, u_n>$, $r_2(\alpha) = <v_1, v_2, \ldots, v_m>$ and $u_1 = v_1$. Of course one can process strings with nondeterministic automaton, but in this case it loses its advantage of speed. To keep this advantage we use standard procedure of determinization of state automaton [8].

5. Method of rules generation

Algorithm of rules generation can be divided into two main steps:

1. generation of initial rules;
2. generation of complicated rules.

5.1. Initial rules

We call "initial rules" all the rules that can be discovered through rather simple operations. Even so rules themselves can be nontrivial.

The core of rules generation process is association of substrings of original names with corresponding substrings of their translations. In other methods this process is purely statistical, but we use other approach.

Ideally name and its translation should consist of the same phoneme succession to be recognizable. For the initial stage of algorithm let we assume that in every language consonant phoneme is written down with one or more consonant letters and vowel phoneme — with vowel letters. From this assumption we deduce that consonant letters usually transform to consonant letters and vowels — to vowels.

In compliance with our hypothesis we divide each word (both original and translation) into groups of consonants and vowels. Let us define predicate isVowel(l), which returns **true** if l is vowel and **false** otherwise for every $l \in \mathbf{V_I} \cup \mathbf{V_O}$. For every word w = $l_1 l_2 \ldots l_n$ bound of group is placed between all such l_i and l_{i+1} that isVowel(l_i) ≠ isVowel(l_{i+1}).

So each name can be represented as a pair w = <**in**, **out**>, where:

in = <in_1, in_2, … , in_n> — set of nonempty chains of letters from $\mathbf{V_I}$

out = <o_1, o_2, …, o_m> — set of nonempty chains of letters from $\mathbf{V_O}$

Then for every word w from learning set, where |**in**| = |**out**|, we form n rules where n = |**in**| = |**out**|

$\mathbf{r}_i(\mathbf{p}_l) = \mathbf{r}_i(\mathbf{p}_r) = \varnothing$, $\mathbf{r}_i(\alpha) = in_i$, $\mathbf{r}_i(\beta) = o_i$.

In other words, we just match i-th group of original name with the i-th group of translated name, provided that name and translation have equal number of groups and i-th group of original have the same type (consonant or vowel) as i-th group of translation (see ex. 1). These mappings form the set of rules-candidates **R**.

Example 1
R | u | gg | ie | r | o M | a | cch | i
R | u | dzh | e | r | o M | a | kk | i
r → r, u → u, gg → dzh, ie → e, o → o, m → m, a → a, cch → kk, i → i

Example of generation of initial rules from names (Italian-English dataset). Vertical lines denote bounds of groups. Corresponding groups of original names and translations are united into rules

After having generated set of rules-candidates we reduce it. We remove rare rules: rules that occur in our set only once or twice. We consider them to be arbitrary letter combinations. Then we remove too big rules — rules whose left side is longer than 3 symbols. We suppose that it can be later explained with several shorter rules. Of course, this solution isn't always correct as one sound may be written down by four or more letters. So we don't remove rule with left side longer than 3 letters if its right side consists of only one letter.

We also remove rules that can be explained with shorter rules. Formally speaking, if for rule $\mathbf{r}_0$ = <**p**, β> exist $r_1, \ldots, r_n \in \mathbf{R}$ such that $\mathbf{r}_0(\alpha) = r_1(\alpha) + r_2(\alpha) + \ldots + r_n(\alpha)$ and $\mathbf{r}_0(\beta) = r_1(\beta) + r_2(\beta) + \ldots + r_n(\beta)$, then $\mathbf{r}_0$ should be removed.

After these operations the system of rules is rather adequate except of one detail. It contains ambiguous rules. We call rules $\mathbf{r}_1$ and $\mathbf{r}_2$ ambiguous if $\mathbf{r}_1(\alpha) = \mathbf{r}_2(\alpha)$, $\mathbf{r}_1(\beta) \neq \mathbf{r}_2(\beta)$ and $\mathbf{r}_1(\mathbf{p}_l) = \mathbf{r}_1(\mathbf{p}_r) = \mathbf{r}_2(\mathbf{p}_l) = \mathbf{r}_2(\mathbf{p}_r) = Æ$. Such cases can be sometimes explained with ambiguities of reading rules of the source language, but we should try to resolve them with the help of contexts. Up to now all the rules in our set had empty contexts ("$\mathbf{r}\hat{\mathbf{I}}\mathbf{R}$: $\mathbf{r}(\mathbf{p}_l) = \mathbf{r}(\mathbf{p}_r) = Æ$). As all of our rules contain reference to words where they meet we can add contexts. Contexts are letters which precede ($\mathbf{p}_l$, left context) and follow ($\mathbf{p}_r$, right context) substring $\mathbf{r}(\alpha)$ in words where it is met. This approach is useful in many cases (see example 2).

Example 2

After employing contexts we receive from two ambiguous rules c → с and c → к (French-Russian dataset) two rules:

{< a e i}c{a o} → к

{< a e i u y}c{e i} → с,

that illustrate one of French reading rules: c is read as [s] ("с" in Russian) before front vowels (e, i).

5.2. Complicated rules

After the first stage of the algorithm we achieve a system of rules that can already be used for strings conversion. If reading rules of source language are plain enough, system of initial rules can transform strings correctly. But in many cases initial rules are not sufficient.

The second step of the algorithm aims to discover rules that can't be discovered at the first step.

The second step of the algorithm consists of several minor steps:

1. Divide names into syllables
2. Try to transform syllables according to the existing rules, generate new rules
3. Go to step 2

We divide all the names and their translations into syllables. Then we try to convert every syllable from source language to target language. If the conversion was failed, that means that the translation of the syllable can't be explained with the existing system of rules. In this case we add a new rule that can explain current syllable. Then, if there are any unexplained syllables left, we repeat step 2 until all of them are explained.

5.2.1. Syllabification

Term "syllable" in the present work is used not in traditional linguistic meaning.

Syllable is nonempty substring of a given word containing one or more vowels. We use the following rules of syllabification:

- Bound of a syllable in the word w = $l_1 l_2 \ldots l_n$ is between a vowel and a consonant, i.e. after letter l_i such that isVowel(l_i) = **true** and isVowel(l_{i+1}) = **false**;

- Set of elements none of which is vowel isn't separated out, including consonants at the end of word;
- Symbols "<" and ">" marking the beginning and the end of word are considered consonants.

Thereby a syllable is a chain C^*V^+ where C is a consonant and V is vowel. Syllable can have form $C^*V^+C^+$ only if it is a last syllable of a word and final set of consonants can't be separated because there is no vowels among them.

Let we define substitution as pair of strings $\zeta_i \to \eta_i$, where $\zeta_i = <u_1, \ldots, u_n>$ is an i-th syllable of original word and $\eta_i = <v_1, \ldots, v_m>$ is an *i*-th syllable of translated word.

Actually syllable is a combination of consonant group and vowel group, that were described in section 4.1.

5.2.2. Trial transformation

We divide words into syllables, because syllables are shorter and in a syllable it's easier to discover a substring that can't be processed with the existing rules, than in a word.

We try to apply existing system of rules to every syllable ζ of original word. If we get syllable η which is right part of substitution $\zeta \to \eta$, we move to the next syllable. Otherwise we should generate a new rule as any subset of existing rules doesn't give us proper result.

Applying rules to the substitution $\zeta \to \eta$ from left to right and from right to left we can represent the substitution as $<u_1, \ldots, u_k, \lambda, u_{k+1}, \ldots, u_n> \to <v_1, \ldots, v_q, \mu, v_{q+1}, \ldots, v_m>$, where $<u_1, \ldots, u_k> \to <v_1, \ldots, v_q>$ and $<u_{k+1}, \ldots, u_n> \to <v_{q+1}, \ldots, v_m>$. So we have three different situation depending on α and β:

- $\lambda = Æ$, $\mu \neq Æ$. We add a new rule r_i such that $r_i(\mathbf{p}_l) = u_{k-1}$, $r_i(\alpha) = u_k$, $r_i(\mathbf{p}_r) = u_{k+1}$, $r_i(\beta) = v_q + \mu$. — in other words, rule for symbol that precedes α. In some contexts it is substituted with two or more symbols.
- $\lambda \neq Æ, \mu = Æ$. This situation means that some of letters of $\mathbf{V}_I$ in some contexts are not read. We add a rule r_i such that $r_i(\mathbf{p}_l) = u_k$, $r_i(\alpha) = \lambda$, $r_i(\mathbf{p}_r) = u_{k+1}$, $r_i(\beta) = Æ$.
- $\lambda \neq Æ$, $\mu \neq Æ$. We add r_i such that $r_i(\mathbf{p}_l) = r_i(\mathbf{p}_r) = Æ$, $r_i(\alpha) = \lambda$, $r_i(\beta) = \mu$. If r_i conflicts with any of exiting rules we add context to r_i.

6. Experiments

6.1. Generated rules

We evaluated our method of rules generation on parallel collections of names in French, German, Spanish, Swedish, Mongolian, Arabic and Japanese. Target language of all test collections is Russian. Collections contain proper names from various sources. They were transcribed into Russian manually by an expert. We didn't estimate fullness of collections.

We received rules of transliteration from every of listed languages into Russian. Table 1 summarizes information about size of test collections and number of generated rules.

Table 1. Results comparison

Language	Size of collection	Number of rules generated with our tool	Number of rules written by expert
Arabic	1900	63	78
French	1900	160	356
German	4200	102	121
Japanese	7000	52	131
Mongolian	230	41	46
Spanish	1000	88	106
Swedish	1900	105	423

Systems of rules that were generated by our system contain fewer rules than systems written by experts. This fact can be explained in two ways. First of all, although experts relied on test collections while writing rules, they often followed their own knowledge of source language and added rules that could not have been deduced from test examples. Thus expert added rule "{<}ai → э" ("ai" in the beginning of the word should be transliterated as "э") in French-Russian rule system, despite the absence of suitable examples in the collection. Some of such rules were generated by our tool, but then excluded as rare and insignificant. And some others were just not generated. For example, expert added rules "aa → а" and "ya → я" for Arabic as well as machine did, but rule "yaa → я" exists only in human-written rules set. Expert's set of rules for Japanese-Russian transliteration contains rules for syllables "ha", "sha" and "cha", transcribed as "ха", "ся" and "тя", respectively. However, algorithm considers only one-symbol contexts, so it generated rule "a → я" with left context "h".

Secondly, rules that were generated by computer are compressed in comparison with expert's rules. For example, expert's variant of French-Russian rules contains a set of rules for substring "ai", while computer generated only one rule "ai → e", that covers all expert's transliterations (except of above-mentioned case of "ai" in the beginning of the word). Rules of Japanese syllabary transliteration were reduced in rules for particular substrings. Thus instead of three rules for syllables that start with "ch" machine generated one rule "ch → т". We should admit that this approach is more proper, but linguists are not used to such notation.

Aside from mentioned drawbacks the rules are consistent, cover major part of test collection and can be used for transliteration of proper names.

6.2. Finite State Automaton

Rules generated by our system were also used to construct finite-state transducer. We have checked quality of learning of our method. Finite state machine has transduced names from the training set. The method has shown rather good results (see Table 2), low values at the bottom of the table can be explained with inconsistencies or mistakes in training set.

Although the method already applicable to practical tasks, it can be still improved with statistics and more full usage of contexts, so results in Table 2 are not ultimate.

Table 2. Quality of learning

Language	Size of collection	Quality of learning
French	580	97%
Mongolian	232	98%
Tagalog	286	93%
Arabic	1606	87%
Spanish	1041	86%
Polish	1413	79%
Romanian	576	78%
Swedish	1629	74%

References

1. *Akho A. V., Lam M. S., Seti R., Ul'man D. D.* 2008. Compilators. Priciples, Technologies and Instruments [Kompiliatory. Printsipy, Tekhnologii I Instrumentarii].
2. *Al-Onaizan Y., Knight K.* 2002. Machine Transliteration of Names in Arabic Text. Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages.
3. *Ermolovich D. I.* 2005. Proper Names: Theory and Practice of Interlanguage Transmission [Imena Sobstvennye: Teoriia I Praktika Mezh"iazykovoi Peredachi].
4. *Giliarevskii R. S., Starostin B. A.* 1985. Foreign Names in Russian Text [Inostrannye Imena I Nazvaniia v Risskom Tekste].
5. *Graehl J.* 1997. Carmel Finite-state Toolkit, available at: http://www.isi.edu/licensed-sw/carmel
6. *Knight K., Graehl J.* 1998. Machine Transliteration. Computational Linguistics, 24(4) : 599–612.
7. *Practical* transcription of Proper Names in the Languages of the World [Prakticheskaia Transkriptsiia Lichnykh Imen v Iazykakh Narodov Mira]. 2010.
8. *Ravi S., Knight K.* 2009. Learning Phoneme Mappings for Transliteration without Parallel Data. Human Language Technology Conference archive Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics.
9. *Sherif T., Kondrak G.* 2007. Substring-Based Transliteration. Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages.
10. *Sproat R., Tao T., Zhai C.* 2006. Named Entity Transliteration with Comparable Corpora. Proc. of ACL.
11. *Zelenko D., Aone C.* 2006. Discriminative Methods for Transliteration. Proc. of EMNLP.

ФАКТОРЫ РЕФЕРЕНЦИАЛЬНОГО ВЫБОРА: КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ

Н. В. Лукашевич (louk@mail.cir.ru)

Г. Б. Добров (wslc@rambler.ru)

МГУ, Москва, Россия

А. А. Кибрик (aakibrik@gmail.com)

Институт Лингвистики, Санкт-Петербург, Россия

М. В. Худякова (mariya.kh@gmail.com)

А. С. Линник (skylinnik@gmail.com)

МГУ, Москва, Россия

Выбор между различными типами референциальных выражений, таких как дескрипции, имена собственные и местоимения, зависит от большого числа одновременно действующих факторов. В данном исследовании роль и значимость этих факторов моделируется при помощи различных алгоритмов машинного обучения. Работа основана на специальном англоязычном корпусе RefRhet, размеченном по референции.

Ключевые слова: компьютерное моделирование, референциальный выбор, референциальное выражение, факторы, RefRhet.

FACTORS OF REFERENTIAL CHOICE: COMPUTATIONAL MODELING[1]

N. V. Loukachevitch (louk@mail.cir.ru)

G. B. Dobrov (wslc@rambler.ru)

Lomonosov Moscow State University, Moscow, Russian Federation

A. A. Kibrik (aakibrik@gmail.com)

Institute of Linguistics, Russian Academy of Sciences, Moscow, Russian Federation

M. V. Khudiakova (mariya.kh@gmail.com)

Lomonosov Moscow State University, Moscow, Russian Federation

A. S. Linnik (skylinnik@gmail.com)

Lomonosov Moscow State University, Moscow, Russian Federation

Referential choice between various referential expressions, such as descriptions, proper names, and pronouns, depends on a variety of factors. We present recent results of our modeling study into referential choice, based on the RefRhet corpus. The account of additional factors and the employment of mixed machine learning techniques enabled an improvement of referential choice prediction. This applies both to the two-way choice between full NP and pronoun and to the threeway choice "descriptive full NP vs. proper name vs. pronoun". We have demonstrated that the great majority of the factors taken into account are signifi cant for modeling the referential choice.

Key words: computational modeling, referential choice, referential expression, factors, RefRhet.

[1] This study was supported by grant #09-06-00390 from the Russian Foundation for Basic Research.

Introduction

When producing discourse, speakers or writers constantly face the necessity to mention persons or objects, that is, perform reference. When performing reference, a speaker or writer chooses between several major forms of reference, including pronouns, descriptive noun phrases, or proper names. We call this procedure referential choice.

Choosing an appropriate form of reference is apparently crucial for the overall felicity of the created discourse. Modeling referential choice is, therefore, an important part of the technologies of language generation. Referential choice is also related to the technologies of automatic text summarization. For example, Nenkova (Nenkova 2008) notes that among the major difficulties of the modern technologies of summarization, based on the identification of key sentences, is the referential organization of discourse. In particular, according to the results of DUC conference (http://duc.nist.gov/) it was found that over one half of automatically generated summaries mention entities, whose relation to the reported events remains unknown.

There is extensive linguistic literature on referential choice, see e.g. (Givón 1983), (Fox 1987). During the last decades computational models of referential choice have appeared, too, see e.g. (Strube, Wolters 2000). Kibrik (Kibrik 1996, 1999) proposed a calculative model of referential choice that pinpointed a number of factors with certain numerical weights. The sum of the weights was supposed to predict referential choice between a full NP and a pronoun. Grüning and Kibrik (Grüning, Kibrik 2005) attempted a model of referential choice in which the contribution of individual factors was defined automatically, and the interaction between factors was allowed to be non-linear. This model was based on neural networks, a well-known algorithm of machine learning. All of the mentioned Kibrik's studies explored small data sets counting one or two hundred referential expressions. In fact, the use of machine learning requires much greater data sets, which presupposes the creation of large corpora annotated for reference.

One corpus of this kind was formed for the GREC conference, see e.g. http://www.nltg.brighton.ac.uk/research/genchal10/grec/. GREC participants were supposed to demonstrate automatic systems generating appropriate referential expressions for the central entity spoken of in a text. A referential corpus of 2000 Wikipedia articles, describing people, countries, landscapes, etc., was collected and annotated for reference. 90% of the corpus was provided to conference participants for training their systems, and the results were demonstrated on the basis of a test subcorpus (10% of the corpus). Participants were allowed 48 hours for providing their results on the test subcorpus.

In (Kibrik et al. 2010) we described the computational modeling of referential choice, based on the specially designed RefRhet corpus. In this paper we discuss the specifics of referential factors, only briefly mentioned in (Kibrik et al. 2010). We also present the new results of our project.

1. RefRhet corpus: The present stage

The RefRhet corpus is based on the English-language corpus RST Discourse Treebank, created under the direction of Daniel Marcu

(http://www.isi.edu/~marcu/discourse/Corpora.html), see (Carlson et al. 2003). The corpus contains 385 Wall Street Journal articles on economics and politics. These articles contain 176 383 words and 21 789 elementary discourse units. This corpus was chosen as the basis for RefRhet because it contained annotation for rhetorical structure. Rhetorical structure has been shown to be important for reference in discourse (Fox 1987); on its basis Kibrik (Kibrik 1996) proposed the measurement of rhetorical distance that proved significant in the studies of referential choice.

The notion of rhetorical distance (RhD) is based on Rhetorical Structure Theory (Mann and Thompson 1988). This theory describes the hierarchical semantic organization of discourse. Each elementary discourse unit (most typically coinciding with a clause) is a minimal node in a rhetorical net. Terminal nodes are combined into groups in accordance with hierarchical closeness. Nodes, both terminal and complex, are connected by rhetorical relations, either symmetric (sequence, conjunction, contrast) or asymmetric (cause, condition, concession, etc.). Rhetorical distance is the measurement of the path along the rhetorical net from one node to another.

Rhetorical distance between clauses helps to take into account those instances in which the anaphor unit and the antecedent unit are hierarchically close but linearly far apart, and vice versa.

Referential annotation was added to RST Discourse Treebank, and as a result the RefRhet corpus emerged. Referential annotation was performed with the help of the MMAX-2 program, created by a group of German computational linguists specifically for modeling reference (see http://mmax2.sourceforge.net/). MMAX-2 annotation is done with the help of a so-called annotation scheme (Krasavina, Chiarcos 2007). The annotation scheme employed contains a set of annotated parameters, or factors.

An element that undergoes annotation, called markable, is a text constituent that can serve as a referential expression. Coreference relations are posited between markables. A coreference relation connects any non-first mention *n* of a referent (that is, anaphor), with the previous mention *n–1* (that is, antecedent). In addition, each markable contains a number of annotated features (grammatical role, animacy, etc.) that can affect referential choice.

Since all of the annotations are performed manually, a certain number of mistakes is inevitable. In order to exclude such mistakes the decision has been made to annotate each text twice and then compare these annotations automatically. Such comparison results in a list of markables that either appear only in one of the annotations, or have different feature values in the two annotations. Subsequently, annotators from a different group choose the correct analysis out of the two available.

The present-day stage of the RefRhet corpus is as follows: 157 texts are annotated twice, 193 texts are annotated once, and 35 texts are not yet annotated. The RefRhet corpus is among the largest of its kind that exist to date; cf. (Byron and Gegg-Harrison 2004; Ge et al.) 1998; Tetrault 2001; Orasan 2004; GREC corpora. Given that the annotation of a referential corpus is an extremely laborious task, creating a larger corpus would simpler be unpractical. From the statistical point of view, the corpus size is more than sufficient for performing machine learning studies.

2. Factors used in modeling referential choice: The full set of features

We are using the following set of factors of referential choice. Most of these factors were already mentioned in (Kibrik et al. 2010); we italicize below those factors that were added later on. Where appropriate, we indicate in parentheses the technical terms used for the factors in the annotation scheme.

Referent's features:

- Animacy: animate (human) or inanimate (non-human)
- *Gender and number* (agreement)*: masculine, feminine, neuter, plural*
- Protagonism, that is a referent's centrality in discourse (see below)

Antecedent's features:

- Affiliation in direct speech (dir_speech); this feature is relevant both for the anaphor and the antecedent, because particularly important are the situations in which they are located across a direct speech boundary
- Type of phrase (phrase_type): noun phrase, prepositional phrase, other
- Grammatical role (gramm_role): subject, direct object, indirect object, other
- Referential form (np_form, def_np_form): definite NP, with further indication of subtype, vs. proper name vs. indefinite NPs
- *Antecedent length, in words*
- *Number of markables from the anaphor back to the nearest full NP antecedent*

Anaphor's features:

- Introductory vs. repeated mention (referentiality)
- *Number of referent mention in the referential chain*
- Affiliation in direct speech (dir_speech)
- Type of phrase (phrase_type): noun phrase, prepositional phrase, other
- Grammatical role (gramm_role): subject, direct object, indirect object, other

Distances between anaphor and antecedent:

- Distance in words
- Distance in markables; this feature partly accounts for referential competition in a discourse context, that is issues related to potential ambiguity or referential conflict (see Kibrik 1987)
- Linear distance in elementary discourse units, as found in the rhetorical representation
- Rhetorical distance in elementary discourse units, as found in the rhetorical representation
- *Distance in sentences*
- *Distance in paragraphs*.

Recently we have given particular attention to modeling the factor of referent's protagonism in discourse. For this goal referential chains were identified, that is sequences of referential expressions naming the same referent. Each

referential chain has a certain length, that is, the number of referential expressions it contains.

Two models of protagonism were used. In the first one, to each referent corresponds the ratio of its referential chain length to the maximal length of a referential chain in the text. In the second model, to each referent corresponds the ratio of its referential chain to the gross number of markables in the text. In both instances the most frequently mentioned referent is the same, but relative weights of referents may be different.

In order to test to which extent a given model of protagonism corresponds to human text understanding, experiments were undertaken (Linnik 2010). Thirty texts were chosen from the RST Discourse Treebank. The length of the texts varied from 70 to 1344 words. Experiment participants were required to read the text and to identify the central entity (protagonist). Each text was analyzed by three experiment participants.

Experiment participants were thirty native speakers of English, from 20 to 54 years of age. For 50% of the texts, namely 15, all participants were unanimous in choosing the protagonist. Eleven more texts showed the agreement between two (out of three) participants in their protagonist assessment. That is, 26 texts out of 30 (87%) provide relatively reliable information on human-selected protagonists.

A comparison of the experiment results with the results of computational analysis demonstrated that the human assessment and the computer's assessment coincide in 24 instances out of 26. Therefore, the automatic models predict human identification of protagonist 92% of the time.

One more factor that deserves special mention is the factor of rhetorical distance. There are several complications in how this measurement is applied to various rhetorical configurations. These complications were discussed in (Kibrik, Krasavina 2005); in the current project we followed the methods proposed in that study.

3. Interaction between factors: methods of computer learning

In the computational model of referential choice the following two tasks were set:

- to predict whether a given anaphor is a (third person) pronoun or a full noun phrase (two-way task)
- to predict whether a given anaphor is a (third person) pronoun or a descriptive noun phrase or a proper name (three-way task).

From the beginning of this project, several algorithms of machine learning were chosen, belonging to different groups: logical classifiers and logistic regression (Kibrik et al. 2010). The results of the logical algorithms (decision trees C4.5, deciding rules algorithm JRip) lend themselves to natural interpretation. Logistic regression was chosen for the following two reasons. First, the results of this algorithm excel those of logical algorithms in quality. Second, logistic regression allows one to obtain probabilistic estimates of referential options.

More recently, we also used the so-called classifier compositions: bagging and boosting.

The boosting algorithm (Freund, Schapire 1996) uses as its parameter another machine learning algorithm that we will call the base algorithm. The base algorithm undergoes optimization. An adaptation of classifiers is performed, that is, each additional classifier applies to the objects that were not properly classified by the already constructed composition. After each call of the algorithm the distribution of weights is updated. (These are weights corresponding to the importance of the training set objects.) At each iteration the weights of each wrongly classified object increase, so that the new classifier focuses on such objects. Among the boosting algorithms, AdaBoost was used in our modeling with the C4.5 base algorithm.

Bagging (from "bootstrap aggregating"; Breiman 1994) algorithms are also algorithms of composition construction. Whereas in boosting each algorithm is trained on one and the same sample with different object weights, bagging randomly selects a subset of the training samples in order to train the base algorithm. So we get a set of algorithms built on different, even though potentially intersecting, training subsamples. A decision on classification is done through a voting procedure in which all the constructed classifiers take part. In the case of bagging the base algorithm was also C4.5.

In the current set of modeling studies we used 4291 anaphor-antecedent pairs, including 2854 full noun phrases and 1437 pronouns as anaphors. In order to control the quality of classification, the cross-validation procedure was used:

1. The training set is divided into 10 parts.
2. A classifier operates on the basis of 9 parts.
3. The constructed decision function is tested on the remaining part.

The procedure is repeated for all possible partitions, and the results are subsequently averaged. The criterion for choosing both an optimal set of features and an algorithm is **accuracy**, that is the ratio of properly predicted referential expressions to the overall amount of referential expressions.

The results of modeling studies are given in Table 1 (two-way task) and Table 2 (three-way task). In the columns "Accuracy 2010" results are provided for the set of factors included in (Kibrik et al. 2010), whereas the columns "Accuracy 2011" include the new factors incorporated into the model at the more recent stage.

Table 1. Modeling referential choice in the two-way task: full noun phrase vs. pronoun

Algorithm	Accuracy 2010	Accuracy 2011
Logistic regression	85.6%	87.0%
Decision tree algorithm	84.3%	86.3%
Deciding rules algorithm	84.5%	86.2%
Boosting	88.2%	**89.9%**
Bagging	86.6%	87.6%

Table 2. Modeling referential choice in the three-way task: descriptive noun phrase vs. proper name vs. pronoun

Algorithm	Accuracy 2010	Accuracy 2011
Logistic regression	76.0%	77.4%
Decision tree algorithm	74.3%	76.7%
Deciding rules algorithm	72.5%	75.4%
Boosting	79.3%	80.7% — 50 iterations **80.9%** — 100 iterations
Bagging	78.0%	79.5% — 50 iterations 79.6% — 100 iterations

Thus the enlistment of new features in the recent modeling studies, as well as the use of additional algorithms of machine learning, allowed us to noticeably improve the prediction of referential choice.

4. Significance of factors and factor correlations

As was shown in section 2, six different distance measurements were used. In order to find out which of the distances correlate with each other, the Spearman's correlation coefficient was computed that reveals linear dependencies between variables. If two variables have the Spearman's coefficient of 1, they are in a linear dependency. If the coefficient value is −1, there is an inverse dependence. The coefficient values obtained for all pairs of distances are shown in Table 3.

Table 3. Correlations between different anaphor–antecedent distances

Distance in:	Words	Markables	Elementary discourse units (linear)	Elementary discourse units (rhetorical)	Sentences	Paragraphs
Paragraphs	0.6629	0.5617	0.6538	0.6169	0.7734	1.0000
Sentences	0.7663	0.6034	0.7530	0.6569	1.0000	
Elementary discourse units (rhetorical)	0.5864	0.4746	0.6598	1.0000		
Elementary discourse units (linear)	0.8748	0.6753	1.0000			
Markables	0.7051	1.0000				
Words	1.0000					

As can be seen from Table 3, the maximal correlation is observed between the distance in words and the linear distance in elementary discourse units, while the minimal correlation is observed between rhetorical distance and the distance in words. Minimally correlated with other types of distance are rhetorical distance and the distance in markables. Note, however, that the cognitive interpretation of the distance in markables is yet to be determined.

Also, for the three-way task the results of classification were computed with the deduction of certain factors and groups of factors, see Table 4. An analysis of the contribution of newly added factors was also performed.

Table 4. The significance of factors in the three-way task of referential choice

Factors	Accuracy
All factors, including the newly added ones (boosting with 50 iterations)	**80.7%**
without protagonism	80.0%
without affiliation in direct speech, for both anaphor and antecedent	80.6%
without animacy	80.68%
without all distances	73.5%
— except for the distance in words only	79.0%
— except for rhetorical distance only	74.9%
— except for the distances in words and paragraphs	79.0%
— except for the distances in words and sentences	79.5%
— except for the distances in words, sentences, and paragraphs	79.4%
— except for rhetorical distance and the distances in words and sentences	79.7%
— except for the rhetorical distance and the distances in words and markables	79.9%
— except for the distances in words, markables, and paragraphs	80.47%
without the anaphor's grammatical role	79.3%
without the antecedent's grammatical role	80.2%
without grammatical role	79.2%
without the antecedent's referential form	77.0%
Old factors (Kibrik et al. 2010) (boosting with 50 iterations)	**79.3%**
plus referent number and gender	79.7%
plus number of markables to the nearest full NP plus chain length	78.9%
plus antecedent length	78.7%
plus distance in sentences	79.5%
plus distance in paragraphs	79.25%
plus antecedent gender plus distance in paragraphs plus distance in sentences	80.3%

Table 4 makes explicit the significance of various factors, such as different distance measurements, protagonism, grammatical role, antecedent's referential form, etc. Note that the inclusion of the distance in markables leads to the improvement of classification (underscored in Table 4). Perhaps this is due to the fact that this factor indeed helps to take into account referent competition or referential conflict.

The analysis of the data in Table 4 demonstrates that the great majority of the factors are significant and cannot be easily removed from the model. Even the numerous distance measurements do not lend themselves to substantial reduction.

Conclusion

In this paper we have presented the recent results of our modeling study in referential choice, based on the RefRhet corpus. The account of additional factors and the employment of compositions of machine learning techniques have led to an improvement of referential choice prediction. This applies both to the two-way choice between full NP and pronoun and to the three-way choice "descriptive NP vs. proper name vs. pronoun". We have demonstrated that the great majority of the factors taken into account are significant for modeling referential choice.

References

1. *Belz A., Kow E., Viethen J., Gatt A.* 2008. The GREC Challenge: Overview and Evaluation Results. Proceedings of the Fifth International Natural Language Generation Conference : 183–191.
2. *Breiman L.* 1994. Bagging Predictors Technical Report 421, Department of Statistics.
3. *Byron D. K., Gegg-Harrison W.* 2004. Eliminating Non-referring Noun Phrases from Coreference Resolution. Proceedings of the Discourse Anaphora and Anaphora Resolution Confenrence (DAARC2004) : 21–26.
4. *Carlson L. D., Marcu D., Okurowski M. E.* 2003. Building a Discourse-tagged Corpus in the Framework of Rhetorical Structure Theory. Current Directions in Discourse and Dialogue : 85–112.
5. *Fox B.* 1987. Discourse Structure and Anaphora in Written and Conversational English.
6. *Freund Y., Schapire R.* 1996. Experiments with a New Boosting Algorithm. Mashine Learning: Proceedings of the Thirteenth International Conference.
7. *Ge N., Hale J., Charniak E.* 1998. A Statistical Approach to Anaphora Resolution. Proceedings of the Sixth Workshop on Very Large Corpora : 161–170.
8. *Givón T.* 1983. Topic Continuity in Discourse: An Introduction. Topic Continuity in Discourse: A Quantitative Cross-language Study : 1–42.
9. *Grüning A., Kibrik A. A.* 2005. Modeling Referential Choice in Discourse: A Cognitive Calculative Approach and a Neural Networks approach. Anaphora Processing: Linguistic, Cognitive and Computational Modelling : 163–198.

10. *Kibrik A. A.* 1987. Mechanisms of Referential Conflict Removal [Mekhanizmy Ustraneniia Referentsial´nogo Konflikta]. Modelirovanie Iazykovoi Deiatel'nosti v Intellektual'nykh Sistemakh :128–145.
11. *Kibrik A. A.* 1996. Anaphora in Russian Narrative Discourse: A Cognitive Calculative Account. Studies in Anaphora :255–304.
12. *Kibrik A. A.* 1999. Reference and Working Memory: Cognitive Inferences From Discourse Observation. Discourse Studies in Cognitive Linguistics : 29–52.
13. *Kibrik A. A., Dobrov G. B., Zalmanov D. A., Linnik A. S., Loukachevitch N. V.* 2010. Referential Choice as a Multi-factor Probabilistic Process [Referencial'nyi Vybor kak Mnogofaktornyi Veroiatnostnyi Protsess]. Computational Linguistics and Intellectual Technologies. Papers from the Annual International : 173–181.
14. *Krasavina O., Chiarcos Ch.* 2007. PoCoS — Potsdam Coreference Scheme. Proceedings of the Linguistic Annotation Workshop (LAW) :156–163.
15. *Linnik A. S.* 2010. Linguistic Support for Computational Analysis of the Corpus of Texts, Annotated with Respect to the Referential Theory.
16. *Mann W. C., Thompson, S. A.* 1988. Rhetorical Structure Theory: Toward a functional theory of text organization, 8(3): 243–281.
17. *Nenkova A.* 2008. Entity-driven Rewrite for Multi-Document Summarization. Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP) : 118–125.
18. *Orasan C.* 2004. The Influence of Personal Pronouns for Automatic Summarization of Scientific Articles. Proceedings of the Discourse Anaphora and Anaphora Resolution Confenrence (DAARC2004) :127–132.
19. *Strube M., Wolters M.* 2000. A Probabilistic Genre-independent Model of Pronominalization. Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics : 18–25.
20. *Tetreault J. R.* A Corpus-Based Evaluation of Centering and Pronoun Resolution. Computational Linguistics, 27(4) : 507–520.

НОМИНАЦИИ ХАРАКТЕРОВ В ОНТОЛОГИЧЕСКОЙ ПЕРСПЕКТИВЕ

Н. Ю. Лукашевич (natalukashevich@mail.ru)

И. М. Кобозева (kobozeva@list.ru)

МГУ, Москва, Россия

Данная работа посвящена проблеме описания слов, называющих либо человека по свойствам его характера, либо сами характерологические свойства, — номинаций характера, отражающих наивную психологию носителей языка. В работе рассматривается, как знания в этой области можно представить в онтологии.

Ключевые слова: свойства характера, номинации характера, наивная психология, онтология.

CHARACTER NOMINATIONS IN ONTOLOGICAL PERSPECTIVE

N. Iu. Lukashevich (natalukashevich@mail.ru)

I. M. Kobozeva (kobozeva@list.ru)

Lomonosov Moscow State University, Moscow, Russian Federation

The focus of this research is on ways to represent the meaning of character nominations –words naming either a person according to the person's traits of character, or the characteristic itself, and providing an insight into naive psychology. An important feature of this lexical semantic group is that we attribute characteristics denoted by them to a person by generalizing from specifi c cases of the person's behaviour. Therefore the meaning of such words can be understood correctly only when both linguistic and extralinguistic information is taken into account. The paper analyses how knowledge in this sphere can be represented in an ontology.

Key words: character nominations, character traits, naïve psychology, ontology.

The focus of this research is on ways to represent the meaning of character nominations –words naming either a person according to the person's character traits, or the characteristic itself, and providing an insight into naïve psychology. An important feature of this lexical semantic group is that we attribute characteristics denoted by them to a person by generalizing from specific cases of the person's behaviour.

The fact that adequate representation of character nominations' meanings is problematic can be explained by several features of their semantics. Firstly, all character nominations have a significant subjective part shaped by personal experience in their meaning besides an invariant part (Cherneyko 1997).

Secondly, very often the meaning of character nominations includes an evaluative component, which is of greater weight in more peripheral nominations than in the central representatives of the field.

Yet another feature of this group of words is that such lexis is often culture-specific. The case is that, even for lexemes referring to the same characteristic of a person in different languages and believed to be quasi-equivalent, volumes of their meanings may differ substantially. This can be proved, for example, by the results of an experiment carried out by a Portuguese linguist J. Pinto de Lima (Pinto de Lima 1994). This experiment aiming to find semantic prototypes of Portuguese and German words denoting 'honest' and 'liar' showed that the relevant rules behind the actions of a person with such character traits were different in Portuguese and German. While for the Portuguese *honesto* the following relevant rule applied: "Do not keep for yourself, or do not get hold of, that which does not belong to you"(p.13), for the German *ehrlich* the rule could be formulated as: "Always tell the truth about your own behaviour, even if this may bring you some disadvantage" (p.14). The Russian *chestny* proved to be closer to the Portuguese *honesto* rather than the German *ehrlich* and fell somewhere in between the Portuguese *honesto* and *sincero* ('sincere') despite the fact certain features were the same for all the three characteristics (*chestny*, *ehrlich* and *honesto*) (Lukashevich N. Yu. 2004). What is important is that in none of the pairs were the word meanings exactly equivalent to each other.

The above said explains why the meaning of such words can be understood correctly only when both linguistic and extralinguistic information is taken into account. From this point of view ontologies, which have been gaining popularity in natural language processing recently, can provide previously unavailable resources for dealing with this issue.

It should be noted, however, that character nominations present certain difficulties for any attempt to represent their meanings formally. Of the possible reasons for such difficulties as discussed in (Loukaschevitch N. V. 2010), two are relevant for character nominations: fuzziness of elements constituting their meaning and the fact that this lexical group consists mostly of rows of near-synonyms. Such near-synonyms are particularly difficult for defining taxonomic relations because they usually demonstrate family resemblance (Wittgenstein 1953). It is also usually difficult to find their equivalents in other languages because in a different language as a rule they are matched with another row of near-synonyms with its different set of distinctive parameters (Loukaschevitch N. V. 2010).

If one considers how character nominations are represented in some of the existing ontologies, one can say that none of them seem to be able to account for semantic features of this group of words properly. This can be partly explained by inappropriate approaches to representing near-synonyms (Loukaschevitch N. V. 2010). However, existing ontologies also seem to show very few links between such concepts and the rest of the ontology.

For example, for such characteristics as *sincere, frank and candid* WordNet predictably provides lists of synonyms such as < **sincere** (open and genuine; not deceitful) > , <earnest, **sincere**, solemn>, <blunt, candid, forthright, **frank**, …> and <**candid**, open, heart-to-heart>. These synsets demonstrate (but in no way explicate) distinctions between *sincere* and *candid* and, unlike other sources, some distinctions between *sincere*, *frank* and *candid*, leaving the distinction between *frank* and *open* unaccounted for.

FrameNet resource relates the three lexemes to the Candidness frame listing such elements of the general situation relevant to all three characteristics as Speaker, Message, Topic, etc.

In MikroKosmos ontology the characteristics in question are related to concepts representing abstract qualities, e.g. *chestny* ('honest') is related to HONESTY-ATTRIBUTE meaning "the degree of honesty with which a person or group conducts himself or themselves" and belonging to scalar-human-attribute. All traits of character which are covered in the ontology fall into two subclasses of SCALAR-OBJECT-ATTRIBUTE. Such concepts as HONESTY-ATTRIBUTE, MODESTY, SHYNESS and KINDNESS belong to SCALAR-HUMAN-ATTRIBUTE ("scalar-attributes involving social-roles"). At the same time DECISIVENESS and SERIOUSNESS-ATTRIBUTE are in SCALAR-SOCIAL-ATTRIBUTE subclass ("an attribute with a numerical range which describes some socially-related phenomena such as salary cost-of-living … etc").

It is therefore clear that the only relation linking these concepts with other concepts in MikroKosmos ontology is the relation existing between a quality and its possessor. Other meaningful connections like, for instance, a connection between HONESTY-ATTRIBUTE and the principles involved in this type of behaviour are not represented at all. Besides that near-synonymic rows of character nominations would inevitably raise the issue of what information should be ontologised in the concept (or concepts) for a particular character trait cluster, and what should be dealt with in lexical entries. It would not seem feasible to introduce a separate concept for each particular characteristic named by one word from a list of near-synonyms, as such concepts would obviously be highly language-specific. But it is not clear how the existing HONESTY-ATTRIBUTE should be distinguished from hypothetically possible CANDIDNESS- ATTRIBUTE (should there be a separate concept introduced in this case).

In lexico-semantic studies aimed at describing the meaning of the words of a discussed semantic field and / or the structure of the field or some of its subsections we can find insights that can be used for building this part of the ontology (either universal or culture-specific). Thus, some traits of character are described as marking an attitude of a person toward some sort of entity (see e.g. an analysis of the group of words like *stingy*, *generous* and the like as denoting one's attitude towards material welfare, assets at one's disposal in (Lomtev 1969)). In the ontology such a semantic feature may be reflected by allowing conceptual relations between PROPERTY concepts

and other classes of concepts, e.g. a special relation MARKS-ATTITUDE-TO with a domain including the class of SCALAR-HUMAN-ATTRIBUTE and a range that can cover different classes of OBJECTS and EVENTS. In case we want to introduce an attribute, say, GENEROSITY into our ontology, and we already have an object concept ASSET in it (as it is in MicroKosmos ontology), we could describe the place of a new attribute not only by connecting it with IS-A relation to the class as SCALAR-HUMAN-ATTRIBUTE but by adding a slot MARKS-ATTITUDE-TO(SEM(ASSET)). In a similar way the attribute COURAGE can be related to the event-concept HAVE-FEAR (both are present in the ontology). In the analysis of a Russian word *spravedlivyj* 'just, fair' Shmelev A. D. insightfully mentioned that such a trait may be attributed to a person only in a special situation, when this person is in charge of distributing resources or inflicting punishment to other people (Shmelev 1999). For such a conceptual dependency a relation connecting a character trait to a situation in which it could be manifested, e.g. MANIFESTED-IN-EVENT with a domain including the class of SCALAR-HUMAN-ATTRIBUTE and a range over EVENTS. There also exists an obvious conceptual relation between some character traits and some concepts corresponding to «core human values» either universal or culture-specific, e.g. the concept FAIRNESS belonging to the SCALAR-SOCIAL-ATTRIBUTE class can be related to the concept JUSTICE belonging to the ABSTRACT-IDEA class in the MicroKosmos ontology. Another obvious conceptual property of some character trait concepts is their evaluative modality: polar values of attributes like FAIRNESS, HONESTY-ATTRIBUTE, LOYALTY and many others (but not WEALTH-ATTRIBUTE or SHYNESS) are associated with polar axiological evaluation (i.e. the corresponding mappings onto the good-bad scale). The axiological aspect of such concepts should also be captured in this domain of the ontology. Connecting character trait attributes to other types of ontological concepts by conceptual relations would certainly make them «more visible» in the ontology. More insight into the ontology of human characters can be achieved from the cognitive standpoint.

Within the framework of cognitive lexical semantics an approach to representing meanings of character nominations was suggested in (Lukashevich N. Yu. 2002, Lukashevich N. Yu. 2004) which seems to provide a better account for specific features of this group of words than traditional methods of semantic analysis. This approach is based on the following ideas. It is crucial that a trait of character is considered to be a behavioural stereotype which is realized with high probability in a situation typical for this trait (e.g., for *otkrovenny* 'frank, candid' the behavioural stereotype can be roughly put as 'to say something about themselves which puts them at a disadvantage' and the typical situation as 'when it is not necessary to mention a fact about themselves or say what they think or feel'). Therefore the most appropriate way to represent the meaning of a word denoting a character trait would be to set a typical situation and a stereotype of behaviour in such conditions. This can be done using behaviour pattern, a notion introduced by Yu. Martemianov and G. Dorofeyev in their works on automatic language processing ((Martemianov, Dorofeyev 1969); (Martemianov 1999)). It is based on a generalized implicative scheme establishing an association between the initial typical situation and the stereotyped behavioural response of a person with this trait of character. It is suggested in (Lukashevich N. Yu. 2002, Lukashevich N. Yu, 2004) to provide such behaviour patterns with prototypical

('best') examples of real-life situations and specific behaviour in them (e.g. *otkrovenny* 'a frank, candid person' would say that he or she often acts carelessly).

How can this approach help to represent this area of knowledge in an ontology?

Our general suggestion is that character nominations should be related to ontological concepts representing actions instead of attributes or abstract qualities, as seems to be the case with present-day ontologies. (For example, for the group of character traits describing candidness INFORM can be considered as such main action (possessors of these traits communicate something to the addressee).)

The reason for doing so is that character nominations are quite dissimilar to many other kinds of attributes (like e.g. size, material or nationality) at least in one aspect. As it has already been said above, any character trait presupposes certain actions of its possessor in certain circumstances. Because of that all character nominations invoke references to various connections reaching across sentence boundaries and existing among elements of a text in a way «routine», «typical» sequences of events (often called scripts (Schank and Abelson 1977), scenarios or complex events (Nirenburg, Raskin 2004)) do.

Although scripts generally involve multiple agents and multiple actions, while character nominations refer mostly to one action performed by one agent, the latter group of words still requires some generalized episodic knowledge in order to be understood correctly. A certain set of conditions has to be satisfied for this action to be triggered and to be classified as a manifestation of a particular character trait (e.g., a person may be sociable, shy, outgoing, etc. when with particular company (specific people or types of people); a person will be called candid in expressing their opinion only if this person does not intend to get any personal benefit from doing so, etc.).

In view of that our second suggestion is that typical situations relevant to character traits should be somehow accounted for in the ontology. Introducing such information would also link concepts representing such human characteristics with various other concepts in the ontology.

A typical situation involved in some character trait manifestation can be represented as a list of statements about the general state of affairs related to a possessor of this character trait. What is important is that such lists of statements seem to be more or less the same for words naming characteristics which belong to one cluster of character traits and forming a row of near-synonyms, with differences between them mostly expressed by different values taken by such statements for different near-synonyms. (This is illustrated below in the table showing a possible list of statements about typical situations for words naming character traits belonging to candidness group.) Each particular near-synonym would then be associated with a unique set of values taken by statements from the general list. In this case, while only one concept related to each row of such near-synonyms would be introduced in the ontology, the list of statements describing typical situations would allow to distinguish between various near-synonymic characteristics. (It may even prove possible to talk of one and the same list of such statements for all character nominations which might also be universal for different languages. This, however, requires substantial further research.)

As for the way typical situations should be represented in ontology, it remains an open question. For example, in MikroKosmos ontology this can possibly be done with the help of precondition and effect ontological properties (an attempt at representing

character traits pertaining to the concept of CANDIDNESS in this manner is shown in the table below). However, in this case it will be necessary to allow for different degrees of significance of sence elements forming precondition and effect for specific characteristics. Otherwise, such information will have to be included into respective lexical entries in the lexicon.

Another possible solution may be to use an additional level of representation as suggested in (Edmonds, Hirst 2000) in order to account for fine-grained distinctions between near-synonyms. Besides the traditional two levels of representation, a conceptual-semantic level and a syntactic-semantic level, the authors introduce a third subconceptual/stylistic level. Near-synonyms are regarded as explicitly related to each other not at a conceptual, but at a subconceptual level. A cluster of near-synonyms is believed to have internal (language-dependent) structure and is situated within a conceptual model (the ontology) on the one side and a linguistic model on the other. It is suggested that near-synonyms should be clustered under a shared coarse-grain concept rather than linked each to a separate concept. The essential shared denotational meaning of near-synonyms is represented as a core denotation on the conceptual-semantic level. As for semantic, stylistic and expressive distinctions between near-synonyms within a cluster, they are represented in terms of peripheral concepts (defined in terms of concepts in the ontology) on the subconceptual/stylistic level. While all near-synonyms in the cluster convey the concepts in the core denotation, the peripheral concepts to be conveyed depend on a particular near-synonym.

Using this approach would make it possible both to use only one concept for a row of near-synonymic characteristics and to represent typical situations in terms of concepts of the same ontology, while leaving the opportunity to account for language-dependent differences between such rows.

Another open question is how to account for the best example of a category. Some way to introduce this information is definitely desirable. This would be important not only for character nominations, but for other lexical groups of words as well, such as the ones denoting emotional relations (*love*, *friendship* etc.) or expressing interpretative notions (e.g. *help*, *heroic deed*, *betrayal*).

The problem here is that in the case when ROBIN is specified as the best example (prototype) for BIRD in English, two ontological concepts can be linked, while the same cannot be done for character nominations as it is likely there would be very few separate concepts for particular characteristics if any. Using an additional subconceptual level of representation might provide a solution in this matter as well.

To illustrate the suggested approach the table below compares much simplified representations of meanings for such Russian characteristics as *iskrenny* 'sincere', *otkrovenny* 'frank, candid', *otrkyty* 'open', *pryamoi* 'straightforward' and *pryamolineiny* 'straightforward'.

(All elements of meaning and distinctive parameters for Russian lexemes are given on the basis of analysis done within a candidate's thesis written and defended at the Department of Theoretical and Applied Linguistics of Philological Faculty of Moscow State University (Lukashevich N. Yu. 2004). For each of the words from the list an analysis of the way they are used in journalistic texts and fiction was carried out. Two text corpora were used for this purpose: the Computer corpus of Russian

newspapers which included some of the full issues of 13 Russian newspapers dated 1994–1997 (provided by the Laboratory for General and Computational Lexicology and Lexicography, the Faculty of Philology, Moscow State University) and later the Russian National Corpus. Initially the analysis was carried out using only classical literature texts of the XIXth century and the newspaper corpus. The results obtained were later revised when Russian National Corpus, covering a much wider variety of texts in genre and style, became available.

Another source of information for Russian lexemes was an experiment similar to the one conducted by Pinto de Lima as mentioned above. In this experiment subjects were asked to write short stories describing what they believed to be optimal instances of human behaviour denoted by character nominations presented to them.)

The aim of this comparison was to see what part of meaning is shared by characteristics describing the candidness cluster of character traits (i.e. may be ontologised in a concept/concepts) as it was not clear from the start how many concepts would be needed to represent these characteristics. (The fact that various dictionaries group these lexemes differently and none list all of the characteristics in one synonymic row supports the idea that the choice would not be obvious here.) Another purpose was to check if distinctions between these words can be formulated in terms of features of typical situations.

(Though expressed in a rough and informal way, all elements of meaning are of a general nature and can supposedly be formulated in terms of specific ontology's concepts.)

As it follows from the table, the characteristics describing the candidness cluster of character traits share the main action — for all of them it is 'X informs Y of Z' where 'Z corresponds to the real state of affairs'. (It should be mentioned here that not all of these characteristics manifest themselves only in speech. Thus *iskrenny* has other channels of expression, therefore in this case INFORM would only be one of the possible actions.)

Important differences lie in the topic of the communicative message a person with such character traits is making. While *iskrenny* tells the addressee something about their inner world, *otkrovenny* и *otrkyty* are disclosing not only that but also facts about themselves; besides, what *otkrovenny* is saying puts the speaker at a disadvantage. As for *pryamoi* and *pryamolineiny*, they are talking of some general state of affairs.

It is obvious that some elements of typical situations are the same for all the five characteristics. Such conditions as 'There is nothing which makes X act this way', 'X does not intend to get any personal benefit' and 'X wants Y to know Z' all need to be realized for any of these characteristics to manifest itself. It can also be noted that such behaviour is not quite standard in a sense that it is a certain deviation from the behaviour of an average person, i.e. people do not usually act this way.

Some elements of typical situations are present in some characteristics and not present in others. Thus, for *pryamoi* и *pryamolineiny* there exists a certain behaviour rule, which prohibits telling such things to other people. (This can be explained by the fact that the speaker says something which can lead to negative consequences both for the speaker and for other people.) When acting this way, *pryamoi* and *pryamolineiny* understand that they are violating this rule, but

they believe this to be a necessary and right thing to do. As far as *iskrenny*, *otkrovenny* and *otrkyty* are concerned, there is no rule which prohibits acting this way (which is apparently so because even if such behaviour may lead to negative consequences for other people, they are insignificant or unlikely). However, there are still some tacit common-sense guidelines recommending not to do such things (because this way the speaker may harm himself to some extent), which is revealed by the fact that such behaviour is perceived as not quite standard, and *iskrenny*, *otkrovenny*, *otrkyty* are aware that they are violating such common-sense rules. This can be proved comparing these characteristics with *neposredstvenny* 'unaffected, straightforward': in the latter case a person with such quality is also violating some behaviour rules, but he or she is doing it because they do not know these rules or have forgotten about their existence.

There are also features where it is possible to talk of various degrees of manifestation of some sense element in different characteristics. Thus for *otkrovenny* it is a distinctive feature that the speaker is harming himself. As for *iskrenny*, there are also cases when what such a person is saying hurts the person and/or other people, but they are not so typical as for *otkrovenny*. Who suffers more from the negative consequences — the speaker or other people — is crucial for choosing between *pryamoi* and *pryamolineiny*: what the speaker says is harmful to both parties, but *pryamoi* is used when the harm is mostly to the speaker, while *pryamolineiny* indicates that the damage is mostly done to other people.

The table below also shows that for the English group of nominations describing candidness (*sincere, frank, candid, open*) the set of conditions is very similar and the differences lie in the same zones which distinguish Russian lexemes from each other. (It should be noted here that the distinctions suggested in the table need to be further confirmed by additional research, as British National Corpus used for the analysis proved to be of insufficient size to provide enough material to define them.)

The above analysis therefore proves that a significant part of their meaning is shared by all character nominations belonging to the candidness group. On this ground to introduce one concept representing the whole cluster in the ontology would be fully justifiable. It is also shown that typical situations can be represented as lists of statements about the general state of affairs related to a possessor of such character traits and these statements are also rather similar for different characteristics.

To make a conclusion it can be said that the suggestions discussed above seem to provide a more appropriate and effective approach to treating character nominations for the purposes of natural language processing.

References

1. *Cherneiko L. O.* 1997. Lingvo-Phylosophical Analysis of Abstract Noun [Lingvo-Filosofckii Analiz Abstraktnogo Imeni].
2. *Dobrov B. V., Loukachevitch N. V.* 2005. Ontologies for Automatic Texts Processing: Concepts and Lexical Meanings Description [Ontologii dla Avtomaticheskoi Obrabotki Tekstov: Opisanie Poniatii I Leksicheskikh Znachenii]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2005" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2005") : 138–142
3. *Edmonds P., Hirst G.* 2000. Reconciling Fine-grained Knowledge and Coarse-Grained Ontologies in Representation of Near-Synonyms. Proceedings of workshop on Semantic Approximation, Granularity and Vagueness.
4. *Lomtev T. P.* 1976. Principles of Differential Semantic Signs Detection [Printsipy Vydeleniia Differentsial'nykh Semanticheskikh Priznakov]. Obshchee I Russkoe Iazykoznanie.
5. *Loukachevitch N. V.* 2010. Near-Synonyms in Linguistic Ontologies [Kvazisinonimy v Lingvisticheskikh Ontologiiakh]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2010" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2010").
6. *Lukashevich N. Iu.* 2002. Characterologic Predicates and Behavior Patterns [Kharakterologicheskie Predikaty I Shablonu Povedeniia]. Vestnik MGU, seriia Filologiia, 5 : 131–141.
7. *Lukashevich N. Iu.* 2004. Cognitive-Semantical Analysis of Predicates with the Meaning of Human Character Traits [Kognitivno-Semanticheskii Analiz Predikatov, Oboznachaiushchikh Cherty Kharaktera Cheloveka].
8. *Martem'ianov Iu. S.* 1999. Meta-languages of Sentence and Text Description [Metaiazyki Opisaniia Predlozheniia I Teksta]. Obrazovanie I Kognitivnye Tekhnologii, 3 : 124–129.
9. *Martem'ianov Iu. S., Dorofeev G. V.* 1969. Logical Conclusion and Sentences Relations in the Text Detection [Logicheskii Vyvod I Vyiavlenie Sviazei mezhdu Predlozheniiami v Tekste]. Mashinnyi Perevod I Prikladnaia Lingvistika, 12 : 36–60.
10. *Nirenburg S., Raskin V.* 2004. Ontological Semantics.
11. *Pinto de Lima J.* 1994. Exploring the Concept of Paradigm in Lexical Semantics: an Experiment on Portuguese and German Evaluative Words.
12. *Schank R. C., Abelson R. P.* 1977. Scripts, Plans, goals, and Understanding: An Inquiry into Human Knowledge Structures.
13. *Shmelev A. D.* 1999. Functional Stylistics and Moral Concepts [Funktsional'naia Stilistika I Moral'nye Kontsepty].
14. *Wittgenstein L.* 2001.Philosophical Investigations.

– — the element is not present in this character trait
(+) — the element is present to a small degree in this character trait
+ — the element is present to a high degree in this character trait

		iskrenny	**otkrovenny**	**otkryty**	**pryamoy**	**pryamolineiny**
action	X informs Y of Z	+	+	+	+	+
	Z corresponds to the real state of affairs	+	+	+	+	+
		information Z is about what X feels and thinks	Z is any negative information about X	Z is any information about X:	Z is any information	Z is any information
preconditions	There is nothing which makes X act this way;	+	+	+	+	+
	X does not intend to get any personal benefit;	+	+	+	+	+
	People usually do not tell such things to other people;	+	+	+	+	+
	There is a behaviour rule which prohibits telling such things to other people;	–	–	–	+	+
	X is aware that he is violating a behaviour rule by acting this way;	–	–	–	+	+
	X is aware that he is violating a common-sense rule by acting this way;	+	+	+	–	–
	X wants Y to know Z;	+	+	+	+	+

		iskrenny	**otkrovenny**	**otkryty**	**pryamoy**	**pryamolineiny**
effects	if Y knows Z, it may lead to negative consequences to X;	(+)	+	+	+	(+)
	if Y knows Z, it may lead to negative consequences to other people	(+)	(+)	?	(+)	+

		sincere	**frank**	**candid**	**open**
action	X informs Y of Z	+	+	+	+
	Z corresponds to the real state of affairs	+	+	+	+
		information Z is about what X feels and thinks or intends to do	Z is any (often negative) information	Z is any information about X or X's (mostly negative) opinion on some state of affairs	Z is any information
pre-conditions	There is nothing which makes X act this way	+	+	+	+
	X does not intend to get any personal benefit	+	+	+	+
	People usually do not tell such things to other people	+	+	+	+
	There is a behaviour rule which prohibits telling such things to other people	–	?	?	?
	X is aware that he is violating a behaviour or common-sense rule by acting this way	+	+	+	+
	X wants Y to know Z	+	+	+	+
effects	if Y knows Z, it may lead to negative consequences to X	(+)	+	+	+
	if Y knows Z, it may lead to negative consequences to other people	(+)	+	+	+

АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ СПОНТАННОЙ УКРАИНСКОЙ РЕЧИ (НА МАТЕРИАЛЕ АКУСТИЧЕСКОГО КОРПУСА УКРАИНСКОЙ ЭФИРНОЙ РЕЧИ)

Т. В. Людовик (tetyana.lyudovyk@gmail.com)

В. В. Пилипенко (valeriy.pylypenko@gmail.com)

В. В. Робейко (valya.robeiko@gmail.com)

Международный научно-учебный центр
информационных технологий и систем
НАН Украины и МОН Украины, Киев, Украина

Работа посвящена исследованию особенностей распознавания спонтанной речи. Основное внимание уделено настройке (обучению) акустической и лингвистической моделей, а также словарю словоформ с транскрипциями, отражающими спонтанное произнесение.

Ключевые слова: спонтанная речь, распознавание речи, акустический корпус, спонтанное произнесение.

AUTOMATIC RECOGNITION OF SPONTANEOUS UKRAINIAN SPEECH BASED ON THE UKRAINIAN BROADCAST SPEECH CORPUS

T. V. Liudovyk (tetyana.lyudovyk@gmail.com)

V. V. Pylypenko (valeriy.pylypenko@gmail.com)

V. V. Robeiko (valya.robeiko@gmail.com)

RAS of Ucraine, Kiev, Ucraine

The paper focuses on automatic recognition of spontaneous Ukrainian speech, introducing the Acoustic Corpus of Ukrainian Media Speech (ACUMS) Three confi gurations of a speech recognition system are considered. Special attention is paid to training basic and thematic acoustic and linguistic models as well as to the lexicon that contains word transcriptions

refl ecting spontaneous pronunciation. The basic acoustic model was trained on recordings from approximately 2,000 speakers (52 hours). The basic language model was trained on ACUMS texts and on texts taken from Internet (400 Mb). Spontaneous variants of word transcriptions were obtained automatically based on standard Ukrainian pronunciation. Experimental results show that clear normative speech is recognized 50 % better than less intelligible speech with hesitations and reductions. Errors are due mainly to erroneous speech corpus annotation, non-vocabulary words (proper names in particular), spontaneous manner of pronunciation, short reduced words (conjunctions and prepositions), and a strong impact of language model on the algorithm searching for the best word sequence.

Key words: spontaneous speech, speech recognition, speech corpus, spontaneous pronunciation.

1. Введение

Многие прикладные задачи в области речевых технологий связаны с распознаванием спонтанной речи. Однако, точность ее распознавания, достигаемая современными системами распознавания речи (automatic speech recognition, ASR-системами), далека от точности распознавания подготовленной (прочитанной) речи, а тем более от точности распознавания изолированно произносимых слов.

Еще в 90-х годах в [1] было проведено сравнение распознавания спонтанной и подготовленной речи на материале одних и тех же 20 дикторов и одних и тех же текстов (сначала дикторы спонтанно вели диалоги, а затем читали тексты-записи своей спонтанной речи). Этот эксперимент показал, что стиль речи является главным фактором, влияющим на точность распознавания: пословная точность распознавания прочитанных текстов составила 62,4%, тогда как точность распознавания разговорной речи — всего 47,4%.

Спонтанная речь труднее поддается автоматическому распознаванию в первую очередь из-за ее вариативности, моделированию которой уделяется большое внимание [2]. Вариативность проявляется как на аллофонном, так и на фонемном уровнях.

Особенности спонтанной речи проявляются также в нарушении плавности речи, выраженном в виде пауз хезитации (например, «а-а», «э-э»), повторов слов или их начальных частей, оговорок, а также нарушение синтаксического оформления высказывания. Не менее важен характер лексики спонтанной речи (использование социальных диалектов, «суржика»).

Следует подчеркнуть, что у разных дикторов спонтанная речь характеризуется разными особенностями: одним мало свойственны паузы хезитации (дипломаты, актеры), другим свойственно хезитативное удлинение, третьим — редуцированное произношение. Широко применяемые в распознавании речи статистические методы «усредняют» дикторов.

Как правило, в ASR-системах используются методы, основанные на скрытых Марковских моделях (hidden Markov models (HMMs) [1, 3]).

Акустические модели фонем, составляющие общую акустическую модель (АМ), получаются путем предварительного обучения системы на большом массиве данных, включающем несколько десятков или сотен часов звучащей речи вместе с ее транскрипцией [4–7]. Акустические модели учитывают аллофонную вариативность произношения (в пределах фонемы). Не зависящее от диктора (дикторонезависимое, многодикторное) распознавание речи требует для обучения речь многих сотен дикторов [8].

Лингвистическая модель (ЛМ) задает возможные последовательности слов либо в явном виде, либо в виде вероятностей следования одних слов за другими. В последнем случае ЛМ получается (обучается) путем предварительного анализа большого массива текстов.

Третьим компонентом ASR-системы является словарь словоформ с транскрипциями (словарь распознавания), используемый непосредственно в процессе распознавания [7, 9]. Именно в этом словаре должна быть отражена вариативность произношения на фонемном уровне, свойственная спонтанной речи. Однако, простое расширение словаря транскрипций за счет добавления вариативных произнесений иногда приводит не к повышению, а к понижению точности распознавания из-за того, что разные слова представлены одинаковыми или похожими транскрипциями [2, 9–11]. Тем не менее, удачный выбор количества вариантов транскрипций одного слова позволил повысить точность распознавания с 78 % до 85,7 % [4].

Обученные АМ и ЛМ, а также словарь распознавания используются в процессе распознавания речи для поиска наиболее вероятной последовательности слов, соответствующей входному речевому сигналу.

В данной работе основное внимание уделено:

А) акустической модели;

Б) лингвистической модели;

В) словарю, используемому при распознавании.

Исследуется распознавание спонтанной украинской речи, в частности, применительно к конкретной предметной области (судебные заседания).

2. Цель исследования

Основной задачей исследования было повышение точности распознавания украинской спонтанной речи.

Были поставлены следующие цели:

- провести эксперимент с использованием базовой системы распознавания украинской речи;
- проанализировать ошибки, предложить и реализовать меры по повышению точности распознавания с учетом спонтанного характера речи и сужения ее тематики; провести соответствующие эксперименты;

- сравнить результаты распознавания спонтанной речи актеров в роли судьи на материале телепередач и речи реального судьи на материале выступлений в ходе судебных заседаний.

3. Материал для исследований

3.1. Речевой материал

Речевой материал для исследований был взят из Акустического корпуса украинской эфирной речи (АКУЭМ) [12]. В АКУЭМ представлена как украинская, так и русская речь, как подготовленная, так и спонтанная. Русская речь в данной работе не анализировалась.

Для экспериментов по распознаванию речи, относящейся к судебной тематике, использовалась только часть аудиофайлов. Это в основном записи телепередач «Судові справи» («Судебные дела»). Речь, звучащую в этих телепередачах, можно назвать спонтанной по форме, но не по содержанию, поскольку дикторы говорили в рамках соответствующих ролей. Кроме этого, часть аудиофайлов содержит записи реальных судебных заседаний, в которых присутствует как спонтанная речь судьи, так и неподготовленное (и, таким образом, приближенное к спонтанному) чтение протоколов.

Речевой материал, использованный для построения АМ, состоял из аудиозаписей (длительностью около 52 часов), в которых содержится речь около 2000 дикторов. Распределение неравномерное: большинство дикторов представлено короткими записями, однако, у 150 дикторов длительность записей составляет более 10 минут.

3.2. Текстовый материал

Текстовый материал, использованный для построения лингвистических моделей, состоит из текстов корпуса АКУЭМ, и текстов, загруженных из Интернета (400 Мбайт).

3.3. Контрольная выборка

Контрольная выборка для всех экспериментов использовалась одна и та же. Для распознавания использовались записи длительностью 3,74 часа, в которых встретилось 29 500 реализаций слов. Всего в контрольной выборке присутствовала речь 34 дикторов. Темп произнесения — средний и быстрый.

В таблице 1 представлены характеристики речи некоторых дикторов, чья речь вошла в контрольную выборку. Внимание было обращено на речь «главных действующих лиц» судебных заседаний: судей, прокуроров, адвокатов, судебного секретаря, судебного пристава.

Таблица 1. Характеристики речи некоторых дикторов контрольной выборки

Дикторы	Пол	Профессия	Степень нормативности	Степень разборчивости	Склонность к хезитации	Склонность к редукции
Окис	м.	актер в роли судьи	средняя	средняя	слабая	сильная
Калинская	ж.	актриса в роли судьи	средняя	средняя	слабая	очень сильная
Ш.	ж.	судья	средняя	средняя	слабая	слабая
Антонюк	ж.	актриса в роли прокурора	средняя	средняя	сильная	слабая
Наум	м.	актер в роли прокурора	низкая	низкая	сильная	сильная
Бойко	м.	актер в роли прокурора	низкая	средняя	сильная	сильная
Бевз	м.	актер в роли адвоката	средняя	средняя	сильная	сильная
Жуковская	ж.	актриса в роли адвоката	средняя	средняя	сильная	слабая
Бабич	ж.	актриса в роли адвоката	средняя	средняя	сильная	сильная
Бузаджи	м.	актер в роли адвоката	средняя	средняя	сильная	сильная
Солодко	м.	актер в роли адвоката	средняя	средняя	сильная	сильная
Сологуб	ж.	актриса в роли судебного секретаря	высокая	высокая	слабая	слабая

4. Система распознавания речи

Для исследований использовался инструментарий HTK [13]. На его основе была создана многодикторная система распознавания речи [5]. На рисунке 1 представлены базовая конфигурация системы распознавания речи и две ее модификации, отличающиеся комбинациями АМ, ЛМ и словарей.

4.1. Описание базового варианта системы

Предварительная обработка речевого сигнала описана в [5].

В качестве АМ используются скрытые Марковские модели, обученные на всей украинской речи корпуса АКУЭМ всех дикторов. 56 украинских контекстно-независимых фонем (включая фонему-паузу) моделируются тремя состояниями Марковской цепи без пропусков. Используется диагональный вид Гауссовских функций плотности вероятности. Редко встречающиеся фонемы моделируются 64 смесями Гауссовских функций плотности вероятности, более часто встречающиеся фонемы моделируются большим числом смесей, наиболее часто встречающиеся фонемы используют 1024 смесей.

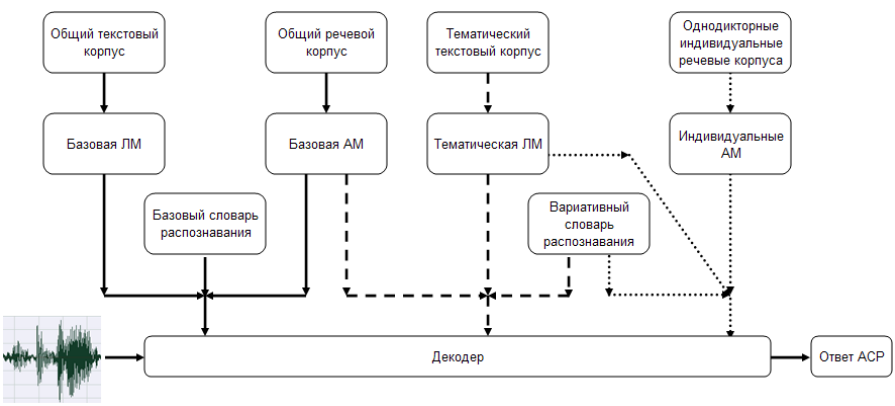


Рис. 1. Различные конфигурации системы распознавания речи: базовая (сплошные стрелки) и ее возможные комбинации: тематическая (пунктирные стрелки) и индивидуальная (точечные стрелки)

В корпусе АКУЭМ отмечены такие паралингвистические явления как вдох-выдох, кашель, смех, плач, причмокивание, а также паузы хезитации («а-а-а», «е-е-е», «м-м», …). Различаются фоновые паралингвистические явления (наложенные на речь) и изолированные. Последние при построении АМ рассматриваются как отдельные слова, состоящие из одной фонемы. Распознанные слова-»паралингвизмы» впоследствии удаляются из окончательного ответа.

Для построения ЛМ тексты, загруженные из Интернета, и тексты корпуса АКУЭМ были модифицированы с целью удаления служебной информации, записи чисел в текстовом виде, а также отделения украиноязычных фрагментов от русскоязычных. На основе полученного материала была построена биграммная модель языка, заданная вероятностями появления пар слов. Поскольку в текстах, на которых вычислялись статистики, встретились далеко не все пары слов, входящие в словарь распознавания системы, для аппроксимации ненайденных пар слов использовались обратные (back off) коэффициенты.

Словарь распознавания базовой системы насчитывает 42598 словоформ. Произнесение каждой словоформы представлено транскрипцией, несколько

отличающейся от канонической (литературной). А именно, односложные словоформы представлены двумя транскрипциями (ударный и безударный варианты), а также упрощены некоторые сочетания согласных в соответствии со спонтанным произнесением (например, «дч» → «чч» вместо канонического «джч»).

4.2. Результаты распознавания базовым вариантом системы

Одним из главных показателей работы систем автоматического распознавания речи является точность (надежность) распознавания. В проведенных экспериментах точность распознавания измеряется путем сравнения орфографических транскрипций, имеющихся в корпусе, с текстами, получаемыми на выходе системы распознавания речи.

Обычно измеряется пословная точность распознавания. Ошибками считаются вставки, пропуски, замены слов. Вставки, пропуски и замены на уровне лексем являются серьезными, поскольку не позволяют восстановить смысл произнесенной фразы. Замены, вставки и пропуски на уровне словоформ являются незначительными, поскольку, как правило, не искажают смысл произнесенной фразы, а лишь нарушают ее грамматическое оформление.

Таблица 2. Результаты распознавания речи контрольной выборки базовым вариантом системы

Дикторы	Профессия	Точность распознавания (%)
Окис	актер в роли судьи	73,47
Калинская	актриса в роли судьи	58,65
Ш.	судья	59,47
Антонюк	актриса в роли прокурора	63,90
Наум	актер в роли прокурора	59,10
Бойко	актер в роли прокурора	57,76
Бевз	актер в роли адвоката	55,93
Жуковская	актриса в роли адвоката	66,38
Бабич	актриса в роли адвоката	51,64
Бузаджи	актер в роли адвоката	60,28
Солодко	актер в роли адвоката	46,95
Сологуб	актриса в роли судебного секретаря	81,26

В таблице 2 представлены результаты распознавания речи контрольной выборки базовым вариантом системы.

Ниже приведены примеры в виде пар «произнесено диктором — распознано системой».

Диктор-мужчина — судья Окис:

А) произнесено: «***введення наркотичних засобів*** *в організм* ***людини***»

Б) распознано: «***введення наркотичних засобів*** *організму* ***людини***»

Диктор-мужчина — адвокат Бевз:
А) произнесено: «*то* ***що ми*** *передивилися* ***файл***»
Б) распознано: «*тому* ***що ми*** *подивилися* ***файл***»

4.3. Анализ ошибок распознавания базовой системой

Обнаруженные ошибки можно разделить на несколько групп:

1) Ошибки, допущенные стенографистами и экспертами на этапе аннотирования корпуса АКУЭМ и словаря распознавания. Обычно это орфографические ошибки, а также необозначение таких явлений как звучащие паузы, вдох, смех, оговорки, фоновые звуки (например, музыка) и речь посторонних дикторов. Ошибки, допущенные на этапе аннотирования корпуса, приводят в дальнейшем к ошибкам распознавания. Примеры: «*слухается*» (правильно «*слухається*»), «*пятнадцять*» (правильно «*п'ятнадцять*»).
2) Ошибки, связанные со словарем распознавания. Это или отсутствие в словаре слов, содержащихся в речи (OOV — out of vocabulary ошибки), или ошибки в орфографии и/или транскрипции. Наиболее часто в разряд OOV попадают фамилии и географические названия. Примеры:фамилия «*гуріна*» распозналась как «*горі на*», фамилия «*фещук*» распозналась как «*те що*».
3) Ошибки, связанные с вариативностью спонтанного произношения. В спонтанной речи некоторые слова теряют ударение, наблюдается редукция (сокращение, упрощение произнесения и выпадение отдельных звуков и звукосочетаний). Как правило, сильно редуцируются при произнесении часто употребляемые слова и выражения. Например, «*слово*» было произнесено как [слО] и распозналось как «*село*» (по-украински произносится «*сэло*»). Аналогично, «*як ви бачите*» [йакубАчити] распозналось как «*я побачити*», «*знаємо*» [знАйми] распозналось как «*з нами*».
4) Ошибки, связанные с распознаванием коротких слов (предлогов, союзов), присутствие которых в речи скорее «угадывается», чем слышится на самом деле.
5) Ошибки, связанные с алгоритмом поиска правильной последовательности слов. Правильная гипотеза может быть отброшена из-за ее малой вероятности, предсказываемой ЛМ.

Высокая степень нормативности и разборчивости речи, а также отсутствие склонности к хезитации и редукции положительно сказываются на точности распознавания (в среднем 81,26% точности). Речь с низкой степенью нормативности и разборчивости распознается плохо (58–59%). Склонность диктора одновременно к хезитации и редукции отрицательно сказывается на точности распознавания (47–60%).

Результаты распознавания речи базовой системой показали, что речь актеров, исполняющих в телепередачах роли судей, прокуроров и адвокатов, распознается лучше, чем речь реального судьи, записанная во время судебного заседания. Это может быть объяснено тем, что в реальных обстоятельствах средний темп речи более быстрый.

5. Влияние тематической (судебной) ЛМ и индивидуальных АМ на точность распознавания речи

Целью экспериментов была проверка того, как влияют на точность распознавания речи а) ограничение тематики; б) максимальное ограничение количества дикторов. В последнем случае речь идет фактически об однодикторной системе распознавания речи.

5.1. Тематическая (судебная) ЛМ

ЛМ, ориентированная на судебную тематику, строилась на текстах из трех источников:

1) Тексты из Интернета (400М);
2) Тексты на судебную тематику, выделенные из АКУЭМ путем автоматической кластеризации;
3) Искусственно созданный текст, в который вошли, в частности, последовательности числительных и обозначения дат (например, «тридцяте січня дві тисячі одинадцятого року»).

5.2. Индивидуальные АМ

Были созданы две индивидуальные АМ на основе речи отдельных дикторов. Для одной из них обучающий материал составил 1,91 часа речи актера-мужчины, исполняющего роль судьи. Вторая индивидуальная модель была обучена на речи диктора-женщины, играющей роль прокурора (1,4 часа).

5.3. Словарь распознаваниия, учитывающий спонтанное произнесение

Все слова можно условно разделить на классы, встречающиеся в речи с разной частотой и подвергающиеся разной степени редукции [9]. В связи с этим было произведено разбиение общего словаря, используемого на этапе распознавания, на подсловари, отличающиеся количеством вариантов произнесения, приходящихся на одну словоформу.

Наибольшее число вариантов транскрипций имеют подсловари «наиболее частотные общеупотребительные слова» (1000 словоформ), «наиболее частотные слова судебной тематики» (1000 словоформ, без пересечения с общеупотребительными), «имена, отчества и фамилии», а также «числительные».

Меньшим количеством вариантов транскрипций представлены «имена собственные географические», «аббревиатуры», «социальные и территориальные диалекты, «суржик» и «устойчивые словосочетания».

Таким образом, более часто встречающиеся словоформы представлены большим числом транскрипций. Большая часть дополнительных транскрипций была получена автоматически с использованием правил, выведенных на основе анализа речевого материала [5]. Меньшая часть дополнительных транскрипций была написана вручную.

Приравнивание устойчивых словосочетаний к отдельным словоформам («ваша честь», «будь ласка») позволяет частично учитывать фонемную и аллофонную вариативность на стыках словоформ.

Объем тематического словаря, учитывающего вариативность спонтанного произнесения, составляет 22 947 словоформ, в среднем 1,35 транскрипций на одну словоформу. В таблице 3 приведены примеры словоформ и их транскрипций.

Таблица 3. Примеры вариантов транскрипций словоформ

Словоформа	**Литературная фонемная транскрипция**	**Спонтанные фонемные транскрипции**
виявлено	в И й а в л е н о	в И й а л е н о в И й л е н о в И й л и н и
п'ятнадцять	п й а т н А дз' ц' а т'	п й а т н А ц' а т' п' а т н А ц' а т' п' а т н А ц'
в'ячеславовича	в й а ч е с л А в о в и ч а	в' а ч е с л А в о в и ч а в' а ч е с л А в и ч а

Результаты распознавания контрольной выборки конфигурацией системы, включающей базовую АМ, судебную ЛМ и вариативный словарь, ориентированный на спонтанное произнесение (рис. 1) в целом не отличаются от результатов, достигнутых базовым вариантом системы.

Конфигурация из индивидуальной АМ, судебной ЛМ и вариативного словаря, как и следовало ожидать, показала лучшие результаты, чем базовая модель. Обучение АМ только на речи актера, исполняющего роль судьи, повысило точность распознавания его речи на 3% (с 73,47% до 76,84%). Обучение АМ только на речи актрисы, играющей роль прокурора, привело к повышению точности распознавания ее речи на 5% (с 63,90% до 69,24%).

6. Направления будущих исследований

Для повышения точности распознавания спонтанной речи многодикторной системой представляются целесообразными исследования в следующих направлениях:

- Совершенствовать АМ путем адаптции к отдельным дикторам или группам дикторов, а также к темпу речи.

- В случае отсутствия распознаваемого слова в словаре выдавать его фонетическую транскрипцию.
- При различении омофонов учитывать их частотность в конкретной предметной области.
- Сбалансировать набор правил, порождающих варианты транскрипций спонтанного произнесения, и адаптировать эти правила к произношению отдельных дикторов.
- Автоматизировать выявление часто встречающихся устойчивых словосочетаний (multi-words) в рамках предметных областей.
- Предложить более гибкий критерий оценки правильности распознавания, в частности, замены одной словоформы лексемы другой ее словоформой считать менее грубой ошибкой, чем замену одной лексемы другой лексемой.

7. Выводы

Базовая система распознавания речи обеспечивает 59,61 % точности распознавания спонтанной украинской речи. Наилучшие результаты достигнуты при распознавании речи диктора, отличающегося высокой степенью нормативности произношения, отсутствием склонности к хезитации и редукции.

Ограничение тематики распознаваемой спонтанной речи не привело к повышению точности.

Индивидуальные АМ позволили значительно (на 3–5 %) повысить точность распознавания.

Система распознавания спонтанной речи может быть использована для автоматизации документооброта в судах.

References

1. *Amdal I., Fosler-Lussier E.* 2003. Pronunciation Variation Modeling in Automatic Speech Recognition. Telektronikk : 70–82.
2. *Burdic J.* 2004. Building a Regionally Inclusive Dictionary for Speech Recognition. http://surj.stanford.edu/2004/pdfs/burdick.pdf
3. *Byrne B., Finke M., Khudanpur S., McDonough J., Nock H., Riley, M., Saraclar M., Wooters C., Zavaliagkos G.* 1997. Pronunciation Modeling for Conversational Speech Recognition: a Status Report from WS97. Automatic Speech Recognition and Understanding :26–33.
4. *Hillard D., Hwang M., Harper M., Ostendorf M.* 2008. Parsing-Based Objective Functions For Speech Recognition In Translation Applications. ICASSP : 5109–5112.
5. *Niklolenko S. I., Korenevskii M. L., Ponomareva I. A., Levin K. E.* 2010. Double Recognition Based on Speech Thematic Classification. Chetvertyi Mezhdistsiplinarnyi Seminar “Analiz Razgovornoi Russkoi Rechi” : 28–32.

6. *Ostendorf M., Byrne B., Bacchiani M., Finke M., Gunawardana A., Ross K., Roweis S., Shriberg E., Talkin D., Waibel A., Wheatley B., Zeppenfeld T.* 1996. Modeling Systematic Variations in Pronunciation Via a Language-Dependent Hidden Speaking Mode. Proc. Intl. Conf. on Spoken Language Processing.
7. *Pilipenko V. V., Robeiko V. V.* 2008. Automatic Stenographer of Ucrainian Speech [Avtomatizirovannyi Stenograf Ukrainskoi Rechi]. Iskustvennyi Intellekt, 4 : 768–775.
8. *Rabiner L. R., Juang B. H.* 1986. An Introduction to Hidden Markov Models. IEEE ASSP Mag : 4–16.
9. *Strik H., Cucchiarini C.* 1998. Modeling Pronunciation Variation for ASR: Overview and Comparison of Methods. Proc. ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition:137–144.
10. *Vasil'eva N. B., Pylypenko V. V., Raduts'kii O. M., Robeiko V. V., Sazhok M. M.* 2010. Creation of the Acustic Spoken Ucrainian Speech Corpus [Stvorennia Akustichnogo Korpusu Ukrains'kogo Efirnogo Movlennia]. Obrabotka Signaliv Izobrazhen' ta Rospiznavannia Obraziv: 10 Vseukrains'ka Mizhnarodna Konferentsiia : 55–58.
11. *Weintraub M., Taussig K., Hunicke-Smith K., Snodgrass A.* 1996. Effect of Speaking Style on LVCSR Performance. Proc. Intl. Conf. on Spoken Language Processing. Philadelphia :16–19.
12. *Young S.* et al. 2009. The HTK Book (for HTK Version 3.4), available at: http://htk.eng.cam.ac.uk/
13. *Zulkarneev M. Iu., Satunovskii P. S.* 2009. Variability of Pronunciation Modeling using Hidden Markov Models [Modelirovanie Variativnosti Proiznosheniia s Ispol'zovaniem Skrytykh Markovskikh Modelei]. Tretii Mezhdistsiplinarnyi Seminar "Analiz Razgovornoi Russkoi Rechi" : 74–78.

ИЛЛЮСТРАТИВНЫЕ ЖЕСТЫ КАК КЛЮЧ К МАКРОСТРУКТУРЕ ДИСКУРСА

Ю. Николаева (lis_julia@list.ru)

МГУ, Москва, Россия

Ключевые слова: жесты, иллюстративные жесты, структура дискурса, макроструктура.

ILLUSTRATIVE GESTURES AS MARKERS FOR DISCOURSE MACROSTRUCTURE

Iu. Nikolaeva (lis_julia@list.ru)

Lomonosov Moscow State University, Moscow, Russian Federation

This article explores interrelations between discourse structure and gestures accompanying oral narration. It shows how illustrative gestures reveal discourse macrostructure. Also it discusses some issues of speech production and comprehension and the role gesture play in it.

Key words: gestures, discourse structure, macrostructure, illustrative gestures.

Discourse as a text in communication acquires additional dimensions. One of them is nonverbal component. According to A. Mehrabian, in face-to-face communication visual modality transfers more than 50% of information. Visual means are poses of interlocutors, their face expressions, appearance and the most important — gestures.

Gestures add new possibilities to verbal modality, having much different characteristics. Language is grammatically determined, so the proposition is built according to presupposed rules, and gestures are free of any structure. Language has paradigmatic and syntagmatic oppositions, but in gestures such contrasting appears only if it approaches language (such as sign languages). Arbitrary mappings are determined by the necessity of distinguishing and contrasting. Gesture form is determined by its meaning (McNeill 1992: 23). These features let the gesturer to express with gestures only those meanings he considers relevant.

Russian language, as well as English, has very delimited possibilities to express discourse structure. It can be e.g. conjunctions (so, then), referential means (full

NP or pronoun), pauses in speech and paragraphs in written text. One can suppose that gestures have their own means to express discourse structure and, maybe, these means are more elaborated and more commonly used.

In this article we investigate gesture characteristics, which can reveal discourse structure. Also we use Van Dijk's (1978) model of macrostructure.

Macrostructure is a number of macropropositions, i. e. propositions built accordingly to special rules out of original text. These rules are:

— Generalization (several propositions are generalized by a super-concept)
— Deletion (for unnecessary information)
— Construction (when the new proposition is a condition, a component or a consequence of replaced propositions).

The text built upon these rules should still remain coherent. The rules can be applied to the new text, and so on recursively. It's worth to note that building macrostructures is similar to storing information in long-term memory. Also macrostructure reveals one of strategies of discourse understanding.

Illustrative gestures, accompanying oral speech, have no specified form and are created spontaneously in the moment of communication. This distinguishes them from emblematic gestures, such as "to put one's forefinger to one's temple and twist it" or "to cock a snook". Emblematic gestures, or emblems, are specific in every culture; they are described in dictionaries and quite well studied by linguists, maybe, because they are very mush similar to language. Illustrative gestures, having great prevalence in speech (one hundred to one, approximately), are studied far less, although many people use them every day, explaining the route to a stranger, talking to a foreigner and in many more common situations. In everyday life illustrative gestures are used very often and they perform some important functions. Some of these functions are discussed below.

Illustrative gestures are divided in four types:

1. Deictic gestures are usually performed by a hand or a finger. They are referred to a point in a space around the speaker. The referent of a deictic gesture may be within eyeshot, may be located somewhere far away, and may be fictional or abstract. Mostly these gestures accompany noun phrases (87 %). Other uses fall on time and place adverbs.
2. Graphic gestures are complicated movements that directly or metaphorically depict some ideas, "draw" in air an illustration to speaker's words.
3. Illustrative regulators relate to accompanied word as metatext . The most often are conduit metaphors (Lacoff, Johnson 1980), when the speaker turns his palm up and moves his hand towards a listener, like giving him the story.
4. Beat gestures are divided into two groups — single and multiple. They are simple moves, depicted only by two vectors. They can be short cutting strokes or more smooth side moves (see Крейдлин 2003).

We will scrutinize graphic gestures meaning their special role in discourse. They mark key phrases in narration following clauses with new and least predictive information.

Graphic gestures help the speaker to process visual information (Goldin-Meadow 2005), and speaker — to make his/her own visual presentation for depicted events (Cassell et al. 1999). These gestures play a special role in discourse: they determine about 55% of acquired information, but remain almost unnoticed by a listener (ibid.).

Our study uses the case formed the retelling of the "Pear film" (Chafe 1980) by MSU students.

We propose a hypothesis that graphic gestures reveal discourse macrostructure. They mark the moments, which the speaker considers to be the crucial. Analysis of some similar narrations can show common tendencies and individual distinctions on the use of graphic gestures.

He is the example, how taking the clauses accompanied by graphic gestures we can get the discourse macrostructure.

The most speakers retold the film following this scheme (each item was present at half stories at least).

1. Introduction (describing the film or the listener's task)
2. Scenery
3. Appearance of a gardener
4. The gardener picks pears on a tree
5. A man with a goat passes by
6. The gardener continues to pick pears
7. A boy on a bike rides by
8. He stops
9. Takes a basket
10. Puts it on the bike
11. And goes away
12. He rides further
13. And meets a girls on a bike
14. He is lost in contemplation of her
15. And loses his hat
16. He falls down
17. The pears scatter
18. There go three other boys
19. They help him to stand up
20. Then they notice his hat on the road
21. They give him the hat
22. He gives them pears
23. The three boys go past the tree, where the gardener picks the pears
24. The gardener climbs down
25. And reveals the absence of a basket
26. At the moment three boys pass by
27. The gardener is surprised
28. Coda ("That's all").

Here are two examples composed from only the clauses accompanied by graphic gestures. Number before each line relates to the plane's items.

(1)

2. {.. (0.7) so mountains} [1]
3. {(... 0.5) some bushes}
4. (...0.8) who(.. a 0.5) staying on a wooden steps (...0.8)a l- ladder,
Puts in his apron,
That guy so slowly picks them,
{then climbs down the ladder},
(...a 0.8) puts these pears in a basket.
{there are (...0.5) three baskets,
(...0.5) so, he fills them gradually.
I mean one is filled already,
..0.2) well, so {leisurely},
Takes them from his apron,
Puts in the basket,
(..0.3) {then climbs back up the ladder},
(...0.9) also {so_
All so slowly,
5. A donkey so looks at the pears,
Passes by (laugh),
6. (...a 1.7) this man still is up on the ladder,
Picks the pears (....1.1) from the pear tree_
8. Looks anxiously at this (... 0.8) man,
(...aa 1.3) aand seemingly wants (... 0.6) the pears so to take,
Thinks to take or not to take,
9. (.. 0.2) well and then he sees that this man does not notices him,
(..0.2) then he {(...0.7) Ve}ry {calmly ta|kes the whole bas|ket of |pears},
10. {puts on his bi|ke} {at the front},
11. (...0.6) well, and goes.
12. (..a 0.6) he rides (..0.3) on (....1.0) a field,
13. (...0.7) also such a typical country girl with long plaits,
14. ...0.7) well, and he {at her looks},
She also rides a bike},
16. (..0.2) { well in general they |collide,
(..0.3) and} { mhm | (.... 1.1) | the boy | falls |from the bike},
17. The bike falls,
{the pears scatter}.
18. he {sits among the pears,
(...0.6)} and rubs hs leg.
15. (....a 1.6) he also} {let the hat fall down.

[1] The symbols used are following: underlined words were accompanied by lengthy gestures. Braces mark the words with any kind of hand movements: often preparative and concluding parts of gesture take some time. The gesture stroke, if it was remarkable, is pointed by vertical bar. About oral discourse notation see Кибрик, Подлесская 2009.

19. (...0.5) in general}{when the boys helped him to pick the pears,
20. he (..0.3) {leaves already,
They} {whistle him so,
Like "you} forgot {the hat»,
21. give him the hat},
(..0.2) {in general (...0.5) | the first boy on the bike} {goes with his plunder},
23. E|at these |pears.
(... aa .. 2.1) So they go| and go|,
{and appear} {in a moment near this man},
24. (...0.6) he (..0.5) wants to put them in a basket,
25. (...0.7) and at| the mo|ment appe|ar these three boys,
And eat |pears with | such a gusto.
26. (..0.4) well, (.. a 0.6) this man {so perplexedly} {looks at them,
Sees there a basket} is absent,
(...0.9) and the pears they eat.
27. The boys go away,
(...0.5) and this man stays with |his pears.

(2)

4. And puts them in his (..0.2) apron,
(..0.4) and then in a bas|ket.
9. (....1.0) took one basket,
11. (..0.2) and stole it.
(.0.3) put on his luggage rack,
12. And went further.
15. (...0.5) aand he lost his hat.
16. And fell down.
17. (..0.3) scattered all the pears
22. For that he gave them {three pears},
24. (...0.9) By that time the farmer climbs down {(...0.5) the tree},
25. (..0.3) sees {that there is no}one bas|ket,

The examples reveal that the speaker follows the line he considers the plot of the story, although details can differ a lot.

Another interesting nuance shown by these examples is that clauses with graphic gestures, marking the key events, are oriented on description of actions and state changes, so the characters in these clauses are often named by pronouns or are not mentioned at all. This corresponds to Vygotsky's ideas about internal predicate, which is actually the newest in the sentence (Выготский 2005).

There is also dependence between number of gestures and accuracy of the retelling. Our case is not great for detailed quantitative analysis, so these observations remain within the limits of a hypothesis. In the whole, we can suppose, that if the number of clauses with gestures is less than 50 %, there is higher possibility of speaker's mistakes (when the characters or their actions are depicted

inaccurately, there are a lot of self-corrections and returns to the already told) or listener's misunderstanding (expressed, e. g., in questions such as "who rode away?"). It's not a rule, just a tendency. For example, in the narration with minimum gestures (only 7 % clauses were with visual illustrations) there were no such mistakes and listener's questions.

Whish of the factors (deficit of gestures or vagueness of narration) is dominant, is not yet clear. We have an example which can point out a possible answer.

Here is a part of this narration (bold are listener's remarks)

(3)

Then three boys pass by,
From somewhere
(.. 0.2) they help him to collect the pears,
Shake him off,
(... aaa .. 1.5) and the boy quickly leaves,
(...0.5) then these boys=
— **With the pears.**
— Yes, | with the pears.
He went further already,
Then these boys whistle,
(... 0.7) (Like=
(..0.2) A sort of= this story is without words_)
But(...0.5) They wanted to give him his hat.
They returned,
Took three pears,
(... aam .... 2.0) started eating them,
— **You mean he came back**,
Gave them pears_
— No, he went further,
(...0.9) they whistled,
And he stopped,
They approached him,
Took the pears.
(... 0.5) Well, they go=
(aam0.7) they go,
And pass by the tree,
Where the man picks the pears.

This example shows, that after listener's questions the number of speaker's gestures increases. It's obvious, that the purpose of his gesticulation is to explain clearer who and where moves in the film. On the other hand, visual signs undoubtedly help the speaker to recall the plot and to process spatial-dynamic information, so it's easier for him to convey his ideas verbally.

Upon these observations we can suppose the following conclusions.

1. Graphic gestures mark the points in the narration the speaker considers to be key or turning for the story.
2. Their appearance, generally, correlates with lower number of speaker's mistakes.
3. Usually the listener understands the story better when there are enough illustrative gestures.

References

1. *Cassell J., McNeill D., McCullough K. E.* 1999. Speech-Gesture Mismatches: Evidence for One Underlying Representation of Linguistic and Non-Linguistic Information. Pragmatics and Cognition, 7(1): 1–33.
2. *Chafe W.* 1980. The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production.
3. *Goldin-Meadow S.* 2005. Hearing Gesture: How Our Hands Help Us Think.
4. *Kibrik A. A., Kibrik A. A., Podlesskaia V. I.* 2009. Stories about Dreams. Corpus Research of Russian Spoken Discourse [Rasskazy o Snovideniiakh. Korpusnoe Issledovanie Ustnogo Russkogo Doskursa].
5. *Kreidlin G. E.* 2003. Man and Woman in a Dialogue I: Nonverbal Gender Stereotypes [Muzhchina I Zhenshchina v Dialoge I: Neverbal'nye Gendernye Stereotipy]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2003" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2003").
6. *McNeill D.* 1992. Hand and Mind: What Gestures Reveal About Thought.
7. *Nikolaeva Iu. S.* 2009. Segmentation of Spoken Narration and Graphic Gestures: Kinetic Signs of Boundaries and Relations between the Discourse Segments [Segmentatsiia Ustnogo Narrativa I Izobrazitel'nye Zhesty: Kineticheskie Priznaki Granits I Sviazei mezhdu Segmentami Diskursa]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2009" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2009"), 8 (15).
8. *Van Dijk T., Kintsch W.* 1978. Towards a Model of Text Comprehension and Production. Psychological Review, 85 : 363–394.
9. *Vygotskii L. S.* 2005. Thinking and Speaking [Myshlenie I Rech].

ЗНАЧЕНИЯ, ДИАТЕЗЫ И ОНТОЛОГИЧЕСКИЕ КАТЕГОРИИ СЛОВА *ВПЕЧАТЛЕНИЕ*

Е. В. Падучева (elena.paducheva@yandex.ru)

Всероссийский институт научной и технической информации РАН, Москва, Россия

Слово *впечатление* в современном языке морфологически не мотивировано, характеризуется уникальной сочетаемостью и нерегулярной многозначностью. В. В. Виноградов трактует имя *впечатление* как производное от глагола *впечатлеть*, который существовал в русском языке до начала XIX века. Национальный корпус русского языка позволяет подтвердить гипотезу о том, что *впечатление* мотивировано глаголом *впечатлеть*, систематизировать его значения и объяснить сочетаемость.

Ключевые слова: впечатление, значения, сочетаемость, онтологические категории, диатезы.

MEANINGS, DIATHESES AND ONTOLOGICAL CATEGORIES OF THE RUSSIAN WORD *VPECHATLENIE* ‘IMPRESSION’

E. V. Paducheva (elena.paducheva@yandex.ru)

VINITI Russian Academy of Science, Moscow, Russia

The Russian word vpechatlenie ‘impression’ is usually included in the class of emotions, as well as the verb vpechatljat ‘to make impression’. But derivational relationship between the noun and the verb remains unclear: dictionaries explicate the meaning of the verb vpechatljat with the help of the verb phrase proizvodit’ vpechatlenie ‘produce impression’, which does not help. The noun vpechatlenie is characterized by an idiosyncratic combinability (non-attested by other nouns of emotion) and an irregular polysemy. In this paper vpechatlenie is treated as motivated not by the verb vpechatljat, but by the verb vpechatlet’ ‘to produce an imprint’, which existed in the Russian language up to the beginning of the 19th century but later disappeared. This verb belongs to the class of image creation verbs, such as depict (something as something), represent (something as something), etc. It used to have an uncommon diathesis: Avpechatlel na /v Y-e obraz

Z X-a = 'A created on /in Y the image <imprint> Z of X'. Or, take a non-agentive variant: X vpechatlel na /v Y-e svoj obraz Z = 'X created on /in Y its image <imprint> Z'. The participant X is, as a rule, the consciousness of a human being. The verb vpechatlet' makes all the relationships transparent. It becomes possible (i) to reveal the derivational patterns corresponding to the different meanings of vpechatlenie and to assign ontological categories to these meanings; (ii) to describe combinability of the word as an effect of its ontological categories; (iii) to uncover semantic relationships between different meanings. In this way we get an account of the unique position of the word vpechatlenie among the nouns of emotion. Still the language of the Internet demonstrates that the word vpechatlenie experiences a pressure from its neighbors and gradually acquires the combinability characteristic of prototypical nouns of emotion, namely, of the names of states. In particular, the verb phrase ispytat'vpechatlenie, lit. 'experience impression', becomes frequent, by analogy with ispytat' udovol'stvie 'pleasure', ispytat' radost' 'joy', etc.

Key words: impression, meanings, combinability, ontological categories, diatheses.

Русское слово *впечатление* с трудом поддается истолкованию. В. В. Виноградов [4: 110] цитирует Льва Толстого, который выбрал именно слово *впечатление* для иллюстрации своей идеи о том, что значение слова нельзя описать через другие. И опровергнуть Толстого не так-то просто.

1. Загадки слова *впечатление*

С синхронной точки зрения слово *впечатление*, действительно, выглядит своего рода изолятом. По своей формальной структуре *впечатление* — это отглагольное имя. Однако единственный глагол в словарях современного русского языка, с которым его можно морфологически соотнести, — это *впечатлить*, несов. вид — *впечатлять*; а *впечатлять* определяется как 'производить впечатление'. Из этого описания мы не получаем информации о семантике глагола, а значит и о семантической связи имени с мотивирующим глаголом.

Семантически, слово *впечатление* должно принадлежать к тематическому классу эмоций: что-то произвело на меня впечатление — значит, привело меня в какое-то эмоциональное (или ментально-эмоциональное, — во всяком случае, психическое) состояние. Впечатления и чисто синтаксически встают иногда в один ряд с эмоциями: *Впечатления меняются одно за другим: недоумение, удивление и, наконец, восхищение лихой напористостью авангардных художников*.[«Вокруг света», 2004.07.15][1] Как и эмоции, впечатления могут быть положительными и отрицательными, ср. *Книга оставляет положительное впечатление*. [«Рекламный мир», 2003.03.31]. Однако по своему языковому поведению слово *впечатление* сильно отличается от имен эмоций.

[1] Здесь и далее датированные примеры — из Национального корпуса русского языка [сокращенно — Корпус], сайт в Интернете — www.ruscorpora.ru.

Начнем с того, что сочетание *произвести впечатление* не позволяет выявить **словообразовательную модель**, по которой имя *впечатление* было бы образовано от глагола *впечатлить*. В самом деле, большая часть имен эмоций порождает вполне регулярные пропорции: *удивить = вызвать удивление*, *смутить = вызвать смущение*, *встревожить = вызвать тревогу*, и т. д. Еще в XIX веке в роли «вербализатора» при именах эмоций активно употреблялся и глагол *производить*[2]. Корпус дает во множестве сочетания типа *произвело удивление, смущение, смятение, волнение, страдание, панику, тревогу* и пр., многие из которых синонимичны мотивирующему глаголу:

> Эта речь *произвела удивление*, не менее предшествующей. [П. В. Анненков. Записки о французской революции 1848 года (1848)]

> Ужели мой друг думал что чин Патриция во мне *произвел радость*. [А. Н. Радищев. [Положив непреоборимую преграду…] (1790)]

> Он ожидал, что его заявление просто *произведет тревогу*, но оно не произвело ничего. [Н. С. Лесков. На ножах (1870)]

Позднее *производить* во всех таких сочетаниях, если и употребляется, то допускает замену на *вызывать*:

> Боже мой! — какой мы *произведем перепуг*! В искусстве нами любуются. [К. А. Федин. Первые радости (1943–1945)] [= *вызовем*];

> Когда картина была выставлена, она *произвела смятение* среди лондонцев. [К. Г. Паустовский. Золотая роза (1955)] [= *вызвала*];

> Они нас не ждут с этой стороны, наше появление *произведет панику*. [Н. С. Гумилев. Записки кавалериста (1914–1915)] [= *вызовет*].

Между тем в контексте *впечатление* заменить *произвести* на *вызвать* никак нельзя: *Книга произвела* (**вызвала*) *хорошее впечатление*.

Дело в том, что имена типа *удивление, смущение* и пр. обозначают состояние, и состояние, действительно, может быть *вызвано* чем-то. А *впечатление* по своей онтологической категории никак не может быть отнесено к именам состояния: в МАС первое значение для *впечатление* определяется как ‘образ, оставляемый в сознании’. Отсюда и следует, что впечатление не может быть вызвано.

---

[2] Вербализаторами я называю слова, которые, подобно лексическим функциям типа Oper, Func и Labor («лексико-функциональным» глаголам), используются как «синтаксические оформители описания ситуации с помощью существительного, ее называющего» [Мельчук 1974 /1999: 93]. См. о вербализирующих операторах в [Ляшевская, Падучева 2011 (в печати)].

У типичных имен эмоций связь с мотивирующим глаголом вполне прозрачна, ср. *радуется — испытывает радость*; *беспокоится — испытывает беспокойство*; *разочаровался — испытал разочарование*. Для слова *впечатление* это не так.

Хотя нормативные словари не включают глагола *впечатлиться* (возвратно-медиальный к *впечатлить*), его можно считать существующим в русском языке. Корпус дает 14 абсолютно приемлемых употреблений, например:

> *Впечатлившись* услышанным и увиденным, С. В. Степашин пообещал, что … . [«Встреча» (Дубна), 2003.04.23]

> Но более всего ревнительские издания *впечатлились* результатами суда в г. Приозерске [«Церковный вестник», 2002.11.10]

Аналогично, хотя сочетание *испытать впечатление* не допускается существующими нормативами[3], оно встречается в Интернете. Однако *впечатление* здесь понимается в значении ‘ощущение’, которое пока не зафиксировано в словарях:

(1.1) Бывают такие *впечатления*, которые не описать словами — их нужно прочувствовать. Они запоминаются навсегда, и *испытать* их хочется каждому (из Интернета);

(1.2) Какие *впечатления* вы *испытали*, встав первый раз на лед в этом году (из Интернета).

В Корпусе это сочетание встречается ровно два раза, оба у С. Т. Аксакова и оба не нормативны; одно — из-за неуместного мн. числа, другое — из-за подчиненного род. падежа (о котором см. ниже):

(1.3) а. <…> *я испытал впечатления* мучительного страха, о котором долго не мог забыть [С. Т. Аксаков. Детские годы Багрова-внука (1858)]

б. Ровно через три года представлялся мне случай снова испытать *впечатление* дальней летней дороги. [С. Т. Аксаков. Детские годы Багрова-внука, служащие продолжением семейной хроники (1858)]

Так что пару (i) *впечатлиться — испытать впечатление* составить можно. Однако соотношение в паре (i) совершенно не такое, как, например, в паре (ii) *разочароваться — испытать разочарование*. В паре (ii) глагол и имя обозначают эмоцию, и *испытать разочарование*, как и *разочароваться*, означает ‘начать

[3] В справочно-информационном портале «Русский язык» (gramota.ru) на письмо человека о том, что сочетание «испытывать впечатление» вызвало у него сомнение, ответ был: — Правильно усомнились. «испытывать впечатление» — однозначно плохо, надо «быть (находиться) под впечатлением».

быть в разочаровании'. А в паре (i) имя *впечатление* обозначает чувство /ощущение (значение, которого оно вне этого контекста не имеет), а глагол *впечатлиться* означает испытать воздействие — значение, которое имя *впечатление*, в принципе, может иметь, см. (1.4), но не в контексте *испытать впечатление*:

(1.4) Супер молчал, очевидно, впечатленный услышанным. *Впечатление усугубил* временно не пьющий поэт Дозморов, который сел за наш столик и припомнил возмутительный случай. [«Октябрь», 2003].

Итак, первая проблема со словом *впечатление* — словообразовательная: какова та деривационная модель, по которой имя *впечатление* соотносится с мотивирующим глаголом?

Вторая проблема состоит в том, что слово *впечатление* имеет **сочетаемость**, которая отличает его от типичных имен эмоций.

Возьмем следующий ряд имен эмоций: *беспокойство, боль, горечь, муки, мученье, наслаждение, неудобство, неудовольствие, огорчение, отдохновение, радость, страдания, удовлетворение, удовольствие, успокоение* (не *спокойствие*!); он подробно рассмотрен в работе [Ляшевская, Падучева 2011 (в печати)]. В [Булыгина, Шмелев 1997] про *удовольствие, радость, огорчение* говорится, что это чувства, которые черпаются из внешнего мира, — чувства, которые нам доставляет внешний мир. Отсюда сочетаемость этих слов с глаголом *доставить*. На самом деле, нечто /некто может *доставить* не только *удовольствие, радость, огорчение*, но и весь ряд в целом, от *беспокойство*, до *успокоение*.

Чувства, которые сочетаются с *доставить*, названы в [Булыгина, Шмелев 1997] впечатлениями. На наш взгляд, названы неудачно, поскольку слово *впечатление* как раз категорически не сочетается с *доставить*: нельзя сказать **Спектакль доставил впечатление* — надо, опять-таки, сказать *произвел впечатление*.

Наш анализ имеет целью показать, что сочетание *произвести впечатление* — это не просто идиоматичность, которая в модели «Смысл–Текст» [Мельчук 1974/1999] описывается аппаратом лексических функций. Несочетаемость слова *впечатление* с *вызвать* и *доставить* и сочетаемость с *произвести*, как мы увидим, имеет ясное семантическое объяснение.

Итак, вторая проблема — семантико-синтаксическая: можно ли приписать слову *впечатление* (в разных его значениях) онтологические категории, которые бы предопределяли его сочетаемость и отличали от других, более «типичных» имен эмоций[4].

Отметим еще третью проблему: **структура многозначности** у слова *впечатление* иная, чем у типичных имен эмоций. У типичного имени эмоции многозначность регулярная (повторяющаяся во многих парах) — чувство /событие, чувство /состояние, чувство /изменение состояния и под.:

[4] Обоснование сочетаемости слова его онтологической категорией дано, на примере имен эмоций, в Ляшевская, Падучева 2011.

(а) *неловкость* — чувство: Она же *испытывала неловкость*, потому что не могла вспомнить, где она его прежде видела. [Л. Улицкая. Путешествие в седьмую сторону света //Новый Мир, № 8–9, 2000]; и событие: Здесь *произошла* небольшая *неловкость* (Yandex); И я опять не знал, как тут поступить, опять *возникла неловкость*. [И. Э. Кио. Иллюзии без иллюзий (1995–1999)]

(б) *разочарование* — чувство: Решение доверчивого горсобрания *доставило разочарование* учителям; и состояние: Он положил сверток к ногам летчицы Зины и ушел *в разочаровании*, потому что была у него идея пригласить ее на танцы. [Галина Щербакова. У ног лежачих женщин (1995)]

(в) *потрясение* — чувство: Ольга пережила тяжелое *потрясение*; и изменение состояния: Страна приближается к большим внутренним *потрясениям* (МАС).

У слова *впечатление* набор значений уникальный, и он требует объяснения. Набор значений естественно соотнести с онтологическими категориями.

2. Мотивирующий глагол

Согласно словарю Фасмера, русское слово *впечатление* — это калька с фр. *impression*, которое, в свою очередь, является калькой с нем. *Eindruck*.

Более длинную и интересную историю слова *впечатление* прослеживает В. В. Виноградов [4]. Согласно В. В. Виноградову, имя *впечатление* произошло от глагола *печатлеть* и производных от него *запечатлеть*, *напечатлеть*, которые вошли в русский язык из старославянского. Аналогичную историю В. В. Виноградов предполагает для *впечатлеть*, у которого наряду с прямым, конкретным значением ‘оттиснуть печать’, развилось переносное значение ‘внедрить, вкоренить’.

Глагол *впечатлеть* еще в XVIII — начале XIX веке был вполне живым — он имел несов. вид (*впечатлевать*) и возвратно-медиальную форму (*впечатлеться*):

> Через несколько недель получил я ответ — он *впечатлелся* навсегда в моем сердце. (Н. Карамзин. Письма русского путешественника).

От глагола *впечатлеть* и образовано существительное *впечатление*, которое еще в начале XIX века сохраняло свое первоначальное конкретное значение имени результата. Слово *впечатление*, по мнению Виноградова, служило для выражения значений лат. *impressio* ‘вдавление’, ‘выразительное произношение’, ‘впечатление’, ‘вторжение’ (от imprimere ‘вдавливать’, ‘ставить клеймо /печать’, ‘оттиснуть — например, на воске’), что способствовало его сближению с фр. *impression* и развитию у него абстрактных значений.

Существенно, что в XVIII — начале XIX века у слова *впечатление* было не только значение результата — ‘отпечаток, оставляемый печатью’, но и значение действия — ‘наложение печати’; т. е. оно имело регулярную многозначность, нормальную для имен на *-ение*, образованных от глагола [Апресян 1974: 193–203]. Пример употребления слова *впечатление* в значении действия:

> Это меня чрезвычайно заняло, и я, для лучшего *впечатления* этих предметов в моей памяти, вздумал перевести всю статью на русский язык. (Н. И. Греч. Записки о моей жизни.)

И словообразовательная модель, связывающая *впечатление* с *впечатлеть*, и система значений слова *впечатление* проницательно описаны В. В. Виноградовым. Тем не менее, имеет смысл вернуться к этой теме — с более развитым теоретическим аппаратом и с дополнительным материалом, который можно почерпнуть из Корпуса.

3. Диатезы мотивирующего глагола

Итак, слово *впечатление* семантически мотивировано не существующим в современном языке глаголом *впечатлять /впечатлить* (ср. *огорчение* от *огорчить, изумление* от *изумить* и мн. др.), а выпавшими из языка глаголами *впечатлеть, впечатлеться* ‘оставить отпечаток, знак’, которые в XVIII — начале XIX века еще существовали:

(3.1) Боже отмщений! Тако ли и я казнюся, как был казним Каин? *Впечатлел* ли ты на челе моем знаки моего злодеяния? [Д. И. Фонвизин. Иосиф (1769)]

(3.2) Пан Меховецкий, друг первого обманщика, сделался руководителем и наставником второго; *впечатлел ему в память* все обстоятельства и случаи Лжедимитриевой истории, [Н. М. Карамзин. История государства Российского: Том 12 (1824–1826)]

(3.3) Лишь бы только мрачная злоба людей *не впечатлела* <…> *в мягкое его сердце недоверчивости*, ненависти к людям (А. И. Тургенев; цит. по ССРЛЯ)

(3.4) Творец <…> *впечатлел в нем* [человеке] *образ* и подобие свое [архиепископ Платон (Левшин). Слово на день Рождества Пресвятыя Богородицы (1780)]

(3.5) Сия картина так сильно *впечатлелась* в его юной душе, что он через двадцать лет после того, не мог без особливого радостного движения видеть большой реки, плывущих судов, летающих рыболовов. [Н. М. Карамзин. Рыцарь нашего времени (1803)]

(3.6) Великодушная государыня ужаснулась и <…> произнесла слова, которые хотя не могли перейти к нам во всей точности, но глубокий смысл их *впечатлелся* в сердцах многих. [Н. В. Гоголь. Портрет (1835)]

Теперь прежде чем говорить о значении слова *впечатление*, надо истолковать глагол *впечатлеть*, выявив набор участников ситуации, обозначаемой этим словом и возможные для него диатезы.

Глагол *впечатлеть* имеет две диатезы — с Агенсом и без Агенса.

Агентивная диатеза глагола *впечатлеть*:

A *впечатлел на* Y-*е образ* Z X-*а* = ‘A создал на Y-е отпечаток (Z) X-а’.

Здесь Z — это отпечаток, т. е. результат действия «впечатления» X-а на Y-е. Так,

(3.1) = *Впечатлел* ли ты (A) на челе (Y) моем знаки (Z) моего злодеяния (X)?

Обратим внимание на то, что участник Y может быть оформлен как предложным падежом, см. пример (3.1), так и предлогом *в* + вин. п., см. *ему в память* в примере (3.2).

Итак, впечатление не приходит к Y-у извне, и не вызывается; оно «впечатывается» в Y (или в Y-е) в виде Z-а. Это впечатывание может произвести Агенс A, как в примере (3.1). Но участие Агенса факультативно — X может и сам впечататься в Y, так что возникнет впечатление Z. Отсюда

Неагентивная каузативная диатеза глагола *впечатлеть*:

X *впечатлел в* Y <свой> *образ* Z [область значений аргумента Y при этом сужается: обычно Y — это *душа, сердце, память* человека; не *чело*].

Место участника A в роли подлежащего здесь заменяет X; но Z остается образом, отпечатком X-а, как и в диатезе с Агенсом:

(3.3) = Лишь бы только мрачная злоба людей (X) *не впечатлела* <…> *в мягкое его сердце* (Y) *недоверчивости* (Z) [здесь *недоверчивость* Z — это отпечаток *злобы* (в переносном смысле; буквально *недоверчивость* — это результат воздействия *злобы*)].

Теперь мы можем определить исходное значение слова *впечатление* через глагол *впечатлеть* как имя результата от глагола *впечатлеть* (в терминах модели «Смысл–Текст» — имя второго актанта, S2):

впечатление = ‘то, что A (или X) впечатлел на /в Y-е / в Y как образ X-а’.

Значение ‘отпечаток’ у слова *впечатление* вполне сохраняется еще у Карамзина:

(3.7) Красота Лизы при первой встрече сделала *впечатление* в его сердце. [Н. М. Карамзин. Бедная Лиза (1792)]

Валентность на Z пропадает — она заполняется самим словом *впечатление*.

Возможна, впрочем, и другая интерпретация — *впечатление* в контексте *произвести* можно трактовать как имя действия от *впечатлеть*: *произвести впечатление* — как *произвести прием* (= ‘принять’), *произвести проверку* (= ‘проверить’) и т. д. В первом случае, если *впечатление* — имя результата, участник Z, по условиям словообразовательной модели, пропадает; а во втором он сохраняется:

А впечатлел на /в Y-е / в Y образ Z X-а =

А произвел на/в Y-е или в Y впечатление X-а в виде Z-а.

Итак, возникает следующая пропорция: *впечатлеть = произвести впечатление = создать впечатление*; как *вклеить = произвести вклейку = создать вклейку.*

Следует отметить одну важную особенность актантной структуры глагола *впечатлеть*. Дело в том, что он относится к классу глаголов создания образа [Падучева 2003]. У этих глаголов есть участники Образ и Прототип, которые могут быть плохо различимы, поскольку оба выражаются вин. падежом. Так, *рисовать* можно *генерала* и *портрет* <генерала>; аналогично, *впечатлеть* можно *пустыню мрачную*, как в (3.8а), и *образ*, как в (3.8б):

(3.8) а. Ты живо *впечатлел* в моем воображеньи Пустыню мрачную, поэта заточенье, Туманный свод небес, обычные снега И краткой теплотой согретые луга. [А. С. Пушкин. К Овидию: «Овидий, я живу близ тихих берегов…» (1821)]

б. Пребудет *образ* ввек во мне, Она который *впечатлела*! [Г. Р. Державин. Видение Мурзы: «На темно-голубом эфире…» (1783–1784)]

Перейдем теперь к самому слову *впечатление* и к современному русскому языку.

4. Значения слова *впечатление*

Словари различают у *впечатление* три значения. Задача в том, чтобы понять, как эти значения связаны друг с другом. Мы покажем, что эта связь осуществляется через глагол *впечатлеть*.

Значение 1 (*впечатление* — это образ)

впечатление от X-*а у* Y-*а* = ‘образ (Z), который X впечатлел в сознание Y-а’.

Это значение слово *впечатление* имеет в примерах (1)–(6); при этом Y часто является не индивидным сознанием, а сознанием релевантного коллектива и опускается. Валентность на образ (Z) у слова *впечатление* утрачивается, поскольку именем образа является само слово *впечатление*:

(4.1) общее *впечатление от команды* становится целостным и радующим глаз. [«Известия», 2003.02.09]

(4.2) <…> который, впрочем, *впечатления* от «Спартака» не *попортил*, а только напомнил, кто в избе хозяин. [«Известия», 2002.10.23]

(4.3) Я помню себя рано, но первые мои *впечатления* разрозненны (В. Г. Короленко, цит. По МАС).

(4.4) *Впечатления* от похорон могут вызвать серьёзный регресс в развитии ребёнка. [//«Домовой», 2002.08.04] [имеется в виду — отпечатки в сознании того, что ребенок видел на похоронах, воспоминания]

(4.5) ей нравилась ранняя утренняя пустота Москвы, даже ноябрьская мокрая мгла не портила *впечатления*. [Анна Берсенева. Полет над разлукой (2003–2005)] [= не портила картины в сознании].

(4.6) Незабываемое *впечатление* осталось от голоса Анны Литвиненко. [«Российская музыкальная газета», 2003.04.09] [незабываемым является образ в сознании]

Участник Y может быть выражен именной группой с предлогом *у* или притяжательным местоимением (*У меня осталось незабываемое впечатление от ее голоса*; *мое впечатление от ее голоса*).

Итак, внешний мир не «доставляет» впечатление нашему сознанию — он именно «производит, создает» его: *впечатлеть* — это глагол создания.

Значение 1 самое вещественное: *впечатление* 1 — это как бы отпечаток. Семантически, в сочетании *впечатление от спектакля у ребенка* реализована та же модель, что, скажем, в сочетании *изображение кремля на гобелене*. Так что глаголы с конкретным значением предполагают именно лексему *впечатление* 1:

> Но с годами флер очарования его героизмом и писательским талантом рассеялся, и мои *впечатления отфильтровались* во вполне четкую картину. [Нина Воронель. Без прикрас. Воспоминания (1975–2003)]

Значение 2 (*впечатление* — это воздействие).

Значение 2 возникает у слова *впечатление*, прежде всего, в контексте глагола *произвести*:

> X *произвел на* Y *впечатление* Z-*а* = ‘X произвел впечатление Z-а в сознание Y-а, т. е. внедрил в сознание Y-а Z как свой образ’.

Примеры:

> Молодая, приветливая блондинка, веселая хохотушка, она (X) *произвела* на меня *впечатление* совершенно несерьезной женщины (Z). [В. Запашный. Риск. Борьба. Любовь (1998–2004)]

> Их действия, поведение (X) *произвели впечатление* отрепетированного спектакля (Z). [Л. Гурченко. Аплодисменты] = ‘X создал образ Z-а’.

> Он (X) *производил впечатление* очень интеллигентного, обременённого жизненным опытом и глубокомысленного человека (Z). [Запись LiveJournal (2004)] = ‘X отражался в сознании людей как Z’

Содержание впечатления-воздействия может быть передано прилагательным:

На Андре Жида архитектура Москвы произвела *удручающее* впечатление: [«Неприкосновенный запас», 2003.07.14]

Вид нежилого неразрушенного города произвел *гнетущее* впечатление. [«Искусство кино», 2003.06.30]

Но может быть так, что воздействие охарактеризовано только с точки зрения существования и силы, а его содержание остается нераскрытым, т. е. участник Z за кадром:

Лизавета Ивановна *произвела впечатление* на майора;

Деловитость *произвела впечатление* на Зимина, и он взял в свою команду нового человека. [«Знание — сила», 2003]

Ср. примеры (4.7) и (4.8). В (4.7) участник Z выражен (*жалкое впечатление* = ‘впечатление жалкости’), а в (4.8) — нет (*сильное впечатление* = ‘сильное воздействие’):

(4.7) Подарок производил *жалкое впечатление*;

(4.8) Подарок произвел *сильное впечатление*.

Участник X в контексте примеров (4.9), (4.10) мог бы быть выражен предлогом *от*, но опущен:

(4.9) Мороз так *притупляет впечатления*… [В. Г. Короленко. Мороз (1900–1901)]

(4.10) Это впечатление *достигает кульминации* в «экстатическом дуэте» Данте и Беатриче ближе к концу сочинения, где Беатриче на разные лады повторяет фразу: «И ты тоже умрешь», а Данте точно так же твердит: «Сладчайшая смерть». [Vita nova (2003) // «Российская музыкальная газета», 2003.05.14]

Иной способ выражения участника X — в (4.11):

(4.11) Причём, помню, предвкушал *впечатление*, какое произведёт подарок. [Юрий Трифонов. Предварительные итоги (1970)] = ‘заранее предвкушал воздействия, которое произведет подарок’.

У лексемы *впечатление* 2 имеется вторая диатеза:

Y *находится под впечатлением от* X-*а* = ‘Y находится в ментально-эмоциональном состоянии, вызванном воздействием X-а’.

В этой конструкции участник Z начисто отсутствует. Примеры:

> <Из этого Милий Алексеевич заключил, что> *майор* всё ещё *под впечатлением от* Лизаветы Ивановны. [Юрий Давыдов. Синие тюльпаны (1988–1989)]

> На следующий день *я* гулял по берегу моря, всё ещё находясь *под впечатлением свидания*, вспоминая его волнующие подробности и, главное, чувствуя себя на голову выше, чем до него. [Фазиль Искандер. Письмо (1969)] = ‘Свидание оказало на меня эмоциональное воздействие: после него я чувствовал себя на голову выше’.

Двоякая диатеза обуславливает конверсное соотношение: *Ваша деловитость произвела впечатление на Зимина — Зимин находился под впечатлением от вашей деловитости*. В обоих случаях *впечатление* — это воздействие.

Значение 3 (*впечатление* — это мнение):

<*от* X-*а*> у Y-а *впечатление*, что Z = ‘<от X-а> у Y-а создалось мнение, что Z’.

В этом значении участник X факультативный, а обязательный участник — Z, мнение, ср. типичную для мнений способность управлять придаточным с *что*. Ситуация X — это то, что Y наблюдал и что послужило источником впечатления, т. е. другой ситуации, Z, которая в данном случае выражается пропозицией. Участник X может быть не выражен. Примеры:

> У Тани *создалось впечатление, что* они играют в какую-то взрослую игру — делят что-то понарошку… Но делили взаправду… [Л. Улицкая. Путешествие в седьмую сторону света (2000)]

> А у меня *осталось впечатление*, *что* упущена реальная возможность предотвратить серьезную ошибку. [Г. Арбатов. Человек Системы (2002)]

> У меня *впечатление*, *что* я телеграфирую в пустоту. [Л. Смирнова. Моя любовь (1997)]

В (4.12) за выражение участника X можно принять сочетание *по общению с Кириллом*:

(4.12) Но у меня по общению с Кириллом *не сложилось впечатления*, *что* ребята поняли логику происходящего. [Освобождение от условностей (блог) (2008)]

В примерах (4.13), (4.14) *что* отсутствует, но то, что *впечатление* имеет значение мнения, подтверждается подчиняющим глаголом:

(4.13) Мои первоначальные *впечатления* подтверждались;

(4.14) Это он не свое *впечатление* выразил в рецензии.

«X производит впечатление Z» может означать ‘кажется, что X есть Z’; впечатление переходит в подозрение: *Его забота производит впечатление надзора*.

5. Динамическая семантика слова *впечатление*

Итак, у слова *впечатление* три значения, каждому из которых соответствует свой гипероним — образ, воздействие, мнение. Так что *впечатление* не имеет ни одной из онтологических категорий, свойственных именам эмоций согласно [Ляшевская, Падучева 2011 (в печати)] (таких, как состояние, отношение, чувство). Других имен эмоций с таким набором значений безусловно нет. Глагол *впечатлеть* позволяет вывести все три значения из одного источника.

Значения 1 и 2 слова *впечатление* происходят непосредственно от глагола *впечатлеть*, одно является именем результата, другое — именем действия. Переход от значения 1 к значению 3 сродни метафоре (т. е. смене концепта); это переход от образа ситуации-каузатора к пропозиции, ее выражающей.

Как правило, три значения слова *впечатление* различаются достаточно четко. Неоднозначность может возникнуть в контексте косвенного вопроса; так, в (5.1) *впечатление* — это скорее воздействие; между тем, в (5.2) это может быть и воздействие, и образ:

(5.1) Легко вообразить, какое <большое> *впечатление* Алексей должен был произвести в кругу наших барышень. [А. С. Пушкин. Барышня-крестьянка (1830)]

(5.2) Вы, надеюсь, понимаете, какое *впечатление* он на меня произвёл. [Анатолий Рыбаков. Тяжелый песок (1975–1977)]

В контексте прямого вопроса постулат информативности Грайса требует понимания слова *впечатление* в значении 1:

Какое впечатление у вас *осталось после* её [кабины загара] посещения? [Красота, здоровье, отдых: Красота (форум) (2005)]

Какое впечатление *произвела* на вас его квартира?

Принципиально важным является пример (5.3), где *впечатление* выступает одновременно в двух своих значениях (*вынес* семантически согласуется с ‘образ’, *огромное* — с ‘воздействие’):

(5.3) Я помню, какое огромное *впечатление вынес* от этого произведения. [«Вестник США», 2003.10.29]

Все значения у русского слова *впечатление* (в отличие от его переводных эквивалентов в англ., нем., фр.) находятся строго в идеальной сфере. Хотя связь внутренней формы слова *впечатление* со словом *печать* прослеживается и в русском языке:

> На какой «внутренний субстрат» «печатается» то, что оказывает на нас впечатление? — задает вопрос психолог, изучающий связь между впечатлениями и эмоциями [Лэнгле 2004].

Итак, слово *впечатление* в его трех значениях произведено по стандартным словообразовательным моделям, но не от *впечатлить*, а от *впечатлеть* — глагола, который на протяжении XIX века выпал из языка. То, что глагол *впечатлеть* выпал, не вызывает удивления — он явно противоречил сложившейся системе, в которой глаголы на *-еть* являются непереходными.

Глагол *впечатлеть*, с его неординарной диатезой, позволил: а) выявить связи между значениями слова *впечатление*, т. е. связать значения имени и глагола естественными семантическими переходами, и б) описать сочетаемость слова как вытекающую из его онтологической категории. Все это объяснило уникальное положение слова *впечатление* среди имен эмоции. Как показывают, однако, примеры (1.1), (1.2), с глаголом *испытывать*, слово *впечатление* подвергается массированному давлению со стороны своих соседей по тематическому классу и начинает обретать сочетаемость, свойственную прототипическим именам эмоции, а именно, именам состояния.

Отметим еще, что слово *впечатление* — характерный пример следующего явления. Его третье значение прозрачным образом связаны с первым. А первое и второе, значения получаются в соответствии с продуктивной словообразовательной моделью, но от глагола (*впечатлеть*), которого в современном языке не существует. Это явление стало сейчас предметом внимания в лексической семантике (см. о продуктивных дериватах от неактуальных значений, например, в [Бабаева 1998], [Урысон 2005]). Оно безусловно является массовым.[*]

References

1. Apresian Iu.D. 1974. *Lexical Semantics: Synonymic Language Means [Leksicheskaia Semantika: Sinonimicheskie Sredstva Iazyka]*.
2. Babaeva E. E. 1998. Who Lives *in* the Den ['Vertep']? *Or* Experiment *of* Semantic History *of a* Word [Kto Zhivet *v* Vertepe, ili Opyt Postroeniia Semanticheskoi Istorii Slova]. *Voprosy Iazykoznaniia*, *3* : *94–106*.
3. Bulygina T. V., Shmelev A. D. 2000. Space Movement *as* Emotional Metaphor [Peremeshchenie *v* Prostranstve kak Metafora Emotsii]. *Logicheskii Analiz Iazyka. Iazyki Prostranstva*.

[*] Выражаю радостную благодарность двум моим анонимным рецензентам. Автор может только мечтать о таких вдумчивых и проницательных читателях.

4. Langle A. 2004. Introduction into the Existential Analytic Theory *of* Emotions: *A* Touch *on* the Value [Vvedenie *v* Eksistentsial'no-Analiticheskuiu Teoriiu Emotsii: Prikosnovenie *k* Tsennosti]. *Voprosy Psikhologii*.
5. Liashevskaia O. N., Paducheva E. V. 2011. Ontological Categories *of* Emotions Names [Ontologicheskie Kategorii Imen Emotsii]. *NTI*, 5, available at: http:// lexicograph.ruslang.ru/05News.htm .
6. Mel'chuk I.A. 1999. *Experiment of the Theory of Linguistic Models "Meaning - Text" [Opyt Teorii Lingvisticheskikh Modelei "Smysl - Tekst"]*, 1.
7. Paducheva E. V. 2003. Image Creation Verbs: Lexical Meaning and Semantic Derivation [GLagoly Sozdaniia Obraza: Leksicheskoe Znachenie *I* Semanticheskaia Derivatsiia]. *Voprosy Iazykoznaniia*,*6* : *30–46*.
8. Uryson E. V. 2005. Logical Structure *of* the Polysemy and its Realization (the Word 'SLIAKOT" *in* the Language System) [Logicheskaia Struktura Polisemii *I ee* Realizatsii (Slovo 'SLIAKOT" *v* Sisteme Iazyka)]. *Russkii Iazyk v Nauchnom Osveshchenii*, *2* (10) : *87–120*.
9. Vinogradov V. V. 1994.*The History of Words [Istoriia Slov]*.

МЕТОД ОПРЕДЕЛЕНИЯ ЭМОЦИЙ В ТЕКСТАХ НА РУССКОМ ЯЗЫКЕ

А. Г. Пазельская (pazelskaya@i-teco.ru)

А. Н. Соловьев (a.solovyev@i-teco.ru)

ЗАО «Ай-Теко», Москва, Россия

В работе рассматриваются методы автоматического определения эмоциональной составляющей (тональности) в тексте и описывается опыт осуществляемой в данный момент практической реализации системы для текстов СМИ на русском языке, в основе которой лежат словари лексической тональности и набор комбинаторных правил объединения отдельных слов и словосочетаний.
В работе впервые предложен метод определения тональности, основанный на предикационных отношениях в пропозиции. В связи с этим нами предложена классификация глаголов в зависимости от их эмотивного воздействия и местоположения объекта тональности.

Ключевые слова: эмоции, тональность, лексическая тональность, метод определения тональности.

A METHOD OF SENTIMENT ANALYSIS IN RUSSIAN TEXTS

A. G. Pazel´skaia (pazelskaya@i-teco.ru)

A. N. Solov'ev (a.solovyev@i-teco.ru)

"I-Teco", Moscow, Russian Federation

This paper presents an overview of methods of sentiment analysis. It also describes our experience of building a system for detecting sentiment in natural Russian texts (mass media). The system uses rule-based approach, calculating sentiment within a simple clause on the basis of word sentiment, output of a Natural Language Processing (NLP) module, and rules of sentiment combination. Word sentiment is determined in sentiment dictionaries created and regularly updated by experts (more than 15000 words and collocations by now). The system uses separate dictionaries for different parts of speech: nouns, verbs, adjectives, adverbs, verbal and nonverbal collocations. Every word and collocation in the dictionary is marked for its sentiment polarity and sentiment strength. The NLP module provides

morphological and syntactic information (NPs, complex verbs, syntactic roles, clause types and boundaries, etc.). This information is further used to combine word sentiment and to identify sentiment of subject and object within a clause, as well as of the clause as a whole and of the monitored object within the clause. The system is regularly tested by experts on new mass media texts, it shows about 80 % recall and 90 % precision.

Key words: emotions, sentiment, sentiment analysis, method of analysis.

1. Эмотивная составляющая в тексте

В данной статье рассматривается один из методов определения эмоционального компонента в тексте. Эта задача относится к обширному кругу задач анализа и обработки различных функций коммуникации на естественных языках. Сегодня в современных информационных технологиях широко применяются системы обработки коммуникационной (или информационной) и метаязыковой функций[1] коммуникации. Наряду с этим возникает необходимость обработки и других функций: фатической, апеллятивной и эмотивной (в том числе оценочной). Информационная функция коммуникации применяется при взаимодействии человека с компьютером, когда нужно получить или уточнить необходимую информацию (например, справочные системы). В автоматических системах перевода используется метаязыковая функция — кодирование и изоморфное преобразование языковой информации. На фатической функции основываются различные развлекательные системы, поддерживающие диалог с пользователем, в том числе с применением речевых технологий (см., например, [Соловьев и др., 2003]). Эмотивную функцию коммуникации пытаются использовать в автоматических системах оценки и сравнения объектов, например, новых продуктов и брендов известных компаний, для выявления отношения людей к событиям в политической жизни страны и т. п.

Эмоциональная составляющая коммуникации пока не столь активно применяется в системах обработки текстовой информации не только ввиду трудностей выделения «нужной» (т. е. относящейся к рассматриваемому объекту) эмоциональной лексики в текстах, но и сложности определения самого эмотивного пространства, количества и состава его измерений[2]. К сожалению, теория эмоций в лингвистике еще недостаточно развита.

Исследования в области теории лингвистических эмоций начались не так давно. В 50-х годах прошлого века Чарльз Осгуд с помощью метода семантического дифференциала пытался определять эмотивное пространство

---

[1] Согласно классификации функций коммуникации, данной Р. Якобсоном в своей работе «Лингвистика и поэтика» [Якобсон,1975].

[2] Например, при оценке электронных продуктов *большое разрешение монитора* — это хорошо, но *большой вес* — плохо; прилагательное *большой* ведет себя по разному с точки зрения эмоциональной оценки, в зависимости от того, о чем идет речь.

различными наборами парных слов [Осгуд и др., 2007]. В настоящее время исследования эмоций лежат в основном в области психологии, нейрофизиологии и психолингвистики. В лингвистике разработаны психометрические инструменты и методы для таких исследований (см., например, OpinionFinder [Wilson & al. 2005] или Profile of Mood States (POMS-bi) [Norcross & al., 2006]).

В современных системах автоматического определения эмоциональной оценки текста чаще всего используется одномерное эмотивное пространство: позитив-негатив, то есть хорошо-плохо. Однако известны успешные случаи использования и многомерных пространств [Bollen & al., 2010]. Более подобный обзор современного состояния в области анализа тональности текста представлен в книге [Pang & Lee, 2008].

В нашем методе при эмоциональной оценке рассматриваемого текста мы используем эмотивное пространство, содержащее негативную-позитивную составляющую плюс силу эмотивности.

2. Понятие лексической тональности и тональности предложения

Эмоциональная оценка, выраженная в тексте, также называется тональностью, или сентиментом текста (от англ. sentiment — чувство; мнение, настроение). Эмоциональная составляющая, выраженная на уровне лексемы или коммуникативного фрагмента[3], называется лексической тональностью (или лексическим сентиментом). Тональность текста в целом определяется лексической тональностью составляющих его единиц и правилами их сочетания.

Автоматическое определение тональности текста подразумевает выделение тех фрагментов текста, которые выражают позитивную или негативную эмоциональность по отношению к объекту эмоциональной оценки (объекту тональности). Таким объектом может быть имя собственное, название продукта, организации, услуги, профессии и т. п., по отношению к которому анализируется текст. Объект эмоциональной оценки может быть задан как один в целом для текста (с учетом его синонимических и анафорических употреблений), так и определяться в предложениях как любое имя собственное или даже нарицательное.

Таким образом, тональность текста определяется тремя факторами: 1) субъект тональности; 2) собственно тональная оценка (позитив/нейтрально/негатив); 3) объект тональности. Под субъектом тональности подразумевается автор статьи (автор цитаты, прямой или косвенной речи), под объектом тональности — тот, о ком он высказывается и под тональной оценкой — эмоциональное отношение автора к такому объекту.

---

[3] «Коммуникативные фрагменты (КФ) — это отрезки речи различной длины, которые хранятся в памяти говорящего в качестве стационарных частиц его языкового опыта и которыми он оперирует при создании и интерпретации высказываний». [Гаспаров, 1996]. Коммуникативный фрагмент обычно больше, чем слово, но меньше, чем предложение.

3. Методы определения тональности текста

Существуют три основных метода определения тональности текста.

1. Анализ текста методами векторного анализа (часто с применением n-граммных моделей), сравнение с ранее размеченным эталонным корпусом по выбранной мере близости и отнесение (классификация) текста к негативу или позитиву на основании полученного результата сравнения.
2. Поиск эмотивной лексики (лексической тональности) в тексте по заранее составленным тональным словарям (спискам паттернов) с применением лингвистического анализа. По совокупности найденной эмотивной лексики текст может быть оценен по шкале, отражающей количество негативной и позитивной лексики. Этот метод может использовать как списки паттернов, подставляемые в регулярные выражения, так и правила соединения тональной лексики внутри предложения.
3. Смешанный метод (комбинация первого и второго подходов).

Первый метод (см., например, [Pang & al., 2002; Pang & al., 2005; Gamon, 2004]) работает достаточно быстро, но требует наличия предварительно размеченного эталонного корпуса, на основе которого происходит обучение алгоритма сравнения. Существенными недостатками такого подхода оказываются увеличение трудоемкости и ограничение разнородности корпуса (т. е. неполнота лексического покрытия), что приводит к потере точности. К тому же данный метод не позволяет провести глубокий анализ текста, то есть выявить и показать эмотивность на уровне предложения.

Второй метод [Nasukawa, 2003; Yi, 2003] не менее трудоемок в составлении тональных словарей (или получения списка тональных паттернов), но в сочетании с синтаксическим и морфологическим анализом более гибок: он позволяет не только показать цепочки тональной лексики, но и получить синтаксически корректные эмоциональные выражения. При хорошем наполнении тональных словарных списков этот метод позволяет достичь хорошей полноты (покрытия эмотивной лексики).

Недостаток этого метода в том, что с помощью него сложно дать количественную оценку негативности-позитивности текста. Чтобы избежать недостатков первого и второго метода, используют смешанный подход [Prabowo & al., 2009; Konig, 2006], частично включающий в себя два первых.

Мы опишем методику и опыт использования второго метода определения тональности текста с использованием правил объединения слов в цепочки и определения тональности у объекта на основе предикационных отношений в пропозиции. Создаваемая нами система предназначена для обработки новостных текстов общероссийских СМИ.

4. Определение тональности с использованием тональных словарей и лингвистического анализа

Анализ тональности текста, реализуемый нами в настоящий момент, состоит из нескольких этапов. Сначала отрабатывает отдельный лингвистический

модуль, автоматически производящий морфологический анализ текста, лематизацию всей лексики и определяющий части речи каждого слова, его морфологические характеристики (падеж, лицо, число, активность-пассивность для глаголов), роль этого слова в предложении (для существительных: подлежащие, обстоятельство, дополнение; для глаголов: причастие, деепричастие, глагол; и др.), его тип (например, для существительных: физическое лицо, юридическое лицо, географическое название и др.).

Затем все слова (существительные, глаголы, прилагательные и наречия) и некоторые словосочетания (коллокации) размечаются по заранее подготовленным словарным спискам тональной лексики. Каждому слову приписывается два атрибута, указывающие на тональность и/или силу тональности. Если слово не нашлось в списках тональной лексики, то оно считается нейтральным.

После этого запускается первичный синтаксический анализ: слова и словосочетания объединяются в тональные цепочки, в предложении выделяются субъект, предикат и объект, идентифицируются причастные и деепричастные обороты, подчинительные предложения, анафорические связи и пр. Естественно, не каждое предложение русского языка можно представить в виде триады субъект-предикат-объект. Учитываются также безличные, неопределенно-личные и обобщенно-личные предложения, предложения с нулевой формой глагола, сказуемые, выраженные неглагольной формой.

На последнем этапе в предложении выделяется объект тональности и определяется его сентимент в зависимости от местоположения и роли этого объекта в предложении.

4.1. Тональные словари

Таким образом, необходимое условие для анализа тональности — составление словарного списка тональной лексики. Мы использовали тональные словари, разделенные по четырем частям речи (существительные, глаголы, прилагательные и наречия), плюс глагольные и неглагольные коллокации[4]. Использование коллокаций было вызвано тем, что далеко не все сочетания слов при объединении их по общим правилам дают в результате правильный сантимент (например, *общество с ограниченной ответственностью*, *взрыв смеха* и пр.).

Все части речи разделяются на разные подклассы в зависимости от лексической тональности. Например, словарь глаголов состоит из одиннадцати подклассов (см. подробнее соответствующий раздел). Тональные словари заполняются экспертно; в начале работ был размечен лексический сантимент наиболее частотных слов разных частей речи, извлеченных из составленного специально для этой цели на основе информационных русскоязычных порталов Интернета корпуса текстов СМИ (около 100 млн. словоупотреблений). В процессе

[4] Под коллокациями понимались любые устойчивые и достаточно часто встречающиеся сочетания слов, как с идиоматическим значением, так и с неидиоматическим.

тестирования и отладки системы тональные списки постоянно пополняются и сейчас содержат более 15 000 тональных слов и коллокаций. В словари попадают только слова и словосочетания, несущие какую-либо тональность или усиливающие тональность связанных с ними единиц.

Каждое слово или словосочетание может попасть только в один из классов по частям речи и тональности. Естественно, при таком подходе мы сталкиваемся с проблемой омонимии (одно и тоже слово может иметь разный сантимент или даже принадлежать разным частям речи). Эту проблему мы частично снимаем с помощью увеличения списка коллокаций и учета глагольного управления (ср. *болеть за что-л.* и *болеть чем.-л.*). Слова, тональность которых зависит от тематики текста, размечались согласно тому, в каком качестве их эмотивность или сила эмотивности чаще употребляется в корпусе СМИ.

4.1.1. Неглагольные лексемы и коллокации

Наречия, прилагательные и неглагольные коллокации делятся на позитивные, негативные и усиливающие эмоциональность, то есть такие слова или словосочетания, которые сами по себе не несут сантимента, но при этом могут усиливать эмоциональность того, к чему присоединяются (например, наречия *круто*, *ужас*; прилагательные *эксклюзивный*, *потрясающий* и коллокации *коренным образом*, *решающая роль*). Сила тональности определялась экспертами по трехбалльной шкале.

Имена существительные также могут быть позитивными (например, *благотворительность* или *зарплата*) и негативными (*налог* или *война*). Однако не все существительные имеют однозначную эмоциональную нагрузку, тональность многих зависит от окружения. Поэтому целесообразно вводить классы потенциально негативных и потенциально позитивных слов — так, потенциально позитивные слова позитивны в позитивном окружении и нейтральны во всех остальных. Например, слово *план* само по себе не несёт в себе тональности, но сочетание *план по выходу из кризиса* должно давать позитив.

Особую роль играют отглагольные существительные: они могут менять тональность следующего за ним существительного. Например, отглагольное существительное *прекращение* меняет её на противоположную. Если за ним следует позитивная цепочка связанных существительных, например, *прекращение поставок угля*, то объединенная цепочка будет негативной. Если за данным отглагольным существительным следует негативная цепочка, например, *прекращение военных действий*, то в целом новая цепочка получит позитив. Поэтому отглагольные существительные выделялись в два отдельных класса: меняющие тональность зависящего от них слова (как *прекращение* или *спад*) и сохраняющие её (*рост* или *проведение*).

4.1.2. Глагольные лексемы и коллокации

Особое внимание при разработке мы уделили классификации и тональной разметке глаголов. В нашем методе именно предикация (элементарная единица текста, состоящая из глагола и его зависимых) является ключевой составляющей при определении тональности объекта. Иными словами, тональность определяется (в общем случае) тремя составляющими: тональностью самого

объекта, действием, производимым объектом или над объектом, и тональностью остальных участников описываемой ситуации.

Любое упоминание объекта в предложении характеризуется двумя параметрами: его окружением и его ролью относительно глагола. Каждый из этих параметров может как влиять, так и не влиять на итоговую тональность объекта в предложении — это определяется глаголом. Соответственно, в зависимости от влияния этих двух параметров на тональность объекта и тональности глагола как такового мы выделили восемь классов глаголов:

- 1 и 2 класс — негативные и позитивные глаголы, определяющие тональность объекта в зависимости от окружения, но независимо от его роли (негативные *уносить*, *стереть*, *освободить от*; позитивные *защищать*, *предрекать*, *болеть за*);
- 3 и 4 класс — негативные и позитивные глаголы, определяющие тональность объекта независимо от окружения, но в зависимости от его роли (например, глаголы *сдаться* и *проиграть* приписывают негатив субъекту и позитив объекту, а глаголы *обуздать* и *повергнуть*, наоборот, приписывают позитив субъекту и негатив объекту);
- 5 и 6 класс — негативные и позитивные глаголы, определяющие тональность объекта в зависимости от его окружения и роли (в основном в эти классы вошли возвратные глаголы; примеры негативных глаголов: *жаловаться*, *испугаться*, *замерзать*; позитивных глаголов: *окупаться*, *влечь*, *согреться*);
- 7 и 8 класс — чисто негативные и чисто позитивные глаголы, определяющие тональность объекта вне зависимости от его роли и окружения (например, *расследовать* и *улучшать* всегда приписывают позитив, а *грабить* и *злоупотреблять* — негатив);

Кроме того, отдельно выделяются три дополнительных класса глаголов:

- 9 класс — глаголы, соединяющие или приравнивающие тональность объекта и субъекта (так называемые связочные глаголы: *являться*, *олицетворять*, *относиться*).
- 10 и 11 классы — позитивные и негативные глагольные коллокации, например, негативные *наложить руки*, *освободить от должности*, *пробивать насквозь* и позитивные *поразить противника*, *заострить внимание*, *заливаться смехом*.

Списки глаголов составлялись с учетом глагольного управления: глаголы, тональность которых менялась в зависимости от глагольного управления, попадали в разные классы (например, *высказаться за* и *высказаться против*). Также каждому глаголу и глагольной коллокации по трехбалльной шкале была приписана сила тональности (это наиболее важно для 7–8-го классов и списков коллокаций).

4.2. Правила сочетаемости

Следующий подготовительный этап тонального анализа текста — составление правил сочетаемости лексем и коллокаций. Слова и словосочетания соединяются по этим правилам между собой, причём сначала объединяются

соседние неглагольные элементы, затем они присоединяются к глаголу, определяя, таким образом, сантимент внутри предикации (простого предложения). Не все правила можно задать корректно: например, сочетание негативного и позитивного существительного в общем случае не определено. Наиболее частотные из таких словосочетаний включались в списки коллокаций, остальные обрабатывались в зависимости от глагола, возглавляющего предложение.

Правила представляют собой комбинации различных членов предложения между собой. Учитываются слова-инверторы, например, *не, нет, без, вне* и пр.

Разрешаются анафорические ссылки, выраженные местоимениями и местоименными словами. Сложные предложения разбиваются на простые, некоторые типы придаточных предложений включаются в родительское, причастные обороты присоединяются к определяемому слову, деепричастные — к субъекту родительского предложения. Придаточные предложения определительного типа с разрешенной анафорией соотносятся с определяемым словом.

В итоге предложение приводится к одному из типов синтаксической структуры из субъекта, предиката и объекта, где каждый член структуры в общем случае представлен цепочкой словоформ с определенной тональностью.

В случае нулевого глагола тональность определялась по тональному окружению объекта, его качественным признакам.

4.3. Определение сантимента по отношению к объекту тональности

На последнем этапе выделяется объект тональности. Он задается пользователем или определяется автоматически: в каждом предложении ищется так называемая именованная сущность, например, имя собственное, одушевленное существительное и т. п.

Единица текста, на которой подсчитывается тональность — предикация, и согласно принятому в системе ограничению в каждой предикации тональность считается относительно только одного объекта. Это значит, что в предикации не может быть двух объектов тональности, и при наличии нескольких именованных сущностей исследуемого типа выбирается одна из них.

На основании роли и местоположения объекта тональности ему по определенным правилам приписывается сантимент, и пользователю предъявляется предложение с выделенным объектом тональности и выявленными тональными цепочками.

Всего было составлено более 20 таких правил приписывания сантимента объекту тональности.

Этапы обработки предложения в системе представлены на Рис. 1.

Для тестирования устойчивости системы к добавлениям и изменениям комбинаторных правил и словарей мы создали небольшой тестовый тонально размеченный корпус, охватывающий различные комбинации тональных словоформ и синтаксиса предложений (около 400 предложений). С помощью него мы оцениваем улучшение или ухудшение модуля при каждом значительном изменении системы.

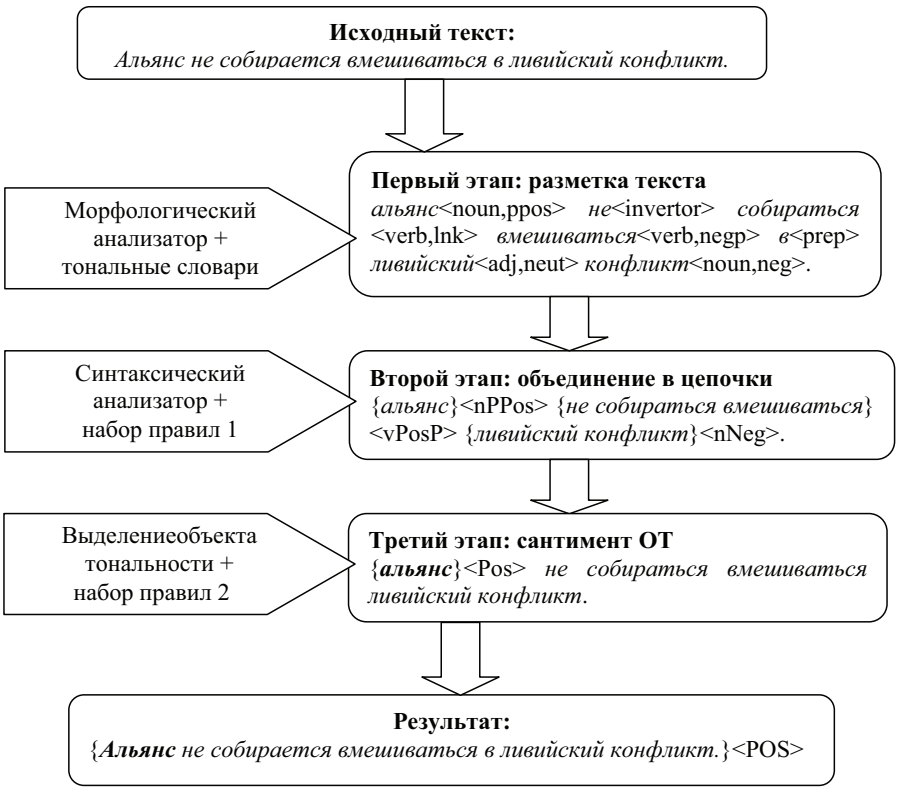


Рис. 1. Последовательность обработки предложения для определения его тональности. Сокращения: noun —существительное, adj — прилагательное, verb — глагол, prep — предлог, invertor — инвертор, pos — позитивный, neg — негативный, negp — чисто негативный, lnk — глагол-связка, ppos — потенциально позитивный, neut — нейтральный, posp — чисто позитивный

4.4. Оценка результата

В настоящее время методов объективного тестирования систем тональной разметки текстов еще не разработано. Поэтому применяемый в настоящее время нами метод тестирования основывается на периодических субъективных оценках небольших текстовых подборок экспертом. Тестирование проводится один раз в неделю на произвольных текстах СМИ, а именно, используются первые 5–7 новостных текстов с сайта rbc.ru за понедельник или вторник каждой недели, что составляет в среднем по 70 предложений в неделю. Таким образом, за период с января по конец марта 2011 г. система была протестирована на 762 предложениях.

Эксперт получает тональную разметку текстов при помощи системы и затем оценивает, насколько он в каждом конкретном случае согласен или

не согласен с результатом. В случае несогласия отмечается тип ошибки: пропуск тональности, неправильный знак тональности (позитив вместо негатива или наоборот), нетональное предложение, размеченное как тональное. Затем на основании этих оценок считается полнота и точность тональной разметки.

Поскольку основной единицей определения тональности является предикация, содержащая не более одного объекта тональности, то полноту и точность тональной разметки также разумно оценивать на основе количества предикаций, на которых тональность сработала правильно или же допустила ошибку какого-либо типа. Таким образом, количество предикаций с верно выделенным объектом тональности и верно определенной тональностью будет соответствовать количеству верных срабатываний системы (A).

Среди ошибок нужно разделять пропуски тональных предложений, содержащих объект тональности (B), и ложные срабатывания — случаи, когда система неправильно определила знак тональности (C) или же выделила как тональное предложение, не содержащее эмоциональной оценки и/или объекта тональности (D).

Эксперт заносит свою оценку тональной разметки, данной системой, в таблицу, строки которой соответствуют предложениям исходного текста, а столбцы — типам срабатываний. Тем самым, в клетках ставится количество выделенных в данном исходном предложении предикаций с верным срабатыванием, с пропусками, с неверным знаком и с «лишней» найденной тональностью (см. Табл. 1).

Табл. 1. Фрагмент таблицы с экспертной оценкой результатов работы модуля тональности. Жирным шрифтом выделены найденные модулем объекты тональности, фигурные скобки означают границы тональных предикаций, <POS> — позитив, <NEG> — негатив

№	Предложение	ОК (A)	Пропуск (B)	Знак (C)	Лишнее (D)
42	{В результате взрыва на **АЭС «Фукусима-1»** поврежден реактор}<POS>.			1	
43	{На четвертом реакторе **АЭС «Фукусима-1»** в 11: 53 по местному времени (05:53 мск) произошел взрыв водорода}<NEG>, передают японские **СМИ**.	1			
44	В 11: 14 по местному времени (05:00 мск) в зоне четвертого реактора начался **пожар**, сообщили в компании-операторе станции Tokyo Electric Power(**TEPCO**).		1		

Для подсчёта точности и полноты число предикаций с верными срабатываниями и с ошибками разных типов по всем текстам суммируется, и общая оценка вычисляется по следующим формулам. Число тональных предикаций в тексте составит A + B + C, а число предикаций, определённых системой как тональные — A + C + D. Тогда полнота определения тональности будет равняться A / (A + B + C), а точность — A / (A + C + D).

Система определения тональности постоянно дорабатывается во всех своих частях: словарей, правил, лингвистической базы, программной реализации, поэтому качество определения тональности на текстах, близких к тем, на которых проводится тестирование (новостных СМИ), растёт. Изменение качества определения тональности с января по конец марта 2011 г. представлено на Рис. 2.

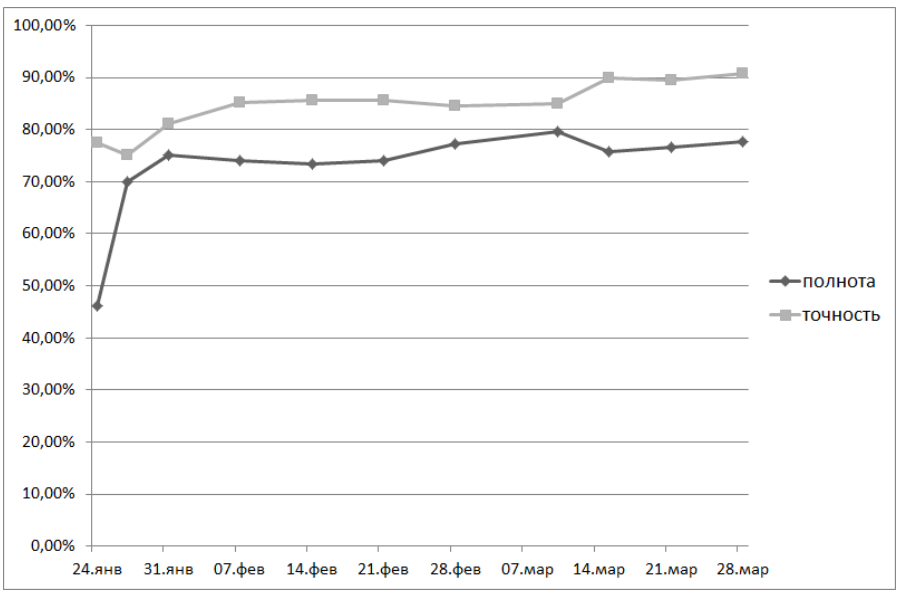


Рис. 2. Изменение качества тональной разметки с января по март 2011 г.

Хорошо заметно, что появление в тестовом массиве большого количества не вполне обычных для системы текстов приводит к замедлению роста качества. Так, 14.03 два из пяти текстов в тестовой выборке оказались биржевыми сводками, в которых много специфической лексики, отсюда потеря в полноте.

Можно выделить три класса ошибок, возникающих при определении тональности:

1. Ошибки работы модуля морфологической и синтаксической разметки текста (около 5–7 %).
2. Ошибки правил комбинаторики (не более 3 %)
3. Ошибки тональных словарей, вызванные их неполнотой и «тональной» омонимией (не более 5 %)

Демонстрационная версия модуля тональности доступна по адресу http://x-file.su/tm/. Напоминаем, что система рассчитана на работу с грамматически правильными текстами СМИ.

5. Заключение

Представленный метод определения тональности относится к так называемому глубокому сентимент-анализу (deep sentiment analysis), основывающемуся на лингвистическом анализе текста на естественном языке (NLP). Как показывают результаты, с помощью этого метода можно достичь достаточно высокой (85–90 %) точности на текстах определенной тематики (в нашем случае — новости СМИ). Тем не менее, остается ряд неисправимых ошибок (не учитывая ошибки внешних модулей, такие как ошибки морфологического и синтаксического анализаторов). По нашему мнению, одной из причин такого рода ошибок является ограниченность используемого эмотивного пространства: часть лексики не попадает (или только частично попадает) в наше эмотивное пространство хорошо-плохо плюс сила эмотивности[5]. Определение размерности — открытый исследовательский вопрос, решение которого лежит в области понимания и восприятия информации мозгом человека. Таким образом, качественное улучшение выбранного нами метода определения тональности нуждается в дальнейших фундаментальных исследованиях не только в области лингвистики, но и в области когнитивных наук, таких как психология, психо- и нейролингвистика.

References

1. *Bollen J., Mao H., Zeng X.-J.* 2010. Twitter Mood Predicts the Stock Market. Technical Report arXiv:1010.3003, CoRR. Http: http://arxiv.org/pdf/1010.3003v1
2. *Gamon M.* 2004. Sentiment Classification on Customer Feedback Data: Noisy data, Large Feature Vectors, and the Role of Linguistic Analysis. Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004) : 841–847.
3. *Gasparov B. M.* 1996. Language, Memory, Image. Linguistics of Language Existence [Iazyk. Pamiat'. Obraz. Lingvistika Iazykovogo Sushchestvovaniia]. Novoe Literaturnoe Obozrenie.
4. *Iakobson R. O.* 1975. Linguistics and Poetics [Lingvistika I Poetika]. Strukturalizm: "Za" i "Protiv".
5. *Konig A. C., Brill E.* 2006. Reducing the Human Overhead in Text Categorization. Proceedings of the12th ACM SIGKDD Conference on Knowledge Discovery and Data Mining : 598–603.

[5] Такое пространство можно назвать полуторамерным — мы не учитывали оппозицию сильный — слабый.

6. *Nasukawa T., Yi J.* 2003. Sentiment Analysis: Capturing Favorability using Natural Language Processing. Proceedings of the 2nd International Conference on Knowledge Capture : 70–77.
7. *Norcross J. C., Guadagnoli E., Prochaska J. O.* 2006. A Visual Map of Public Mementos and Conjectures. Journal of Clinical Psychology, 40 : 1270–1277.
8. *Osgud Ch., Susi J., Tannenbaum P.* 2007. Application of the Semantic Differential Method to the Researches on Aesthetics and Adjacent Problems [Prilozhenie Metodiki Semanticheskogo Differetsiala k Issledovaniiam po Estetike I Smezhnym Problemam]. Iskusstvometriia. Metody Tochnykh Nauk I Semiotiki : 278–297.
9. *Pang B., Lee L.* 2008. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 2 (1–2) : 1–135.
10. *Pang B., Lee L.* 2005. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL) : 115–124.
11. *Pang B., Lee L., Vaithyanathan S.* 2002. Thumbs up? Sentiment Classification Using Machine Learning Techniques. Proceeding of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2002) : 79–86.
12. *Prabowo R., Thelwall M.* 2009. Sentiment Analysis: A Combined Approach. Journal of Informetrics, 3(2).
13. *Solov'ev A.N., Razumikhin D. V., Viktorova K. O.* 2003. "And What Do You Think?" (On the Use of Non-informative Functions in Spoken Communication Models) ["A Sam-to Ty Kak Dumaesh?" (Ob ISpol'zovanii Neinformativnykh Funktsii v Modeliakh Rechevoi Kommunikatsii)]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2003" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2003") : 653–657.
14. *Wilson T., Hoffmann P., Somasundaran S., Kessler J., Wiebe J., Choi Y., Cardie C., Riloff E., Patwardhan S.* 2005. OpinionFinder: A System for Subjectivity Analysis. Proceedings HLT/EMNLP : 34–35.
15. *Yi J., Nasukawa T., Niblack W., Bunescu R.* 2003. Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques. Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003) : 427–434.

РОДОВЫЕ ПОНЯТИЯ В БЫТОВОЙ ЛЕКСИКЕ КАК ОБЛАСТЬ ТОНКИХ РАЗЛИЧИЙ МЕЖДУ СЕРБСКИМ И ХОРВАТСКИМ ЯЗЫКОМ

А. Ч. Пиперски (apiperski@gmail.com)

МГУ, Москва, Россия

В бытовой лексике сербского и хорватского языков есть заметные различия. Различия видовых наименований (слов с такими значениями, как ‘ложка’, ‘очки’, ‘паспорт’) хорошо осознаются носителями, а различия родовых наименований (‘посуда’, ‘столовые приборы’, ‘канцелярские принадлежности’) менее заметны.

Ключевые слова: быт, бытовая лексика, род, родовые понятия, родовые наименования.

GENERIC TERMS IN EVERYDAY VOCABULARY AS A SPHERE OF SUBTLE DIFFERENCES BETWEEN SERBIAN AND CROATIAN

A. Ch. Piperski (apiperski@gmail.com)

Lomonosov Moscow State University, Moscow, Russian Federation

There are signifi cant differences in the everyday vocabulary of Serbian and Croatian. The speakers are aware of diverging specifi c terms (e. g., words for ‘spoon’, ‘glasses’, ‘passport’), but they fail to notice some diverging generic terms (words for ‘kitchenware’, ‘cutlery’, ‘writing supplies’). This is explained by the fact that generic terms show considerable amount of variation even within one language and cannot serve as markers of identity.

Key words: everyday items, everyday vocabulary, generic terms, identity marker.

It is typical for everyday vocabulary to exhibit considerable variation. To study this variation, a group of researchers under the lead of Boris Iomdin organized a survey for speakers of various languages (see Iomdin et al., this volume). The participants were shown 33 pictures and were requested to provide a name for each object in the picture (a specific, or subordinate term) and a word denoting the group to which this object belongs (a generic, or superordinate term). For example, a picture of the chair could be described with the words *chair* and *furniture*.

There were 5 speakers of Croatian and 6 speakers of Serbian who participated in the survey. Though this sample cannot count as representative, their responses were extremely instructive and made it possible to put forward some hypotheses about the lexical differences between Serbian and Croatian. These hypotheses were verified using other methods.

There is a number of well-described divergences between Serbian and Croatian literary languages (for references see Greenberg 2004, Tošović 2009)[1]. They do not impede mutual understanding, even though they are present on all levels of the language system — in phonetics, morphology, word formation, vocabulary, and syntax. But lexical differences are the most striking ones, and no wonder that they receive the greatest attention.

Among the responses to our survey there were some well-known pairs distinguishing the two languages:

S *naočare* (5) / *naočari* (1) vs. Cr *naočale* (4) / *naočali* (0) 'glasses'[2]

S *kašika* (6) vs. Cr *žlica* (5) 'spoon'

S *karmin* (5) vs. Cr *ruž (za usne)* (5) 'lipstick'

S *lenjir* (6) vs. Cr *ravnalo* (5) 'ruler'.

S *pasoš* (6) vs. Cr *putovnica* (3), *pasoš* (2) 'passport'

Known phonological and morphological differences between Serbian and Croatian could also be observed:

S *odeća* (11) vs. Cr *odjeća* (8) 'clothes' (about socks, gloves and high-heeled shoes)

S *prtljag* (1) (masc.) vs. Cr *prtljaga* (4) (fem.) 'luggage' (about trunks and cosmetics bags).

More important is that the survey showed some new facts that have never been considered as markers of differentiation between Serbian and Croatian.

High-heeled shoes were described by all 6 Serbian respondents as *cipele*, but 4 speakers of Croatian called them *štikle* (only 1 Croatian-speaking person wrote *ženske cipele* 'ladies' shoes'). The word *cipele* actually has the meaning 'shoes' according to all the dictionaries of Serbian and Croatian, but the word *štikla* is described as having only the meaning 'heel' even in the voluminous and up-to-date Croatian

[1] The other two languages closely related to Serbian and Croatian, namely Bosnian and Montenegrin, will not be discussed in this paper because no speakers of these languages took part in the survey and I lack reliable data about everyday vocabulary in these languages.

[2] The number of relevant responses obtained in the survey is given in parentheses. If the total of Croatian responses lies under 5 or the total of Serbian responses lies under 6, this means that the other responses were irrelevant for the topic under discussion (e. g., they contain absolutely different words). The Serbian words are rendered in Roman script.

dictionary by Vladimir Anić (2003). Thus, a metonymy 'heels' > 'high-heeled shoes' has occurred in Croatian, but not in Serbian, and this fact has not yet been reflected in the dictionaries.

Another semantic difference could be observed among the generic terms. The word *vaza* 'vase' was categorized as *nameštaj* by 3 Serbian respondents out of 6, while none of the Croatians used the words *namještaj* or its typical Croatian synonym *pokućstvo* (two of them could not even think of a suitable category for vases and left this field blank). Furthermore, *stolnjak* 'tablecloth' is categorized as *nameštaj* by 2 Serbians, and 1 Serbian regards *pokrivač* 'blanket, comforter' as *nameštaj*, too. It can be inferred that the S *nameštaj* has a more general meaning 'furniture, home décor or accessories', while the meaning of Cr *namještaj/pokućstvo* is limited to 'furniture'. In a somewhat outdated Serbo-Croatian-Russian dictionary by I. Tolstoy (1957) the word *nameštaj* is translated as 'мебель, обстановка'. It seems that the first meaning applies to Croatian and the second one to Serbian.

When examining the differences between Serbian and Croatian, scholars are usually concerned with absolute differences of the type "the word *X* exists in one of the languages but not in the other one" and pay less attention to statistical differences of the type "the word *X* exists in both languages, but it occurs significantly more frequently in one of them"[3]. The results of our survey allowed me to posit some hypotheses about differences of the second type, which could be then verified using greater amount of data from the Google search engine (www. google.com; the results were retrieved on 31.01.2011). For each word or phrase 3 queries were made: in Roman script in the domain *.rs*, in Roman script in the domain *.hr* and in Cyrillic script in the domain *.rs*[4]. The results were presented in form of tables.

	site:.rs	**site:.hr**
Roman script	s_{cyr}	h
Cyrillic script	s_{lat}	
Total (Σ)	$s = s_{cyr} + s_{lat}$	h

A coefficient $k = s : h$ was calculated for each word/phrase. This coefficient reflects the frequency of a word/phrase in Serbian texts relative to its frequency in Croatian texts.

Given the proximity of the two languages, I assume that the words that are not specifically Serbian or specifically Croatian (which is the case with the most words)

[3] There are in fact some studies of lexical divergences between Serbian and Croatian that use simple statistical methods (cf. Grčević 2002), but most scholars tend to rely on their own impressions and not on statistical data. However, highly elaborated statistical methods have been proven useful for analyzing lexical divergences between closely related language systems (cf. Berdicevskis *forthcoming* for comparison of the Russian language in Russia and Latvia).

[4] Surely there are texts written in Serbian in the domain *.hr* and texts written in Croatian in the domain *.rs*, but their amount is negligible.

have similar frequencies in both languages. This means that k does not vary significantly for such words, and its mean value ($\bar{k}$) reflects only the size of the corpus for each language. It turns out that $\bar{k} \approx 0.2$ (in other words, the Google database contains 5 times more Croatian texts than Serbian). For words/phrases which are characteristic only for one of the languages, k will greatly diverge from the mean value. If there is a pair of synonymous words/constructions w_1 and w_2, for which k_1 is significantly less than $\bar{k}$ and k_2 is significantly greater than $\bar{k}$ (or vice versa), the usage of the members of this pair constitutes a statistical difference between Serbian and Croatian.

In the responses to the survey 4 pairs of this type occured:

Generic term for ruler: S ***pribor za crtanje*** (1) vs. Cr ***crtaći pribor*** (1)

w_1	site:.rs	site:.hr		w_2	site:.rs	site:.hr	
pribor za crtanje	45 300	67 300		crtaći pribor	102	3760	
прибор за цртање	1 900			цртаћи прибор	3		$k_1 : k_2 = 25{,}11$
Σ	47 200	67 300		Σ	105	3760	
	$k_1 = 0{,}70$				$k_2 = 0{,}028$		

Generic term for ruler, eraser, pencil: S ***pribor za pisanje*** (ruler — 1, eraser — 2, pencil — 2) vs. Cr ***pisaći pribor*** (ruler — 1, eraser — 1, pencil — 2)

w_3	site:.rs	site:.hr		w_4	site:.rs	site:.hr	
pribor za pisanje	635 000	88 800		pisaći pribor	9160	81 300	
прибор за писање	2 320			писаћи прибор	64		$k_3 : k_4 = 63{,}26$
Σ	637 320	88 800		Σ	9224	81 300	
	$k_3 = 7{,}18$				$k_4 = 0{,}11$		

Generic term for spoon: S ***escajg*** (2) vs. Cr ***beštek*** (2)

w_5	site:.rs	site:.hr		w_6	site:.rs	site:.hr	
escajg	180 000	4 350		beštek	59	11 400	
есцајг	319			бештек	3		$k_5 : k_6 = 7621{,}94$
Σ	180 319	4 350		Σ	62	11 400	
	$k_5 = 41{,}45$				$k_6 = 0{,}0054$		

Generic term for pot, spoon, vase, teapot, wineglass: S ***posuđe***[5] (pot — 1, spoon — 1, teapot — 4, wineglass — 4) vs. Cr ***suđe*** (pot — 3, spoon — 1, vase — 1, teapot — 4, wineglass — 4)

w_7	site:.rs	site:.hr		w_8	site:.rs	site:.hr	
posuđe	458 000	1 340 000		suđe	23 500	241 000	
posudje	5 520	2 790		sudje	9 420	53 600	$k_7 : k_8 = 3{,}07$
посуђе	2 220			суђе	363		
Σ	465 740	1 342 790		Σ	33 283	294 600	
	$k_7 = 0{,}35$				$k_8 = 0{,}11$		

The following conclusions can be made:

1) Deverbal adjectives in *-ući* meaning 'intended for smth.' are more widespread in Croatian, while in Serbian the construction *za* 'for' + deverbal substantive in *-nje* is more frequently used (S *pribor za crtanje* 'instrument for drawing' vs. Cr *crtaći pribor* 'drawing instrument', S *pribor za pisanje* 'instrument for writing' vs. Cr *pisaći pribor* 'writing instrument'). It is probably not a coincidence that in the entry *pribor* in a Serbo-Croatian-Russian dictionary by I. Tolstoy (1957) which is rather Serbian-oriented one can find exactly the examples *pribor za pisanje* and *pribor za crtanje*.

2) There is a lexical divergence between Serbian and Croatian which to my knowledge has never been properly described before[6]: S *escajg* vs. Cr *beštek* 'cutlery'. It is noteworthy that both words are loanwords from German (*Esszeug* resp. *Besteck* 'cutlery'), but only the second word is in use in modern German. This was also corroborated by the survey: about 60 % of German-speaking respondents used the word *Besteck* for categorizing spoon, and none of them wrote *Esszeug*.

3) The differences in frequency of the word for 'kitchenware' (*posuđe/suđe*) are not as striking as in the other cases. But some speakers of Serbian and Croatian make notice of the fact that the pair *posuđe* vs. *suđe* distinguishes the two languages[7]. It is interesting that in the Croatian dictionary by V. Anić (2003) the word *posuđe* is given an explication, and the entry *suđe* contains only the reference to *posuđe*. The author of the dictionary was probably influenced by Serbian-oriented dictionaries and was not aware that it is not in accord with the actual Croatian usage.

Our survey proves that everyday vocabulary exhibits differences even in closely related languages. The present study has shown five discrepancies of this kind that had not been sufficiently accounted for in the scholarly literature:

---

[5] One of the Croatian respondents also categorized teapot and wineglass as *posuđe*.

[6] In Brodnjak (1992) *escajg* is listed as a Serbian word, but no one-word Croatian equivalent is given. It is translated (or rather explained) as *pribor za jelo (žlica, vilica, nož)* 'eating utensil (spoon, fork, knife)'.

[7] http://forum.ffzg.hr/viewtopic.php?p=108806 (retrieved on 31.01.2011)

1) S *štikle* 'heels' vs. Cr *štikle* 'heels; high-heeled shoes'; 2) S *nameštaj* with a broader meaning than its Cr equivalent; 3) S *za -nje* vs. Cr *-aći*; 4) S *escajg* vs. Cr. *beštek*; 5) S *posuđe* vs Cr *suđe*.

It is striking that most of these newly observed facts (4 out of 5) concern generic terms and not specific ones. The question arises why speakers and linguists are less aware of differences in superordinates than of differences in subordinates.

A probable explanation is that there is less variation among subordinates even within the same language. In our survey, trunk, gloves, umbrella, slippers, teapot, tablecloth and wineglass were unanimously described as *kofer, rukavice, kišobran, papuče, čajnik, stolnjak*[8] and *čaša* respectively by all Croatian and Serbian speakers. For many other words responses were pretty similar, e. g. *gumica* or *gumica za brisanje* for eraser (literally: 'rubber' or 'rubber for erasing'). For no generic term there were 11 identical responses nor 5 identical responses among Croatians were found, and there were only 2 cases where all 6 Serbians agreed with each other (they classified buiding blocks as *igračke* 'toys' amd socks as *odeća* 'clothes'). The uncertainty about generic terms that exists even within one language prevents the speakers from noticing that the speakers of a closely related language use other generic words in their everyday vocabulary.

References

1. *Anić V.* 2003. Veliki Rječnik Hrvatskoga Jezika.
2. *Berdicevskis A.* Predictors of Pluricentricity: Lexical divergences between Latvian Russian and Russian Russian.
3. *Brodnjak V.* 1992. Rječnik Razlika Između Hrvatskoga i Srpskoga Jezika.
4. *Grčević M.* 2002. Some Remarks on Recent Lexical Changes in the Croatian Language. Lexical Norm and National Language: Lexicography and Language Policy in South Slavic Languages after 1989 :150–163.
5. *Greenberg R. D.* 2004. Language Identity in the Balkans: Serbo-Croatian and Its Disintegration.
6. *Iomdin B. L.* 2009. Everyday life Vocabulary. Search of Standard [Terminologiia Byta. Poiski Normy]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2009" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2009"), 8 (15) : 127–135.
7. *Iomdin B., Piperski A., Russo M., Somin A.* 2011. How Different Languages Categorize Everyday Items. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2011" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2011"), 10 (17).
8. *Tolstoi I. I.* 1957. Serbian-Croatian-Russian Dictionary [Serbsko-Khorvatsko-Russkii Slovar'].
9. *Tošović B. 2009. Korrelative Grammatik des Bosni(aki)schen, Kroatischen und Serbischen. Teil 1: Phonetik — Phonologie — Prosodie.*

[8] One of the Serbian speakers wrote *stoljnjak*.

RELATIVE CLAUSES IN SPOKEN RUSSIAN AND ELSEWHERE: A CORPUS APPROACH

V. I. Podlesskaia (podlesskaya@ocrus.ru)

Russian State University for the Humanities, Moscow, Russian Federation

The paper addresses the problem of discrepancy between syntactic and prosodic grouping in Russian relative clauses. Basing on oral corpora systematically annotated for prosodic details, the paper demonstrates structural and prosodic "autonomy" of relative clauses from their heads, which previously remained unnoticed in the literature on relativization based mainly on written data.

Key words: relative clauses, syntactic grouping, prosodic grouping, prosodic details.

1. Introduction

In the paper, I discuss the systematic mismatch between syntactic and prosodic boundaries in Russian relative clauses[1]. An example below demonstrates the pattern in question, square brackets show syntactic constituents, bold curly brackets show intonation phrases:

(1)

{*Mne nravitsja* [*plat'je***}** **{**[*kotoroe ona iz sitca sšila*]]**}**
I like dress REL she of printed.cotton made
'I like the dress that she made of printed cotton '

Syntactically, the head noun 'dress' and the relative clause are within one constituent (are governed by the same maximal projection), but the most common prosodic boundary of this sentence comes between the head noun and the relative clause. And what is still more important; the intonation phrase **{***Mne nravitsja* [*plat'je***}** 'I like the dress' doesn't form a constituent, that is, it is not a full grammatical unit. Bill Croft in his paper on the relation between intonation units and grammatical units (Croft 1995: 847–848) paid attention to the similar phenomenon in spoken English and made the following statement:

[1] The research is supported by the Russian Foundation for Fundamental research, grant #10-06-00338a

> "Final relative clauses, embedded as well as adjoined, often make up their own intonation unit. ... This means that in the standard phrase-structure analysis, the clause preceding RC is not a complete constituent: it is missing a modifier of a subconstituent... In contrast, clause-internal RCs are NEVER separated from their head NPs, although the NP+RC grammatical unit is itself often split off from the rest of the clause... None of other syntax-prosody mismatches exhibits this categorical yet schizophrenic behavior."

I must say that in Russian this mismatch is not as categorical as in English, but it is schizophrenic enough to have this Croft's statement as a headline for the rest of my paper.

Basing on corpus data, I will put and try to answer the following research questions:

- What are possible prosodic phrasings in Russian relative clauses?
- What types of phrasing are actually attested in discourse? The answer will be — that most commonly attested prosodic phrasing place a boundary between the head noun and the relative clause (RC).
- How this most common prosodic phrasing functions in natural discourse?
- What are the prosodic symptoms of the boundary between the head noun and the relative clause? I will look at pausation patterns to view one of such cues.
- What are possible syntactic operations that can aggravate the separation of the relative clause from the head noun?
- How can Russian data be mapped to what we know about the interrelationship between syntax and prosody?

The paper is based on the following data:

1. Two oral corpora systematically annotated for prosodic details, incl. pausation and pitch movements: "Night Dream Stories", 1h 50 min of sounding (Kibrik, Podlesskaya 2009) and "Stories about presents and skiing", 35 min of sounding (Xurshudjan 2006)
2. National Corpus of Russian (www.ruscorpora.ru)

2. Three possible prosodic phrasings in Russian relative clauses

Russian allows three main types of prosodic phrasing in relative clauses described below as patterns A, B and C.

Phrasing A. The prosodic boundary comes between the head noun and the RC, this is the phrasing demonstrated above, in (1).

Phrasing B. No mismatch between syntax and prosody — the prosodic boundary comes in front of the head noun. There is only one (!) example with a full noun head in our oral corpora:

(2) Night Dream Stories[2]
..(04) i /uvidel iz /temnoty, mal'en'kuju ten' kotoraja \molitsja
and saw from darkness small shadow REL prays
'{And [he] saw from the darkness a small shadow that was praying}

Phrasing C. Also no mismatch between syntax and prosody — the whole sentence is pronounced as a single intonational phrase without an internal prosodic boundary. This is possible, for instance, for short RCs with contrastive focus. No such cases are attested in our oral corpora, but (3) is a constructed example with the focused intensifier *sama* 'herself' (shown in bold):

(3)
Mne nravitsja plat'je kotoroe ona ***\sama*** *sšila*
I like dress REL she herself made
{'I like the dress that she made herself'} <and not the one she bought>

In sum: actual prosodic phrasings attested in the corpus follow only the pattern A — putting the break between the head noun and the RC — neglecting the two other possibilities.

3. Discourse functions of pattern A

In natural discourse, the prosodic phrasing pattern A is used to accomplish one of the three following functional tasks:

Task 1. To detach the head noun and wrap it in a topical constituent of the sentence, while the RC forms the comment constituent:

(4) Night Dream Stories
i my uvideli ..(0.2) /ploščadku,
and we saw ground
..(0.3) na kotoroj bylo množestvo raznyx \červjakov.
on REL were lots various worms
'{And we saw the ground}$_{\text{Topic}}$ {on which there were lots of various worms}$_{\text{Comment}}$

Information structure of (4) is very close to that of:

(5)
{On the ground}$_{\text{Topic}}$ {there were lots of various worms}$_{\text{Comment}}$

[2] Breaking into intonational phrases is shown in the original text — by breaking into lines, and also in translation — by bold curly brackets; slashes iconically show the direction of the pitch movement, the nuclear pitch of the intonational phrase is shown by underlying the respective syllable

In relativization of this type, the matrix clause introduces an entity (referenced to by the head noun), while the RC contains a statement "about" this entity. RCs of this type comprise 30 % of the total number of RCs in our corpora.

Task 2. To produce the RC as a separate utterance added to the matrix clause as an afterthought:

(6) Night Dream Stories

...(1.4)Togda	*..(0.1)*	*moj*	*/komandir*	*/menja*	*/nagradil*	*zolotoj* *\medalju.*
then		my	commander	me	awarded	golden.INSTR medal.INSTR
··(0.3)	*Kotoraja*	*stoila*	*dvesti*		*dollarov*	
	REL	costed	two.hundred		dollars	

'Then my commander awarded me with a gold medal. That cost two hundred dollars.'

In relativization of this type, the matrix clause can be articulated with the falling pitch movement in the nuclear pitch, i.e. as an independent clause projecting no continuation in the unfolding discourse. Hence, the relative clause is added after the speaker recognizes the just produced portion of discourse as insufficient and requiring elaboration.

RCs of this type comprise 45 % (!) of the total number of RCs in our corpora. The figure 45 % being unexpectedly high by itself , becomes even more significant, if we look into other types of postnominal subordinate clauses and see how often they are produced as fragmented afterthoughts. For the Night Dream Stories corpus, the counts published in [Korotaev 2009] show the ratio of clauses used an afterthought to the total number of postpositional subordinate clauses of the given type:

Of the total amount of

- complement clauses — 26.3 % are produced as afterthoughts
- adverbial clauses — 50 % are produced as afterthoughts
- relative clauses — 44.9 % are produced as afterthoughts

So, RCs appear as afterthoughts much more often than complement clauses, and almost as often as adverbial clauses.

Task 3. To produce RCs as parentheticals — semantically, these RCs provide background information; prosodically, they are articulated with lowered pitch, narrower pitch range, reduction in loudness, increased tempo etc.:

(7) Stories about presents and skiing (the parenthetical RC is shown in bold curly brackets)

...(0.5)	*On*	*..(0.1)*	*priexal*	*v /\ avtosalon,*	
	he		came	to	car.shop
uh(0.2)	*vybral*	*bol'šuju*	*krasivuju*	*/–mašinu*	
	chose	big.ACC	beautiful.ACC	car.ACC	
***{**kotoruju*	*rešil'*	*podarit'*	*svoej*	*— žene**}**,*	
REL	decided	to.present	his.DAT	wife.DAT	

...(0.6) /vot,
well
\no um(0.3) kogda uh(0.2) on uznal /cenu etoj mašiny,
but when he found.out price.ACC this.GEN car.GEN
on /užasnulsja,
he was.shocked
i rešil čto \net.
and decided that no
'He went to a car shop, chose a big beautiful car **{**which [actually] he decided to give his wife as a present**}**, but when he found out the price, he was shocked and decided "NO".'

RCs of this type comprise 25 % of the total number of RCs in our corpora.

In sum: in our spoken corpus, RCs are used (1) as comment constituents, (2) as afterthoughts, or (3) as parentheticals. These uses are distributed as 30 % : 45 % : 25 %. Thus, in natural discourse, RCs tend to be desubordinated, and show symptoms of communicative autonomy.

4. Prosodic symptoms of the boundary between the head noun and the relative clause: pausation patterns

For the Night Dream Stories corpus, Table 1 shows how the distribution of pauses depends on the type of postnominal subordinate clause (counts are based on [Korotaev 2009])

Table 1. Number and length of pauses on the left edge of (=in front of) postpositional subordinate clauses

	complement clauses	**adverbial clauses**	**relative clauses**
Number of postpositional subordinate clauses	182	53	52
Mean duration of pauses on the left edge	0.14 sec	0.20 sec	0.24 sec
The ratio of zero pauses to the total number of left edges	74.7 %	66.0 %	55.8 %
The ratio of pauses longer than 0.5 sec to the total number of left edges	11.5 %	15.1 %	19.2 %

As Table 1 shows, pauses that come in front of RCs are much longer than pauses in front of postpositional complement clauses, and even longer than pauses in front of postpositional adverbial clauses. Then, the percentage of cases with no pauses at all is maximal for complement clauses and minimal for relative clauses. Here again, relative clauses appear to be even more strongly detached than adverbials. Finally,

we consider medial and long pauses (for oral stories, these are, typically, half-second or longer). The table shows that medial and long pauses occur more often in front of relative clauses than in front of two other types of relative clauses.

In sum: pausation patterns convincingly demonstrate the strong prosodic break on the left edge of RSc.

5. Syntactic operations that can aggravate the separation of the relative clause from the head noun

The default most common word order in relative clauses requires that the head noun is immediately followed by the relative pronoun. There are however at least two processes that result in intervening material between the head noun and the relative pronoun; they are schematically shown in (8):

(8)

	Matrix clause	||	**Relative clause**
Default RC:	**head noun**	||	**relative pronoun**
Discontinuous RC:	**head noun ←**	||	**relative pronoun**
RC with pied-piping:	**head noun**	||	**➔ relative pronoun**

The first process moves the head noun inside the matrix clause to the left from the boundary; the resulting phenomenon is known as "extraposed", or "discontinuous" relatives

The second process moves the relative pronoun to the right from the boundary inside the RC, as a result of a phenomenon known as pipe-piping.

An example of a discontinuous relative given in (9), which is actually a modified pattern example (1):

(9)

{*Mne*	*/plat'je*	*nravitsja****}***	***{{****kotoroe*	*ona*	*iz*	*\sitca*	*sšila****}***
I	dress	like		REL	she	of	printed.cotton made

'I like the dress that she made of printed cotton'

Lit. '**{**I dress like**}**$_{\text{Topic}}$ **{**that she of printed cotton made**}**$_{\text{Comment}}$'

The "detached" head noun (in (9), it is *plat'je* 'dress') usually retains the main phrasal accent (typically, with the rising pitch movement) responsible for signaling the topical status of the first intonation phrase and its non-final status in the sentence, i.e. signaling the continuation of the unfolding discourse. The RC forms a comment constituent and retains its prosodic autonomy as a separate intonation phrase.

Extraposition of RCs is optional in Russian. In the National corpus of Russian approximately 1 % of RCs are discontinious.

Pipe-piping of RCs, on the other hand, is forced in Russian, especially with prepositions, in the National corpus of Russian approximately 30 % of RCs are pipe-piped.

An example of a pipe-piped relative is given in (10):

(10)

{*Mne*	*nravitsja*	*/plat'je***}**$_{\text{Topic1}}$	
I	like	dress	
{*árukava*	*/kotorogo*ñ$_{\text{Topic2}}$	*ásšity iz*	*\sitca*ñComment2**}**$_{\text{Comment1}}$
sleeves	REL	made of	printed.cotton

Lit. '{I like dress} {< sleeves [of] which> < are made of printed cotton>} '

Of pipe-piped RCs in the National corpus of Russian: 74% are within prepositional phrases, 21.9% are within NPs, like the one in example (10), other 4.1% include groups headed by infinitives, comparative forms of adjectives and even finite verb forms, like in (11):

(11)

{*Eto*	*byla*	*/armija}*	***{****<komandoval*	*/kotoroj>*	*<general*	*\Samsonov>}*	
this	was	army	lead	which	general	Samsonov	

'This was the army which general Samsonov lead'

Lit. '{this was army} {<lead **which>** <general Samsonov>}'

Previous studies (Zaliznjak, Padučeva 1979, Ljutikova 2009 inter alia) have convincingly demonstrated the effect of pied-piping in Russian RCs: if a relative pronoun appears inside a smaller constituent within the RC, the whole this constituent is fronted rather than the single relative pronoun. What remained unnoticed so far is that the fronted constituent and the rest part of the relative clause can be articulated as separate intonation phrases (shown with angle brackets in (10) and (11)). Within the fronted constituent the pied-piped relative pronoun, otherwise strictly atonic, may acquire the main phrasal accent. The accent (typically, with the rising pitch movement) signals the topical status of this intonation phrase and its non-final status within the relative clause, which thus acquires its internal information structure. The internal intonation and information structuring, in its turn, increases the autonomy of the RC and levels it with an independent predication.

In sum: discontinuous RCs and RCs with the pipe-piping effect can aggravate the separation of the relative clause from the head noun both syntactically and prosodically.

6. Conclusions: How can Russian data be mapped to what we know about the interrelationship between syntax and prosody

Sun-Ah Jun in her paper on "Prosodic Phrasing and Attachment Preferences" (Jun 2003) reports an experiment in which native speakers of English, Greek, Spanish, French, Japanese and Korean were asked to produce a sentence meaning 'John chased/saw the dog that bit the cat' to check the most natural prosodic phrasing. The results are summarized in (12):

(12) based on Jun 2003

English:	*{John chased the dog} {that bit the cat}*				
Greek:	*{O Giannis kinigise to skilo} {pu dagose ti gata}*				
Spanish:	*{Juan vio al perro} {que persiguió al gato}*				
French:	*{John a poursuivi le chien} {qui a mordu le chat}*				
Japanese:	*John-ga*	*neko-ni*	*kamitzuita*	*inu-o*	*oikaketa*
	John-nom	cat-at	bit	dog-acc	chased
	{Johnga} {nekoni kamitzita inuo} {oikaketa}				
Korean:	*John-i*	*koyangi-lul*	*mun*	*kangaji-lul*	*ccochatta*
	John-nom	cat-acc	bit-that	puppy-acc	chased
	{Johni} {koyangilul mun} {kangajilul ccochatta}				

As shown in (12), speakers of English, Greek, Spanish and French put a prosodic break between the head noun and the relative clause, exactly as in Russian. Japanese, on the other hand, shows a different pattern: the prosodic break comes after the head noun (*inu-o*), so that the relative clause and the head noun form one intonation phrase.

English Greek, Spanish, French and Russian are right-branching with postpositional relatives, while Japanese is a left-branching language with prepositional relatives. So, one could hypothesize that the difference in prosody is conditioned by word-order. But this hypothesis makes a wrong prediction for Korean: in Korean, which is also left-branching with prepositional relatives, the prosodic break comes before the head noun (*kangaji-lul*), so the head noun and the relative clause appear again in separate intonation phrases.

Jun and Chisato (2008) suggest that this may be due to a morpho-syntactic fact, that, unlike Korean, Japanese has no complementizer marking the boundary of a relative clause. If there is a prosodic break after the RC, the verb could be interpreted as a sentence final verb. In (12), this would result in false understanding, like, 'John bit the cat <and then> chased the dog'.

The rule, however, can be overridden and the prosodic break after the RC is favored when the head noun is complex, having an internal nominal modifier, like in the sentence *John chased his friend's dog // that bit the cat*.

Even this short illustration shows that prosodic phrasing is a multifactorial process influenced by various structural and speech production parameters. It is not clear yet, to what extent these parameters are universal. Regarding relative clauses, so far, very little is known about their cross-linguistic prosodic variation. There is some literature on the prosody of relative clauses in individual languages, but it is mostly restricted to the two topics: first, the prosodic difference between restrictive and non-restrictive relatives, and second, the prosodic phrasing in relative clauses with a complex head noun (otherwise known as the problem of distinguishing between "early" and "late" closure).

The absolute majority of this literature is based on experimental, laboratory material, rather than on natural data. However, only natural corpus data can help not only in discovering actual prosodic patterns, but also in understanding why some possible prosodic phrasing patterns remain underrepresented, while others are favored

in particular discourse settings. In this paper, basing on data from the prosodically annotated corpus, I have shown that Russian systematically demonstrates syntactic and prosodic autonomy of the relative clause, favoring the prosodic break between the head noun and the relative clause, thus systematically desubordinatings RCs.. Our corpus data allows to hypothesize that there might be a strong discourse reason for this — namely, speakers tend to produce discourse in such a way, that each clause appears as a separate intonation phrase. But certainly, the typological validity of this hypothesis is to be further checked against natural data from other languages.

So far, this work is intended as an empirical study, but, hopefully, the one that can form the foundation for further theoretical and typological development, for better understanding the nature of syntax-prosody mapping in sentential embedding constructions.

Notes

*The research is supported by the Russian Foundation for Fundamental research, grant #10-06-00338a

References

1. *Croft W.* 1995. Intonation units and grammatical structure. Linguistics, 33 : 839–852.
2. *Jun, Sun-Ah.* 2003. Prosodic Phrasing and Attachment Preferences. Journal of Psycholinguistic Research, 32 (2) : 219–49.
3. *Jun, Sun-Ah, Koike Chisato.* 2008. Default Prosody and Relative Clause Attachment in Japanese. Japanese-Korean Linguistics, 13 : 41–53.
4. *Korotaev N. A.* 2009. Syntax and Prosodia in the Systems of Discourse Connection Means [Sintaksis i Prosodiia v Sisteme Sredstv Diskursivnoi Sviaznosti Teksta].
5. *Liutikova E. A.* 2009. Relative Sentences with the Word 'KOTORYI'[Otnositel'nye Predlozhenia s Soiuznym Slovom 'KOTORYI']. Korpusnye Issledovaniia po Russkoi Grammatike : 436–511.
6. *Zalizniak A. A., Paducheva E. V.* 1979. Syntactic Characteristics of the Pronoun 'KOTORYI' [Sintaksicheskie Svoistva Mestoimeniia 'KOTORYI']. Kategoriia Opredelennosti-Neopredelennosti v Slavianskikh i Balkanskikh Iazykakh : 289–329.

EXPLORING SEMANTIC ORIENTATION OF ADVERBS

S. B. Potemkin (potemkin@philol.msu.ru)

G. E. Kedrova (kedr@philol.msu.ru)

Lomonosov Moscow State University, Moscow, Russia

Sentiment analysis often relies on a semantic orientation lexicon of positive and negative words. Determining the semantic orientation of words is necessary for correct estimation of the content of statements in the media, Internet, in the writings and speech. Qualitative adverbs expressing evaluation, intensity, direction of action are important as the modifiers of the main sentence predicate. In this paper we propose a method for extracting seed set of adverbs from a collection of pairs of antonym. A model based on the representation of a set of synonyms from the Russian lexicons as a graph, and determination the semantic orientation of the adverbs concerning three main dimensions of the semantic differential also demonstrated. The assessment of performance of the method in comparison with the dictionary data shows effectiveness of the method obtained.

Key words: adverb, semantics, semantic orientation, sentiment analysis.

Introduction

Nowadays, the availability of resources for Natural Language Processing (NLP) remains a hot topic, in particular for Russian especially due to the lack of comprehensive semantic resources, despite efforts made to provide a freely-available Russian WordNet [1]. Ability to establish relativity, similarity, or semantic distance between words and concepts is the basis of computational linguistics. This paper deals with measuring of distance within the syntactic category of adverbs. This set of words is crucial for some applications because adverbs modify or clarify the meaning of other words (verbs, nouns, adjectives). The adverbs are of particular interest to determine the semantic orientation of syntagma containing a main word and its modifier (adverb). Measuring the semantic distance or similarity between the English words most often is based on WordNet [2], and almost exclusively on taxonomic relationships established in this database. So such approach is applicable only to the syntactic categories of nouns and verbs.

The aim of this paper is to extract a list of semantically oriented adverbs and develop the measure of proximity based on dictionaries of synonyms. The article is structured as follows. In Section 1 the problem of extracting the seed set of semantically oriented adverbs from the lexicon of Russian antonyms is discussed. In Section 2 we describe the previously proposed measures of semantic distance between words, as well as an elementary way to map synonyms onto a graph. In Section 3 the

basic characteristics of the subjective understanding of the meaning and the measures based on the distance in a graph of synonyms are discussed. Finally, Section 4 presents some results and conclusions. Additionally, we explore the use of visualization techniques to gain insight into the results obtained.

1. Extracting the seed set of adverbs

A number of approaches have been proposed for creating semantic orientation lexicons in English, most of them are computationally expensive and rely on significant manual annotation and large corpora. Particularly, the General Inquirer [3] created in the beginning of the last century is used as the gold standard for assessment the quality of new-generated lexicons. For Russian language there is no open-source and reliable lexicon with positively and negatively marked entries. We propose some approaches to generate a broad coverage semantic orientation lexicon for Russian adverbs which includes both individual words and multi-word adverbial expressions using only dictionaries of antonyms and synonyms, requiring a small amount of manual pruning and database processing.

First of all we have analyzed a list of antonyms collected from published dictionaries of antonyms [4, 5]. This list contains 7,300+ antonymous pairs (adjectives, nouns, verbs, adverbs and prepositions as well). The semantically oriented words were manually extracted from this list and arranged in 2 separate lists — positive (1,859) and negative (2,229) words. This seed lexicon could be compared with the GI lexicon which contains orientation labels for only about 3,600 entries.

Next step was to extend our seed lexicon to obtain a broad coverage of different texts under consideration concerning sentiment analysis. Automatic approaches to create (English) semantic orientation lexicon and, more generally, approaches for word-level sentiment annotation can be grouped into two kinds: (1) those that rely on manually created lexical resources—most of which use WordNet; and (2) those that rely on text corpora [6]. As a lexical source we use a structured list of Russian synonyms collected from a number of published and Internet-available dictionaries such as [7] and others (11 sources). List of synonyms contains ~600,000 word-pairs including ~10,000 pairs of adverbs. All synonyms $\{s(w_i)\}$ of each seed word w_i receives the same semantic orientation as w_i. The number N of occurrences of a synonym $s(w_i)$ in the extended set contributed by different seed-words w_i, $(i=1…N)$ indicates the confidence of semantic orientation. After manual pruning we have got a list of positively marked (5990, including 731 adverbs) and negatively marked (6853, including 592 adverbs) words. Since the most part of Russian adverbs could be derived as the short form singular neutral or short form plural adjective (3135) the list of semantically orientated adverbs could be expanded.

2. Measures of distance

A number of distance or similarity measures exist for English based (completely or partially) on WordNet. In particular, such measure is defined as the number of edges

of the path through the taxonomic relations (IS-A, Part-of, or WordNet's hyponymy relation). In [8] the concept of bond length was extended for all relations in WordNet by their clustering in the horizontal (synonyms) or vertical (hyponymy) direction and assigning a penalty for changing the direction of the path motion. Overview of five measures and evaluation of their effectiveness using the associations between the words is given in [9]. Exclusive usage of hyponymy delimits the measure of distance or similarity only to the syntactic categories of nouns and verbs, as hyponymy relations in WordNet are established only for these grammatical categories. Therefore, such measures could not be applied to adjectives and adverbs.

The semantic distance between the words could be determined in the similar way as the definition adopted in graph theory [10]. The simplest approach is just to gather all the words from the Dictionary of synonyms and to link each member of a synonymous group with its dominant word as indicated in the Dictionary. Let *G(W,S)* be the undirected graph, with *W* the set of nodes being all the words from the Dictionary with associated part-of-speech, *S* — the set of edges connecting each member of synonymous group with its dominant word. Every group of synonymous words could be connected to each other and form a clique in *G* graph. A path *P* is the sequence of nodes connected by edges of *G* and geodesic is the shortest path between two nodes. Geodesic distance, $D(w_i,w_j)$ between two words w_i and w_j is the length (number of edges) of the shortest path between w_i and w_j. If there is no path between w_i and w_j, the distance between them is infinity. The minimal path-length defines a *metric* on the set of synonyms. All axioms of the metric space are fulfilled in this case. Usually synonymous groups comprises the words of the same grammatical category and entire graph *G* is decomposed into disjoint sub-graphs or networks for nouns, verbs, adjectives and adverbs. (Fig. 1). In each network exists a maximal connected component that contains 70–90% of all nodes of the graph constructed from the Dictionary of synonyms. Maximum component in the class of Russian adverbs contains about 8500 words. The words in this connected component could be analyzed using the metric defined by the length of geodesics.

3. Semantic orientation of adverbs

Classical work on the measurement of emotional or affective values in texts is the theory of semantic differential by Charles Osgood. Word meaning in cognitive psychology, is "a strictly psychological one: those cognitive states of human language users which are necessary antecedent conditions for selective encoding of lexical signs and necessary subsequent conditions in selective decoding of signs in messages." [11]. Semantic differential method was applied mainly to the adjectives measured in such dimensions as *active/passive*, *good/bad*, *positive/negative*, *beautiful/ugly*, etc. Each pair of bipolar adjectives is a factor or an axis in the method of semantic differential. Application of factor analysis to extensive empirical material gave an unexpected result: most of the variance in judgment could be explained by only three major factors including the *evaluative* factor (e. g., *positive/negative*); the *potency* factor (e. g., *strong/weak*); and the *activity* factor (e. g., *active/passive*). Among these three factors, the evaluative factor has the strongest relative weight for determining the semantic orientation.

Turning to the selected Russian adverbs, we note that the vast majority of adverbs is matched with the words of other parts of speech primarily with the adjectives (*cheerful — cheerfully // бодрый — бодро*, *brutal — brutally // жестокий — жестоко)*, so that the semantic differential can be naturally extended to motivated adverbs, which bear semantic meaning and, accordingly, deliver the information on their semantic orientation. All three pairs of bipolar adverbs *negatively/positively* (*плохо/хорошо*); *weakly/strongly (слабо/сильно)*, *passively/actively (пассивно/активно)* are contained in the maximal component of the sub-graph of synonymous adverbs G_{adv}. One can assume that the distance to *positively (хорошо)* is a measure of positive assessment of an adverb. However, it is easy to show that this measure is in fact rather controversial.

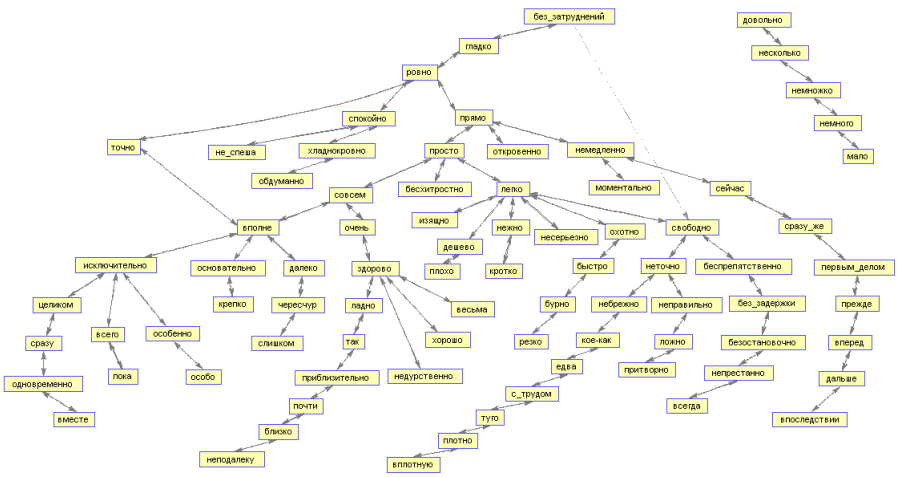


Fig. 1. A fragment of the maximum connected component subgraph of adverbs, $G_{adv.}$ One cannot select a consistent spanning tree

A striking example of this is that the words *positively (хорошо)* and *negatively (плохо)* are closely related through the path of synonyms. There is a sequence of only 5 words in English (*negatively, hardly, tightly, thoroughly, comprehensively, soundly, positively*), and 6 words in Russian (*плохо, дешево, легко, просто, совсем, очень, здорово, хорошо*) — see Fig. 1 — connecting opposites, each pair of words in this sequence is certainly synonymous (at least in one of their meanings). Thus, we find that *d(positively, negatively)* = 6; *d(хорошо, плохо)* = 7. Despite the fact that the adverb *positively (хорошо)* and *negatively (плохо)* have opposite meanings, they are closely related by synonymy path. Of course, this is not due to any error in the Dictionary of synonyms. Partial explanation lies in the wide use of two Russian adverbs *хорошо* (625 ipm), *плохо* (187 ipm) [12]. The other source of uncertainty is the fact that a spanning tree probably could not be chosen perfectly, e. g. the path from *неподалеку (not far from)* to *вплотную (close to)* is 12 arcs length while their meanings are very similar. In addition, we observe 'shift of meaning' while travelling by the path of polysemic

synonyms, i. e. the left arc of path A — B1/B2 — C connects A with B1 meaning while the right arc connects A with the other one, B2. Here we assume that B1/B2 do have some sema in common (if these are not the pure homonyms which could be filtered out automatically). Nevertheless due to the fact that both words *хорошо*, *плохо* are members of the maximum connected component of G_{adv} sub-graph, we can consider not only the shortest distance from any adverb to "*positively*", but the shortest distance to its antonym, "*negatively*". This idea is concretized [13] in the definition of EVA function, which allows to measure the relative distance from the word of two opposites, "*positively*" and "*negatively*":

$$EVA\ (w) = (d\ (w, neg) - d\ (w, pos)) / d\ (neg, pos).$$

Under the assumption that there is no word "worse than *negatively*" or "better than *positively*" the values of EVA lie in the interval [-1,1], for example, the word "honestly" is evaluated by function EVA (*honestly*) gives a value of 1 as follows EVA (*honestly*) = (d (*honestly*, *neg*) — d (*honestly, pos*)) / d (*pos, neg*) = (8-2) / 6 = 1,. The measures for other Osgood's dimensions is defined similarly. For the potency factor the function: POT (w) = (d (w, *weakly*)-d (w, *strongly*)) / d (*strongly, weakly*) is defined; for the activity factor the function: ACT (w) = (d (w, *passively*)-d (w, *actively*)) / d (*actively, passively*) is defined. This fact allows to define measures for any two words belonging to the maximal connected component of the adverbs subgraph.

An assumption on the boundary position of words *negatively/positively* is not entirely justified. Intuitively, *perfectly (превосходно)* is better than positively, *disgustingly (отвратительно)* is worse than negatively. Bearing this in mind and using the geometry of a triangle with vertices {w, *pos, neg*}, we redefine the function of EVA, namely:

$$EVA_1(w) = (d(w, neg) - d(w, pos))*(d(w, neg)+d(w, pos))/d^2(neg, pos).$$

The values of EVA_1 sometimes are beyond the interval [–1,1]. Similarly, we can redesign POT(w) and ACT(w).

For English adjectives (and motivated adverbs) there exists the source for assessing the measure constructed above in comparison with the independently obtained answers to the "General Inquirer" [11], which contains a set of words to assess three Osgood's factors. Word lists were obtained from the Stanford political dictionary, where each of the 3,000 most frequent common words were assessed by three or more experts concerning each Osgood's factor. Thus 765 positive and 873 negative words for the assessment factor were obtained, 1,474 strong and 647 weak word for the potency factor and 1,568 active and 732 passive words for activity factor. Comparison of results obtained with the General Inquirer gave the values of 70–80 % of matches, depending on what words were considered as neutral in terms of EVA function.

In the absence of available data for content analysis we used the Russian dictionaries of antonyms as an independent source. Antonymous pair is a pair of words (or rather, the specific meanings of words), one opposed to the other on semantic grounds, such as *hot — cold fast — slow, present — absent*. We suggest that adverbs belonging to pair of antonyms lie on the "opposite sides" of the entire set of adverbs.

Methods of multidimensional scaling deliver a mapping of multidimensional space with the defined distance between individual points $d(w_i, w_j)$ onto a space of smaller dimension, namely the plane (Fig. 2). Figure 2a, b shows that the pairs of antonyms lie near the diameters of the set of adverbs. For a more profound study of the structure of the space of adverbs we have constructed chains of synonyms connecting antonyms pairs within the sub-graph G_{adv}.

Chain in Fig. 2a is a consistent result, i. e. the chain of synonyms passes on the periphery of the set of adverbs and the distances between the synonyms do not exceed the distance between the antonyms. Unfortunately, the situation is not always as favorable. In Fig. 2b pair of antonyms is close to the diameter, but the chain of synonyms is not at the periphery of the set, but lays in the central part of the set, alternates its direction, and the distances between synonyms is often greater than the distance between antonyms. Probably it is necessary to determine more accurate distance between the words and to choose correctly the axes of the adverb space using the principal components method. These new axes should not coincide Osgood's dimensions.

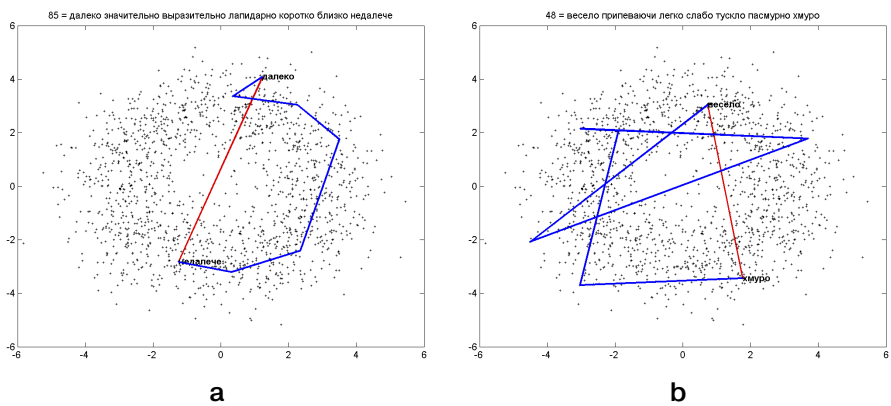


Fig. 2. Two chains of synonyms, joining antonymous pairs of adverbs. a) Left — a consistent path; b) Right — an inconsistent result

4. Discussion and conclusions

In this paper we define a measure of the distance between adverbs using synonyms graph. It seems obvious that the choice of similarity measure, or distance largely depends on the type of the problem. The choice of distance measure on the grounds of synonyms is connected with the goal of determining the semantic orientation of adverbs. In contrast to Osgood's semantic differential associated with the reaction of people on the stimulus — words presented, or the possible emotional impact of words, this model is based solely on the lexical material and is intended to represent relatively objective meanings which are fixed in Dictionaries.

Some inadequate results (as in Fig. 2b) probably arise from the inadequate dimension (3 axes) of Osgood's space.

Further studies will determine the semantic orientation of sentences or the whole text on the basis of the orientation of its constituent words. Our method allows to evaluate other classes of words such as nouns, adjectives and verbs, but this extension will require a significant increase of calculations and special methods for processing large data sets, since an algorithm for computing shortest paths requires $O(n^3)$ operations, where n is the number of words in graph G(W,S).

References

1. *Aleksandrova Z. E.* 2005. Russian Synonyms Dictionary [Slovar' Sinonimov Russkogo Iazyka].
2. *Azarova I. V., Mitrofanova O. A., Sinopal'nikova A. A.* 2003. Computational Thesaurus of Russian Language of the kind of WordNet [Komp'iuternyi Tezaurus Russkogo Iazyka Tipa WordNet]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2003" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2003") : 43–50.
3. *Budanitsky A., Hirst G.* 2001. Semantic Distance in WordNet: An Experimental, Application-Oriented Evaluation of Five Measures. Workshop on WordNet and Other Lexical Resources. Second meeting of the NAACL.
4. *Hirst G., St-Onge D.* 1998. Lexical Chains Representations of Context for the Detection and Correction of Malapropisms". WordNet. An Electronic Lexical Database.
5. *Kamps J., Marx M., Robert J., Mokken M.* 2004. Using WordNet to Measure Semantic Orientations of Adjectives. Proceedings of the 4th International Conference on Language Resources and Evaluation, IV : 1115–1118.
6. *Mohammad S.* et.al. 2009. Generating High-Coverage Semantic Orientation Lexicons From Overtly Marked Words and a Thesaurus. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2009).
7. *Osgood C. E., Succi G. J., Tannenbaum P. H.* 1957. The Measurement of Meaning.
8. *Potemkin S. B.* 2008. Semantic Distance in the Linguistic Database WorldNet [Semanticheskoe Rasstoianie v Lingvisticheskoi Baze Dannykh WorldNet]. Materialy 10 Mezhdunarodnoi Konferentsii "Kognitivnoe Modelitovanie v Lingvistike" (Proc. of the X International Conference "Cognitive Modelling in Linguistics").
9. *Sharov S. 2003.* Frequency Dictionary [Chastotnyi Slovar'], available at: http://www.artint.ru/projects/frqlist.asp
10. *Stone P. J.* 1997. Thematic Text Analysis: New Agendas for Analyzing Text Content. Text Analysis for the Social Sciences.
11. *Vvedenskaia L. A.* 2004. Russian Antonyms Dictionary [Slovar' Antonimov Russkogo Iazyka].
12. *L'vov M. R.* 2006. Russian Antonyms Dictionary [Slovar' Antonimov Russkogo Iazyka].
13. *WordNet: An Electronic Lexical Database. Language, Speech, and Communication Series.1998.*

НЕКОТОРЫЕ ОСОБЕННОСТИ СИНТАКСИЧЕСКОЙ СТРУКТУРЫ РУССКИХ ПОСЛОВИЦ (НА ПРИМЕРЕ МОНОПРЕДИКАТНЫХ ПРЕДЛОЖЕНИЙ)

Е. А. Ренковская (jennyrenk@rambler.ru)

ABBYY

В рамках данной работы рассматриваются особенности синтаксиса и словопорядка в предложениях-пословицах с единственным главным предикатом, выраженным глаголом. Акцент делается на зависимости синтаксической структуры пословицы от общего прагматического назначения этого вида паремий как фольклорного жанра и на тех синтаксических особенностях, которые отличают пословицу от других повествовательных предложений русского языка.

Ключевык слова: пословицы, синтаксис, предложение-пословица, глагольный предикат, паремия, фольклор.

SOME PECULIARITIES OF THE SYNTACTIC STRUCTURE OF RUSSIAN PROVERBS: A STUDY OF ONE-PREDICATE SENTENCES

E. A. Renkovskaia (jennyrenk@rambler.ru)

ABBYY

The paper discusses some peculiarities of the syntactic structure and word order in Russian proverbial sentences with one verb as a predicate. We argue that the syntactic structure of proverbs is dependent on their general pragmatic purposes. The paper focuses on the syntactic features that make proverbs a specifi c type of Russian sentences.

Key words: proverb, proverbial sentence, syntax, paremia.

Введение

В статье перечислены некоторые основные особенности синтаксиса в пословицах. Утверждается, что синтаксические особенности пословицы, а именно синтаксическое наполнение и порядок слов, во многом определяются прагматическим назначением этого вида паремий как фольклорного жанра. В частности, показывается, что прагматической заданностью пословицы как текста определяются: — тип предиката и его место в предложении; — количество и место в предложении актантов и сирконстантов; — условия распространенности актантов.

Материалом для анализа послужили не все возможные синтаксические типы пословиц, а только те, синтаксическая структура которых представляет собой простое предложение с единственным главным предикатом, выраженным глаголом. Это такие пословицы, как *Береженого Бог бережет, Старую собаку новым фокусам не научишь, Цыплят по осени считают, На чужой каравай рот не разевай* и т. д. Выбор такого типа пословиц для исследования не случаен. Во-первых, этот тип пословиц является наиболее частотным. Исследуя частотность синтаксических типов пословичных простых предложений по сборникам пословиц XVIII века В. И. Даля, П. Симони, А.М Жигулева и пословиц XVII века Петровской галереи, И. В. Пауса, В. Н. Татищева, Н. А. Добролюбова, З. К. Тарланов (Тарланов, 1999) указывает именно такой тип предложений как самый частотный. Кроме того, среди самых употребимых пословиц современного русского языка, приведенных в книге Г. Л. Пермякова (Пермяков, 1988), примерно 60% пословиц относятся к этому типу (всего около 80% наиболее частотных пословиц представляют собой простые предложения с одним главным предикатом). Во-вторых, это наиболее простой тип синтаксической конструкции; в порядке гипотезы можно сделать предположение, что особенности синтаксиса, выделенные в конструкциях такого типа, с большой вероятностью будут представлены и в пословицах с более сложной синтаксической структурой.

Предыдущие работы, посвященные синтаксису пословиц, либо претендовали на комплексное описание структуры пословицы в целом (см, например, Глаголевский 1873, Krikmann 1984), либо были посвящены отдельным синтаксическим конструкциям. В докладе делается попытка на примере конструкций одного синтаксического типа показать зависимость синтаксической структуры и словопорядка в пословицах от прагматического назначения пословицы как фольклорного жанра.

1. Типы глагольных предикатов в монопредикатном простом предложении-пословице

В пословицах, представляющих собой простое предложение, можно выделить 5 основных типов предикатов:

1) Глаголы несовершенного и совершенного видов, употребленные в настоящем или будущем времени, в форме третьего лица единственного

и множественного числа. Такие глаголы могут употребляться как с отрицательной частицей «не», так и без нее. По большей части это предикаты несовершенного вида, пословицы с вершинным предикатом совершенного вида составляют очень немногочисленный класс. В основном в таких предложениях присутствует подлежащее, хотя встречаются и безличные предложения. В таких пословицах сообщается некоторый факт действительности, о котором следует помнить.

Примеры пословиц такого вида: *Волка ноги кормят, В тихом омуте черти водятся, Копейка рубль бережет, Дурная голова ногам покоя не дает, На торной дорожке трава не растет, Семеро одного не ждут, Язык до Киева доведет, Два медведя в одной берлоге не уживутся, Отольются волку овечьи слезы, На бедного везде каплет, На зло и дурака хватит.*

2) Глаголы несовершенного вида, употребленные в форме повелительного наклонения, единственного числа. Такие глаголы могут употребляться как с отрицательной частицей «не», так и без нее. Материально выраженное подлежащее при таких предикатах отсутствует. С точки зрения прагматики, пословицы такого типа содержат прямой императив, указание к действию.

Примеры: *По одежке протягивай ножки, По напору и отпор держи, На чужой каравай рот не разевай, Не в свои сани не садись, От сумы да от тюрьмы не зарекайся, Не пугай сокола вороной, хвали утро вечером, знай, кошка, свое лукошко.*

3) Глаголы совершенного вида, употребленные в будущем времени в форме второго лица единственного числа. Такие глаголы за очень редким исключением употребляются с частицей «не». Материально выраженное подлежащее при таких предикатах отсутствует. Прагматически пословицы такого типа являются предупреждением на будущее.

Примеры: *Старую собаку новым фокусам не научишь, Шила в мешке не утаишь, Отрезанный ломоть к хлебу не приставишь, Каши маслом не испортишь, За один раз дерева не срубишь, Попа и в рогоже узнаешь, Ласковым словом многого добьешься.*

4) Глаголы несовершенного вида, употребленные в настоящем времени в форме третьего лица множественного числа. Такие глаголы могут употребляться как с частицей «не», так и без нее. Тем не менее, предикаты с частицей «не» в такого вида пословицах встречаются намного чаще. Материально выраженное подлежащее в таких предложениях отсутствует. Пословицы такого типа отсылают адресата к народному опыту (сообщают о том, что среди людей так принято или так бывает).

Примеры: *За одного битого двух небитых дают, Цыплят по осени считают, Клин клином вышибают, В чужой монастырь со своим уставом*

не ходят, После драки кулаками не машут, Соловья баснями не кормят, Дареному коню в зубы не смотрят.

Три перечисленных выше разновидности пословичных предложений (а именно, пункты 2), 3) и 4)) являются обобщенно-личными.

5) Глаголы совершенного и несовершенного видов в форме инфинитива. Такие глаголы почти всегда употребляются с частицей «не» (исключением, в частности, являются пословицы с глаголом «знать» в качестве предиката, ср. *Знать волка и в овечьей шкуре*). Материально выраженное подлежащее в таких предложениях отсутствует. Пословицы такого типа достаточно редки и представляют собой описание фактов действительности.

Примеры: *Грех не уложить в мех, И большой бадьей реки не вычерпать, Деньгами души не выкупить, Из спасиба шубы не шить.*

Подавляющее большинство глагольных предикатов в пословицах, представляющих собой простое предложение, относятся к одному из вышеперечисленных пяти типов. Но есть и небольшой класс глагольных предикатов пословиц, которые к этим типам не относятся. Изучая типы глагольных сказуемых в пословицах, Тарланов утверждает, что глаголы в прошедшем времени в основном характерны для поговорок, а в пословицах могут появляться только в сложных синтаксических структурах. Такая особенность проистекает из вневременного характера пословиц как наставлений. И все же можно привести крайне немногочисленный класс пословиц, которые, будучи простыми предложениями, содержат предикат в прошедшем времени. Все такие пословицы-предложения являются двусоставными. Например: *Каждая река своим устьем в море впала, Свет клином не сошелся, Москва не сразу строилась, Русский на авось и взрос, Руном с овцы одевались и отцы* а также вопросительные конструкции*: Кто горя не знавал, Кого черт рогами под бока не пырял, Горе да беда с кем не была*. Про предикаты в 1 лице Тарланов пишет, что они возможны только в сложных конструкциях, тем не менее глаголы в 1 лице множественном числе могут встретиться, хотя и крайне редко, в качестве предиката и в простом предложении (*Миром и горы сдвинем, Собором и черта поборем*).

2. Количество зависимых членов в пословице

Рассмотрев вопрос о типах глагольных предикатов в исследуемых нами пословицах, остановимся на вопросе количества членов предложения, синтаксически зависимых от предиката. От количества актантов и сирконстантов главного предиката зависит возможность функционирования предложения как пословицы, так как большое их число является препятствием для запоминания пословицы, а недостающее — обуславливает смысловую неполноту предложения. Согласно имеющемуся в нашем распоряжении материалу, в пословице,

за редким исключением, может быть не больше трех зависимых от предиката членов предложения.

Посмотрим теперь, как количество актантов и сирконстантов зависит от типа предиката в пословице:

1) В пословицах с одновалентным главным предикатом представлено в большинстве случаев материально выраженное подлежащее (ср. *На торной дороге трава не растет*). Обобщенно-личные и инфинитивные предложения здесь почти не встречаются. Этому обстоятельству можно дать объяснение: в обобщенно-личных и инфинитивных предложениях соответственно восстанавливаемое нулевое подлежащее в значении обобщенности обладает недостаточной семантикой, смысловая информация в нем минимальная, поэтому смысловая нагрузка падает на другие члены: актанты и сирконстанты. Напротив, если подлежащее обладает смысловой наполненностью, то на актанты и сирконстанты приходится меньше смысловой нагрузки, это делает их менее обязательными.

2) В предложениях с предикатом 3 лица множественного числа этот предикат может либо иметь три валентности (субъектная и две несубъектные актантные), ср. *Клин клином вышибают*, *Соловья баснями не кормят*, либо быть двухвалентным, но в таком случае в пословичном предложении присутствует также сирконстант. Например, у глагола ‘*считать*’ две валентности, субъектная и объектная, и, в пословицах указанного типа в качестве предиката он присоединяет к себе сирконстант: в пословице *В чужом хлеву овец не считают* — этот сирконстант обозначает место действия, в пословице *Цыплят по осени считают* — время действия.

В остальных трех синтаксических разновидностях пословиц предикаты могут присоединять любое число актантов от одного до трех включительно.

3. Место главного предиката в предложении-пословице

Одной из основных синтаксических особенностей монопредикатных пословиц является то, что главный предикат в них в большинстве случаев стоит в конце предложения. Такая особенность словопорядка является также одним из основных структурных отличий пословиц от обычных повествовательных простых предложений того же синтаксического типа (подробнее о порядке слов см. Йокояма 2005, Ковтунова 2002, Янко 2001). Важно, что это свойство пословиц не зависит даже от количества второстепенных членов, все они ставятся перед предикатом. Примеры можно привести из всех выделенных выше разновидностей пословиц по типу предиката: *В полую воду за рекой не ночуй, От сумы да от тюрьмы не зарекайся, Любви, огня да кашля от людей не утаишь, В чужой монастырь со своим уставом не ходят, Бодливой корове Бог рог не дает, Два кота в одном мешке дружбы не заведут, Без притчи веку не прожить, Каждая река своим устьем в море впала, Собором и черта поборем*. Постоянное расположение предиката в конце предложения-пословицы указывает на его особую важность

в предложении — в частности, предикат в пословице-предложении замыкает рему.

В своей книге (Тарланов, 1999) Тарланов приводит данные, которые говорят о соотношении постпозиции, препозиции и интерпозиции (термины принадлежат Тарланову) различного типа предикатов в обобщенно-личных пословицах-предложениях. Согласно приводимым им сведениям:

А) Глагол 2 лица единственного числа будущего времени занимает следующие позиции:

- Постпозицию (т. е. последнее место в предложении) — в 96% случаев
- Интерпозицию (относительно срединное место в предложении, точнее не первое и не последнее) — 2%
- Препозицию (первое место в предложении) — 2%

Б) Для глагола 3 лица множественного числа настоящего времени соотношение другое:

- Постпозиция — 70%
- Интерпозиция — 21%
- Препозиция — 9%

В) Рассматривая в качестве главного предиката глагол в императиве, исследователь отмечает, что такой предикат «…свободно занимает как препозицию, так и постпозицию» и значительно реже встречается в интерпозиции.

Однако такие данные Тарланов приводит только на основании статистики, без какого-либо анализа, в частности без учета рифмы в пословицах. Если же учитывать рифму и то, что, как и в любых других стихотворных текстах, в рифмованных пословицах синтаксические правила, характерные для данного жанра, отходят на второй план, то большинство случаев препозиции и интерпозиции можно будет не принимать во внимание. И если препозиция предиката в пословицах, не содержащих рифму, еще возможна (стоит, однако, отметить, что препозиция предиката 2 лица единственного числа будущего времени в нерифмованных пословицах встречается крайне редко), то интерпозиция в основном появляется в пословицах, представляющих собой рифмованные двустишия. Ср. *По одежке протягивай ножки, На чужой роток не накинешь платок, Без забора и затвора не спасешься от вора, Полотна не сносишь без пятна, Найдешь келью и под елью, Рыбак рыбака видит издалека, Из угольного мешка не посыплется мука.* Помимо обычной рифмы стоит упомянуть и характерную для пословиц частичную рифму, характеризующуюся несовпадением либо конечных согласных, либо ударных гласных и часто равным количеством слогов в рифмующихся словах (ср. *Не бей Фому за Еремину вину, На весь мир не испечешь блин, Задним умом не ходят вперед, Лишняя говоря доводит до сорома, Без матки пропадут детки, Царевы слуги не жалеют ноги*). Таким образом, можно сделать вывод, что предикат в подавляющем большинстве случаев занимает постпозицию в предложении-пословице.

4. Об «идеальной» структуре пословицы

На синтаксис предложения-пословицы влияют две важные особенности пословицы как фольклорного жанра. С одной стороны, пословица — это самостоятельный, законченный текст, ориентированный на отдельное существование. Следовательно, такой текст должен обладать семантической и синтаксической полнотой. С другой стороны, пословица рассчитана на запоминание и цитирование, поэтому она должна быть лаконичной, лапидарной, краткой. Исходя из этих особенностей, можно предположить, что синтаксическая структура пословицы должна быть завершенной и исчерпывающей, но вместе с тем не слишком распространенной и сложной.

Все это допускает предположение, что для разбираемого в работе типа пословиц идеальной синтаксической структурой можно считать такую структуру, при которой все остальные члены предложения, кроме главного предиката, будут заполнять валентности этого предиката. Пословиц с такой «идеальной» синтаксической структурой достаточно много, ср.:

> *Мышь копны не боится* — глагол «бояться» имеет 2 валентности, субъектную и объектную, обе они заполнены в данном предложении. Или *Счастье в оглобли не впряжешь* — заполнены все 3 валентности глагола «впрягать».

И все же, несмотря на то, что «идеальная» структура полностью удовлетворяет двум противоположным критериям синтаксической полноты и краткости, она встречается в пословицах далеко не всегда. Среди усложняющих структуру элементов, которые могут встретиться в предложении-пословице, можно назвать следующие:

- сирконстанты
- однородные члены
- определения при зависимых членах (другие зависимые у актанов главного предиката и сирконстантов встречаются крайне редко)
- обращения (только в предложениях с предикатом в императиве).

5. Зависимые члены, занимающие препозицию в предложении-пословице

В пословицах с постпозицией предиката первый зависимый член предложения обычно является темой, остальные слова составляют рему. Словорасположение в пословице зависит не только и не столько от семантики единиц, образующих данный текст и его синтаксическую структуру, сколько от назначения данного текста, от его прагматической заданности. Пословица служит для того, чтобы в максимально обобщенном виде и непрямо, в художественной форме представить некоторое суждение применительно к данному лицу или данной ситуации — либо с целью показать, как тому или иному лицу в той или иной ситуации надлежит действовать, либо с целью

относительно данного лица или данной ситуации высказать некоторую аналитическую истину, которую люди в своей жизни должны принимать во внимание. Коль скоро тема пословицы играет роль объекта (в широком смысле слова, включая ситуацию), относительно которого произносится пословица, для нее предпочтительно отводится первое место не только как теме, но и как смысловой теме.

Что же может занимать препозицию или позицию темы в пословице-предложении? Выше было отмечено, что «идеальным» предложением-пословицей может считаться такое предложение, которое состоит только из предиката и заполняющих его валентности актантов. С какой же целью вводятся в пословицу осложняющие элементы — сирконстанты, определения, однородные члены и т. д.? По нашей гипотезе, почти все «лишнее», что появляется в синтаксической структуре пословицы, относится к смысловой теме пословичного текста, т. е. употребляется для описания участника ситуации или самой ситуации, относительно которых произносится некоторое назидание. И именно такие члены предложения оказываются в препозиции.

Итак, обычно препозицию в пословице-предложении занимают:

- **Зависимые члены с определением**. Это правило словопорядка почти не знает исключений. Например: *Изломанного лука двое боятся, На отложенное дело снег падает, От лишнего веселья работа тоскует, За худую привычку и умного дураком обзывают, Осенней озими в закром не сыплют, Большой милостыней в рай не войдешь, От стриженого барана шерстью не поживишься, Кривое полено в поленицу не уложишь, Зяблые семена всегда поздно встают*. Обычно (за редким исключением) в пословице есть только одно определение при каком-нибудь из второстепенных членов.

Особая семантическая функция определений может быть выявлена, если задать к пословице вопрос «почему?» (этот вопрос естественным образом будет отнесен к предикату пословицы) — ответом на этот вопрос часто может служить определение (ср. *Бездонную бочку водой не наполнишь* — «Почему?» — «Потому что бездонная». *Бессовестного гостя пивом из избы не выгонишь* — «Почему?» — «Потому что бессовестный»).

Если же определений в предложении все же два, то они всегда относятся к разным определяемым словам и очень часто оказываются антонимами (*От худого семени не жди доброго племени, Старую собаку новым фокусам не научишь, Мал грех велику вину приносит*).

- **Однородные члены** (*Иглой да бороной деревня стоит, Малому да глупому все с рук сходит, Монаху и попу портной одной меры карманы шьет, С печали да слез в могилу уйдешь, Промеж жизни и смерти и блошка не проскочит, В игре да в попутье людей узнают*.
- **Сирконстанты** (из правила препозитивного расположения сирконстантов исключения составляют обстоятельства длительности действия и обстоятельства, выраженные наречиями образа действия). Роль сирконстантов заключается в том, что они являются *ситуативными уточнениями*, то есть с их помощью сфера истинности пословичного утверждения сводится

до конкретной единичной ситуации, а ситуация эта, находясь в теме, уже прагматически соотносится с контекстом, в рамках которого и произносится пословица. Такими ситуативными уточнениями могут быть:

— собственно ситуация: *В драке волос не жалеют, В нужде и кулик соловьем свищет, После свадьбы в барабаны не бьют*. В кругу ситуативных уточнений этот класс — наиболее частотный.
— время действия: *Временем и дурак умно говорит, Раз в год и палка стреляет*.
— место действия: *Близ норы лиса на промысел не ходит, Выше лба уши не растут, На болоте все гнилью пахнет*.
— компонент ситуации, необходимый для совершения действия (этот компонент выражается конструкцией существительного в родительном падеже с предлогом «без», ср. *Без столбов и забор не стоит, Без труда не вытащишь и рыбку из пруда, Без росы трава не вырастет, Без счастья и в лес по грибы не ходи*) или компонент ситуации, при отсутствии которого, произойдет нечто нежелательное, ср. *Без перевязи и веник рассыпается, Без подпорки и большая стенка падает*.
— причина: *С поклону голова не отвалится, С уменья руки не болят, За стыд голова гинет, От недосмотра хозяйство гинет*.

Отдельно нужно отметить **одушевленные существительные.** Их нельзя назвать однозначно «лишними» в синтаксической структуре пословицы, в большинстве случаев они являются актантами предиката (в тех редких контекстах, когда одушевленные существительные являются сирконстантами, они играют роль того лица, относительно которого истинно утверждение, содержащееся в семантической структуре пословицы, ср. *Лежебоке и солнце не впору всходит*). Категория одушевленности лица очень важна для пословиц, и одушевленные существительные часто отсылают к ситуации, в которой произносится пословица, а следовательно, относятся к смысловой теме. В основном для пословиц характерны три типа одушевленных существительных: (1) одиночные одушевленные существительные (*Охотника погода не держит, На воре шапка горит, Девку веретено одевает, Волка ноги кормят*); (2) имена собственные в сочетании с определениями, на которые возлагается основная смысловая нагрузка (*По бедному Захару всякая щепа бьет, Про горького Егорку поют и песню горьку*), и (3) наиболее частотный в пословицах класс одушевленных существительных — а именно, класс субстантивированных прилагательных и причастий. То, что большинство одушевленных существительных в пословицах являются субстантивированными прилагательными или причастиями, можно объяснить тем, что такие субстантиваты сочетают в себе одновременно характеристику одушевленности и уточняющие свойства прилагательного или причастия. Например: *Береженого Бог бережет, Смелым счастье помогает, На сердитых воду возят, Виноватого пуля найдет, К любящему страх нейдет, Печальному шутка на ум нейдет, Горбатого могила исправит, С бодливого рога сбивают, Рыжему палец в рот не клади*.

6. Зависимые члены, занимающие второе место от начала

После того, как предпочтительное синтаксическое заполнение тематической части было указано, нужно определить, что же может стоять в начале рематической части. В пословицах исследуемого типа вторая позиция от начала открывает рему. Поскольку именно в рематической части пословицы содержится наставление, носящее обобщенный характер, а информативным центром ремы является предикат и на него приходится основная доля смысловой нагрузки, то обобщенность обеспечивается за счет оставшихся зависимых. В случаях с пословицами основное обобщение приходится на первый зависимый член в рематической позиции. Таким образом, второе место от начала в пословицах занимают:

- **Кванторные местоимения и наречия и конструкции с ними**, ср. *Нужда всему научит, За печью ничего не высидишь, Зяблые семена всегда поздно встают, Пословица всем делам кормилица, Рыболова одна тоня кормит.*
- **Конструкции с кванторной частицей «и»,** близкой по значению к частице «даже», ср. *За глаза и царя ругают, Дурака и в алтаре бьют, Острый топор и дуб рубит, По дважды и Бог за одну вину не карает, Впотьмах и гнилушка светит.* В семантике второстепенного члена, вводимого частицей «и», содержится понятие о том, что данный объект является наиболее ярким представителем класса ‘все’ по определенным признакам, которые выделяются из семантики остальных слов в предложении, и чаще всего из семантики глагола-предиката.

Помимо кванторных конструкций вторую позицию от начала в пословице занимают **конструкции с присоединительной частицей «и».** Эта частица не просто вводит в предложение еще одного участника ситуации, но и указывает на то, что его появление является следствием из сказанного ранее, ср. *Блудливая свекровь и невестке не верит, На ловца и зверь бежит, На красный цветок и пчела летит, Крестьянскими мозолями и бары сыто живут.*

Стоит особо отметить, что все второстепенные члены, которые были перечислены выше как обычно занимающие начальную позицию в линейной структуре пословицы (это распространенные группы, сирконстанты, одушевленные существительные), при употреблении с частицей «и» (как кванторной, так и присоединительной) ведут себя как кванторные выражения и занимают в предложении вторую позицию от начала, ср. *Без подпорки и большая стенка падает, К весне и добрую скотину за хвост подымают, На грех и незаряженное ружье выпалит, Бедность и мудрого смиряет, От беды и без вина зашатает.*

7. Зависимые члены, занимающие место рядом с предикатом

В предложениях-пословицах с глаголами действия или состояния в качестве предикатов встречаются второстепенные члены, которые постоянно занимают позицию рядом с главным предикатом — это разного рода обстоятельства длительности действия или состояния и обстоятельства образа действия, чаще

всего выраженные наречиями. Такие второстепенные члены являются непосредственной характеристикой действия или состояния, и этим может быть объяснено их расположение рядом с предикатом. Например: *Кошка с собакой дружно не живут, Испуганный зверь далеко бежит, Большая рыба маленькую целиком глотает, Дурак деньги напоказ носит, Масло всегда поверху плывет, Битая посуда два века живет, Обещанного три года ждут, Скрипучая береза дольше стоит.*

Заключение

В работе была сделана попытка выявить наиболее частотные закономерности в синтаксисе и порядке слов в предложении-пословице. Было показано, что синтаксическое заполнение и порядок слов в пословице во многом зависят от ее прагматической функции.

Все перечисленные особенности отличают пословицу не только от обычного предложения русского языка, но также и от других устойчивых кратких текстов, как фольклорных, так и авторских текстов назидательного содержания (как то афоризмы, афористические цитаты и др.). Таким образом, структурные особенности могут учитываться при автоматизированной обработке пословиц, в частности служить критерием выделения пословиц из корпусов текстов, близких к пословицам по тем или иным параметрам.

References

1. *Glagolevskii P. P.* 1873. The Syntax of Russian Proverbs [Sintaksis Iazyka Russkikh Poslovits].
2. *Ianko T. E.* 2001. Communicative Strategies of Spoken Russian [Kommunikativnye Strategii Russkoi Rechi]. Iazyki Slavianskoi Kul'tury.
3. *Iokoiama O. B.* 2005. Cognitive Discourse Models and Russian Word Order [Kognitivnye Modeli Diskursa I Russkii Poriadok Slov]. Iazyki Slavianskoi Kul'tury.
4. *Kovtunova I. I.* 2002. Modern Russian Language. Word Order and Actual Segmentation [Sovremennyi Russkii Iazyk I Aktual'noe Chlenenie].
5. *Krikmann A.* 2001. Frage zur logischen Struktur der Sprichwörter. Code. Ars Semeiotica 7, 3 (4) : 387–408.
6. *Permiakov G. L.* 1988. Fundamentals of Structural Paremiology [Osnovy Strukturnoi Paremiologii].
7. *Tarlanov Z. K.* 1999. Russian Proverbs: Syntax and Poetics [Russkie Poslovitsy: Sitaksis I Poetika].

ОПРЕДЕЛЕНИЕ ПОЛА АВТОРА КОРОТКОГО ЭЛЕКТРОННОГО СООБЩЕНИЯ

А. С. Романов (alexx.romanov@gmail.com)

Р. В. Мещеряков (mrv@keva.tusur.ru)

ГОУ ВПО «Томский государственный университет систем управления и радиоэлектроники», Томск, Россия

В статье рассматривается проблема определения пола автора короткого электронного сообщения длиной 20–200 символов. Приводится описание экспериментов и их результаты.

Ключевые слова: автор, пол, определение пола, электронное сообщение, короткое элекстронное сообщение.

GENDER IDENTIFICATION OF THE AUTHOR OF A SHORT MESSAGE

A. S. Romanov (alexx.romanov@gmail.com)

R. V. Meshcheriakov (mrv@keva.tusur.ru)

Tomsk State University of Control System and Radioelectronics, Tomsk, Russian Federation

Gender identifi cation of the author of a short message (20–200 characters) is studied. The paper describes a set of experiments with short message texts performed using a support vector machine approach. The task is viewed as a classifi cation problem with two possible alternatives: male and female. Important features of short messages to be considered when determining the author's gender are singled out. The database of electronic communications collected for research included 41780 posts by 15 men and 15women. Experiments used a software system Avtoroved developed by the paper's authors. Altogether, about 50 text attributes at the level of symbols, words, sentences and their combinations were studied. As a result, relevant characteristics of short messages were identified: unigrams and trigrams of symbols,function words, punctuation and emoticons. The total accuracy of gender identifi cations was 0.74.

Key words: author, gender, gender identification, message, short message.

Ежедневно миллионы людей общаются друг с другом посредством передачи коротких электронных текстовых сообщений с помощью системы SMS, электронной почты, интернет-пейджеров, социальных сетей и т.д. Среда и системы передачи сообщений становятся важной частью человеческой жизни и несут в себе важную информацию об интересах, привычках, социальном поведении людей. Мониторинг этой информации в определенные моменты времени и выявление лиц, имеющих целью совершение злонамеренных действий, становится актуальной практической задачей противостояния террористической угрозе и защиты государства. В беседе люди обычно выбирают манеру общения исходя из пола предполагаемого собеседника, поэтому создание интеллектуальной системы для определения потенциального злоумышленника в среде передачи сообщений и определение его психологического портрета целесообразно начать именно с определения пола автора сообщения. Известно, что сообщения мужчин проблемно ориентированы, кратки и содержательны одновременно. Женщины более общительны, их сообщения выразительны и эмоциональны [1].

Задача определения пола автора текста решалась многими зарубежными исследователями. Так в работе [2] для турецкого языка была получена точность правильного принятия решения о поле автора короткого сообщения из различных источников в сети Интернет близкая к 0,9 при использовании линейного дискриминантного анализа, а также лексических, морфологических, синтаксических характеристик текста. В работе [3] на корпусе англоязычных электронных писем с помощью метода опорных векторов и деревьев решений была достигнута точность идентификации пола 0,82 по функциональным словам и характеристикам уровня символов. Авторам работы [4] на примере британских эссе удалось достичь точности 0,81 путем анализа частоты встречаемости тетраграмм слов. Стоит отметить, что подобных исследований для русского языка не проводилось.

Проблема определения автора и пола автора короткого электронного сообщения имеет следующие важные отличия от других классических задач атрибуции текста, решавшихся авторами ранее [5]:

1. Небольшая длина сообщений (в среднем порядка 50–100 символов) по сравнению с другими типами текстов. Однако, как правило, существует возможность собрать большое количество сообщений для исследований.
2. Стиль сообщений одного автора от сообщения к сообщению может сильно меняться в зависимости от адресата: от формального в служебной переписке до неформального в частной.
3. Возраст, образование, сфера деятельности различных авторов-мужчин и авторов-женщин существенно варьируются. Формировать корпуса текстов для исследований следует с учетом этих особенностей. Также необходимо вводить корректирующие коэффициенты в итоговую методику.
4. Авторы могут умышленно скрывать информацию о себе или дезинформировать собеседников, выдавая себя за человека другого пола и гендера. Поэтому использование явных гендерно-окрашенных

признаков, таких как, например, глаголов в прошедшем времени в соответствующем роде, местоимений и т. д., не всегда представляется возможным. Таким образом, итоговый вектор признаков текста, описывающий соответствующий пол, должен состоять из слабоконтролируемых человеком характеристик, чтобы методика определения пола оставалась актуальной и для описанных случаев.

5. В коротких электронных сообщениях появляются дополнительные лингвистические элементы, такие как эмотиконы («смайлики»). Эмотиконы служат для придания написанным словам дополнительной эмоциональной окраски или для того, чтобы выразить эмоции по отношению, например, к предыдущей фразе собеседника. К дополнительным признакам также можно отнести использование Bulletin Board Code (BB-коды) — разметки текста, позволяющей расставить акценты в тексте путем выделении отдельных слов и фраз жирным шрифтом, курсивом и т. д.

Основной задачей, которую необходимо решить при определении пола автора, является получение репрезентативной и хорошо разделимой выборки числовых данных из корпуса текстов.

Формально проблема представляется как задача бинарной классификации произвольного сообщения $t \in T$ к одному из классов множества $C = \{C_1, C_2\}$, где C_1 — мужчины, C_2 — женщины. Целью является построение классификатора, решающего данную задачу, т. е. нахождение некоторой целевой функции $F : T \times C \rightarrow [-1,1]$, определяющей пол автора произвольного сообщения из множества T. При этом каждое сообщение рассматривается как вектор признаков $X = \{X_1, ..., X_n\}$. Обучение классификатора производится на сообщениях, пол авторов которых достоверно известен, т. е. существует множество пар $(t_i, c_j) \in D \subseteq T \times C$, где $t_i \in T$, $c_j \in C$.

Классификатор достаточно обучить один раз и более не переобучать, как это приходится делать при решении задачи идентификации автора текста, где учитываются индивидуально-личностные характеристики каждого автора.

С целью сбора базы данных для исследований была разработана утилита, в автоматическом режиме собирающая комментарии пользователей с интернет-форумов, работающих на основе систем управления сайтом phpBB и Invision Power Board. С её помощью на интернет-форуме http://forum.tomsk.ru для экспериментов было собрано 41 780 сообщений 30 авторов (15 мужчин и 15 женщин). Сообщения авторов выбирались таким образом, чтобы охватить как можно больше тем обсуждений (знакомства, политика, автомобили и т. д.). Все представленные авторы имеют большой «стаж» общения (более двух лет), что подтверждается весомой серией сообщений у каждого из них. Текст сообщений был очищен от всех вспомогательных метаданных, так или иначе способных повлиять на результаты исследования. Однако в дальнейших исследованиях авторами планируется использовать часть данной информации в качестве дополнительных атрибутов при идентификации (в частности BB-коды). Информация о корпусе представлена в табл. 1 и на рис. 1. Стоит отметить, что все собранные сообщения были обращениями к некоторому собеседнику.

Таблица 1. Средняя длина сообщения в символах

Отправитель и получатель	Средняя длина сообщения, символов	Количество сообщений
Мужчин мужчинам	57,3	10 406
Мужчин женщинам	61,5	8 179
Женщин мужчинам	62,9	8 351
Женщин женщинам	65,3	14 844
Мужчины	59,4	18 585
Женщины	64,1	23 195

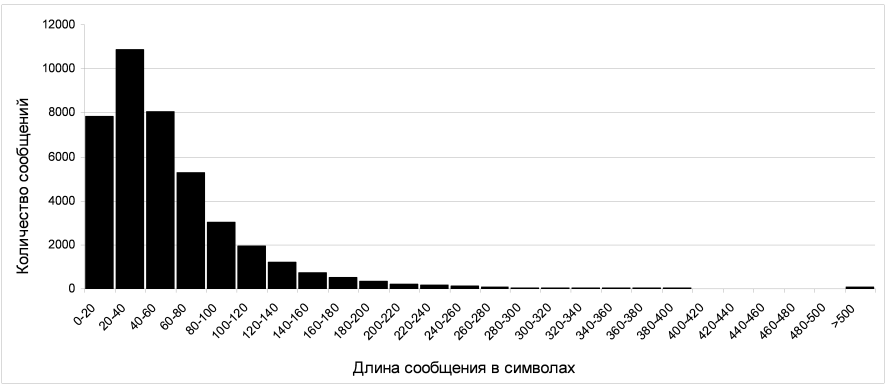


Рис. 1. Распределение длины сообщений в исследуемом корпусе

Из таблицы 1 видно, что средняя длина сообщения мужчин к мужчинам короче, чем сообщения мужчин к женщинам. Женщины пишут более длинные сообщения, чем мужчины независимо от пола получателя, с которым они общаются. Однако если получателем является также женщина, то такие сообщения, как правило, имеют наибольшую длину.

Как видно из графика на рисунке 1, большая часть собранных сообщений имеет длину не более 200 символов, поэтому тексты, состоящие из большего количества символов, не анализировались. Нижняя граница длины анализируемых сообщений была ограничена 20 символами. Этим условиям удовлетворяют 32 555 сообщений, т. е. около 78 % первоначального корпуса.

Для исследований использовалась программная система «Авторовед» [6], разработанная с целью статистического анализа текста на различных уровнях его организации и исследования характеристик текста задачах атрибуции текстов. Программная система успешно применялась в исследовательских целях для решения задачи идентификации автора литературных текстов и коротких сообщений [5, 7], а также для решения ряда частных практических задач. В частности для коротких текстов длиной до 100 символов удалось достичь точности 0,7 путем анализа частот наиболее частых слов русского языка, наиболее частых триграмм русского языка, униграмм символов, знаков препинания (одиночных и составных) и эмотиконов.

В качестве классификатора в настоящем исследовании используется машина опорных векторов (Support Vector Machine, SVM), математический аппарат которой был предложен В. Н. Вапником [8]. SVM может работать напрямую с векторным пространством высокой размерности без необходимости предварительного анализа и снижения количества измерений. Метод SVM изначально предназначен для классификации по двум возможным альтернативам, поэтому, как нельзя лучше, подходит для решения задачи определения пола автора. Для обучения моделей SVM применяется метод последовательной оптимизации (Sequential Minimal Optimization) [9], ядро выбрано линейное, параметр регуляризации C = 1, допустимый уровень ошибки — 0,00001.

Для выявления признаков текста, позволяющих определить пол автора, выполнялась следующая последовательность действий:

1. Из корпуса случайным образом извлекались тексты для обучения в количестве 5000 и тестирования в количестве 2500. Первая группа используется для обучения модели классификатора. Вторая — для проверки точности с помощью обученной модели.
2. Формирование вектора признаков для каждого из сообщений.
3. Приведение значений признаков в единый диапазон с помощью операции нормирования. Использовалось минимаксное нормирование в диапазон [-1..1].
4. Корректировка параметров классификатора, позволяющих обеспечить высокую разделяющую способность, путем обучения классификатора на нормированных векторах признаков группы обучающих текстов и проверки точности обученного классификатора на векторах признаков тестовой группы текстов.
5. Изменение перечня групп характеристик и/или признаков, составляющих группу, в случае, если изменением параметров классификатора достичь приемлемых результатов не удается.

Всего было исследовано порядка 50 различных признаков текста на уровне символов, слов и предложений, а также их сочетаний. Для каждой из характеристик было проведено по 20 описанных выше опытов. В качестве результирующей точности по данному признаку подсчитывалась средняя частота правильных классификаций. Результаты исследований представлены в табл. 2 (представлены характеристики, показавшие наилучший результат).

Таблица 2. Результаты экспериментов

Характеристика текста	**Точность**
униграммы символов	0,57
биграммы символов	0,51
триграммы символов	0,59
условные биграммы символов	0,53

Характеристика текста	Точность
служебные слова	0,58
распределение длин слов	0,57
знаки пунктуации и эмотиконы	0,68
словарный запас	0,52
ансамбль SVM	0,74

В результате исследований были определены характеристики короткого сообщения, имеющие преимущественное значение для использования в методиках определения пола автора. К ним можно отнести употребление человеком определенных сочетаний букв, служебных слов русского языка, знаков пунктуации, придание эмоциональной окраски высказыванию с помощью эмотиконов. Также в вопросе определения пола автора можно ориентироваться на длину слов в сообщениях.

Точность, превышающая 0,5, позволяет сделать вывод о принципиальной возможности определения пола автора короткого электронного сообщения на русском языке. Обучим модели SVM отдельно на каждой из групп признаков и объединим результаты классификации следующим образом — итоговым решением считается автор, выбранный большинством классификаторов. Использование такого ансамбля классификаторов позволило увеличить точность определения пола автора до 0,74. Это позволяет сделать вывод о целесообразности использования ансамблей классификаторов для принятия итогового решения и дальнейшего развития данной группы методов классификации в контексте решаемой задачи.

References

1. *Cheng N., Chen X.* et al. 2009. Gender Identification from E-mails. Proceedings of IEEE Symposium on Computational Intelligence and Data Mining : 154–158.
2. *Doyle J., Keselj V.* 2005. Automatic Categorization of Author Gender via N-Gram AnalySis. Proceedings of The 6th Symposium on Natural Language Processing, SNLP'2005, available at: http://web.cs.dal.ca/~vlado/papers/SNLP05J.pdf.
3. *Köse C., Özyurt Ö., Amanmyradov G.* 2007. Mining Chat Conversations for Sex Identification. Emerging Technologies in Knowledge Discovery and Data Mining Lecture Notes in Computer Science, 4819 : 45–55.
4. *Langer J., Jones V., McNabb M.* Gender Differences in Text Message Content, available at: http://www.jennalanger.com/academic/Langer-Jenna-Gender_dif_SMS_Content.pdf.
5. *Platt J. C.* 1999. Fast Training Support Vector Machines using Sequential Minimal Optimization : 185–208.
6. *Romanov A. S., Meshcheriakov R. V.* 2009. Text Author Identification by Support Vectors Device [Identifikatsiia Avtora Teksta s pomoshch'iu Apparata Opornykh Vektorov]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy

Mezhdunarodnoi Konferentsii “Dialog 2009” (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference “Dialog 2009”), 8 (15) : 432–437.

7. *Romanov A. S.* 2009. Programming System for Written Text Author Identification “Avtoroved” [Programmnaia Sistema dla Identifikatsii Avtora Pis’mennoi Rechi “Avtoroved”]. Khroniki Ob”edinennogo Fonda Elektronnykh Resursov “Nauka I Obrazovanie”, 7 : 7.
8. *Romanov A. S., Meshcheriakov R. V.* 2010. Short Message Author Identification with the Methods of Machine Learning [Identifikatsiia Avtorstva Korotkikh Tekstov Metodami Mashinnogo Obucheniia]. Komp’iuternaia Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii “Dialog 2010” (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference “Dialog 2010”), 9 (16) : 407–413.
9. *Vapnik V.* 1998. Statistical Learning Theory.
10. *Platt J. C.* 1999. Fast Training Support Vector Machines using Sequential Minimal Optimization : 185–208.

КОРПУСНОЕ ИССЛЕДОВАНИЕ ВАРИАНТОВ РОДОВОЙ ПРИНАДЛЕЖНОСТИ ИМЕН СУЩЕСТВИТЕЛЬНЫХ В РУССКОМ ЯЗЫКЕ[1]

С. О. Савчук (savsvetlana@mail.ru)

Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Излагаются результаты корпусного исследования одного из участков вариативности морфологической системы русского языка. Данные, полученные на материале Национального корпуса русского языка, сравниваются с результатами, полученными на основании других источников, отмечаются изменения в тенденциях развития вариантности рассматриваемой группы существительных.

Ключевые слова: имя существительное, род, родовая принадлежность, вариативность.

A CORPUS-BASED STUDY OF MORPHOLOGICAL VARIABILITY: VARIATION IN GENDER FORMS OF RUSSIAN NOUNS

S. O. Savchuk (savsvetlana@mail.ru)

Russian Language Institute, Russian Academy of Sciences, Mocow, Russian Federation

The paper presents the results of a corpus-based study on gender variation in Russian nouns. The list of variants was composed by analyzing textbooks and dictionaries compiled at the beginning and the 2nd half of the 20th century. The total number of gender variants. including outdated and

[1] Работа выполнена при поддержке: Программы ОИФН РАН «Текст во взаимодействии с социокультурной средой: уровни историко-литературной и лингвистической интерпретации»; проекта «Русский язык XVIII в.: корпусные исследования лексической и морфологической вариативности и словаря» в рамках Программы фундаментальных исследований Президиума РАН «Историко-культурное наследие и духовные ценности России»; РФФИ (грант 08-06-00371-а).

substandard ones, exceeds 600. The variants are classified according to their morphological and semantic features. The next stage of the research is focused on gender variants within the group of indeclinable nouns. The usage of every lexeme from the list was analysed in the texts of the Russian National Corpus, all gender variants was registered in the database and the correlation between variants was determined. The comparison of corpus data with the data derived from dictionaries made it possible to find out the changes in correlation between variants within the studied period and to formulate some trends in variants functioning.

Key words: noun, gender, gender form, variability.

Введение

Изучение вариантов родовой принадлежности имен существительных проводится в рамках большой работы по исследованию нестабильных точек морфологической системы русского языка в синхроническом и диахроническом аспектах на основе Национального корпуса русского языка[2]. Национальный корпус дает пользователю срез современного употребления русского языка, поэтому его в какой-то степени можно рассматривать как результат массового обследования, полученный, однако, не путем анкетирования, как это было в 1960-е годы[3], а путем целенаправленного отбора текстов[4]. Обширный корпусной материал дает возможность не только зафиксировать наличие вариантов, но и оценить их соотношение в статике и динамике, установить их функциональное распределение и зависимость от социолингвистических факторов.

Категория рода — одна из центральных в русском языке, она охватывает именные части речи — имена существительные, прилагательные, часть числительных и местоимений и часть форм глагола (формы причастий, прошедшего времени и условного наклонения). Для существительных эта категория является словоклассифицирующей и тесно связана с другими именными категориями (одушевленностью, числом, падежной системой), для остальных из перечисленных частей речи — словоизменительной. На *синтаксическом уровне* она имеет универсальное выражение, которое проявляется в том, что каждое

---

[2] Перспективы использования НКРЯ в изучении грамматических норм обсуждались в работах Гришина, Савчук 2007; Савчук, Гришина 2008; Савчук 2009; Савчук 2010; Корпусные исследования по русской грамматике 2009.

[3] См.: Русский язык и советское общество: Морфология и синтаксис современного русского литературного языка. М.: Наука, 1968; Русский язык по данным массового обследования. М.: Наука, 1974

[4] Как отмечала Л. К. Граудина, «метод пассивного наблюдения за устной и письменной речью можно считать идеальным с точки зрения безыскусственности условий собирания материала. Однако он не всегда может обеспечить репрезентативность выборки для вариантов с низкой частотой употребления» (Граудина 1980: 77).

существительное в системе языка принадлежит к определенному классу (который традиционно называется родом) и требует соответствующей формы рода зависимого слова при атрибутивной и предикативной связи. На основании объединения двух критериев — распределения существительных по родам и по признаку одушевленности/ неодушевленности (Зализняк, 1964, Зализняк 1967, Грамм) — в русском языке выделяется 7 согласовательных классов: мужской род (*дом*), мужской род одушевленный (*врач*), женский (*стена*), женский одушевленный (*коза*), средний (*окно*), средний одушевленный (*чудовище*) и так называемый парный род, объединивший существительные класса pluralia tantum (*брюки*, *сани*).

Второй аспект формального выражения категории рода — *морфологический* — не имеет универсального характера и по-разному проявляется в разных группах существительных. На этом уровне род тесно связан с распределением существительных по словоизменительным парадигмам склонения. При этом в разных парадигмах родовые противопоставления имеют различное морфологическое выражение. У существительных типа *стена — стол — окно* (традиционные 1-е и 2-е склонения) родовые различия обнаруживаются в форме именительного и винительного падежей, показатели родовой принадлежности — окончания –а, ∅, -о. В косвенных падежах существительные мужского и среднего рода падежными окончаниями не различаются и противопоставлены существительным женского рода. У существительных типа *дверь* (традиционное 3-е склонение) — *зверь* (2-е мягкое склонение) в качестве различителей рода выступают окончания косвенных падежей, в именительном падеже падежные окончания совпадают. У существительных несклоняемых (0 склонения) *кафе*, *пенальти*, *салями* нет никаких морфологических показателей рода, так что родовая принадлежность выражается только синтаксически: *нов**ое** кафе*, *точн**ый** пенальти*, *вкусн**ая** салями* [5].

Третий уровень родовых противопоставлений — *лексико-семантический* — по-разному представлен у одушевленных и неодушевленных существительных. Для одушевленных существительных отнесение к определенному роду мотивировано семантически и связано с полом живого существа или персонажа ирреального мира (сказочного, мифологического, литературного героя и под.). Небольшая часть одушевленных существительных относится к среднему роду (*животное, млекопитающее, насекомое, чудовище, дитя, лицо ‘личность’, существо, божество*). Для неодушевленных существительных отнесенность к мужскому, женскому, среднему роду не мотивирована и условна. Распределение существительных по родам

[5] Это обстоятельство сильно сокращает количество доказательных контекстов при корпусном исследовании, поскольку даже для вполне частотных слов число нужных контекстов, неопровержимо свидетельствующих о родовой принадлежности существительного, может быть невелико (например, для существительного *кантри* на 138 употреблений приходится 3 контекста, выявляющих значение среднего рода, 1 контекст, диагностирующий значение мужского рода, и 5 контекстов, в которых мужской и средний род не различаются; словоформа *кюри* в газетном корпусе встретилась 74 раза, как правило, в сочетании с количественными числительными, при этом ни в одном из контекстов не проявилось значение рода этого существительного).

неравномерно: по данным словарей, существительные мужского рода составляют около 46%, женского рода 41%, среднего рода — 13% (Мучник 1971)[6].

Варианты родовой принадлежности имен существительных

Отсутствие строгой зависимости между значением слова и его морфологическим оформлением создает почву для колебаний в определении родовой принадлежности у части существительных и возникновения вариантов рода. Вариативность характерна для современного состояния русского литературного языка, который унаследовал ее из языка XVIII и XIX вв.[7] В отечественной лингвистике эта тема рассматривалась как в синхронических описаниях (Горбачевич 1978; Зализняк 1967; Мучник 1971; Маринова 2008; РГ 1980; Шанская 1963), так и в историческом (Булаховский 1953, 1954; Марков 1992, Демьянов 2001), и в нормативно-стилистическом аспекте (см. словари и справочники ГПРР, Грамм 1977/2003, Еськова 1994, Розенталь, СТ, ТС, Чернышев 1915 др.). Основные точки вариативности, связанные с родом, хорошо известны и остаются в основном неизменными, но изменяется состав и соотношение вариантов в каждом типе.

1. Колебания в роде, выраженные морфологически

1.1. Сущ. м. р. на твердый согласный и ж. р. на -а: *рельс — рельса*

1.2. Сущ. на мягкий согласный, *ж, ш*: м. р.//ж. р.: *толь, толя — толь, толи*

1.3. Сущ. с суффиксами субъективной оценки: м. р.//ср. р., м. р.//ж. р., ср. р.//ж. р.: *этот домишко — это домишко, огромный домина — огромная домина, маленькое ведерко — маленькая ведерка*.

1.4. Сущ. общего рода: *этот чудила — это чудило*.

2. Колебания в роде, выраженные синтаксически

2.1. Сущ. несклоняемые: *боа* пушистый — *боа* пушистое.

2.2. Аббревиатуры: наша *ЖЭК* — наш *ЖЭК, СОЭ* повысился — *СОЭ* повысилось

2.3. Композиты: *часы-будильник* встали — *часы-будильник* встал.[8]

---

[6] От общего количества существительных, зафиксированного в словарях, которое по данным этой работы составляет 33 952 слова. Близкое соотношение приводится в Зализняк, 1967: существительных мужского морфологического рода 40,5%, женского — 43%, среднего — 16,5% (обследовано 47 700 существительных).

[7] Варианты существовали и в древнерусском языке, особенно широко вариативность была представлена у отглагольных существительных: *перевес* и *перевеса*, *огорода* и *огород*, *обмен* и *обмена*, *оград* и *ограда*, *отрад* и *отрада*, *примес* и *примеса*, *присяг* и *присяга*, *укор* и *укора* (Марков, 1992, 11).

[8] В ГПРР в разделе «Род и смежные с ним явления» рассматривается также употребление вариантов, связанных с существительными — наименованиями лиц: сущ. мужского рода, называющие лиц женского пола (*директор пришел — директор пришла*); субстантивированные прилагательные и причастия, называющие лиц женского пола (*заведующий Иванова — заведующая Иванова*); женские соответствия к мужским наименованиям лиц (*она чемпионка — она чемпион*).

Группы 1.1–1.4 включают существительные, в которых значение рода выражается не только синтаксически, но и морфологически, и варианты представляют собой разные лексемы, отличающиеся окончаниями и принадлежащие к разным парадигмам склонения. Группы 2.1–2.3 включают слова, в которых родовые различия в форме лексем никак не выражены, а проявляются только синтаксически — в формах согласующихся слов.

Исследование на материале Национального корпуса русского языка предпринято с целью фронтального обследования данного участка морфологической системы. *Лингвистический аспект* предполагает установление состава лексем, испытывающих колебания в родовой принадлежности, исследование соотношения вариантов рода, их статистические, семантические, функционально-стилистические характеристики, динамику их употребления по различным периодам развития языка. Сравнение данных, полученных на материале Национального корпуса русского языка, с результатами, полученными на другом материале, позволит отметить изменения в тенденциях развития вариантности рассматриваемой группы существительных. *Нормативный аспект* состоит в изучении рекомендаций грамматических справочников и словарей, отражающих основной фонд вариантов и дающих им оценку с точки зрения действующих норм литературного языка. Сравнение этих рекомендаций с реальным употреблением вариантов в текстах корпуса может дать интересные результаты для специалистов по культуре речи и лексикографов. *Прикладной аспект* заключается в том, что результаты исследования будут способствовать улучшению качества морфологической разметки Национального корпуса.

В настоящей статье будут рассмотрены родовые варианты существительных, относящиеся в основном к первым двум группам и к группе несклоняемых существительных.

1. Состав вариантов

Для установления состава вариантов с морфологическим выражением колебания в роде в современном русском языке были использованы: Грамм 2003, ГПРР, ТС. Для того чтобы исследовать соотношение вариантов в исторической перспективе, были привлечены словари и справочники, описывающие состояние норм языка XVIII–XIX и начала XX века: Долопчев 1909, Чернышев 1915, Еськова 2008, что позволило значительно расширить круг рассматриваемых вариантов. Совокупный словник, составленный по всем источникам, превысил 600 вариативных пар.

В Грамматическом словаре приводится 121 пара вариантов с морфологическим выражением колебания в роде и 66 пар родовых вариантов несклоняемых существительных. В словаре Граудиной (ГПРР) количество вариативных пар для склоняемых и несклоняемых существительных составляет соответственно 170 и 40 пар. Анализ словников показал, что при сопоставимом количественном составе их качественный состав не совпадает. Общая часть составляет 52 вариативные пары. Семантически они объединяются в несколько групп.

Конкретные предметы: *закут//закута, зал//зала и зало (оба — устар.), вольер//вольера, санаторий//санатория (устар.), пазанок//пазанка, просека и просек, проток//протока; вуаль, м (устар.)//вуаль, ж, рояль, м//рояль, ж*

(устар.), табель, м//табель, ж (устар.), псалтырь, м//псалтырь, ж; наргиле, м//с, жалюзи с//мн, бибабо м//мо (кукла), *тамагочи, со//мо* (игрушка).

Внутри группы конкретных предметов можно выделить тематические группы, представленные значительным количеством вариантов.

- **Растения:** *анемон и анемона (реже), георгин//георгина, маниок//маниока, чинар и чинара//чинара, мирта и мирт; тополь, м//тополь, ж (устар.); брокколи с//ж, дурро с 0 [//дурра] ж*.
- **Парные предметы**: *ботинок//ботинка (простореч.), туфель (устар.)// туфля, эполет//эполета; сабо, с//мн, галифе, мн//с.*
- **Употр. преимущественно во множ. числе**: *канделябр//канделябра (устар.), клавиш//клавиша, рельс//рельса, скирда//скирд, ставень//ставня, заусенец// заусеница, туберкул//туберкула, банкнот//банкнота;* названия денежных единиц*: евро м//с, эскудо, м//с, крузейро, м//с, песо, м//с, экю, м//с и др.*
- **Животные:** *лангуст//лангуста, мангуст (устар.)//мангуста, шпрот// шпрота, жираф//жирафа (устар. и простореч.), глист//глиста (разг.); выхухоль, м//выхухоль, ж, лебедь, м//лебедь, ж (нар.-поэт.); жако, кенгуру, колибри, динго, колли, гуанако (предок ламы), окапи, чау-чау (все имеют варианты мо//жо).*
- **Названия марок автомобилей:** *ауди с//м, вольво, с//м, феррари, м//с, шевроле, м//с.* В настоящее время эта группа значительно пополнилась новыми названиями марок автомобилей и других изделий (компьютеров, телефонов и бытовой техники, магазинов и др.).

Термины и абстрактные понятия: *абак//абака, аневризм//аневризма, апофеоза и апофеоз//апофеоза, арабеск//арабеска, катаракт (разг.)//катаракта, кремальер//кремальера, парафраз//парафраза, перифраз//перифраза, спазм//спазма, эпюр//эпюра; габбро, м//с* (горная порода), *статус-кво м//с, бери-бери, ж//с* (болезнь), названия единиц измерения: *га, м//с, генри, м//с, кюри, м//с*, названия, связанные с культурой и искусством: *па-де-де, м//с, па-де-труа, м//с, липси, м//с, сиртаки, м//с, сазандари, м//с, танка ж, За//с 0, граффито//граффити, с//мн* и др.

Вещества: *персоль м//персоль ж, шампунь м//шампунь ж;* среди несклоняемых существительных значительную группу составляют названия напитков и блюд национальных кухонь: *кофе, м//с, мокко, м//с, виски, м//с, шерри, м//с, бренди, м//с, мартини, м//с, боржоми с 0//боржом, м , мацони, с//ж, чили, м//с, ткемали, с//ж, сулугуни, м//с, спагетти, с//мн, хинкали, с//мн.*

Анализ не совпадающих частей словников Грамм и ГПРР показал, что в ГПРР количество вариативных пар увеличено за счет привлечения стилистически отмеченных вариантов — разговорных: *помидор* и *помидора* (разг.), *сандалета* и *сандалет* (реже и разг.); просторечных: *абрикос* и *абрикоса* (прост.), *застенок* и *застенка* (прост. и устар.); устаревших: *координата* и *координат* (устар.), *жилет* и *жилета* (устар.) и пр., в то время как в Грамм отражены варианты, находящиеся в пределах кодифицированного литературного языка.

2. Распределение вариантов по типам склонения

Существует ли формальная и содержательная предрасположенность русских существительных к образованию вариативных пар? Распределение

материала Грамм в соответствии с типами склонения позволило выявить три наиболее активные зоны вариативности, в которых сосредоточена основная масса вариантов:

1 тв м//1 тв ж[9] — 40 пар (*рельс//рельса, анемон//анемона, банкнот//банкнота*), **3 кгх м//3 кгх ж** — 25 пар (*арабеск//арабеска, присосок//присоска, проток//протока*). Всего 37%

2 мяг м//8 ж — 15 пар (*картель, м//картель, ж, шампунь, м//шампунь, ж, перкаль, м//перкаль, ж, персоль, м//персоль, ж*), **1 тв м//8 ж** — 7 пар (*занавес, м//занавесь, ж* (устар.), *подрез, м//подрезь, ж*). Всего 13%.

0 (несклоняемые существительные): м//с (*кофе, виски, эскудо, евро, сиртаки*), с//ж (*ткемали, мацони, брокколи*), со//мо (*тамагочи*) и др. Всего 32%.

Остальные 18% вариативных пар рассредоточены по малочисленным группам: **2 мяг м//2 мяг ж** — 2 пары (*ставень, м//ставня, ж, туфля, ж//туфель, м*), **5 ц** — 4 пары (*заусенец//заусеница, шпорец//шпорца*), **1 мо//жо** — 5 пар (*лангуст//лангуста, шпрот//шпрота*), **1 тв ж//с** — 7 пар (*дурра//дурро, кайла//кайло, кодла//кодло*), **1 тв м//с** — 1 пара (*мыт//мыто*) и др.

Сравнение с данными старых словарей показывает, что некоторые из этих малочисленных групп были значительно более представительными в начале XX века, что отчасти объясняется тем, что пособия учитывали в составе пар устаревшие или областные варианты. Так, представленная в Грамм единственной парой вариантов группа **1 тв м//с** (**мыт//мыто**) в словаре Чернышева включает 14 пар. К середине XX в. эти существительные утратили вариативность, при этом в одних случаях закрепился вариант среднего рода *индиго//индиг* (редко), *контральто//контральто*, *крылец* (стар. и нар.) / *крыльцо*, *начал//начало*, *облако* и *облак* (стар., нар.), *яблоко* и *яблок*, *ярем* и *ярмо*, в других — мужского рода *войлоко* (нар.)*//войлок*, *перло* (стар.)*//перл*, *плес//плесо//плеса* (юж.), *стул//стуло* (юж.).

Группа **1 тв ж//с (титла//титло),** содержащая, согласно Грамм, 7 пар вариантов, в словаре Чернышева также выглядит гораздо более многочисленной**:** *бедра* (стар.)*//бедро*, *берёста//берёсто* (стар.), *богословие//богословия* (стар.), *ботвинья//ботвинье* (нар., в лит. — редко), *ества//ество, яства//яство, коленка//коленко* (редко), *кросна//кросно*.

Среди слов, относящихся к активным зонам вариативности, которые не включены в современные пособия, но входили в справочники начала века, — 55 пар с колебанием **м2//ж8, м1//ж8** и **м4//ж8**.

м2//ж8 (33)**:** *антресоль м//антресоль ж, бутыль, ж//бутыль, м* (неправ.), *виолончель, м* (стар.)*//виолончель, ж, госпиталь, м//госпиталь, ж* (стар), *капель, ж//капель, м* (очень редко), *карусель* (стар.)*//карусель, ж, ковыль, м//ковыль, ж* (стар.), *контроль, м//контроль, ж* (редко), *миндаль, ж//*совр. *миндаль, м*, *модель, м//*совр. *модель, ж*, *опухоль м//опухоль ж* (прав.), *параллель, м* (редко)*//параллель, ж, профиль, м//профиль, ж* (стар.), *сераль, ж//*совр. *сераль, м*, *шаль, м//*совр. *шаль, ж* и др.

[9] Типы склонения приводятся в соответствии с Грамм.

м1//ж8 (14): *диагонал//диагональ*, *одеколон//одеколонь* (стар.), *подклет// подклеть*, *накип / накипь* (прав.), *подпис / подпись, полын / полынь, приме'с// при'месь, проруб* (юж.)*//прорубь*, *протор, протора* (прото'ры и убытки)*//проторь* (про'тори и убытки) и т. д.

м4//ж8 (8): *бреш, м* (редко)*//брешь, ж*, *бреша* (стар.), *брошь//брош* (непр.), *душ, м//душь, ж*, *поташь, ж / поташ, м, светочь//*совр. *светоч*, *харчь, ж//*совр. *харч, м, ераралаш* и *ералашь*, *спич, м//спичь, ж* (редко).

Данные, полученные в результате изучения словарей и пособий начала XX в., имеют большую прикладную ценность для усовершенствования аннотации и поиска в диахронической части НКРЯ. Поскольку тексты XVIII и XIX в. размечены морфологическим анализатором, имеющим в основе грамматический словарь современного русского языка, при грамматическом анализе вариативных форм не учтены вышедшие из употребления варианты, вследствие чего повышается количество ошибочных и лишних гипотетических разборов. Так, например, вариант *диагонал*, представлен в корпусе четырьмя разными словоформами: *диагонал* имеет 8 вхождений, *диагоналы* — 3, *диагоналом* — 2, *диагонала* — 1. *Как было доказано, что призма, которой основание параллелограмм, разделяется на две трехсторонние одинаковые плоскостию, проходящею чрез* ***диагоналы*** *оснований, и как стороны параллелограмма и* ***диагонал*** *могут быть взяты совершенно произвольно, то отсюда следует, что всякая трехсторонняя призма равна по величине с другой, которой основание и высота те же*. [Н. И. Лобачевский. Геометрия (1823)]. *…недалеко от мельницы впадает Бокла в Насягай, который* ***диагоналом*** *с северо-востока торопливо катит свои сильные и быстрые воды прямо на юго-запад.* [С. Т. Аксаков. Семейная хроника (1856)]. Во всех 14 случаях анализатор строит гипотетический разбор этих форм, предлагая рассматривать их как формы имени собственного: *диагонал* — S, persn, anim, m, sg, nom, bastard. Еще хуже обстоит дело с попыткой программы-анализатора опознать формы варианта *карусель, м*. *По написании сих полезных строк Лавид прочитывает сие всему собранию и повелевает, чтобы во всей строгости наблюдать его повеления, с досадою удаляется из собрания, подтверждая однако, чтобы* ***карусель*** *непременно в скором времени был готов.* [Н. И. Новиков. Пословицы российские (1782)]. *Июля 1-го двор в город возвратится для смотрения* ***каруселя****, который будет 2 числа, а потом все приедут опять в Петергоф и, как слышно, пробудут долго….* [Д. И. Фонвизин. К родным (1763–1774)]. Для формы *карусель* анализатор, опираясь на имеющуюся в словаре лемму женского рода, предлагает два ошибочных разбора: S, inan, f, sg, acc, norm / S, inan, f, sg, acc, norm. Для формы *каруселя*, отсутствующей в словаре, предлагается 9 гипотез, из которых лишь одна оказывается верной: S, inan, m, sg, gen, bastard. Внесение в словарь морфологического анализатора сведений о вариантах помогло бы избежать ошибочных разборов и сократить количество избыточных гипотез.

3. Вариативность в текстах

Проведенный анализ словарей свидетельствует о количественном сокращении вариантов рода, что дает исследователям основание говорить о том,

что зона родовой вариативности изживает себя[10]. Имеются ли основания для такого заключения? В поисках ответа на этот вопрос переходим к изучению употребления вариантов в текстах. Материалом исследования служит прежде всего Национальный корпус русского языка — основной корпус письменных текстов, и Большой корпус СМИ 2000-х годов, или газетный корпус. Основной корпус включает как ранние тексты, относящиеся к XVIII, XIX и 1-ой пол. XX в. (около 76 млн словоупотреблений), так и современные тексты, относящиеся ко 2-й пол. XX в. (более 100 млн). Газетный корпус содержит тексты печатных и электронных СМИ начала XXI в. (более 100 млн). В отдельных случаях, при необходимости, привлекаются другие электронные ресурсы.

Пилотное обследование функционирования активных вариативных пар, представленных в современных справочниках, на материале Национального корпуса показало, что ситуация с соотношением вариантов не столь однозначна. В целом количество вариантов с колебанием в роде в современном русском языке сократилось (либо один из вариантов ушел в пассив, либо все слово целиком устарело), однако говорить об исчезновении варьирования по роду рано. Полное исчезновение варианта рода наблюдается у слова *брелок* (в современных текстах не зафиксировано ни одного варианта *брелока, ж*). Близка к нему пара *ботинок//ботинка*: 3 современных контекста являются цитатами и отражают речь 1-й половины XX века[11]. Явно прослеживается сокращение варианта женского рода в паре *апофеоз//апофеоза* (если в ранних текстах соотношение вариантов *м. р./ж. р.* было 115/31, то в текстах 2-й пол. XX в. — 253/3. Может меняться соотношение вариантов в паре: в паре *аневризм//аневризма* в современных текстах преобладает форма женского рода (3м//13ж), в то время как в старых текстах было наоборот (20м/5ж). Интересно, что конкуренция вариантов у существительных даже одной семантической группы может иметь разный исход: в паре *бутс//бутса* (соотношение форм 5м//12ж при 71 форме мн. ч.) явно побеждает вариант женского рода, а в паре *кед//кеда* закрепляется вариант мужского рода (соотношение 14м//3ж при 202 формах мн. ч.). Кроме того, корпусной материал позволяет обнаружить появление новых вариантов, не зафиксированных в словарях (*корректива* ж, *бездарь* м)[12].

Почвой для возникновения колебаний в роде служит отсутствие у говорящих автоматизма в образовании нужной формы, необходимость выбирать

---

[10] Мучник 1971, 192; Горбачевич 1978, 141. Е. В. Маркина отмечает затухание вариативности типа м2//ж8: среди обследованных автором новых заимствований родовое варьирование зафиксировано только у одного слова — гель, м//гель, ж, остальные слова с основой на мягкий согласный сразу оформились как существительные мужского рода (Маркина 2008, 131).

[11] Анализ контекстов дает наглядное представление о том, как происходило вытеснение варианта женского рода, в данном случае, по-видимому, под влиянием причин нелингвистического порядка: выстроив контексты по дате написания текста, можно видеть, как изящная женская ботинка вытеснялась детским, спортивным или в тяжелым солдатским ботинком.

[12] Рассмотрены на материале НКРЯ в Савчук, Гришина 2008; Гришина, Савчук 2007.

их двух или нескольких возможностей, применяя то или иное правило или действуя по аналогии. И наоборот, закреплению какого-либо одного родового варианта способствует употребление слова в конструкциях, выявляющих родовую принадлежность слова, причем скорость распространения варианта напрямую зависит от частотности конструкций. Исходя из этого можно предположить, что наиболее предрасположены к появлению и длительному существованию вариантов несколько групп существительных. Во-первых, имена существительные, которые преимущественно употребляются во множественном числе. Поскольку в современном русском языке во множественном числе сняты все родовые противопоставления, то при образовании значительно менее употребительных форм единственного числа говорящий оказывается перед необходимостью всякий раз заново конструировать эти формы, выбирая из двух (как в случае *кед, м//кеда, ж*) или (реже) трех возможностей. Во-вторых, благоприятная среда для сохранения вариативности — термины. Здесь причина кроется в малой проницаемости сферы общелитературного языка и специальных языков, и если слово закрепилось в терминосистеме и в общем употреблении в разных родовых формах, то параллельное существование двух вариантов может сохраняться долго. В третьих, наиболее вероятно ожидать появления вариантов среди новых заимствований. Вариативность разных типов естественно связана с процессом освоения иноязычного слова русским языком[13], и это наиболее динамичная среда обитания вариантов. Если новое заимствование обладает высокой употребительностью, оно быстро усваивается говорящими, встраивается в лексическую и грамматическую систему русского языка и избавляется от вариантов. Такова судьба большинства вариативных пар, представленных в словарях начала XX в. и исчезнувших в течение десятилетий. Однако поскольку приток заимствований не иссякает, то появление новых вариантов предопределено[14].

Среди иноязычных заимствований выделяется группа несклоняемых существительных, которая представляет особый интерес именно в отношении вариантов родовой принадлежности[15]. Несклоняемые неологизмы, встраиваясь в систему родовых противопоставлений русского языка, испытывают влияние нескольких факторов, которые могут действовать разнонаправленно: например, общее нормативное правило требует отнесения неодушевленного существительного к среднему роду, но одновременно

---

[13] Проблема освоения иноязычных слов в русском языке достаточно хорошо изучена (см., например, библиографию в Маркина 2008), в самой монография Е. В. Маркиной на обширном материале рассматривается функционирование иноязычной лексики в русской речи на рубеже XX и XXI в.

[14] Справедливости ради следует отметить, что в новейшее время происходит пополнение лексического фонда в основном за счет заимствований из английского языка, которые пополняют класс субстантивов мужского рода и не порождают вариантов (Маркина 2008, 130).

[15] Об истории формирования этой группы существительных см. Мучник, 1974, Гловинская 2008.

признается влияние семантических связей, в частности рода соотносительного по значению слова. Так появляются варианты *торнадо* с//м (ураган), *цунами* с//ж (волна), *сиртаки* м//с (танец) и т.д. В XX в. перестал действовать такой важный для языка XIX века фактор, как влияние рода слова в языке-источнике, определявший родовую принадлежность многих заимствованных слов в русском языке и вызывавший появление вариантов. Зато усилилось действие фактора аналогии, формальной и семантической, проявляющееся в том, что новое слово встраивается в систему языка по образцу, в качестве которого выступает русское или ранее заимствованное слово.

Варианты родовой принадлежности в группе несклоняемых имен существительных

Исследование вариативности в группе несклоняемых существительных проводилось следующим образом. Во-первых, был поставлен вопрос, как ведут себя в текстах слова, у которых словари отмечают наличие вариантов рода, сохраняется ли колебание в родовой соотнесенности на протяжении изучаемого периода, происходят ли изменения в соотношении вариантов. Во-вторых, анализировались слова с устойчивой родовой соотнесенностью, для которых словари не отмечают вариантов. В-третьих, включались в рассмотрение новые заимствования, еще не зафиксированные в Грамм. Приведем некоторые результаты.

Интересные результаты дает исследование названий марок машин. В отличие от марок других изделий, названия которых значительно реже употребляются в отрыве от названия самих изделий (например, *купил Тошибу* — ноутбук, телевизор, проектор?), названия автомобилей уже давно функционируют как самостоятельные слова и даже включаются в словари. Словари по-разному описывают их грамматические характеристики: Орф всем названиям марок приписывает помету мужского рода, Грамм допускает здесь вариативность с//м (автомобиль). Однако в реальном употреблении название марки может ассоциироваться и со словом *машина, марка, иномарка, модель* — так появляется третий вариант женского рода.

Она уже не видела ***белой*** *«****ауди****» впереди*. [Дина Рубина. Несколько торопливых слов любви (2001)//«Новый Мир», 2003]. *Приказ всем машинам: блокировать* ***красную*** *«****мазератти****»*. [Виктор Левашов. Заговор патриота (2000)]. Дин загнал свой ***любимый*** *«****порше****» в угол гаража, исчез и вскоре приплыл на «корабле пустыни»: двести пятьдесят лошадиных сил, автоматическая трансмиссия, эр кондишн*. [Василий Аксенов. Круглые сутки нон-стоп//«Новый Мир», № 8, 1976]. *Он ездит туда на простом* ***недорогом*** *«****Шевроле****» — как, впрочем, и я*. [Игорь Свинаренко. Провинциальные куплеты (1997)//«Столица», 1997.10.28]. — *Скажи, Учитель, только честно: если бы у меня были деньги, мне бы это* ***голубое*** *«****Пежо****» без всякой очереди завернули?* [Анатолий Гладилин. Большой беговой день (1976–1981)].

Таблица 1

Название марки			НКРЯ				Газетный			
	Грамм	Орф	ж	м	м//с	с	ж	м	м//с	с
«Ауди»	с//м	м	45	7	5	0	162	15	8	1
«Вольво»	с//м	м	48	9	18	0	36	15	16	2
«Шевроле»	м//с	м	1	9	19	0	7	11	7	0
«Феррари»	м//с	м	3	7	6	0	8	15	8	0
«Рено»	с	м	1	10	14	0	0	12	4	2
«Пежо»	с	м	1	12	12	1	1	13	13	0
«Порше»/«Порш»		м	0	10	2	0	1	12	14	0
«Дэу»			6	0	0	0	15	3[16]	0	0
«Субару»			6	4	1	0	7	1	2	0
«Ламборгини»			2	1	0	0	3	3	0	0
«Мазератти»			5	4	0	0	1	2	3	0
«Мицубиси»		м	2	6	2	0	5	4	3	0
«Инфинити»			0	0	0	0	3	4	2	0

Данные, которые приводятся в таблице 1, показывают, что в современном узусе рекомендуемый словарями средний род для марки машин — редкость, большинство марок предпочитают либо форму мужского рода («шевроле», «рено», «пежо»[17], «порше[18]»), либо женского рода (бесспорные лидеры — марки «ауди» и «вольво»), либо допускают более-менее равноправные варианты.

В таблице наряду с заимствованиями, учтенными в словарях, представлены и лексемы, относительно которых словари и справочники не дают никаких рекомендаций, следовательно, их употребление должно подчиняться какому-то одному из общих правил: либо средний род для неодушевленного предмета, либо аналогия с родом соотносительного слова. Как видим, во всей группе побеждает второе правило. Однако выполнение его осложняется тем, что соотносительных слов не одно, а несколько (*автомобиль* или *машина*, *модель*, *марка*), причем разной родовой принадлежности, что приводит

[16] Два случая зафиксированы в одной статье одного автора [«Комсомольская правда», 2004.07.19].

[17] Единственное вхождение «рено» со значением женского рода зафиксировано в составе «рено-лагуна» и, вероятно, объясняется влиянием второго элемента композита. Ср. «Рено-Модус», «Рено-Символ», «Рено Винд» — все м. р. Это предположение подтверждается значением женского рода у таких сложений, как «Шевроле-Нива», «Субару-Импреза», «Дэу-Нексия» и мужского рода у «Шевроле Блейзер», «Субару-Аутбек» (по 2 вхождения), «Мицубиси-Кольт», «Мицубиси-Галант» и «Дэу-Матиз» Ср. однако, «Шевроле-Люмина» м, «Порше-Карера», м и «широкоплечий шевролет «Импала». Колебания в родовой соотнесенности слов-композитов — проблема, требующая самостоятельного изучения.

[18] Возможно, под влиянием распространенного варианта названия той же марки «порш».

к возникновению вариантов[19]. Можно предположить, что варьирование по роду сохранится в этой группе достаточно долго. Относительно причин, почему говорящие одни названия соотносят со значением женского рода, а другие со значением мужского, можно строить гипотезы. Возможно, у этого факта есть психологическое или какое-то иное объяснение[20].

Другую значительную по объему группу составляют названия напитков и блюд национальной кухни. Результаты обследования словарных лексем приведены в таблице 2.

Таблица 2

Лексема	Грамм	НКРЯ					Газетный				
		ж	м	с	м-с	мн	ж	м	с	м-с	мн
шерри	м//с		1		2						
виски	м//с	2	20	57	72						
бренди	м//с		8	6		3		2	1		
мартини	м//с		18	5	4			7		4	
шабли	м//с		1	5	3					2	
боржоми//	с	2	5	1	4		4	17	1	16	
боржом	м		9					23			
сулугуни	м//с		2		4			2		5	
чили	м//с		2					2	1	2	
брокколи	с//ж	6				1	9				1
ткемали	с//ж		2	1	1		1	2		1	
мокко	м//с		9		6			8		2	
мокка		2	1	1				1			
эспрессо			18		11						
капуччино			8		2						

[19] Симптоматично сосуществование разных вариантов в одном тексте: Если к этому прибавить «антитабак» комплексно, алкоголизм безвозвратно, то при чем здесь **белый** «Вольво»? [Елена и Валерий Гордеевы. Не все мы умрем (2002)]. Евгения могла рассказать многое: и про группу крови, и про резус-фактор, и кем работал, и про **белую** «Вольво», — но не могла. [Елена и Валерий Гордеевы. Не все мы умрем (2002)].

[20] В проезде стояли серый, довольно **старый** «**Опель**» и оранжевая **красотка** «**Ауди**». [Марианна Баконина. Девять граммов пластита (2000)]. Возможно влияние причин фонетического порядка: среди марок, склонных к женскому роду, — слова с последним безударным открытым слогом, который может ассоциироваться с безударным окончанием существительных 1-го склонения, в большинстве своем женского рода. Нельзя исключать и влияния разговорных русифицированных форм: — *Мы ж с тобой на нашей «Вольве» не слабо идем — сто двадцать в час, а они, суки, на своих легковых «мерсах», «беэмвухах» и «поршах» нас как стоячих делают!* [Владимир Кунин. Кыся (1998–2000)]. Марка *Дэу* имеет разговорный аналог *дэушка*, что часто обыгрывается в текстах: ***Дэушка** моей мечты. Тест-Драйв Daewoo Lacetti (Дэу Лацетти)* [http://www.bcetyt.ru/auto/another/18187065.html]. *Очаровательная «**дэушка**»*[АвтоМир, 20 сентября 2005]. На Неве появились «оборотни» и «дэушки» [Комсомольская правда, 2005.10.29]

Лексема	Грамм	НКРЯ					Газетный				
		ж	м	с	м-с	мн	ж	м	с	м-с	мн
мюсли	с//мн	1				8					16
спагетти	с//мн					5				2	22
хинкали	с//мн		3			5		2			4
тамагочи	со//мо	1	8			1					7

Как видим, вариативность м//с сохраняется у слов *виски*, *бренди* и *мартини*. *Виски приятно* ***обжигало*** *горло. Закуски были подобраны со вкусом.* [Андрей Ростовский. По законам волчьей стаи (2000)]. *Подошел официант и спросил, чего я хочу. Я сказал: «****Чистый*** *виски.* ***Двойной****».* [Сергей Юрский. Вспышки. Заключительная глава книги // «Октябрь», 2001]. У слова виски выявлено 2 употребления варианта женского рода, не отмеченного словарями: *Два раза леди Кембл подвигала О'Келли ликер — и два раза О'Келли подливал себе* ***шотландскую*** *виски.* [Е. И. Замятин. Островитяне (1917)]. *Завтра он поедет на Пересыпь — возле Балтского шляха, у него есть один знакомый, который даст такой первач, что лучше* ***всякой*** *виски.* [Аркадий Львов. Двор (1981)]. У слов *шерри* и *шабли* вариантность, по данным корпуса, сходит на нет: у *шерри* отмечен вариант мужского рода (ср. *ликер*), у *шабли* закрепляется вариант среднего рода (по влиянием соотносительного слова *вино*). Следует отметить, что все названия вин устойчиво употребляются со значением среднего рода: *кьянти*, *киндзмараули*, *напареули*, *цинандали* и др.

Для *боржоми, с* в Грамм фиксируется склоняемый вариант *боржом, м,* оба они представлены в корпусе. Однако, во-первых, склоняемый вариант в количественном соотношении явно превосходит несклоняемый, а во-вторых, несклоняемый вариант среднего рода оказывается наименее популярен, в узусе преобладают варианты мужского рода (соотносительно с *напиток* или под влиянием склоняемого варианта) и женского рода (*вода*), в чем опять можно усмотреть влияние соотносительного понятия. *Потом Володя прибежал за нами с дочкой, и мы пошли, ели шашлык, пили кавказское вино и* ***боржом*** *и думали, что мы на Кавказе.* [Нина Катерли. «Сквозь сумрак бытия» // «Звезда», 2002]. *Сколько в этом потоке поддельной продукции, сказать сложно, хотя подчас сами продавцы утверждают, что* ***настоящую*** *боржоми сегодня можно купить только в Грузии.* [Берестов Серафим. Сгинь, нечистая!//Труд-7, 2005.06.24]. *Этикетка воды сильно напоминала* ***запрещенный*** *«Боржоми», чего не скрывали и сами производители клонов.* [Олег Трутнев. «Русский Боржомъ» в законе // РБК Daily, 2007.05.23]. *…в лучшем случае это слегка почищенная водопроводная, в худшем — «боржоми»,* ***приготовленное*** *с помощью соды и поваренной соли грубого помола если не на Малой Арнаутской, то уж точно в гараже у дяди Васи в Мытищах.* [Наука чистой воды (1998) // «Профессионал», 1998.07.01].

Влияние семантического фактора (грамматических признаков соотносительного слова) прослеживается в судьбе вариантов *сулугуни, м* (сорт сыра), *чили, м* (перец и соус), *брокколи, ж* (капуста). Ни одного бесспорного варианта

среднего рода у этих слов в корпусном материале не обнаружено. То же можно сказать о *мокко*, *эспрессо*, *капуччино* (кофе) — вопреки формальному сходству с существительными среднего рода все имеют значение мужского рода[21]. Со словом *ткемали* (сорт алычи и соус из нее) сложнее — из-за разной родовой принадлежности соотносительных слов варианты м//ж сохраняются.

Для названий блюд *спагетти*, *хинкали*, *мюсли* важна не соотнесенность со словом *блюдо*, а то, что эти блюда представляет собой множество мелких предметов. Впрочем, каждое из них можно соотнести с конкретными названиями знакомых блюд — *макароны, пельмени, хлопья*. Семантика множественности поддерживается и формальной аналогией: конечное *-и* воспринимается как окончание множественного числа, что, по-видимому, объясняет отсутствие вариантов среднего рода и резкое преобладание вариантов множественного числа, в том числе и в склоняемой форме: ***Спагеттей*** *побольше, пожалуйста, — твердил свое Пашка. — Если в шкафу не имеется, у нас есть резерв*. [Сергей Каледин. Записки гробокопателя (1987–1999)]. *Скажем, другая из моих многочисленных подруг кормила мужа сплошными* ***мюслями****, сосисками и черным кофе (если он не забывал все это купить), а он остается при ней уже лет десять и никуда уходить не собирается*. [Марина Каминарская. Три веселых супа (2002) // «Домовой», 2002.01.04][22].

Еще один фактор поддержки вариативности можно назвать стилистическим. Он связан с закреплением разных вариантов в разных сферах общения, в литературном языке и субстандарте. В связи с этим следует упомянуть о слове *тамагочи*. Обозначая электронную игрушку (яйцо Тамагочи), очень популярную в 1990-е годы, которая воспринималась как живое существо, это слово получило в Грамм помету со//мо. Однако в реальном употреблении вариант среднего рода фактически не был востребован: в корпусе среди 90 вхождений слова встретилось 11 вариантов мужского рода, 2 — женского рода, а из 13 случаев множественного числа 3 склоняемые формы (конечное *-и* в *тамагочи* воспринимается как флексия). *Мы идем на кухню, я готовлю тамагочи ужин,* ***ее*** *любимый суп из шампиньонов, мы едим и ложимся спать*. [Елена и Валерий Гордеевы. Не все мы умрем (2002)] *…заверещал телефон, не его, незнакомо, настырно, как* ***изголодавшийся*** *тамагочи, — в этом лифте на стене оказался небольшой аппарат, и Виталик, резко вскочил…* [Марина Вишневецкая. Вот такой гобелен (1999)]. *В мое школьное время классная выкидывала «****Тамогоч****» в окно*. [Сегодня в топе блогов история учительницы (блог) (2008)]. *Конечно,*

---

[21] В текстах встретился также зафиксированный в словарях старый вариант мокка, обнаруживший колебание в отношении родовой принадлежности и склоняемости. *Сидеть в мягком кресле, читать последний номер газеты и отпивать небольшими глотками* ***душистый*** *мокка — ничего лучшего Виктор Николаич никогда не желал.* [Д. Н. Мамин-Сибиряк. Приваловские миллионы (1883). *У нее — мокка* ***аравийская****, а у нас — цикорий…* [В. В. Крестовский. Петербургские трущобы. Части 1–3 (1864)]. *Старушка нянька, жившая при его холостой квартире, принесла фарфоровый, дымящийся* ***моккой*** *кофейник.* [А. Н. Толстой. Хождение по мукам (1922)].

[22] О распространении склоняемых вариантов несклоняемых существительных в текстах Интернета см. Гловинская 2008.

*проще всего выдумать свой виртуальный мир и заполнить его человекообразными «**тамагочами**».* [Экология любви//Труд-7, 2002.10.03]. В блогах это слово чаще встречается в русифицированной форме в разных вариантах: *тамогоча, мо-жо, тамогоч, мо, тамагоч, м. Как меня достала эта тамогоча. Мой тамогоча не хочет спать, а я скоро ложиться планирую… Я хосю тамогоча…))) Я хочу ежика, тамогоч, голубые коньки с блестящими снежинками, киндер сюрприз с бегемотиком, лизуна и мир во всем мире)))))))))) Помню мой первый тамагоч я потеряла, мы тогда всей семьей его искали.* В живой разговорной речи, отраженной в текстах электронной коммуникации, слово *тамагочи* быстрее прошло полный цикл грамматической адаптации, чем в книжно-письменной речи, что привело к параллельному существованию несклоняемого и склоняемых вариантов в разных речевых сферах.

Таким образом, изучение на материале корпуса вариантов рода в группе несклоняемых существительных показало, что вариативность в этой группе существительных сохраняется, однако соотношение вариантов в ряде случаев может меняться. Одни лексемы утрачивает свою вариативность, другие, напротив, могут ее приобретать. Среди факторов, которые способствуют адаптации слова к грамматической системе русского языка и угасанию вариантов, прежде всего нужно отметить соответствие формального облика слова признакам того или иного грамматического класса. Например, существительные с основой на согласный неизбежно попадают в класс существительных мужского рода 2-го склонения, иногда после короткого периода «несклоняемости»: так *общение в Интернет* быстро сменилось *общением в интернете*, *коллекция бонсай* вытесняется *коллекцией бонсаев* и др. Влияние категориальной семантики проявляется в распределении названий одушевленных или неодушевленных предметов по классам мужского-женского или среднего рода. В середине XX в. этот фактор считался определяющим при нормативном описании грамматических признаков несклоняемых существительных. Однако не менее важным является и лексико-семантический фактор, который способствует ассоциативному сближению нового слова с уже существующим и «копированию» его грамматических характеристик, в частности, родовой принадлежности. Если все факторы действуют в одном направлении, освоение иноязычного слова протекает гладко. Конфликт факторов, как правило, вызывает появление вариантов, которые могут сосуществовать на протяжении длительного времени.

Если рассматривать всю группу существительных, испытывающих колебание в родовой принадлежности, то можно видеть, что, несмотря на общие свойства, внутри отдельных подгрупп и даже у отдельных лексем обнаруживается специфика взаимоотношения вариантов, что подтвердило корпусное исследование. Поэтому общие выводы относительно тенденций развития вариативности на данном участке морфологической системы можно будет сделать на основании наблюдений за поведением вариантов во всех выделяемых подгруппах.

References

1. *Bulakhovskii L. A.* 1953. Russian Literary Language Course [Kurs Russkogo Literaturnogo Iazyka].
2. *Bulakhovskii L. A.* 1954. Russian Literary Language of the 1st half of the XIX century [Russkii Literaturnyi Iazyk Pervoi Poloviny XIX Veka].
3. *Chernyshev V.* 1915. Correctness of Russian Speech: Experiment of Russian Stylistic Grammar [Pravil'nost' I Chistota Russkoi Rechi: Opyt Russkoi Stilisticheskoi Grammatiki].
4. *Corpus* Researches on Russian Grammar [Korpusnye Issledovaniia po Russkoi Grammatike]. 2009.
5. *Dem'ianov V. G.* 2001. Foreign Vocabulary in the History of Russian Language in the XI-XVII cc. The Problems of Morphological Adaptation [Inoiazychnaia Leksika v Istorii Russkogo Iazyka XI-XVII vekov. Problemy Morfologicheskoi Adaptatsii].
6. *Dolopchev V.* 1909. Experiment of Irregularities of Spoken Russian Dictionary [Opyt Slovaria Nepravil'nostei v Russkoi Razgovornoi Rechi].
7. *Es'kova N. A.* 2003. Brief Russian Difficulties Dictionary. Grammar Forms. Accent [Kratkii Slovar' Trudnostei Russkogo Iazyka. Grammaticheskie Formy. Udarenie].
8. *Es'kova N. A.* 2008. Russian Literary Norms of XVIII-XIX. Accent. Grammar Froms. Words Variants. Vocabulary. Explanatory Articles [Normy Russkogo Literaturnogo Iazyka XVIII-XIX. Udarenie. Grammaticheskie Formy. Varianty Slov. Slovar'. Poiasnitel'nye Stat'i].
9. *Glovinskaia M. Ia.* 2008. Active Processes in Grammar [Aktivnye Protsessy v Grammatike].
10. *Grammar* Correctness of Russian Speech [Grammaticheskaia Pravil'nost' Russkoi Rechi]. 2004.
11. *Gorbachevich K. S.* 2003. Modern Russian Difficulties Dictionary [Slovar' Trudnostei Sovremennogo Russkogo Iazyka].
12. *Graudina L. K., Itskovich V. A., Katlinskaia L. P.* 1971. Grammar Variants: Experiment of Frequency Dictionary [Grammaticheskie Varianty: Opyt Chastotnogo Slovaria].
13. *Graudina L. K.* 1980. Questions of Russian Language Normalization: Grammar and Variants [Voprosy Normalizatsii Russkogo Iazyka: Grammatika I Varianty].
14. *Grishina E. A., Savchuk S. O.* 2008. Russian National Corpus as an Instrument for Grammar Norms Variability Researches [Natsional'nyi Korpus Russkogo Iazyka kak Instrument dla Izucheniia Variativnosti Grammaticheskikh Norm]. Trudy Mezhdunarodnoi Konferentsii "Korpusnaia Lingvistika - 2008".
15. *Marinova E. V.* 2008. Foreign Words in the History of Russian Language in the XX-XXI cc.: Mastering and Functioning Problems [Inoiazychnye Slova v Russkoi Rechi kontsa XX-nachala XXI veka: Problemy Osvoeniia I Funktsionirovaniia].
16. *Markov V. M.* 1992. Russian Historic Grammar. Noun Declension [Istoricheskaia Grammatika Russkogo Iazyka. Imennoi Sklonenie].
17. *Muchnik I. P. 1971. Verb and Noun Grammar Categories in Modern Russian Literary Language [Grammaticheskie Kategorii Glagola I Imeni v Sovremennom Russkom Literaturnom Iazyke].*
18. *Modern* Russian: Active Processes in XX-XXI [Sovremennyi Russkii Iazyk: Aktivnye Protsessy na rubezhe XX-XXI]. 2008

19. *Rozental' D. E.* 2007. Practic Stylistics of Russian [Prakticheskaia Stilistika Russkogo Iazyka]. Russkii Iazyk: Spravochnik-Praktikum.
20. *Rozental' D. E., Telenkova M. A.* 2007. Dictionary of Difficulties of Russian [Slovar' Trudnostei Russkogo Iazyka].
21. *Russian* Grammar [Russkaia Grammatika]. 1980.
22. *Russian* Language and Soviet Society: Morphology and Syntaxis of Modern Russian Literary Language [Russkii Iazyk I Sovetskoe Obshchestvo: Morfologiia I Sintaksis Sovremennogo Russkogo Literaturnogo Iazyka]. 1968
23. *Russian* Language up to Mass Research Data [Russkii Iazyk po Dannym Massovogo Issledovaniia].1974
24. *Russian* Orthography Dictionary [Russkii Orfograficheskii Slovar']. 2005.
25. *Savchuk S. O., Grishina E. A.* 2008. Variability of Russian. Dictionary Project [Variativnost' v Russkom Iazyke. Proekt Slovaria]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2008" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2008"), 7 (14).
26. *Savchuk, S.* 2009. The Russian National Corpus as a Tool for the Research on Grammatical Variability. Proceedings of the Third International Conference Grammar & Corpora Mannheim.
27. *Savchuk S. O.* 2010. Experiment of Corpus Study of the Morphological Variability: Variants of Masculine Nouns Gen. Pl. [Opyt Korpusnogo Issledovaniia Morfologicheskoi Variativnosti: Varianty Roditel'nogo padezha Mnozhestvennogo chisla Sushchestvitel'nykh Muzhskogo roda]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2010" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2010"), 9 (16).
28. *Shanskaia T. V.* 1963. Variants of Noun Gender Forms in Modern Russian Literary Language [Varianty Rodovykh Form Imen Sushchestvitel'nykh v Sovremennom Russkom Literaturnom Iazyke].Vestnik MGU, 6.
29. *Skortsov L. I.* 2007. Large Explanatory Dictionary for Correct Spoken Russian [Bol'shoi Tolkovyi Slovar' Pravil'noi Russkoi Rechi].
30. *Sovremennyi* Russkii Iazyk: Aktivnye Protsessy na Rubezhe XX-XXI Vekov.
31. *Gorbachevich K. S.* 1978. Word Variability and Language Standard [Variativnost' Slova I Iazykovaia Norma].
32. *Valgina N. S.* 2001. Active Processes in the Modern Russian Language [Aktivnye Protsessy Sovremennom Russkom Iazyke].
33. *Word* Usage Difficulties and Variants of Norms of Russian Literary Language: Dictionary-Guide [Trudnosti Slovoupotrebleniia i Varianty Norm Russkogo Literaturnogo Iazyka: Slovar'-Spravochnik].1973
34. *Zalizniak A. A.* 1964. On the Question of Grammar Categories of Gender and Animacy in Modern Russian Language [K Voprosu o Grammaticheskikh Kategoriiakh Roda I Odushevlennosti v Sovremennom Russkom Iazyke]. Voprosy Iazykoznaniia.
35. *Zalizniak A. A.* 1967. Russian Noun Inflexion [Russkoe Imennoe Slovoizmenenie].
36. *Zalizniak A. A. 2003. Grammar Dictionary of Russian Language [Grammaticheskii Slovar' Russkogo Iazyka].*

ИДЕНТИФИКАЦИЯ ОБЪЕКТОВ В ЗАДАЧЕ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ДОКУМЕНТОВ

А. С. Серый (32112.alien@gmail.com)

Е. А. Сидорова (lena@iis.nsk.su)

Институт систем информатики им. А. П. Ершова СО РАН, Новосибирск, Россия

Предлагается подход к автоматизации наполнения информационной системы данными, полученными в результате автоматической обработки текстов. Учитывается устаревание информации, появление неточных и дублирующихся данных, противоречия с уже имеющейся информацией.

Ключевые слова: автоматическая обработка, автоматизация, информационная система, идентификация объектов.

OBJECT IDENTIFICATION IN PROBLEM OF AUTOMATIC DOCUMENT PROCESSING

A. S. Seryi (32112.alien@gmail.com)

E. A. Sidorova (lena@iis.nsk.su)

Institute of Informatics, Systems SB, Russian Academy of Sciences, Novosibirsk, Russian Federation

The paper presents an approach to automation of filling of an information system with the data obtained as a result of automatic document processing. The extracted data must be standardized as a network of information objects of a certain format. The backbone of such technique is to build so called focus set for every information object found in a text. Focus set for a single information object consists of all of the relations between this object and other input entities. There are several separate data processing stages: the search for duplicates, direct search, the search for similars and the search via the focus sets technique. A degree of data reliability is also provided. Thus an obsolescence of information, occurrence of the inexact and duplicated data, and conflict of new data with legacy information is taking into consideration.

Key words: automatic processing, automatization, information system, object identification.

Введение

В связи с быстрым развитием Интернет-технологий стремительно увеличивается количество накапливаемой неструктурированной текстовой информации. Вследствие этого возрос интерес к системам, которые позволяют автоматически извлекать знания из представленных документов и преобразовывать её в такую форму, с которой будет удобнее работать конечному пользователю. Таким образом, одной из основных задач, решаемых разработчиками информационных систем, является сканирование корпуса документов, написанных на естественном языке, и наполнение базы данных выделенной из текста полезной информацией. Для решения этой задачи существуют различные инструменты: достаточно известной является линейка продукции компании RCO (http://www.rco.ru/), также можно упомянуть и систему ИСИДА-Т [1,2] разработки Исследовательского Центра Искусственного Интеллекта.

Современные подходы извлечения информации не предусматривают проверки полученных данных при наполнении БД. Однако это может затруднить поиск конкретной информации в огромных архивах информационных систем. Информация устаревает, появляются копии имеющихся данных, возможно появление противоречий. Чтобы избежать таких ситуаций, снизить загруженность баз данных недостоверной и избыточной информацией, необходима предварительная обработка извлекаемых из текста фактов.

В данной статье предлагается подход к автоматизации наполнения информационной системы данными, полученными в результате автоматической обработки естественно-языковых ресурсов. Извлекаемые из текста данные должны быть унифицированы в виде сети информационных объектов[1] определенного формата. В частности, нами для извлечения данных используется текстовый анализатор, разрабатываемый в ИСИ СО РАН [3]. Формат ИО, в котором он инкапсулирует извлеченные данные, и был взят нами за основу. В процессе добавления в базу данных полученные ИО идентифицируются. Под идентификацией понимается однозначное разрешение *контекстной омонимии*, возникающей в том случае, когда одному входному объекту по его набору атрибутов можно сопоставить несколько объектов из базы данных.

1. Знания и данные в информационной системе

В дальнейшем под информационными системами будем понимать так называемые *информационные системы под управлением онтологии* [3], т. е. системы, предметная область которых ограничена и явно описана на определенном языке (считается, что описание доступно как конечным пользователям, так и внешним программным сервисам системы). Каждый ИО соответствует некоторому понятию онтологии и имеет заданную им структуру. Между

[1] *Информационный объект (ИО)* — описание некоторого объекта предметной области. Наборы разнотипных ИО составляют информационное наполнение системы.

ИО могут существовать связи, семантика которых определяется отношениями, заданными между соответствующими понятиями онтологии. Здесь, онтология — шестерка вида <C, A, T, D, R, F>, где

C — множество классов, описывающих понятия предметной области;
A — множество атрибутов понятий;
T — множество типов данных;
D — множество доменов (домен атрибута определяет множество его допустимых значений);
R — множество отношений, заданных на классах;
F — множество ограничений на значения атрибутов.

Каждый атрибут имеет, по крайней мере, имя и значение, и используется для хранения информации, специфичной для объекта и привязанной к нему. Значение атрибута может быть сложным типом данных.

Мы рассматриваем текстовый анализатор как внешний сервис информационной системы, необходимый для автоматической обработки текстовых документов и наполнение БД системы. В качестве анализатора может выступать любая программная система обработки текста, результат работы которой приводится к «понятному» для информационной системы формату (т. е. сети ИО). Рассматриваемый в статье модуль является универсальным (с указанными ограничениями) передатчиком, решающим задачу контроля данных. Однако предлагаемый подход позволяет решать не только эту задачу, но и улучшать результаты анализа текста, предоставляя доступ к глобальному контексту, т. е. знаниям не представленным непосредственно в тексте, поскольку для идентификации объектов необходимо обращение к онтологии системы и ее информационному наполнению. Информационной системой, на которой будет демонстрироваться данный подход, является портал знаний по компьютерной лингвистике (КЛ) [4]. В контексте данного портала имеются следующие основные классы извлекаемых объектов: *раздел науки, персоны, организации, географическое место, событие, деятельность, результат (продукт) деятельности,* а также связей между ними: *Работает-в, Направление-Исследований, Персона-Участник-События* и др.

2. Методика идентификации данных

В качестве входных данных выступают список извлеченных из документа ИО, упорядоченный по встречаемости в тексте, и список связей между этими ИО. Ключевым для предлагаемого метода идентификации данных является понятие *фокусного множества.* Фокусное множество включает все экземпляры отношений, с помощью которых текущий объект непосредственно связан с другими входными объектами. При этом множество отношений разбивается

на подмножества связей с идентифицированными и требующими идентификации объектами.

Основой метода является построение фокусных множеств для найденных в тексте объектов и сопоставление с фокусными множествами объектов, уже содержащихся в базе данных информационной системы.

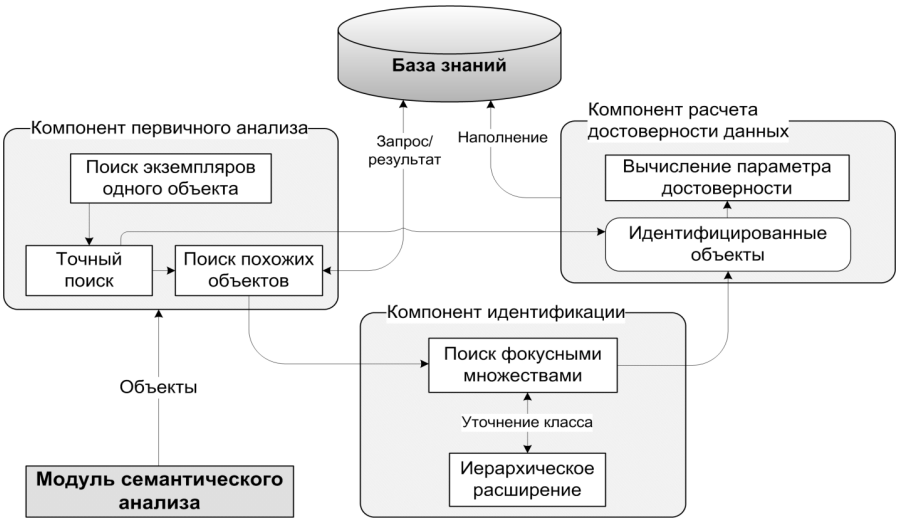


Рис. 1. Схема процедуры идентификации

На Рис.1. представлена общая схема процесса идентификации, который включает:

- Поиск дубликатов объектов[2];
- Точный поиск;
- Поиск похожих объектов;
- Поиск фокусными множествами.

Теперь подробнее о каждом этапе.

2.1. Поиск различных экземпляров одного объекта

В процессе обработки текста в случае референции [5] к упомянутому ранее ИО могут порождаться дубликаты объекта. Чтобы объединить всю

[2] Под *дубликатами объекта* подразумеваются объекты, возникшие вследствие многократного упоминания одного и того же объекта в тексте документа. Они могут содержать в себе различные непересекающиеся части сообщаемой в тексте информации о различных свойствах объекта.

информацию об ИО в одном месте, следует установить *кореферентность* дубликатов.

Дубликаты могут быть обнаружены по нескольким внешним признакам. Во-первых — это наличие связей и неопределенность ключевых атрибутов, кроме некоторого необходимого для отсылки набора, какого именно — зависит от класса. Например, у человека это чаще всего будут имя и отчество или фамилия, у организации — название, тип или аббревиатура. Во-вторых, потенциальными дубликатом также может считаться ИО, не имеющий связей с другими объектами. Например:

(1) *ABBYY — компания, производящая электронные словари и программное обеспечение для распознавания документов. Наиболее известные продукты* ***компании*** *— система распознавания документов FineReader и электронные словари Lingvo.*

(2) *Ю. Д. Апресян — российский лингвист, академик РАН. … Юрий Дереникович Апресян родился в 1930 году в Москве.*

Пример (1) иллюстрирует случай упоминания компании для установления связи с производимыми продуктами, а пример (2) — случай дальнейшего уточнения информации после краткого вступления.

Объект, удовлетворяющий одному из признаков, считается потенциальным дубликатом и сравнивается с объектами классов того же иерархического дерева. Поиск производится и вправо и влево, но, в силу правил построения предложений и текстов в русском языке, приоритет отдается объектам, упомянутым ранее, т. е. объектам «слева».

2.2. Точный поиск

Следует отметить, что для применения основного алгоритма необходим некоторый «стартовый» список идентифицированных объектов. Этот список может быть получен с помощью процедуры точного поиска. По входному объекту в базе данных проводится поиск объектов, имеющих идентичный набор ключевых атрибутов[3]. Этот набор не обязательно должен быть полным. Если был найден лишь один объект, то входной объект идентифицирован, и дальнейший его анализ уже не требуется.

Стоит добавить, что объект может быть признан идентифицированным и без достижения однозначного соответствия с объектом БД, хотя в этом случае он не участвует в формировании фокусных множеств. Это возможно если объект имеет полностью определенный набор ключевых атрибутов

[3] Под *ключевыми атрибутами* понимается набор атрибутов, однозначно идентифицирующий объект в информационном пространстве. Значения всех ключевых атрибутов определены у каждого объекта информационного пространства.

и не достигается однозначное соответствие ни с одним из имеющихся в БД. Такой объект является новым для БД и вносится в информационную систему как есть, а не как уточнение одного из старых объектов.

Для иллюстрации дальнейшего анализа рассмотрим пример из [6], где требовалось выявлять упоминания в сообщениях известных персон и научных организаций, а также извлекать информацию о том где, когда и в какой должности эти персоны работали:

(3) *Александр Михайлович является директором Института русского языка им. В. В. Виноградова РАН с 1997 года.*

В данном примере содержатся объекты классов *Персона* и *Организация*, и имеет место локальная неоднозначность (наименование персоны vs фрагмент наименования организации). В данном случае омонимия снимается на уровне сборки лексических шаблонов объектов: подстрока *В. В. Виноградова* входит в лексическую конструкцию, реализующую шаблон наименования объекта класса *Организация*. С помощью точного поиска можно идентифицировать организацию (*Институт русского языка им. В. В. Виноградова РАН*) в БД.

Персона *Александр Михайлович* задана недостаточно точно (отсутствует ключевой атрибут *Фамилия*), поэтому запускается поиск похожих объектов. Обозначим организацию $\boldsymbol{a}$, а персону — $\boldsymbol{b}$.

2.3. Поиск похожих объектов

Алгоритм предназначен для поиска объектов базы данных, наиболее похожих на объект, найденный в тексте. При построении списка похожих объектов участвуют только атрибуты. Список можно рассматривать как нулевой шаг последовательности фильтраций, осуществляемых алгоритмом поиска фокусными множествами.

Список строится путем сравнения текущего объекта из входного списка с объектами базы данных по различным подмножествам атрибутов. Так i-й шаг представляет собой выбор объектов БД, имеющих совпадения по любому набору из i атрибутов с анализируемым объектом.

Мощность списка найденных объектов БД может становиться только меньше от шага к шагу. За окончательный принимается результат i-го шага при условии $i < n$ и результат шага $(i+1)$ — пустое множество. Если этого не происходит, то после n-го шага выбирается список минимальной положительной мощности. Это и есть наиболее похожие объекты. Если на любом шаге список похожих объектов сузился до одного элемента, то анализируемый объект объявляется идентифицированным и его дальнейшее рассмотрение прекращается. Отметим также, что поиск похожих объектов производится только среди экземпляров одного класса (либо среди экземпляров классов одного иерархического дерева в случае учета иерархии).

Вернемся к примеру (3). Допустим, что в базе данных несколько человек с именем и отчеством *Александр Михайлович*, возможно работающих в той же организации.

Построение списка для объекта ***b:***

Шаг 1: выбираются персоны с именем *Александр* или отчеством *Михайлович*.

Шаг 2: выбираются персоны с именем *Александр* и отчеством *Михайлович*.

Согласно предположению, в системе существует больше одной персоны с такими значениями атрибутов. Они и сформируют список наиболее похожих объектов.

2.4. Поиск фокусными множествами

Основной алгоритм процедуры идентификации. Здесь основными фигурантами выступают уже связи объектов. Общий принцип работы кратко описывается следующими шагами:

a. Входной список объектов делится на два подсписка, **A** и **B** — идентифицированных и неидентифицированных объектов соответственно.
b. Для каждого объекта $\mathbf{b}_i \in \mathbf{B}$ строится фокусное множество $F_i = \langle b_i^I, b_j^{II} \rangle$ — пара множеств отношений, связывающих $\mathbf{b}_i$ с объектами подсписков **A** и **B** соответственно.
c. Из списка похожих объектов поочередно удаляются объекты, имеющие l связей из множества b_i^I, до тех пор, пока мощность его не станет равной 1 ($l = 0, 1, 2, \ldots, |b_i^I|$). В случае если этого не произошло, объект bi не может быть идентифицирован по имеющейся о нем информации.
d. Пусть объект $\mathbf{b}_i$ был идентифицирован на предыдущем шаге. В этом случае он переносится в подсписок **A**, все связи вида $\langle b_i, b_j \rangle \in b_j^{II}$, $\mathbf{b}_j \in \mathbf{B}$, переносятся во множество b_j^I, а объекты b_j анализируются даже в том случае, если ранее уже были отброшены за недостатком информации.

Теперь посмотрим, как это будет выглядеть применительно к нашему примеру (3).

$A = \{a\}$, $B = \{b\}$, $F_i = \langle b^I, b^{II} \rangle$ — *фокусное множество объекта* ***b****.*

$b^I = \{\langle a,b \rangle\}$ — *отношение «работает-в» со значением «директор» атрибута «должность».*

Из списка похожих объектов удаляются все, не связанные с ***a*** и имеющие другую должность. Остается либо один, либо ни одного объекта, соответствующего объекту ***b***. В нашем случае это был *Александр Михайлович Молдован.*

Выполнение алгоритма продолжается до тех пор, пока в подсписке **B** есть хоть один объект, доступный для анализа, после чего база данных редактируется в соответствии с полученными результатами.

$b^{II} = \varnothing$

2.5. Использование иерархических отношений

Понятия онтологии находятся в иерархической связи друг с другом («общее-частное»). Если объект не был идентифицирован, то можно сделать предположение о неточности указания его онтологического класса и расширить ареал поиска на экземпляры всех классов его иерархического дерева, Необходимо определить как классы-наследники, так и классы-родители. Глубина поиска по иерархии понятий может регулироваться в зависимости от количества входных объектов и требований производительности.

Использование иерархии по отношению «часть-целое» возможно в случае, когда объект подчинен другому объекту и имеет сложную структуру, представленную линейными цепочками наименований, совокупность которых образует дерево (множество деревьев) информационных объектов. Для идентификации такого объекта нужно восстановить иерархию вложенности объектов.

3. Наполнение базы данных

При редактировании объекта в системе могут возникать противоречия между старыми и новыми значениями его атрибутов. Это не относится к ключевым атрибутам[4]. Существуют несколько способов разрешения подобных противоречий:

1. Замена старых значений атрибутов на новые. Считается, что новая информация более достоверна.
2. Сохранение старых и новых значений с указанием даты внесения. Потерявшие актуальность данные удаляются экспертом вручную.
3. Введение параметра достоверности значений. По сути, это автоматизация предыдущего способа. Для этого требуется сохранение старых и новых данных, но по мере изменения параметра достоверности, система автоматически избавляется от недостоверных данных.

Как было сказано, третий способ предполагает введение специального параметра, количественно выражающего достоверность того или иного значения или связи. В данной работе предлагается следующая формула расчета такого параметра, выражающая зависимость от трех основных факторов, могущих послужить причинами противоречий с информацией, хранимой в базе данных: времени, рейтинга (авторитета) документа, из которого получены данные, и вероятности ошибки семантического анализатора, обрабатывавшего документ:

$$a\ (\boldsymbol{a}\colon v) = s(\boldsymbol{D}) \cdot h(T) \cdot G$$

где $\boldsymbol{a}\colon v$ — атрибут $\boldsymbol{a}$ в значении v, $\boldsymbol{D}(\boldsymbol{a}\colon v)$ — документ, в котором встретилось $\boldsymbol{a}\colon v$, T –текущая дата, как дата встречи $\boldsymbol{a}\colon v$.

[4] В противном случае, объект не мог бы быть идентифицирован.

Коэффициент $s(\boldsymbol{D})$ отражает зависимость $\boldsymbol{a}$ от степени доверия к документу. Документ — первый из источников противоречий с БД. Чем более авторитетный документ — тем больше доверия к данным, добытым из него.

$$s(\boldsymbol{D}) = \frac{R(\boldsymbol{D}(\boldsymbol{a}{:}\,v))}{\sum_j R(\boldsymbol{D}(\boldsymbol{a}{:}\,v_j))}$$

В числителе стоит суммарный рейтинг документов, в которых встретилось значение v, а в знаменателе — сумма рейтингов документов, в которых встретилось любое из имеющихся значений данного атрибута. Таким образом, $s(\boldsymbol{D})$ представляет собой «удельный вес» документов, упоминающих значение v, в массе всех документов, упоминающих данный атрибут (или связь). $s(\boldsymbol{D}) \leq 1$.

$$R(\boldsymbol{D}(\boldsymbol{a}{:}\,v)) = \sum_{i=1}^{n} r(\boldsymbol{D}_i(\boldsymbol{a}{:}\,v))$$

где $r(\boldsymbol{D}_i(\boldsymbol{a}{:}\ v))$ — рейтинг i-го документа, в котором встретился атрибут $\boldsymbol{a}$ в значении v. Общая формула рейтинга отдельного документа выглядит следующим образом:

$$r(\boldsymbol{D}) = R_{res} \cdot \left(1 - \frac{\sum_{j=1}^{N} \sum_{v_j^i \in D} k_j^i}{\sum_{j=1}^{N} \sum_{i=1}^{n_j} k_j^i}\right)$$

где R_{res} — сумма рейтингов ресурсов документа $\boldsymbol{D}$, N — количество атрибутов, встретившихся в документе $\boldsymbol{D}$ и имеющих больше одного альтернативного значения[5], $\boldsymbol{v}_j^i$ — i-е значение j-го атрибута, $\boldsymbol{k}_j^i$ — количество встречаемостей i-го значения j-го атрибута. Задача классификации ресурсов нами не рассматривается. Предполагается, что либо рейтинги ресурсов, из которых получены документы, вычислены и представлены в информационной системе, либо рейтинг каждого ресурса полагается равным 1.

Следующий коэффициент $h(T)$ отвечает за зависимость достоверности от времени, в том числе и за «старение» информации со временем. Изменения в реальном мире — второй источник противоречий с БД.

$$h(T) = h(T(\boldsymbol{a}{:}\,v)) = 1 + \ln\left((T - t_{last}) + 1\right)$$

Здесь t_{last} — ближайшая к T дата встречи отличного от v значения атрибута $\boldsymbol{a}$.

Третий коэффициент введен для учета ошибки анализатора при извлечении фактов. Ошибка при извлечении фактов — третий источник противоречий с БД.

[5] Учитываются только атрибуты, значения которых не могут быть множественными.

Коэффициент G применяется в случае наличия информации о принципе работы текстового анализатора, в частности, веса, выставленные экспертом схемам сборки фактов [3]. Обозначим эти веса w_l, $l = 1, 2, ..., F$. Также, ошибка при извлечении фактов может возникнуть вследствие неверного толкования понятий, т. е. омонимии. Поэтому G зависит еще и от количества альтернативных в данной позиции значений.

$$G = \frac{c \sum w_l}{L}$$

Здесь c — константа, L — количество альтернативных (омонимичных) значений, $\sum w_l$ — сумма весов схем фактов, участвовавших в формировании $\boldsymbol{a}$: v.

4. Данные эксперимента

В таблице 1 представлены результаты анализа трех документов, предназначенных для портала КЛ.

Таблица 1. Время работы алгоритмов на разных типах документов

К-во слов в документе	К-во ИО	К-во связей	Макс. к-во связей одного ИО	Макс. к-во атрибутов в одном ИО
413	17	14	12	4
65	6	4	1	2
532	11	5	5	5

К-во дубликатов	Поиск дубликатов (мс)	Поиск наиболее похожих объектов (мс)	Общее время работы (мс)	Идентифицировано объектов
3	312	2140	6125	4
0	63	105	394	4
8	453	179	850	4

Как можно видеть, документы имеют довольно небольшой размер. Как правило, это заметки, новостные сообщения или короткие статьи.

В первом случае длительное время обработки обусловлено в основном не количеством извлеченных ИО, а числом связей между ними, что увеличивает время построения и прогонки фокусных множеств, а также количеством ИО с наибольшим числом атрибутов, для которых требуется построить список наиболее похожих объектов. Как видно из таблицы 1, поиск похожих объектов занял более трети общего времени.

Второй документ является обыкновенным новостным сообщением с сайта компании ABBYY и содержит небольшое количество ИО.

Третий документ является отрывком из биографии, он был взят для иллюстрации работы на наборе ИО с высоким содержанием дубликатов, поскольку герой биографии упоминается в ней постоянно. Тем не менее, время анализа

невелико. Это обусловлено тем, что основной ИО — персона, являющаяся предметом статьи, был идентифицирован алгоритмом прямого поиска. В итоге большую часть времени занял поиск дубликатов.

Заключение

Описанный метод применяется при разработке сервисов анализа документов для информационного ресурса «Хроники СО АН» [6] и портала знаний по компьютерной лингвистике [4]. Проведенные эксперименты показали, что алгоритм поиска наиболее похожих объектов дает ощутимые расходы по времени, в случае объектов с десятью и более атрибутами. Это происходит вследствие значительного усложнения генерируемых запросов к базе данных, что приводит к увеличению временных затрат на их оптимизацию и выполнение. Для решения этой проблемы планируется ввести фильтрацию рассматриваемых атрибутов (например, не учитывать полнотекстовые атрибуты типа *Комментарии*, *Толкование, Описание* и др.), и осуществлять их упорядочивание по значимости. Также, возможно искусственно ограничивать время работы алгоритма, а за результат принимать список, построенный к моменту остановки (например, в случае большого количества входных объектов). Будут продолжаться работы по оптимизации механизма расчета коэффициента достоверности и его параметров, вполне вероятно увеличение доли участия эксперта в этой процедуре.

References

1. *Aleksandrovskii D. A., Kormalev D. A., Kormaleva M. S., Kurshev E. P., Suleimanova E. A., Trofimov I. V.* 2006. The Development of Means of Text Analytic Processing in the System ISIDA-T [Razvitie Sredstv Analiticheskoi Obrabotki Teska v Sisteme ISIDA-T]. Trudy 10 Natsional'noi Konferentsii po Iskusstvennomu Intellektu (Proc. of the X National Conference on Artificial Intellect), 2 : 555–563.
2. *Borovikova O. I., Zagorul'ko Iu. A., Zagorul'ko G. B., Kononenko I. S., Sokolova E. G.* 2008. Designing of a Portal of Knowledge on Computational Linguistics [Razrabotka Portala Znanii po Komp'iuternoi Lingvistike]. Trudy 11 Natsional'noi Konferentsii po Iskusstvennomu Intellektu (Proc. of the XI National Conference on Artificial Intellect), 3: 380–388.
3. *Kononenko I. S., Sidorova E. A.* 2009. An Ontology-Based Facts Extraction Approach [Podkhod k Izvlecheniiu Faktov iz Teksta na osnove Ontologii]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2009" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2009") : 451–457.
4. *Kormalev D. A., Kurshev E. P.* 2006. The Development of Language for Infromation Extraction Rules in the System ISIDA-T [Razvitie Iazyka Pravil Izvlecheniia

Infromatsii v Sisteme ISIDA-T]. Trudy Mezhdunarodnoi Konferentsii “Programmnye Sistemy: Teoriia I Prilozheniia” (Proc. of National Conference "Program Systems: Theory and Applications"), 1 : 365–377.

5. *Lebedev M. V., Cherniak A. Z.* 2001. Ontological Problems of Reference [Ontologicheskie Problemy Referentsii].
6. *Sidorova E. A., Zagorul’ko Iu. A., Kononenko I. S.* 2006. Semantic Approach to Document Analysis basing on the Ontology of Object Area [Semanticheskii Podkhod k Analizu Dokumentov na osnove Ontologii Predmetnoi Oblasti]. Komp’iuternaia Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii “Dialog 2006” (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference “Dialog 2006”) : 468–473.

THE PROPER PLACE OF MEN AND MACHINES IN LANGUAGE TECHNOLOGY. PROCESSING RUSSIAN WITHOUT ANY LINGUISTIC KNOWLEDGE

S. Sharov (s.sharoff@leeds.ac.uk)
University of Leeds, UK

J. Nivre (joakim.nivre@lingfil.uu.sv)
Uppsala University, Sweden

The paper describes several experiments aimed at designing tools for processing Russian texts, namely for Part-Of-Speech tagging, lemmatisation and syntactic parsing, exploiting exclusively statistical approaches without coding any linguistic rules specifically for Russian. While not claiming any new ground for machine learning research, the results demonstrate the possibility to create state-of-the-art tools for Russian in very short time using only machine learning and no hard-coded linguistic knowledge. One of the results of this study is a set of publicly available resources which can be used in standard pipelines for processing Russian. However, they also demonstrate hidden costs associated with the use of purely statistical methods and the need to integrate linguistic parameters into statistical procedures.

Key words: language technology, processing texts, machine, machine learning.

1. Introduction

The title of this paper refers to a famous research report produced by Martin Kay in the 1980s, "The proper place of men and machines in language translation", finally published in (Kay, 1997), in which Kay argued for the proper distribution of labour between the human translators and the Computer-assisted Translation systems. Another reference appropriate to the topic of the paper presented here is a statement attributed to Fred Jelinek "Every time I fire a linguist the results of speech recognition go up", i. e. explicit linguistic knowledge is dispensable.[1] This sentiment is related to a paradigmatic shift that happened in the computational linguistics in the beginning of the 1990s: with more and more data available and with the advance in the methods of machine learning, more approaches switched from careful encoding of linguistic phenomena to finding statistical correlations in texts (either annotated or raw). The vast majority of publications at major conferences on computational linguistics belong to this paradigm. However, to the best of our knowledge relatively few attempts have been made to apply entirely statistical methods to building tools for processing

[1] However, this story is not entirely correct, see (Jelinek, 2005).

Russian, e. g., (Sokirko and Toldova, 2005, Nivre et al., 2008, Sharoff et al., 2008). Purely statistical approaches to language processing are also very infrequent in the proceedings of Russian conferences (like this one).

The paper describes three experiments on designing Russian NLP tools, respectively for Part-Of-Speech (POS) tagging, for lemmatisation and for syntactic parsing. Thus, they cover the basic tools needed for doing NLP and corpus linguistics in Russian. The experiments did not exploit any prior knowledge of the Russian language, i. e. we did not use any rules for dealing with any specific Russian phenomenon. Each experiment can be described in the following lines:

1. take an annotated Russian corpus;
2. design a simplified representation of annotations to convert the corpus into the format suitable for the learning tool to be used;
3. learn a model in several iterations to tune the learning parameters.

In this approach the human efforts are invested into creating annotated corpora, representing data and designing machine learning algorithms, while the machine is able to learn the links between the data. In the end, linguistic knowledge is induced from annotated corpora rather than explicitly hand-crafted by linguists. In a similar way, development of corpora is possible without manual selection of texts from a range of sources. It can be facilitated by crawling or using the API of a search engine and automatically annotating them with respect to their domains and genres (Baroni et al., 2009, Sharoff, 2010).

The automatically induced rules also do not take the form of hard constraints, separating the possible from the impossible, but rather as graded constraints, distinguishing the more probable from the less probable. This makes the automatically acquired models more robust to noise.

In the sections below we briefly outline the statistical methods used in each of the three tasks (Section 2), ways of representing corpus phenomena (Section 3) and the results obtained using our tools (Section 4)

2. Methods used

2.1. Statistical part-of-speech tagging

POS tagging is aimed at assigning a POS label (tag) to each word in the input stream. Until the end of the 1980s this task had been usually performed by sets of carefully crafted rules for disambiguating the contexts, e. g., for detecting contexts in which the form *стали* is a noun ('steel') or a verb ('become'), cf. one of the earliest descriptions of this sort (Nikolaeva, 1958). Ken Church was one of the first researchers to show the possibility of abandoning the rules and relying exclusively on POS-annotated data (Church, 1988). This led to proliferation of statistical approaches to tagging, either using automatic derivation of decision trees, e. g., TreeTagger (Schmid,

1994), Hidden Markov Models (HMM), e. g., TnT (Brants, 2000), or machine learning, e. g. SVMTool (Giménez and Màrquez, 2004).

Probably, the most widely used approach is based on HMM for estimating the probability of a tag from the distribution of words over tags (which tag is more likely for this word), as well as over $N-1$ adjacent tags, with N often fixed at 3 (a trigram model). For example, given a sentence like:

\gll Это была гравюра на стали
this was engraving on steel
\glt `It was a steel engraving',

the sequence of tags *Noun Preposition Verb* is much less likely than the one for *Noun Preposition Noun*, hence the word стали in this sentence receives the tag *Noun*. Still the probability of the sequence *Noun Preposition Verb* in Russian is greater than zero because of such constructions as шутки ради позвонили…

This study uses the TnT tagger (Brants, 2000). In addition to standard HMM tagging it employs several useful methods for approximating the probabilities of unseen tag sequences (smoothing) as well as for guessing possible tags of unseen words. The latter is done by computing the probability of the last m characters of an unseen word form co-occurring with a given tag. For example, when such forms as *vociferation, votazione*, конъюгация, 自由主义 are missing in respective training corpora, they are still more likely to receive the noun tag on the basis of POS tags for words with the same ending.

2.2. Learning lemmatisation rules

Lemmatisation rules can be also derived automatically from a list of word forms paired with their possible lemmas and POS tags obtained from an annotated corpus (Erjavec and Džeroski, 2004, Jongejan and Dalianis, 2009). The CST lemmatiser used in our experiments tries to find for each pair the longest shared part, e. g., for the pair близкий-поближе the inner part is бли, this leads to the rule *зкий->по*же (the asterisk indicates any character). The training process then tries to apply the new rule across all pairs with the same POS tag. If lemmatisation is successful, nothing needs to be done, e. g., for низкий-пониже. However, if an applicable rule from the rule base produces incorrect lemmatisation, e. g., for the pair плохой-похуже, the rule *зкий->по*же produces хузкий, which does not match the target lemma, then a new lemmatisation rule is generated to cover more specific cases (there is a special strategy to determine which rules are retained as more general and which cover specific cases). The rule generated in this case *лохой->*охуже, since п is shared. Even though the rule is not entirely correct, it is quite unlikely to cause problems in processing real texts, since it fires only when we have a form ending with охуже which gets the tag of a comparative adjective. The training stage runs until all forms in the training set are successfully mapped to their lemmas.

2.3. Syntactic parsing

Syntactic parsing aims at computing a complete hierarchical representation of an input sentence. Statistical methods for parsing has until recently focused on phrase structure parsing for English, resulting in a series of increasingly accurate parsers trained on the Penn Treebank (Magerman, 1995, Collins, 1997, Charniak, 2000, Charniak and Johnson, 2005). However, dependency parsing has emerged as an interesting alternative, especially for languages with more flexible word order than English, as seen in the CoNLL shared tasks on dependency parsing (Buchholz and Marsi, 2006, Nivre et al., 2007). In fact, for decades dependency parsing was the standard approach in the Soviet/Russian linguistic tradition (Mel'čuk, 1988).

Most recent approaches to statistical dependency parsing can be characterized as either *graph-based* or *transition-based* (McDonald and Nivre, 2007). A graph-based parser learns a model for scoring entire dependency graphs and performs exhaustive search for the highest-scoring graph at parsing time; a typical example is MSTParser (McDonald, 2006). A transition-based parser instead learns a model for predicting the next parser action — or transition — and performs greedy search for best transition sequence at parsing time; a typical example is MaltParser (Nivre et al., 2006). Both approaches can give state-of-the-art accuracy, but the transition-based method is potentially much more efficient, which is useful when parsing large amounts of data. The transition-based MaltParser system has previously been applied to Russian with promising empirical results (Nivre et al., 2008).

3. Russian corpora and their representation

3.1. Annotated corpora used for training

Information about the training corpora is given in Table 1. The Russian National Corpus contains a component with morphosyntactic annotation (Plungian, 2005), which is commonly known as снятник (disambiguated). Originally it contained only fiction, but it has been expanded to cover a range of genres, such as newspapers, informal communication (jokes and forums), scientific&technical texts, etc. For training the parsing tool, we used SynTagRus, a Russian corpus with dependency annotation for every sentence (Boguslavsky et al., 2000). This has been produced by using the output of ETAP (Apresian et al., 2003) with manual correction of incorrect analyses.

Table 1. Annotated corpora used in this study

Disambiguated RNC			**SynTagRus**		
Tokens	Orth words	Sentences	Tokens	Orth words	Sentences
5 801 316	5 115 016	432 611	719 957	635 524	41 186

3.2. Adapting the Russian tagset

Zalizniak's Grammatical dictionary (Zalizniak, 1977) is a formalisation of Russian morphology, which is commonly used in NLP tools for automatic morphological analysis, e. g., (Segalovich, 2003, Sokirko, 2004). The tagset used in the disambiguated RNC is also largely based on the Zalizniak categories (with few expansions, such as the use of the vocative case).

The problem with using statistical taggers is that they usually operate with atomic labels, e. g., NNS in the English Penn tagset stands for 'plural common noun', NP stands for 'singular proper noun', while the output of morphological analysis is traditionally represented by a set of features, e. g., for mystem (Segalovich, 2003):

шлепают: шлепать=V,несов=непрош,мн,изъяв,3-л,пе

which corresponds to 'to slap=Verb,imperfective=nonpast,plural,indicative,3rd person,transitive'.

It is possible to produce a tagset by concatenation of the feature set for each word. However, this results in a fairly large number of tags, for example, concatenation of features for all words in the disambiguated RNC produces 4,592 tags, which is too much for trigram tagger learning on a corpus of five million words. The total number of tags reported in (Sokirko and Toldova, 2005) in an experiment, which also used the disambiguated RNC, is 829 tags. This indicates some kind of tagset design, though it is not described in the report.

MTE is a project aiming at standardising the tagset for a range of language (Erjavec, 2010), it covers many other Slavonic languages, so the added advantage of using it was the possibility to create a unified tagset.

The tagset is positional, i. e., for a major POS (Noun, Verb, etc) there are fixed positions with values for features. For example, Ncfsgn stands for 'Noun, common, feminine, singular, genitive, inanimate', while Vmis-sfp stands for 'Verb,main,indicative,past,-,singular,feminine,perfective', with the hyphen occupying the place of the person value (which is not detected for the Russian verbs in the past tense). The prepositions are marked for the case of the noun phrase they govern. Example exAmbig receives the following analysis:

\gll Это была гравюра на стали
P--nsnn Vmis-sfa Ncfsnn Sp-l Ncfsln

SynTagRus is also a part of the Russian National Corpus, but because of the differences in its morphological categories, it uses a separate query interface. The SynTagRus tagset has been also mapped to a subset of MTE. Given that SynTagRus does not contain the category of pronouns (the personal pronouns in it are coded as nouns, possessive pronouns as adjectives, etc), its mapping to MTE produces a smaller tagset in comparison to the RNC. So the extra task in this case was to map the RNC-based output of the tagger to the SynTagRus-based set of tags.

4. Results

4.1. Tagging

Out of the 5 million orthographic words of the disambiguated RNC 10% was kept in the held-out portion used for evaluation. The tagger was trained on the remainder of the disambiguated RNC, and the overall accuracy on the held-out portion was 95.28% (with punctuation excluded).

We also measured the performance of TnT on a reduced tagset of Russian (only codes in Table 2). The accuracy reached 97.09%, which is only slightly better than the performance of the tagger on the detailed tagset, while the detailed tagset is more beneficial for many NLP tasks.

Table 2. Incorrectly assigned POS tags

Code	Explanation	Error rate	Relative error	Coverage
N	Nouns	2.08%	7.21%	28.80%
A	Adjectives	0.86%	9.05%	9.51%
P	Pronouns	0.65%	7.82%	8.28%
V	Verbs	0.50%	4.89%	10.16%
C	Conjunctions	0.14%	2.37%	5.84%
R	Adverbs	0.13%	4.69%	2.81%
S	Prepositions	0.13%	0.89%	14.62%
M	Numerals	0.13%	4.60%	2.81%
Q	Particles	0.10%	4.03%	2.59%
I	Interjections	0.01%	26.42%	0.02%

The types of errors produced by the tagger on the full tagset are illustrated in Table 2 and Table 3. The error rate in Table 2 refers to the total count of errors for this category, this is a measure of how important this type of errors is for tagging a text (the table is sorted by this column). It is also interesting to know the amount of word forms *within* each category tagged incorrectly. This is the relative error rate, which reflects how difficult the category is for the tagger, e.g. 7.21% rate for nouns means one out of 14 nouns gets a tag which is incorrect in at least one position, while only one out of 112 prepositions (0.89%) gets a wrong tag (the preposition is not recognised or the case is not assigned correctly). The coverage refers to the total amount of such POS tags in the held-out portion of the RNC, this indicates the relative importance of the category.

The evaluation on individual categories reveals that the most difficult POS category is the category of nominals, which includes adjectives and nouns, as well as pronouns, which is a fringe member, including nominal pronouns (P-----n) and attributive pronouns (P-----a) with nominal inflection, as well as adverbial pronouns (P-----r). The apparently high relative error rate for interjections is explained by the

fact that the two most common interjections are ‘а’ and ‘о’ (ambiguous with a common conjunction and preposition respectively), and their low frequency does not influence the overall error rate much.

Table 3. Most common incorrectly tagged words

0.0932%	TnT	как	C
0.0920%	RNC	как	P-----r
0.0788%	TnT	что	C
0.0682%	TnT	ЭВМ	Ncfsgn
0.0682%	RNC	ЭВМ	Ncfpgn
0.0507%	RNC	что	P--nsnn
0.0444%	TnT	это	P--nsnn
0.0438%	TnT	как	P-----r
0.0413%	TnT	судов	Ncnpgn
0.0413%	RNC	судов	Ncmpgn
0.0413%	RNC	как	C
0.0363%	TnT	все	P--nsnn
0.0357%	RNC	это	Q
0.0350%	RNC	все	R
0.0338%	RNC	что	P--nsan
0.0325%	TnT	его	P-3msan
0.0300%	RNC	их	P-3-pgn
0.0288%	TnT	то	P--nsnn
0.0288%	RNC	когда	P-----r
0.0288%	TnT	когда	C
0.0269%	RNC	то	C
0.0263%	TnT	же	Q
0.0263%	RNC	же	C
0.0244%	RNC	что	C
0.0244%	RNC	лиц	Ncnpgy
0.0244%	TnT	лиц	Ncnpgn
0.0238%	TnT	что	P--nsnn
0.0238%	TnT	ли	Q
0.0238%	RNC	ли	C
0.0219%	TnT	право	Ncnsan
0.0219%	TnT	их	P-----a
0.0206%	RNC	право	Ncnsnn
0.0188%	TnT	его	P-----a
0.0181%	RNC	все	P--nsan

A more detailed look at the sources of errors presented in Table 3 reveals the following problems:

1. distinguishing between closely related POS classes, such as pronouns and conjunctions (как, когда, что, то), similarly for particles (же, ли);
2. dealing with long-distance dependencies, especially in distinguishing between the nominative and accusative cases (все, право, это);
3. domain mismatch, when the training corpus and the held-out one referred to different domains (судов, masculine or neuter, лиц, animate or inanimate);
4. guessing the full tag for abbreviations (ЭВМ, which was plural genitive in the held-out portion of the RNC, but got the tag of singular genitive in the absence of other indicators of plurality);
5. distinguishing between adverbs and short adjectives (e. g., удобно).

In spite of the number of problems in statistical tagging, a recent comparison of several Russian disambiguation tools in (Ljashevskaja et al., 2010) demonstrated its reasonable performance against other disambiguation and lemmatisation tools (our tagger and lemmatiser are reported there under the names of Peru and Pine). The accuracy of POS tagging achieved on that corpus was 97.3 %, which was considerably better than the majority of other (rule-based) systems. In addition to this, the worst performing component of the tagger was the rule-based tokeniser, which incorrectly identified token boundaries and thus decreased the overall performance.

4.2. Lemmatisation

These are the rules generated for the tag Ncmsgy for nouns ending in -ц:
iwonac

ец	еца
иц	ица
заяц	зайца
ец	йца
я-муромец	и-муромца
ринц	ринца
ртц	ртца
ец	ьца
ец	ца

The model for Zalizniak's Index 5 (masculine nouns ending in -ц) is well-represented, including the regular forms with and without morphological alternation (кузнец-кузнеца, фриц-фрица, европеец-европейца, принц-принца, владелец-владельца, чеченец-чеченца), as well as some exceptions, including the irregular заяц-зайца and the occasional forms артц-артца (used in Vasily Grossman's "Life and fate") and Ильи-Муромца, which came from the inability of the lemmatiser to deal with the hyphenated nouns.

The statistical lemmatiser depends on the output of tagging, but it is moderately tolerant to tagger errors. For example, irrespectively of the error in getting the animacy of лиц in Table 3 it still gets the right lemma. However, the error in getting the gender of судов leads to incorrect lemmatisation.

Table 4. Parsing results on development set of SynTagRus; labeled attachment score (LAS) and unlabeled attachment score (UAS)

	LAS	UAS
SynTagRus tags, poly-SVM	83.4	89.4
MTE tags, poly-SVM	82.8	88.8
MTE tags, linear SVM	82.2	88.0

4.3. Syntactic parsing

Because of the need to tune the parameters during parsing, SynTagRus was split into three parts, the training set (507 986 words), the development set for tuning the parameters (64 196 words) and the test set for the final evaluation (63 342 words). Table 4 shows results on the development set for three different settings with the standard evaluation metrics: labeled attachment score (LAS), the proportion of words that are assigned the correct head *and* dependency label, and unlabeled attachment score (UAS), the proportion of words that are assigned the correct head (regardless of label).

The first experiment replicates the settings from (Nivre et al., 2008) exactly, using the original part-of-speech tags from the SynTagRus treebank and using SVMs with a polynomial kernel to predict the next parser transition.[2] The results obtained are slightly better than the ones reported by (Nivre et al., 2008) (LAS 82.3, UAS 89.0), which is probably due to a larger training set. The second experiment uses the same features and the same type of classifier (poly-SVM) but replaces the SynTagRus part-of-speech tags with the MTE tags. This results in slightly lower parsing accuracy, about 0.6 percentage points for both metrics.

Using SVMs with a polynomial kernel is rather inefficient during both training and parsing. For example, parsing the development set of 68,314 tokens takes about three hours. In the third experiment, we therefore used a linear SVM, together with a slightly extended set of features to compensate for the lack of the polynomial kernel. The result is a much faster parser, which parses the development set in under two minutes, although with slightly lower accuracy. This parsing model will be applied to the Russian Web corpus of about 3 billion words, and it is expected to complete parsing in under two months.

[2] Besides part-of-speech tags, the parser uses word forms, lemmas and morphosyntactic features as a basis for prediction; see (Nivre et al., 2008) for more details.

5. Conclusions

This paper presents a fairly radical stance: it is redundant to encode linguistic knowledge explicitly; a completely automatic machine learning procedure can quickly produce a fast and reliable NLP component, which rivals (and in some cases exceeds) the performance of hard-coded linguistic rules requiring the efforts of many person-months (if not years). Hence, the efforts of linguists need to be spent on creating data rather than writing rules.

Nevertheless, this claim needs to be taken with a pinch of salt. First, the approach was reasonably successful since it implicitly utilised some information about the language. The methods for unknown word guessing as well as lemmatisation used in this study rely on the fact that Russian is a flective language. Statistical tagging and lemmatisation are known to be more difficult for agglutinative languages, like Turkish (Dincer et al., 2008). For an isolating language, like Chinese, there is no problem with lemmatisation, but the greater average ambiguity of the POS tags for known words and the lack of reliable prediction of the POS tag for unknown words makes the accuracy of knowledge-free methods considerably lower.

Second, data representation in terms of tag labelling is sufficiently simple and efficient, but a tag label lacks information about the internal structure of linguistic phenomena. For example, when the system learns the structure of Russian noun phrases, it does not take into account the agreement in case, number and gender. It only learns the fact that `Afpmsg` is normally followed by `Ncmsgn`, `Ncmsgy` or `Npmsgy`, while `Afpfsd` is followed by `Ncfsdn`, etc. However, if the set of training examples does not contain a proper masculine *inanimate* noun (`Npmsgn`) in this sequence, the tagger will fail to treat the sequence of `Afpmsg Npmsgn` as a noun phrase, even if the concept of animacity is not relevant to the noun phrase construction.

Yet another problem in using purely statistical methods is the reliance on patterns present in training data. Each training set has its own peculiarities, which do not necessarily match the peculiarities of the application domain. For example, the impressive accuracy of 97–98 % for HMM tagging is obtained on well-controlled newspaper texts (*The Wall Street Journal* for English and *Frankfurter Rundschau* for German), but the accuracy of taggers trained on these corpora drops dramatically on other text genres, down to 85.7 % on Internet forums, i. e., every seventh word is tagged incorrectly (Giesbrecht and Evert, 2009). This does not indicate any inferior status of Internet forums, just the fact that the trigram model trained on newspaper texts does not approximate them well. Annotating texts in the application domain to obtain more training data is expensive, so the tools are often used in new domains without formal evaluation of their accuracy, e. g., ukWac (Baroni et al., 2009) has been tagged and lemmatised with the default TreeTagger model. This problem is partly addressed by new approaches to machine learning using domain adaptation, which uses a training corpus from the source domain (with available annotated data), a small number of annotated examples from the target domain and a large number of unlabelled examples from the target domain (Daumé III et al., 2010).

In addition to the known problem of unknowns in the domain mismatch, there is a problem of unknown knowns, namely when peculiarities inherent in the annotated set are not obvious, while machine learning is likely to emphasise them for making classification decisions. In the end, the system might achieve reasonably good accuracy on the held-out portion of the annotated set (since it is drawn from the same distribution), while this accuracy could be irrelevant outside of the annotated set alone. For example, in the field of automatic genre classification it has been shown that a large number of texts on a particular topic within a genre heading can considerably affect the decisions made by the classifier, e. g., by treating texts on hurricanes and taxation as belonging to FAQs (Wu et al., 2010). At the same time, a classifier based on POS trigrams is much less successful, but it suffers less from the transfer from one annotation set to another (Petrenz and Webber, 2010).

Finally, there are problems with correcting the results. An error produced by a rule-based tagger can be corrected by debugging, finding the incorrectly fired rule, modifying it and testing the performance again. A statistical model can be amended by modification of the learning parameters or by providing more data, but this is only indirectly related to the performance of the system in the case of an individual problem.

In either case, the main contribution of the paper is two-fold. First, we describe the baseline for natural language processing for Russian using only statistical methods and minimal adjustment to the representation of source data. In spite its minimalism, the baseline outperforms the majority of the rule-based systems (Ljashevskaja et al., 2010). Second, the tools reported in this paper are available for linguistic research.[3] This defines the entire pipeline, which starts with POS tagging of pre-tokenised texts, proceeds to lemmatisation and ends with syntactic parsing.

Acknowledgements

Research reported in this paper was partly funded by European Community's Seventh Framework Programme (FP7/2007–2013) under Grant Agreement no 248 005 (TTC)[4] and partly by European Community's Life Long Learning Programme (project Kelly, Keywords for Language Learning for Young and adults alike).[5]

[3] They can be downloaded from http://corpus.leeds.ac.uk/tools

[4] http://www.ttc-project.eu

[5] http://su.avedas.com/converis/contract/321

References

1. *Apresian J., Boguslavskii I., Iomdin L., Lazurskii A., Sannikov V., Sizov V., Tsinman L.* 2003. ETAP-3 Linguistic Processor: a Full-fledged NLP Implementation of the MTT. First International Conference on Meaning-Text Theory : 279–288.
2. *Baroni M., Bernardini S., Ferraresi A., Zanchetta E.* 2009. The WaCky Wide Web: a Collection of Very Large Linguistically Processed Web-Crawled Corpora. Language Resources and Evaluation, 43(3):209–226.
3. *Boguslavskii, I., Grigor'eva S., Grigor'ev N., Kreidlin L., Frid N.* 2000. Dependency Treebank for Russian: Concept, Tools, Types of Information, 2 : 987–991.
4. *Brants T.* 2000. TnT — a Statistical Part-of-Speech Tagger. Proc. of 6th Applied Natural Language Processing Conference : 224–231.
5. *Buchholz S., Marsi E.* 2006. CoNLL-X Shared Task on Multilingual Dependency Parsing. Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL) : 149–164.
6. *Charniak E.* 2000. A Maximum-Entropy-Inspired Parser. Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL) : 132–139.
7. *Charniak E., Johnson M.* 2005. Coarse-to-fine N-best Parsing and MaxEnt Discriminative Reranking. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL) : 173–180.
8. *Church K.* 1988. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. Proceedings of the Second Conference on Applied Natural Language Processing : 136–143.
9. *Collins M.* 1997. Three Generative, Lexicalised Models for Statistical Parsing. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL) and the 8th Conference of the European Chapter of the Association for Computational Linguistics (EACL) : 16–23.
10. *Daumé III H., Kumar A., Saha A.* 2010. Frustratingly Easy Semi-Supervised Domain Adaptation. Workshop on Domain Adaptation for Natural Language Processing at ACL2010.
11. *Dincer T., Karaoglan B., Kisla T.* 2008. A Suffix Based Part-of-speech Tagger for Turkish. Third International Conference on Information Technology: New Generations : 680–685.
12. *Erjavec T.* 2010. Multext-east Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10).
13. *Erjavec T., Džeroski S.* 2004. Machine Learning of Morphosyntactic Structure: Lemmatizing unknown Slovene words. Applied Artificial Intelligence, 18(1) : 17–41.
14. *Giesbrecht E., Evert S.* 2009. Part-of-Speech (POS) Tagging — a Solved Task? An Evaluation of POS Taggers for the Web as Corpus. Proceedings of the Fifth Web as Corpus Workshop (WAC5) : 27–35.
15. *Giménez J., Màrquez L.* 2004. SVMTool: A General Pos Tagger Generator Based on Support Vector Machines. Proceedings of the Forth Language Resources and Evaluation Conference.

16. *Jelinek F.* 2005. Some of My Best Friends are Linguists. Language Resources and Evaluation, 39(1) : 25–34.
17. *Jongejan B., Dalianis H.* 2009. Automatic Training of Lemmatization Rules that Handle Morphological Changes in Pre-, In- and Suffixes alike. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP.
18. *Kay M.* 1997. The Proper Place of Men and Machines in Language Translation. Machine Translation, 12(1–2):3–23.
19. *Liashevskaia O., Astaf'eva I., Bonch-Osmolovskaia A., Gareishina A., Iu., G., D'iachkov V., Ionov M., Koroleva A., Kudrinski M., Litiagina A., Luchina E., Sidorova E., Toldova S., Savchuk S., and Koval' S.* 2010. Evaluation of Automatic Text Parsing Methods: Morphological Parsers in Russian [Otsenka Metodov Avtomaticheskogo Analiza Teksta: Morfologicheskie Parsery Russkogo Iazyka]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2010" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2010") : 318–326.
20. *Magerman D. M.* 1995. Statistical Decision-tree Models for Parsing. Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL) : 276–283.
21. *McDonald R.* 2006. Discriminative Learning and Spanning Tree Algorithms for Dependency Parsing.
22. *McDonald R., Nivre J.* 2007. Characterizing the Errors of Data-driven Dependency Parsing Models. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL): 122–131.
23. *Mel'chuk I.* 1988. Dependency Syntax: Theory and Practice.
24. *Nikolaeva T.* 1958. Soviet Developments in Machine Translation: Russian Sentence Analysis. Mechanical Translation, 5(2):51–59.
25. *Nivre J., Boguslavskii I. M., Iomdin L. L.* 2008. Parsing the SynTagRus Treebank of Russian. Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008) : 641–648.
26. *Nivre J., Hall J., Kübler S., McDonald R., Nilsson J., Riedel S., Yuret D.* 2007. The CoNLL 2007 Shared Task on Dependency Parsing. Proceedings of the CoNLL Shared Task of EMNLP-CoNLL 2007 : 915–932.
27. *Nivre J., Hall J., Nilsson J.* 2006. Maltparser: A Data-driven Parser-Generator for Dependency Parsing. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC) : 2216–2219.
28. *Petrenz P., Webber B.* 2010. Stable Classification of Text Genres. Computational Linguistics, 34(4).
29. *Plungian V. A.* 2005. What do We Need Russian National Corpus for? [Zachem Nuzhen Natsionalnii Korpus Russkogo Iazyka?]. Natsionalnii Korpus Russkogo Iazyka : 6–20.
30. *Schmid H.* 1994. Probabilistic Part-of-speech Tagging Using Decision Trees. International Conference on New Methods in Language Processing.

31. *Segalovich I.* 2003. A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine. Proc. of MLMTA-2003.
32. *Sharov S.* 2010. In the Garden and in the Jungle: Comparing Genres in the BNC and Internet. Genres on the Web: Computational Models and Empirical Studies.
33. *Sharov S., Kopotev M., Eriavets T., Feldman A., Diviak D.* 2008. Designing and Evaluating a Russian Tagset. Proceedings of the Sixth Language Resources and Evaluation Conference, LREC 2008.
34. *Sokirko A.* 2004. Morphological Modules on the web-site www.aot.ru [Morphologicheskie Moduli na saite www.aot.ru]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2004" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2004").
35. *Sokirko A., Toldova S.* 2005. Sravnenie Effektivnosti Dvukh Metodik Sniatiia Lexicheskoi i Morfologicheskoi Neodno znachnosti dlia Russkogo Iazyka. Internet-matematika.
36. *Wu Z., Markert K., Sharov S.* 2010. Fine-grained Genre Classification using Structural Learning Algorithms. Proc. of ACL 2010.
37. *Zalizniak A.* 1977. Russian Grammar Dictionary [Grammaticheskii Slovar' Russkogo Iazyka. Russki Iazyk].

REFLECTING ACCENTUATION IN THE RUSSIAN MORPHOLOGICAL DICTIONARY OF THE MULTIFUNCTIONAL LINGUISTIC PROCESSOR ETAP-3[1]

V. G. Sizov (sizov@iitp.ru)

O. Iu. Podlesskaia (olga@iitp.ru)

Laboratory of Computational Linguistics, Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russian Federation

Our work is aimed at the introduction of accentual information into the morphological dictionary of the multifunctional linguistic processor ETAP-3. A special formal description language has been created, and special rules for most of the basic accentual schemes have been designed. Special algorithms have been written for morphological analysis and synthesis.

Key words: ETAP, ETAP-3, accent, accentuation, morphological dictionary.

1. Introduction. Problem statement

Accentual information (information about the location of accent, or stress, in word forms) in morphological dictionaries of text processing systems allows solving important and useful tasks. First, this information is necessary for high-quality speech synthesis, especially combined with means of homonymy disambiguation (disambiguation of word forms that have the same form but different readings, e, g., *vse* 'everybody' vs *vsjo* 'everything'). Accentual information may also be used for automatic disambiguation in texts tagged with stress diacritics and in accentuated corpora.

Most text processing systems that operate with accentual information for Russian words are based on the grammatical dictionary of Russian language (GD) by acad. A. A. Zaliznyak [1]. Among these systems are morphological processors Dialing [2], Starling (Starostin˚S. A., [3]), text-to-speech synthesis system «Multifon» (Lobanov B. M, [4]), and some others. In modern corpus Slavistics the stress-tagging problem was stated in the Russian national corpus (RNC) (see [5]). RNC has accentual tagging: disambiguated texts have been automatically tagged with stresses[2].

[1] This work has been supported by Russian foundation for Basic Research (grant № 10-07-90001 Bel_a).

[2] The accentual tagging was done using programs Mystem (Yandex, [6]) and Dialing.

While working on the international project "Intellectual speech synthesis model based on deep linguistic text analysis", it was decided to enrich the morphological dictionary of multifunctional linguistic processor ETAP-3 with accentual information [7]. It was found that in order to reduce the number of errors during speech synthesis, linguistic processor ETAP-3 should be able to disambiguate homonymy, and place accents during morphological synthesis. So, among the goals of the projects was a specific task — reflecting accentuation in the morphological dictionary of ETAP-3.

2. Main features of the formal morphological model of ETAP

The morphological part of text processing systems is used for morphological analysis of the input text and for morphological synthesis of the output text. The aim of the morphological analysis is to find all possible morphological sets (lexeme names and corresponding morphological characteristics) for all word forms in the input text. The aim of the morphological synthesis is to generate the word forms from lexeme names and their morphological characteristics. To improve the performance of analysis and synthesis components, the dictionary compiler generates all possible word forms with all possible characteristics for the forms that have entries in the dictionary[3], after that the pairs «lexeme + word form with morphological features» that form the paradigm are stored in the final-state transducer [10].

Russian is well-known for its rich morphology, and manual introduction of all word forms with stresses is hardly possible. Therefore, inflection description in the morphological dictionary assumes that for each word form there should be specified (a) non-changeable **base** — a part of word that is common for the word forms in the paradigm and (b) inflectional parts of the word: **prefixes**, **themes**[4], **suffixes**, **endings** and **particles** (*-sja/-sj*). Regular sequences of inflectional parts are described using standard objects (STO). A simple STO may be a part of one or more complex STO. In most cases, to create a paradigm of a word one needs to specify the base and provide a reference to one or more standard objects. At present, Russian morphological dictionary of ETAP-3 contains about 130,000 dictionary entries and nearly 1000 standard objects.

The introduction of accentual information in the dictionary assumes highlighting of the stressed vowels with the sign of reversed accent (`), if the stress is strong, or with the tilde (~), if the stress is weak, and also the interchange between «e» and «jo». This task is far from trivial because there are three different types of stress alternations in Russian:

---

[3] In the ETAP-3 system each dictionary entry describes one lexeme.

[4] The theme is the first verbal affix after the root in morphological dictionary of ETAP-3 system, for example: *ris-**u**-ju 'drawing', ris-**ova**-nn-yj 'painted', tolk-**nu**-t' 'push'*.

- Alternations in STO, for example, in the endings:(*mope`d-**ami** 'by motorbikes' / stol-**a`mi** 'by tables'*); in suffixes (*umn- **e`jsh**-ij 'the most clever' / razu`mn-**ejsh**-ij 'the wisest'*);
- Alternations within the paradigm of the lexeme: ***kra`sn**-yi 'red' / krasn-**ejsh**-ij 'the reddest' / krasn-**a`** short feminine form for 'red'*.
- Alternations within a morpheme: *risk-**ova`**-tj 'risk' / risk-**o`va**-nn-yj 'risky'*.

Among typical Russian alternations is also «e» / «jo» alternation in morphemes (*kon-jo`m 'by horse' / lo`s-em 'by elk'*) and in paradigms (*jozh 'hedgehog' / ezha` '(without) hedgehog'*). It is almost always connected with the stress location: accentuated vowel is pronounced as *jo* and vowel without accent as *e*.

Due to diversity of such alternations (and *e/jo* alternations), introduction of the stress (and also the letter *jo*) to the bases and inflectional parts of the words will require that the existing STO should be divided and, consequently, necessitate the changes in every dictionary entry that may need years to be completed. To avoid this, a description of stress alternations that requires minimal changes in existing STO and MD entries should be used.

In GD, where declination / conjugation and accentual schemes are described separately, regularities of morphological paradigms and regularities of stress alternations within the paradigm are, generally speaking, uncorrelated. This fact allowed us to compile the morphological entry in two steps: at the first step the paradigm is constructed with no regard to the stress while at the second step the accents are assigned. The analysis of the accentual patterns has shown that in order to place the stress one needs:

- The accentual pattern of the word;
- Morphological characteristics of the word form;
- Morphemic structure of the word form (e. g., in the word form *doroga* 'road' the first five letters make up the base, and the sixth letter is the ending);
- Stress location for morphemes consisting of more than one syllable (*doro`g-a*);
- (in some cases) Letters that build up a morpheme (*vetr-y* 'winds' vs. *vetr-a* 'winds');
- (in some cases) The type of declination / conjugation according to GD.

This implies that the current set of STO can practically remain unchanged, only accents in polysyllabic morphemes should be set, and e should be replaced by jo, where needed. Then the final accent setting and the choice of e/jo will be made in the already built paradigm, so that all information required for this will be obtained from this paradigm.

3. Accentuation rules

The conditions that have to be met in order to assign the correct accentual scheme are checked by the special rules. These rules are language-independent, which allows using them for accentuation in other languages. Formal language

of linguistic rules ETAP-3 FORET was taken as a base for the formal language needed for the accentual rules.

As rules of linguistic processor ETAP-3, the accentual rules contain logical expressions that check truth of the conditions in them and the instructions that are fulfilled if the check has shown the truth of the conditions. The logical expressions contain predicates united in conjunctions, disjunctions and parentheses expressions. Rules may be united in blocks of rules. A block of rules consists of head word and then (optionally) name of block and the list of rules. Blocks that have names are called named blocks, other blocks are called unnamed blocks.

We will look at rule specification and the working algorithm in more detail.

3.1. Instructions

Accentual rules' **instructions** are of two types: stress location and stress location shift. The instructions of the first type set preliminary accents in word forms on the morpheme specified by instruction. Their names (*pref:*, *osn:, tm:, sf:, ok:* and *chs:*) correspond to the type of the morpheme that bears the stress. Stress location may be specified in the instruction by pointing out the number of the stress-bearing syllable. If the stress location within the morpheme is not pointed out explicitly, then the instruction preserves the original stress location in the morpheme. If the stress location is pointed out explicitly, then the instruction deletes the stress and sets the new stress in the position mentioned. The stress occurring in all other morphemes of the word form is deleted. If the morpheme mentioned in the instruction lacks vowels or is absent in this word form altogether, then stress is located on the last syllable of the preceding morpheme. The instruction may perform additional actions that change the appearance of the word forms, such as weak stress instead of strong stress or change of accentuated vowel (for example, replacement of the stressed *e* by *jo*[5]. The stress location change instructions make corrections to the previously located stress — they shift it a number of syllables to the left or right. If the number is positive, there is a right shift, if it is negative, then the shift is to the left.

Each instruction has a **priority —** an integer-valued characteristic that is assigned explicitly or implicitly. If during compilation a word form fulfills the conditions of several rules, then only those rules are valid that have the instructions with higher priorities.

The accentual system allows that several rules operate on the same word form, and the stress is located in various positions. In this case several copies of one word form are created, and each copy is treated by a separate rule. Such mechanism describes alternative accentuation (for example, *tvoro`g / tvo`rog 'curds', kazaki` / kaza`ki 'Cossacks'*). Each word form of the lexeme *tvorog* is treated by two rules: one sets stress

[5] In most cases *e / jo* alternations obey the following rule — «*jo* is stressed, *e* is without stress». The option "change symbol" allows describing non-standard alternation in such word forms as *izrjok — izrekshij*, where *e* is also stressed although it is not converted into *jo*.

on the ending, and the other on the base. Word form *tvorog* in nominative singular has a zero ending, that is why an accent ascribed by the second rule is moved to the last syllable of the base.

3.2. Predicates

Predicates that are part of conditions in accentual rules are divided into predicates that check morphological characteristics and predicates that check the string of letters in the morphemes. Predicates of the first type coincide with the name of the checked characteristic and are true if this characteristic is in the list of characteristics of this word form. Predicates of the second type check if there is a certain morpheme in the word form, and (if there is one) — they check a symbol string that is in the regular expression in the predicate.

3.3. The scope of the rules

The scope of the rules may vary from a unique dictionary entry to the whole dictionary. To simplify the description of the scope, the rules were divided into **general**, **template**[6] and **dictionary rules**.

General rules are applied to word forms of all dictionary entries and stored in a special file along with other language specificities.

Template rules are applied to subsets of entries, from twenty up to several thousand elements (as a rule these are entries with the same accentual schemes). They are stored in the file of standard objects under the standard objects class named *acct*. These subsets mostly correspond to the main accentual schemes from the Grammatical dictionary by acad. A. A. Zaliznyak. The entries to which the template rules are applied contain references to these templates.

Dictionary rules are applied to the specific dictionary entries and stored directly in the entries.

3.4. The algorithm of processing the accentual rules

General, template and dictionary rules are applied to the paradigm of the lexeme, consequently to each word form: at first the stress location rules, then stress location shift rules. The priority of the rules is also taken into account: at first instructions with the highest priority are applied.

[6] This type of rules has been named in the same way as the syntactic rules of a similar type in the ETAP-3 processor [3],[4].

3.5. Accentual rules. Examples

We will illustrate how the rule functions by the exampled of the lexeme *TRAKTOR* 'tractor'. This lexeme has a stress on the first vowel of the base in singular and in plural for the form *traktory*, while for the alternative plural form *traktora* the stress is on the ending, and the stress location varies in other plural forms.

The dictionary entry of the accentuated morphological dictionary for this lexeme looks as follows:

(1) ENTRY:TRAKTOR acct:c_a
base:tra`ktor f:1,end:'a'nom,pl,'a'acc,pl t:6

At the first compilation stage, the lexeme paradigm will look like:

(2) *tra`ktor — S,SG,MASC,NOM,INAN*
tra`ktor|a — S,SG,MASC,GEN,INAN
tra`ktor|u — S,SG,MASC,DAT,INAN
tra`ktor — S,SG,MASC,ACC,INAN
tra`ktor|om — S,SG,MASC,INS,INAN
tra`ktor|e — S,SG,MASC,LOC,INAN
tra`ktor|y — S,PL,MASC,NOM,INAN
tra`ktor|a — S,PL,MASC,NOM,INAN
tra`ktor|ov — S,PL,MASC,GEN,INAN
tra`ktor|am — S,PL,MASC,DAT,INAN
tra`ktor|y — S,PL,MASC,ACC,INAN
tra`ktor|a — S,PL,MASC,ACC,INAN
tra`ktor|a`mi — S,PL,MASC,INS,INAN
tra`ktor|ah — S,PL,MASC,LOC,INAN
tra`ktor|o — S,MASC,INAN,COMP

(the last line corresponds to the word form with the tag COMP, which is used in composite words such as *traktorostroenie* 'manufacturing tractors').

After that for each word form of this paradigm the conditions of those rules are checked that correspond to the nouns, in particular:

(3)
- *base:(0,"*~")=COMP+^V;* [The stress in COMP is always on the base and weak]
  acct:c_a [*tra`ktor, traktora`, tra`ktory*]
- *base:=S;*
  end:=PL+^(NOM|ACC+INAN);
  end:{2}=PL+(NOM|ACC+INAN)+search("a",end:)

The word form *tra`ktor|o* — S,MASC,INAN,COMP fulfills the conditions of the template rule *base:=S;* and the basic rule *base:(0,"*~")=COMP+^V;* While the

priority of the basic rules is higher, the word form is constructed with the instruction *base:(0,"*~"),* that changes strong stress in the base to the weak one:

tra~ktoro — S,MASC,INAN,COMP

The word forms *tra`ktor|a — S,PL,MASC,NOM,INAN, tra`ktor|a — S, PL, MASC, ACC, INAN* fulfill the conditions of the template rule *base:=S;* and the dictionary rule *end:{2}=PL+(NOM|ACC)+search("a",end:).* The priority of the dictionary rule is higher, therefore the stress in these forms will be on the endings:

traktora` — S,PL,MASC,NOM,INAN
traktora` — S,PL,MASC,ACC,INAN

The remaining word forms with characteristic PL satisfy the conditions of the template rule *base:=S;* and dictionary rule *end:=PL+^(NOM|ACC);* while the word forms with characteristic SG only satisfy the rule *base:=S*. Therefore word forms with characteristic PL will be duplicated, the stress in one copy will be placed on the first vowel of the base, and the stress in the second copy will be placed on the ending. Word forms marked with SG fulfill only the rule "base:=S", and the stress will be placed on the first vowel of the base.

4. Automatical introduction of accentual information to MD of ETAP-3

Introduction of accentual information to MD will be done mostly automatically. Rules for accent setting in accordance with accentual schemes used in GD were written and tested. The correspondences between MD and GD entries were established. The last step of introduction consists in supplying MD entries with accentual rules that correspond to the accentual schemes in GD entries.

4.1. Working on main accentual rules

While writing the rules describing accentual schemes used in GD, some problems had to be solved. Most of them are concerned with systematic discrepancies between MD and GD dictionaries.

First, the different aspect verb forms in MD are usually merged into one entry, while in GD they are always separated. To place the accents correctly two rules were used. Each rule was applied only to the word forms of one aspect (it was done so by mentioning the aspect in the rule conditions).

The second systematic difference between the dictionaries is the lack of the comparative and superlative degrees in the paradigm of the adjectives and adverbs in GD.

Accentual schemes for adjectives in GD were supplied with additional accentual rules for the comparative and superlative degrees and adverbs were also fully accentuated.

When we studied the group of these word forms, we could establish a new regularity: adjectives with the stress located on the base (this is one of the accentual schemes) retain it in the forms of superlative. The only exception seems to be *boga`t-yj — bogat-e`jsh-ij*. Special complementary rules were added to the adjective templates for correct accentuation in the comparative and superlative degree forms in MD.

More problems arise in case of several GD entries being merged in one MD entry because of synonymy (GD entries *chitat' 'read'/ prochitat' 'have read'/ prochest' 'have read'* correspond to MD entry *CHITAT'*, GD entries *povorachivat' 'turn'/ povertyvat' 'turn'/ povernut' 'have turned'/ povorotit' 'have turned'* correspond to MD entry *POVORACHIVAT')*. In such entries the word forms with the same sets of morphological characteristics often correspond to different accentual schemes. This may complicate the conditions of the rules in accentual schemes. To solve this problem we have inserted a special separator that divides morphological dictionary entry into parts that correspond to GD. These parts include corresponding STO and stress location rules, therefore a separate compilation and stress location is possible, and after that separate word forms are merged into one paradigm.

4.2. Working on dictionary entry correspondence tables

At present, the table of corresponding accentual schemes and rules and the table of corresponding entries have been made. For the latter table a program was created that finds the correspondence between MD and GD entries, between such morphological characteristics as part of speech, gender, animacy, and so on. There are nominal lexemes in GD that have both masculine and feminine and animated and non-animated forms in the (*zanuda 'bore'* ma//fa, *mikrob 'microbe'* m//ma), while in MD it is not so. Conversely, in MD many verb entries merge verbs of two aspects. The found pairs were combined into larger groups.

Morphologically homonymous MD entries (for example, DERZHATEL' 'a man / a device', or USTANOVKA — 'installation' — 'action' vs. 'object') were combined in a table of «many-to-many» correspondence. The created tables need post-editing for making the correspondences «one-to-one».

Accentual information from GD was transferred to those entries in MD that have been put in the «one-to-one» correspondence table. During this transfer in the dictionary entry of MD: 1) an accentual rule was assigned that corresponded to the accentual scheme from GD; 2) a base was accentuated, and the stress location was defined according to the entry in GD. As a result, 65 000 entries were given accentual information (there are about 129 000 entries in MD altogether). The transfer of accentual information from other tables is also in process now.

Words that have not been put in the tables were accentuated manually. Among these words are 10 000 adverbs, the most part of which is absent in GD.

4.3. Creating the set of accentual rules for adjectives

The main principles of accentual rules may be shown on the example of accentual rules for adjectives. In GD there are 12 most frequent accentual schemes: *a, a', a/b, a/b', a/c, a/c', a/c"; b, b', b/c, b/c', b/c"*. The first letter shows accentual scheme for full forms, the second letter — for short forms. If the second letter is the same as the first, it must be omitted.

Accentual scheme *a* means that word forms have stress on the base and accentual scheme *b* means that stress is always on the ending, excluding comparative and superlative that are not part of the paradigm in GD. Scheme *c* (only for short forms) means that accent falls on the ending in feminine singular forms. According to these schemes accentual rules may be built.

At first accents for full and short forms as in GD are created:

Rule adj_a (corresponding to scheme *a*): the accent falls on the base

(4) *acct:adj_a [suro`v-yj 'severe', udo`bn-yj 'comfortable']*
base:=A;

This rule is applied to adjectives in MD that have the scheme a in GD.

Rule adj_a1 (corresponding to scheme a'): the accent falls on the base while for BREV+FEM accent also falls on the ending:

(5) *acct:adj_a1 [vla`stn-yj 'powerful' (vla`sten, vla`stn-a/vlastn-a`, vla`stn-o, vla`stn-y) end:=BREV+FEM;*

base:=A;

Rule adj_ab (scheme a/b) prescribes that the accent is placed on the ending in short forms and on the base in all other forms:

(6) *acct:adj_ab [zdoro`v-yj (zdoro`v, zdorov-a`, zdorov-o`, zdorov-y`)]*
end:=BREV;
base:=^BREV;

Other rules are listed without detailed description:

(7) *acct:adj_ab1 [sve`zh-ij 'fresh' (sve`zh, svezh-a`,svezh-o`, sve`zh-i/svezh-i`); scheme **a/b']***
end:=BREV;
base:=^(SG+BREV);

*acct:adj_ac [tse`l-yj 'whole' (tse'l, tsel-a`, tse`l-o, tse`l-y); scheme **a/c]***
end:=BREV+FEM;
base:=^(BREV+FEM);

*acct:adj_ac1 [ми́л-ый 'dear' (mi`l, mil-a`, mi`l-o, mi`l-y/mil-y'); scheme **a/c']***

end:=BREV+(FEM|PL);
base:=^(BREV+FEM);

*acct:adj_ac2 [be`l-yj 'white' (bel, bel-a`, be`l-o/bel-o`, be`l-y/bel-y`) scheme **a/c"**]*
end:=BREV+^MASC;
base:=^(BREV+FEM);
*acct:adj_b [smeshn-o`j 'funny' (smesho`n, smeshn-a`, smeshn-o`, smeshn-y`); scheme **b**]*
end:=A;

*acct:adj_bc [zhiv-o`j 'alive' (zhiv, zhiv-a`, zhi`v-o, zhi`v-y) ; scheme **bc**]*
base:=BREV+(NEUTR|PL);
end:=^(BREV+(NEUTR|PL));

*acct:adj_bc1 [skup-o`j 'stingy' (skup, skup-a`, sku`p-o, sku`p-y/skup-y`); scheme **bc'**]*
base:=BREV+(NEUTR|PL);
end:=^(BREV+NEUTR);

*acct:adj_bc2 [dryann-o`j 'bad' (drya`nen, dryann-a`, dryann-o`/ drya`nn-o, dryann-y`/drya`nn-y) ; scheme **bc"**]*
base:=BREV+^FEM;
end:=^(BREV+MASC);

For the comparative and superlative word forms, general rules are written (because they are applied to all schemes excluding ***a***):

- a rule that places stress on the base in the comparative forms (*bo`l'-she 'bigger', glu`b-zhe 'deeper'*)

(8) *base:=COMPAR+search("^e$",sf:);*

- a rule that places stress on the suffix in the comparative forms (*smel-e`e 'more boldly', vesel-e`j 'funnier'*)

(9) *(9)* sf:=COMPAR+search("^é?[eй]$",sf:);[7]

- a rule that places stress on the suffix in the superlative forms (*velich-aj`sh-ij 'the greatest', umn-e`jsh-ij 'the cleverest'*)

(10) *sf:=SUPER*.

[7] The regular expression "^é?[eй]$" means that the first character of the line has to be *e* followed by an optional stress symbol, while the last character of the line is *e* or *ŭ*.

These rules are applied to almost all adjective accentual schemes (11 in GD), except scheme *a*, where the stress is always on the base. For this reason, a template rule for *a*-adjectives should have a higher priority (*acct:adj_a base:{2}=A*).

By default, the general rules have a higher priority than the template rules. That is why the instructions *adj_a — adj_bc2* are not applied to the comparative and superlative word forms, even if these word forms meet the conditions of these rules. This is true for adjectives with accentual schemes ***a' — b/c"***, and not true for adjectives with accentual scheme ***a,*** where the accent falls on the base also in the comparative and the superlative. To cancel the action of the general accentual rules mentioned above, the template rule for scheme ***a*** is assigned a higher priority than that of the general rules:

(11) *acct:adj_a [суро́в-ый, удо́бн-ый]*
*осн**:{2}**=A;*

However, among the adjectives of accentual scheme a, there is an exception: the word *bogaty 'rich'* that has a superlative *bogat-e`jsh-ij* with the stress on the suffix, rather than on the base. For a correct description of the word forms in the superlative, the dictionary entry:

(12) *ENTRY:BOGATY acct:adj_a **acct:sf:{3}=SUPER;***
base:boga`t no:compar t:211 har:A,compar,osn:boga`che
har:A,compar,att,base:poboga`che

This rule has a higher priority yet, 3, whereby the accent set for scheme a of a word *bogatyj* by the rule *acct:adj_a* is suppressed.

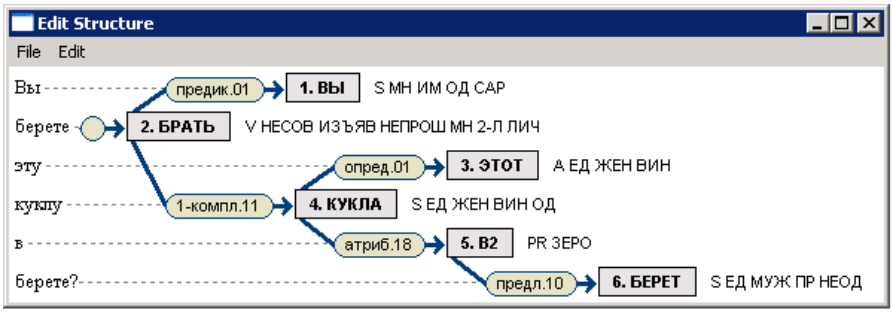
5. *Morphological analysis and synthesis using accentual information*

New requirements to the morphological component of ETAP-3 must be met while adding accentual information:

- morphological analyzer must recognize word forms from the input text both with accents and *jo* and without;
- morphological analyzer must disambiguate strictly and operate with texts that have consequent *jo* distinction. This mode should not confuse *e* and *jo*: (*on osel 'he has collapsed'* ≠ *on osjol 'he is an ass'*);
- morphological synthesizer must generate output text both with *jo* and accentuation and without it;
- when word forms with alternative accentuations are produced (for example, *profe`ssoram / professora`m '(to) professors'*), it should be possible to have the most useful word form (implicitly) or the word form you need.

Let us look at the new possibilities of ETAP-3 with accentual information in more detail. The syntactic analyzer is used for disambiguation also in a special pre-processing mode of speech-synthesis. Phrases are assigned with syntactic structures and disambiguated in this mode. They are also accentuated and sent to speech synthesizer.

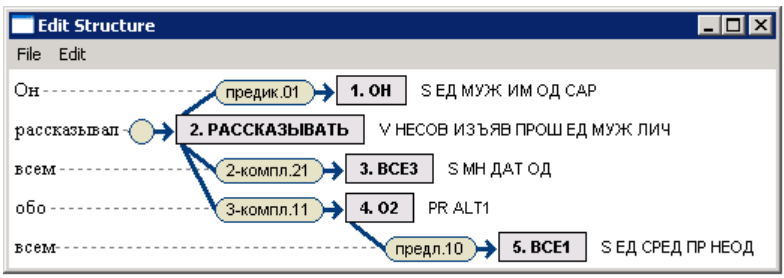
For the phrase «*Vy berete etu kuklu v berete*?» 'Are you taking this doll in a beret', the text-to-speech synthesizer Multifon will mistake the second occurrence of the word form *berete* 'beret' for the verb *berjote* 'take' But with the ETAP-3 preprocessing mode activated we will have the correct syntactic structure of this sentence:



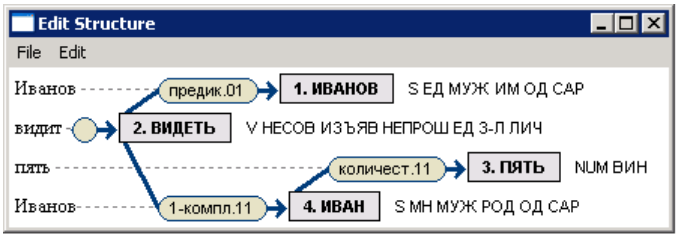
Pic. 1. Phrase "Vy berete etu kuklu v berete?"

Then the morphological synthesizer produces accentuated and disambiguated sentence from this structure: *"Vy` berjote e`tu ku`klu v bere`te?"*.

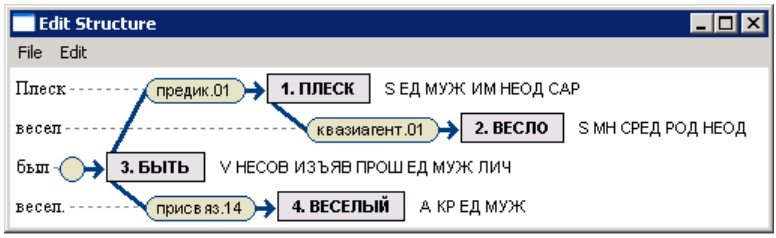
Below are several additional examples of homonymy in phrases: 1) *On rasskazyval vsem obo vsem* 'He told everybody about everything'*;* 2) *Ivanov vidit pjatj Ivanov* 'Ivanov sees five Ivans'*;* 3) *Plesk vesel byl vesel* 'The splash of the oars was merry'. ETAP-3 produces correct syntactic structures:



Pic. 2. Phrase "On rasskazyval vsem obo vsem"



Pic. 3. Phrase "Ivanov vidit pjatj Ivanov"



Pic. 4. Phrase "Plesk vesel byl vesel"

The morphological synthesizer produces accentuated disambiguated texts for these structures: *"O`n rasska`zyvajet vse`m o`bo vsjo`m"* / *"Ivano`v vi`dit dvu`h Iva`nov"* / *"Ple`sk vjo`sel by`l ve`sel"*. This text is then sent to the speech synthesizer.

6. Conclusion

As a result, in place of the morphological dictionary of ETAP-3 system we will have a fully accentuated large-size Russian morphological dictionary supplied, among other things, with completely formalized rules for creating accentual paradigms of new lexemes. Another important feature of this dictionary is a sufficiently full accentuation coverage of degrees of comparison for adverbs and adjectives. The accentual data on these word forms in other dictionaries is rather scarce. The accentual information from the dictionary will allow one to use the syntactic analyzer for the disambiguation during speech synthesis. The information on the accents may also be used for accentual tagging of the Syntactic corpus of Russian language (SynTagRus).

The main factor that facilitated the introduction of accentual information into the large dictionary was the creation of a formal language for describing this information. This language is entirely independent from the formalism used for the description of paradigms. The mechanism of automatic transfer of the data from GD into MD was essential, too. Such formal language may be effectively used for the introduction of accentual information into morphological systems for Russian that use the formalism different from that of GD. Due to its maximum linguistic

independence, this formal language may be used to create accentual rules for other natural languages besides Russian.[8]

References

1. *Apresian Iu. D, Boguslavskii I. M., Iomdin L. L., Lazurskii A. V., Mitiushin L. G., Sannikov V. Z., Tsinman L. L.* 1992. Linguistic Processor for Complex Informative Systems [Lingvisticheskii Processor dlia Slozhnykh Informatsionnykh Sistem].
2. *Apresian Iu. D., Boguslavskii I. M., Iomdin L. L., Lazurskii A. V., Pertsov N. V., Sannikov V. Z., Tsinman L. L.* 1989. Linguistic Supply for ETAP-2 [Lingvisticheskoe Obespechenie Sistemy ETAP-2].
3. *Grishina E. A.* 2009. The "History of Russian Accent" Corpus [Korpus «Istoria Russkogo Udarenia»]. Natsionalyi Korpus Russkogo Iazyka. Novye Rezultaty i Perspektivy.
4. *Iomdin L. L., Lobanov B. M.* 2009. Syntaxic Correlates of Prosodic Marked Sentence Elements and its Role in the Synthesis of Text-to-speech [Sintaksicheskie Korreliaty Prosodicheski Markirovannykh Elementov Predlozheniia i ikh Rol' v Zadachakh Sinteza Rechi po Tekstu], available at: http://www.dialog-21.ru/dialog2009/materials/html/23.htm
5. *Kazennikov A. O.* 2008. The Use of Final Automatons dor Morphological Analysis and Synthesis basing on ETAP Dictionaries [Ispol'zovanie Konechnykh Avtomatov dlia Morfologicheskogo Analiza i Sinteza na osnove Slovarei Sistemy ETAP]. Sbornik Trudov 31 Konferentsii Molodykh Uchenykh i Specialistov IPPI RAN "Informatsionnye Tekhnologii i Sistemy " (Proc. of the 31 Conference "Information Technologies and Systems"): 201–205.
6. *Lobanov B. M.* 2007. «Muiltifon» — a Personalized Text-to-speech Synthesis System for Slavic Languages. Linguistic Polyphony : 849–866.
7. *Segalovich I.* A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine. MLMTA-2003, available at: http://download.yandex.ru/company/iseg-las-vegas.pdf
8. *Sokirko A.* 2001. A Short Description of Dialing Project, available at: http://www.aot.ru/docs/sokirko/sokirko-candid-eng.html
9. *Zalizniak A. A.* 2007. Russian Grammar Dictionary [Grammaticheskii Slovar' Russkogo Iazyka].
10. *www.starling.rinet.ru*

[8] It is evident that the introduction of accentual information into the morphological dictionary is on the one hand useful for the language orthographies in which the letter structure is close to the phonemic structure. In this case, the information on the word's letter structure and the stress location is sufficient for speech synthesis On the other hand, identifying the stress location in this language should not be too trivial, like in Spanish, where the stress location is determined by means of simple rules and is explicitly marked in exceptions. Among such languages with non-trivial stress location are besides Russian, Byelorussian, Ukrainian and most probably German.

РАЗРЕШЕНИЕ АНАФОР ЛИЧНЫХ МЕСТОИМЕНИЙ ТРЕТЬЕГО ЛИЦА В ТЕКСТАХ УЗКИХ ПРЕДМЕТНЫХ ОБЛАСТЕЙ С ГРАММАТИЧЕСКИМИ ОШИБКАМИ И ОПЕЧАТКАМИ

Д. С. Скатов (ds@dictum.ru)

С. В. Ливерко (liverko@dictum.ru)

Dictum Ltd, Нижний Новгород, Россия

Ключевые слова: анафора, разрешение анафор, местоимение, ошибка, опечатка.

ANAPHORA RESOLUTION OF THE THIRD-PERSON PRONOUN IN TEXTS FROM NARROW SUBJECT DOMAINS WITH GRAMMATICAL ERRORS AND MISTYPINGS

D. S. Skatov (ds@dictum.ru)

S. V. Liverko (liverko@dictum.ru)

Dictum Ltd, Nizhnii Novgorod, Russian Federation

The third personal pronoun anaphora resolution in texts from the Internet sources (forum comments, opinions) with a given subject domain (cars, household appliances etc) is being discussed. A concrete solution to the task is offered. High precision with acceptable recall (and vice versa) is shown by an example of opinions about mobile phones.

Key words: anaphora, anaphora resolution, pronoun, error, mistyping.

1. Introduction

The problem of the third personal pronoun anaphora resolution discussed in this paper consists in the replacement of pronouns such as *"he", "his", "her", "it", …* with nouns (antecedents) that these pronouns were used instead. Its solution is needed firstly in text mining applications, such as opinion mining (about goods, people) or fact extraction. Without resolved anaphoras those applications lose in recall of their results. The loss degree depends on the type of proceeded texts: e. g., in opinions about goods the density of *"it"* (masculine gender in Russian) pronoun is 1,5 times higher than in news[1].

The known methods of anaphora resolution can be divided into two groups — (1) statistical and (2) syntactical. Methods from class (1) [3] are based on the results of machine learning and are potentially applicable to texts of significantly different nature. Class (2) [1,2] exploits the sentence syntactical parsing tree (or semantic graphs as their derivatives) and as a result the applicability of such methods is limited to relatively «correct» texts (e. g., dossier texts [2]). This article describes a method combining these two approaches in a certain sense.

Texts from «real life» are full of typos and specialized slang with their grammar far from correct one:

(1) Ive got a **whit** case and **butons** peel **gradauly** and they become**s grey** no cleaning helps or anything **likethat**. Weak processor also made upset as well as small memory amount, it works terribly **slo**.

The method of anaphora resolution, offered by the authors, takes mistypings and the results of syntactic parsing of text fragments (with mistypings corrected) into account. It is adapted to process texts from specific subject domains. Method can work with «correct» texts as well as informal ones (such as opinions or notes). To achieve a high processing quality for texts from a selected domain, a preliminary adjustment to the method is needed. It consists in learning on an unmarked corpus and composing the operating terminological dictionaries.

Three modes of the method have been implemented:

(A) good precision (70–80 %) with high recall (90–95 %),
(B) approximately equally good precision and recall (75–85 %),
(C) excellent precision (up to 95 %) with high acceptable recall (40–50 %).

The implementation of the technology is represented by a software module called DictaScope Anaphora. It is adjusted to processing opinions about mobile phones from Internet sources. Within the bounds of the article an estimation of recall-precision ratio for processing such kind of data is carried out. The model is being used in the real application for online opinion monitoring. Modes A, B and C were obtained in the process of looking for a solution effective for this application — i. e. the one with high precision on possibly intentionally reduced nput data.

[1] A random sample of news from [12] (the anaphora density — 0.34 per 1 K) and a sample of opinions about mobile phones from the sources such as [13] (the anaphora density — 0,53 per 1 K) were used to perform measurements, each one of 1 Mb

2. Problem statement

Basic statement. For each pronoun pr_i, $i = 1, ..., N$ from text T choose the resolving pronoun (antecedent) a_i. *Remark.* In certain cases it is impossible to choose a_i, e. g.:

(2) This mobile phone has a sensor screen. It's very inconvenient. *(screen or phone?)*

Resolving of such an ambiguity (which can conditionally be called semantic) is a hard task even for a human, as both variants are of equal possibility. In the current problem statement it is offered either to choose a concrete antecedent or not to resolve the anaphora.

Advanced statement. It sometimes turns out that an acceptable precision of selecting a sole variant is unreachable. Therefore the following task specification is proposed: for each pronoun pr_i, $i = 1, ..., N$ form a list of possible resolving variants (a_i^1, ..., $a_i^{l_i}$) sorted in accordance with their ranks (the first one is the best). Then a_i^1 can be chosen as a_i. In case a requirement of a high recall takes place (e. g., for posterior hand processing of results) it is sufficient to ensure high quality of ranking.

The variants of resolving antecedents can be supplied with real-value weights $w = w(a_i^k) \in (0,1]$, $i \in \{1, ..., N\}$, $k \in \{1, ..., l_i\}$, which correspond to each variant's confidence.

Traits. Let's resort to an example to make the task statement clear:

(3) bought it for business, very useful because [it] {* = 0.652166, business = 0.2371, NULL = 0.168611} supports two sim cards. Nice, big display, no dead spaces fount on [it]{display = 0.466248, * = 0.284525, NULL = 0.0777368, business = 0.0101848}

For pronoun pr_1 = «*it*» the list of variants is formed (a_1^1 = «*», a_1^2 = «*bussiness*», a_1^3 = «*NULL*») with weights $w(a_1^1) \approx 0.65$, $w(a_1^2) \approx 0.237$, $w(a_1^3) \approx 0.1686$ (similarly for pr_2 = «*it*»). There are also special «*» and «*NULL*» designations:

- «*» — «the current object of discourse», so-called «implicit» antecedent. This is typical for opinions and reviews — i. e. for texts representing direct speech in writing. In the example above the word «*phone*» (as well as its concrete model reference) is not found anywhere before pr_1 = «*it*», though the teller means exactly «*this phone*».
- «*NULL*» — a directive «not to resolve pronoun». If «*NULL*» is at first position in the list of variants, the pronoun is left unresolved.

Thus, there are two cases in a basic problem statement in which the anaphora will not be resolved:

1) No variants for pronoun resolution is found;
2) «*NULL*» is the first in the ranged list of variants. It is easy to see that if, in case of semantic ambiguity, the probability of the correct choice of antecedent is less than ½, the precision will not fall on the average. Therefore, in this case the choice of «*NULL*» variant is justified.

In the example (3) the task in the basic statement is resolved correctly by choosing the first variant for each pronoun. A solution in a basic statement will be further estimated.

3. Review

The subject area of this paper is covered in the works of three Russian groups.

1) Ermakov A. E., RCO. In [2] empirical regularities of persons referencing are shown for texts from Russian mass media; they can be used to build a mechanism for anaphora resolution in text sources of this class (with the help of natural language syntactic parser).
2) Tolpegin P., Vetrov D., Kropotov D. Article [3] describes an experience of this group in resolving the third personal pronoun anaphora in news by machine learning methods. The approach is typical for this type of solvers, the precision shown equals 62 % on a control collection.
3) Okatiev V., Erechinskaya T., Skatov D. In the report [1] it is shown how pronoun anaphoras of different types can be resolved with the help of syntax parsing trees analysis. This approach is well applicable to the texts in which most of the sentences allow building correct syntax trees.

The specificity of this article — processing texts from narrow subject domains with mistypings and slang — is not touched upon in the works listed above.

The question discussed is more widely represented in foreign scientific works:

- from English-speaking authors patented system [11] and work [8] (which demonstrates values of basic indicators at a level about 80 % while using probability model) are first to be mentioned;
- authors of [9] use maximum entropy method to resolve the third personal pronoun anaphora in Chinese, with F-measure about 70 %;
- [10] describes an application of machine learning to personal pronouns anaphora resolution in Turkish with recall-precision at about 60–70 %.

The overall impression of these works is the following: competent combination of analysis methods and rather full vocabulary data results in recall-precision not less that 70 %.

4. Solution

4.1. Lists of variants and attributes

After tokenization (when the lists of grammar values of the tokens are supplemented taking mistypings into consideration) and dividing text into “conditional” sentences all the pronouns are looked through in the text from left to right.

A concrete pronoun *pr* is fixed, $i = 1, ..., N$, and list var(*pr*) of possible antecedents is formed:

1) from all the words located within $\mu = 2$ sentences to the left of *pr*, nouns in concordance with pr_i by gender and number are selected;
2) from the same words pronouns which are in concordance with *pr* by gender and number are selected and the list var(*pr*) is supplemented with nouns that resolve these pronouns.

Possible antecedents can also be found **to the right** of *pr*; however, not more than 30 examples of this were found in the corpus, with the correct variant also found to the left of *pr* in ⅓ cases. Therefore, the possible variant location to the right is ignored by the method.

The proposed scheme has a chain character: pronouns on the left of given *pr*, which are close to it and already resolved, add antecedents which are located to the left of the boundary of the window $\mu = 2$ to var(*pr*). The scheme presents a certain compromise: the list can be imprecise but var(*pr*) remains quite compact. Advancing the window border μ up to 5 with the chain scheme disabled has led to a noticeable decrease in the solution precision during the experiments, so the decision was made to reject the varying left border.

For the further ranking of the lists var(*pr*) a vector of attributes $A(a)$ is calculated for each $a \in \text{var}(pr)$. Let us mention the following attributes from the operational ones:

- *IsVoc* $\in \{0,1\}$ — the belonging of a to a terminological dictionary *TermVoc*;
- *Freq* $\in \mathrm{N} \cup \{0\}$ — the number of mentionings of the given word (in any form) to the left of *pr*;
- *Dist* $\in \mathrm{N}$ — the distance between the pronoun *pr* and the position of a inside the text (measured in words);
- *IsVerb* $\in \{0,1\}$ — the presence of direct father in a form of verb in syntax tree for a fragment containing a;
- *NumNodes* $\in \mathrm{N} \cup \{0\}$ — the number of nodes in a bush subordinate to a.

The last two attributes have been introduced based on exploring correlation between numeric properties of a tree and resolving antecedents. For example, greater *NumNodes* were often correspondent to proper variants of resolution. These attributes values are set into null in case the tree was not formed.

The distance is measured in words for a number of reasons: (a) to get a valid syntactical unit (clause, noun phrase) was not possible (at that moment) due to the laboriousness of the adaptation of the syntactical parser to the special features of input texts (e. g. the absence of punctuation); (b) a paragraph is too large for being a unit of measure — the majority of opinions consist of one paragraph; (c) windows are measured in sentences and a two-sentence diapason is considered to be sufficient for the research.

IsVoc attribute implements the following idea: taking a subject domain's specificity into account allows to obtain higher quality of analysis. In fact, *IsVoc* allows to raise the priority of variants relating to subject domain of the text — they are of most interest (not always, though).

4.2. The test corpus

To evaluate the work of the methods a corpus of 3M was built from opinions about mobile phones from the sources like [13, 14, 15]. Due to the specificity of the application the corpus was additionally divided into three groups: positive, negative and neutral opinions, each of 0.8–1.2 M. As a next step it was marked up with the resolved anaphoras according to the following scheme:

- if the correct antecedent could be chosen directly from the text, its occurrence which was closest to the left of the pronoun being resolved was marked in a special way;
- in case of semantic ambiguity the pronoun was marked with *«NULL»* variant;
- the resolving word was written next to the pronoun in the corresponding case.

The statistical characteristics of the corpus were estimated.

- The whole number of 8.3 thousand opinions formed of 37 thousand unique word forms (including mistypings).
- The most frequent opinion length varying from 15 to 35 words; average opinion length — 54 words; the bulk of the opinions containing 10 to 90 words; opinions of more than 100 words are rare. The length scatter — from 2 to 340 words (Pic.1).
- Opinions consisting of one sentence are the most frequent; average opinion length — 4 sentences. The majority of opinions include 1 to 16 sentences; lengths more than 24 sentences are very rare (Pic.2).
- The corpus contains about 6.2 thousand third personal pronouns, including 4.5 thousand ones of masculine gender, 0.8 thousand of feminine gender, 0.7 thousand of plurals. The reason for a great number of masculine pronouns is the subject of the opinions (mobile phones).
- Less than 50% of the opinions do not contain any of the pronouns under research. 35% contain only one pronoun, about 10% — two of them. The maximum is 9 pronouns per opinion (Pic.3).

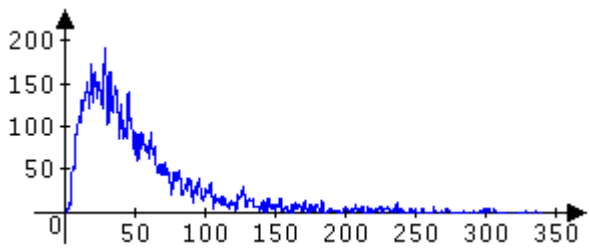


Fig. 1. Distribution of opinions lengths in words

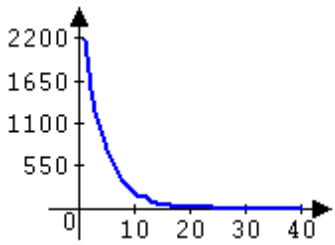


Fig. 2. Distribution of opinion lengths in sentences

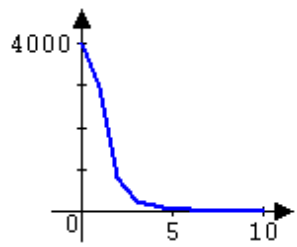


Fig. 3. Opinions distribution by a number of pronouns

4.3. Lexicographical analysis method

At the initial stage of studying a heuristic method for the options ranking was implemented:

- a system of priorities is formed on the set of attributes, which were listed in subparagraph 4.1;
- attribute values for each option are sorted according to the priorities;
- options are sorted lexicographically according to their sets of attributes.

The method resolves all the anaphoras for which it has found variants to the left with precision rate not more than 60%. The experiments in introducing new attributes and varying their priorities were not efficient. This has led the authors to the idea of filtration of the input data in order to achieve higher precision rate.

4.4. SVM-method based on machine learning

Let there be a general set of objects Ω, divided into previously unknown classes, and a sample set $O \subset \Omega$, for each element of which its class is known. The task of classification is to answer the question: *"which class does each object* ω *from* Ω *belong to"*, knowing only the sample set O (or the probabilities of belonging).

Let us fix a list $\mathrm{var}(pr_i)$ for one specific pronoun pr_i. In this case $O_i = \{A(a) \mid a \in \mathrm{var}(pr_i)\}$, $i = 1, ..., N$, and two classes are of interest — "are antecedents" and the inverse to it. Then the first class distance can be taken as $w(a)$.

Now we need to generalize the approach for N pronouns. Each set O_i represents an independent group, each of which consists of two classes — *"is the antecedent for* pr_i*"* and the inverse one, $2N$ classes for the whole training set. It is impossible to use this classification in practice with a different number $Q \neq N$ of *other* pronouns. In order to get exactly two classes for any number of pronouns, it is necessary to construct an acceptable combination of these groups. For this purpose, the authors propose adding attributes characterizing the group to each set $\omega_i \in O_i$. Thus within the same group all its members are additionally provided with

the same set of numbers describing the group. The centroid can be taken as these numbers.

After expanding of the group members a sample set $\overline{O} = \mathbf{U}_{i=1}^{N} O_i$ with the corresponding universe $\overline{\Omega}$ and a fuzzy classifier $K(\omega) \in (0,1]$ which determines a distance between ω and the class *"are antecedents"* are constructed.

$K(\omega)$ is constructed in a form of so-called probabilistic decision function as described in [5,6] based on a classical C-SVM with a nonlinear kernel [7]. Selection of the core and the constants for the SVM was performed by minimizing the overtraining on the parameters grid while verifying the recall-precision ratio on the training and control samples. In the end, the kernel was chosen to be a polynomial one with a small degree.

Centroids raised the precision of the SVM-method from 70% to 80% (mode A).

4.5. Recall-precision regulator

To reach the precision rate of 90% linear discriminative analysis [4] was used: its aim is to find a line between classes, in the projection on which they are most discernible. With the help of discriminant, pronouns which may be not resolved (for the purpose of rising the precision rate) were identified. The combination of this filtration and SVM-method allowed to reach the desired result (mode C). Along the way, it was managed to derive mode B in which basic rates are balanced in the region of 75–85%.

5. Analysis of the results

5.1. Quality requirements and evaluation

Processing of the input set containing L third personal pronoun anaphoras is carried out in 2 steps.

Filtration of anaphoras. From the total number of L objects those for which the algorithm: (1) failed to form the set of variants, (2) put *«NULL»* in the first place in the list of variants or (3) eliminated from the examination due to regulator work are deleted. As a result, N anaphoras are left, for each of them the algorithm can choose an antecedent (not necessarily the correct one). If the whole of L anaphoras resolved *correctly* are considered as relevant, the recall rate of this step is N/L while the precision is equal to 1, as all chosen objects (N) are included in the relevant (L).

Resolution of the left anaphoras. In this step the whole of N anaphoras resolved correctly are considered as relevant. The algorithm attempts to resolve them, succeeding in K cases. Due to the coincidence between the volumes of relevant objects and those being resolved, the precision and recall rates are both equal to K/N.

Two out of four rates mentioned above (precision and recall for each step) are informative:

- recall is a portion of pronouns for which the algorithm succeeded in finding an antecedent;

- precision is equal to a percent of this portion containing correctly identified antecedents.

To the writers' opinion, this approach to evaluation conforms to the quality requirements. In addition, the estimations do not depend on the mechanism of anaphora resolution (including the size of variant lists).

5.2. The quality of SVM-method and sensitivity to the sample volume

Opinions containing at least one of the pronouns under research (4 thousand altogether) were selected from the corpus. To evaluate the SVM-method sensitivity to the sample volume this set of opinions underwent the procedure of q-fold cross validation.

Verification was carried out for $q = 1, \ldots, 300$, i. e. $q = 1$ means verification of the model for the whole 4 thousand opinions, $q = 300$ — for a sample of 13 opinions. For each $q$the mean of recall and precision was calculated for each iteration as well as their minimum and maximum for the diagrams reflecting the dependency between quality and the volume of input data.

Measuring was conducted for modes A, B and C (Pic. 4, abscissa corresponds to q).

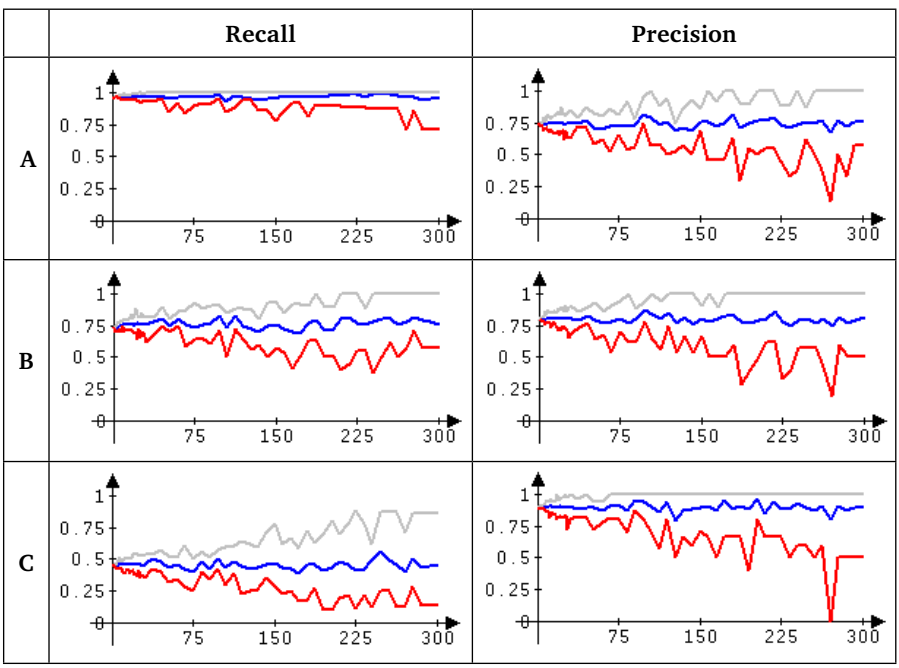


Fig. 4. Results for SVM-method cross-validation in A,B,C modes

It can be seen that all the means are stable even for small-sized samples.

Table 1. Averaged quality measures for SVM-method

	Recall	Precision
A	97.3%	74.2%
B	75.4%	80.7%
C	45.6%	90.3%

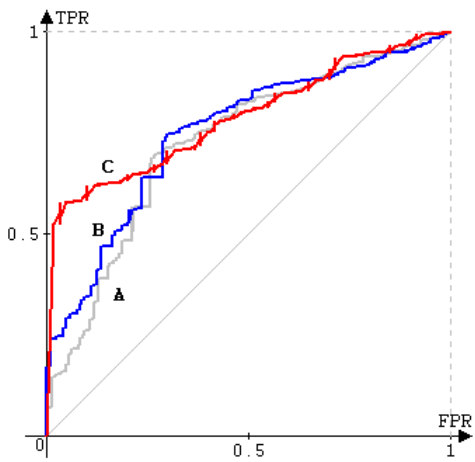


Fig. 5. ROC-curves for SVM-method in A,B,C modes

The area under **A** curve is 0.74, under **B** one — 0.76, which is considered as "good" according to the expert scale. The area under **C** curve is 0.81 with this mode considered as "very good".

5.3. The SVM-method independence of the sentiment of the corpus

It was additionally verified in empirical way that the SVM-method is independent of the sentiment of the texts processed, since it cannot be forgotten that anaphoras in negative opinions might be different from those in positive opinions.

The "negative" corpus was used as a training set, the "positive" one as a control set.

Table 2. Check for SVM-method independency from sentiment

(Recall %, Precision %)	(A)	(B)	(C)
Negative *(training)*	(95.1, 80.2)	(77.8, 86.7)	(43.1, 93.2)
Positive *(control)*	(96.3, 78.7)	(79.1, 83.4)	(56.2, 89.9)

5.4. Significance of the factors

Discriminative analysis provides an estimation of contribution of the attributes to the common decision — the judgment can be made based on the coefficients for the corresponding attributes in the linear discriminant and the range of attribute values. It is also possible to estimate how much influence components of the centroid bring to the solution.

According to the Table 3, the frequency is two times more important than the distance, the presence of a father-verb is more important than the number of nodes in the bush (even if correcting this by a wide range of *NumNodes* — sometimes up to 10–15 knots). Picture according to the centroid is consistent on a whole, except for *IsImp* and *IsVoc*, so their contribution can be estimated to be approximately equal.

Table 3. Valuing the attributes significance according to the results of discriminant analysis

Attribute	Coefficient in linear discriminant	Corresponding coefficient near the component of the centroid
IsImp ∈ {0,1}	– 2.9	18.8
IsVoc ∈ {0,1}	9.3	1.1
HasVerb ∈ {0,1}	– 7	35.8
NumNodes ∈ N ∪ {0}	– 0.5	18.9
Freq ∈ N ∪ {0}	– 21.5	-1.6
Dist ∈ N	– 10.6	0.1

Compiling vocabularies for *IsVoc* is rather laborious. The authors have discovered that the main coefficients in modes A and C (recall and precision respectively) reduce from about 90 to 70 % when this attribute is not used; in mode B both coefficients reduce by ~10 %. It can be stated that it is precisely *IsVoc* attribute that allows to achieve the precision rate of 90 % and higher.

5.5. Evaluation of lexicographical method

The advantage of this method is that no marked-up corpus is needed for its initialization. The practical use of the SVM-method has shown that a trained classifier copes with texts from domains different from that of the training set with the rates declining by several procents (with the exception of *IsVoc* attribute — new vocabularies are needed).

Table 4. Estimation of the lexicographical method quality

	With IsVoc	Without IsVoc
(Recall %, Precision %)	(93.7, 51.9)	(93.7, 42.4)

The main error f the method is an excessively strong influence of an attribute with the highest priority. E.g. using *IsVoc* attribute often results in an incorrect choosing a vocabulary word while not using it — in choosing the word closest to the left.

6. Conclusion

This paper offers a solution to the problem of the third personal pronoun anaphora resolution. The software complex called DictaScope Anaphora was implemented based on the models and methods discussed in this paper. It has the following characteristics:

- there are three modes, which allow to achieve both recall and precision rates of 80 % or to give preference to one of them and achieve the result of 95 %;
- it is possible to take mistypings and grammatical errors into account, which is important for processing texts from online sources (such as reviews);
- in this case an adjustment of the parameters for a specific subject area is needed.

The features of the internal structure of the system and the mathematical foundation are described; the detailed evaluation of the test data and the quality of its processing is carried out.

Among the shortcomings it is a drop in accuracy on the masculine pronouns that should be noted. It is caused by the choice of the subject of opinions (a mobile phone). It is mentioned very often (including implicit mentionings) and the main part of malfunctions consists in choosing an implicit antecedent «*». In authors' opinion, the problem can be solved by taking new attributes connected with the result of syntactical parsing into consideration.

The development plans include the application of the system to other domains and improving the recall-precision ratio by introducing new attributes and refining the adjustment of the coefficients.

References

1. *Ermakov A. E.* 2005. The Reference of Persons and Organizations Definitions in Russian Media Texts: Empiric Regularities for Computational Analysis [Referentsiia Oboznacheniia Person I Organizatsii v Russkoiazychnykh Tekstakh SMI: Empiricheskie Zakonomernosti dla Komp'iuternogo Analiza]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2005" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2005").
2. *Lin Hsuan-Tien, Lin Chih-Jen, Weng Ruby C.* A Note on Platt's Probabilistic Outputs for Support Vector Machines.
3. *Michael P., Kazuhide Y., Eiichiro S.* 2002. Anaphora Analyzing Apparatus Provided with Antecedent Candidate Rejecting Means using Candidate Rejecting Decision Tree.
4. *Ning Pang, Jun-feng Shi.* The Third Personal Pronoun Anaphora Resolution in the Paroxysmal Text of the Chinese Web.

5. *Niyu G., Hale J., Charniak E.* 1998. A Statistical Approach to Anaphora Resolution. Proceedings of the Sixth Workshop on Very Large Corpora. COLING-ACL'98.
6. *Okat'ev V. V., Gergel' V. P., Alekseev V. E., Talanov V. A., Barkalov K. A., Skatov D. S., Erekhinskaia T. N., Kotov A. E., Titova A. S.* 2008. R&D Report n the Theme "Designing of the Russian Language Syntactic Analysis System, Trial Version" [Otchet o Vypolnenii NIOKR po teme: "Razrabotka Pilotnoi Versii Sistemy Sintaksicheskogo Analiza Russkogo Iazyka"].
7. *Oldenderfer M. S., Bleshfield R. K.* 1989. Factor, Discriminant and Cluster Analysis.
8. *Platt John C.* 1999. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods.
9. *Tollegin P. V., Vetrov D. P., Kropotov D. A.* 2006. The Algorithm of Automatic Third Person Anaphora Settlement basing on Machine Learning [Algoritm Avtomatizirovannogo Razresheniia Anafory Mestoimenii Tret'ego Litsa na osnove Metodov Mashinnogo Obucheniia]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2006" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2006") : 504–507.
10. *Vapnik V.* 1998. Statistical Learning Theory.
11. *Yıldırım S., Kılıçaslan Y.* 2007. A Machine Learning Approach to Personal Pronoun Resolution in Turkish. Proceedings of 20th International FLAIRS Conference, FLAIRS-20.
12. *http://www.allnokia.ru.*
13. *http://www.novoteka.ru.*
14. *http://market.yandex.ru.*
15. *http://zoom.cnews.ru.*

ПРОГРАММА АНАЛИЗА ФОНЕТИЧЕСКИХ СТАТИСТИК В ТЕКСТАХ НА РУССКОМ ЯЗЫКЕ И ЕЕ ИСПОЛЬЗОВАНИЕ ДЛЯ РЕШЕНИЯ ПРИКЛАДНЫХ ЗАДАЧ В ОБЛАСТИ РЕЧЕВЫХ ТЕХНОЛОГИЙ

Н. С. Смирнова (nsmirnova@speechpro.com)

П. Г. Чистиков (chistikov@speechpro.com)

Центр речевых технологий (http://speechpro.ru)

В статье представлена реализация статистического анализатора текста TextAnalyser, описаны основные возможности его использования в сфере речевых технологий и приведены некоторые результаты статистической обработки в анализаторе большого текстового корпуса, в частности, ранжированные по частотности списки аллофонов русских фонем и наиболее частотных бифонемных сочетаний. Программа TextAnalyser и получаемые с ее помощью статистические данные могут быть полезны при разработке систем автоматического синтеза и распознавания речи.

Ключевые слова: статистика, фонетическая статистика, речевые технологии, статистический анализатор.

SOFTWARE FOR AUTOMATED STATISTICAL ANALYSIS OF PHONETIC UNITS FREQUENCY IN RUSSIAN TEXTS AND ITS APPLICATION FOR SPEECH TECHNOLOGY TASKS

N. S. Smirnova (nsmirnova@speechpro.com)

P. G. Chistikov (chistikov@speechpro.com)

Speech Technology Center (http://www.speechpro.com)

Currently the development of most speech technology applications is based on the use of pre-recorded speech data produced by one or several speakers. The principal requirement to the speech corpus is suffi cient coverage of speech units involved in a specifi c task. The type of units may differ depending on the approach adopted. The most popular way of obtaining

speech material is through making speakers read some text, since read speech allows strict control over unit coverage (phonetic, prosodic and the like). For the purpose of automating and facilitating the acquisition of text corpora of desired phonetic composition and coverage, a special tool "TextAnalyser" has been developed. The software is primarily intended for the development of automatic speech recognition and synthesis systems. It makes use of an electronic dictionary containing 180 000 Russian word forms and is based on an automatic transcription tool developed for the Russian TTS system. It allows the generation of texts with required phonetic coverage, the assessment of several types of phonetic unit frequencies in Russian texts (monophones, diphones, triphones, syllables) and the reduction of data redundancy. TextAnalyser was applied for statistical analysis of a large text corpus in Russian comprising 460 965 words (2 500 288 phonemes). As a result of text processing, frequencies of occurrence were obtained for all relevant kinds of Russian-language phonetic units. In the paper we present ordered monophone and diphone frequency lists. The obtained monophone statistics is compared to previously published data.

Key words: statistics, phonetic statistics, speech technology, statistical analyzer, statistical analysis.

Введение

Многие современные направления разработки речевых технологий связаны с использованием речевых баз данных, полученных на основе текстов в произнесении одного или нескольких дикторов. Критерием пригодности речевой же базы служит, прежде всего, полнота представления в ней речевых элементов, избранных в качестве основных или релевантных для решения тех или иных прикладных задач. Так, например, для разработки систем синтеза речи состав таких элементов может быть различным в зависимости от типа избранной базовой единицы — чаще всего это дифоны или аллофоны (трифоны). Для дифонного синтеза требуется база данных, содержащая все возможные для заданного языка двучленные комбинации фонем (аллофонов), тогда как при аллофонном синтезе необходим учет всех возможных сочетаний левого и правого контекста (как правило, различные типы контекстов при этом объединяются в классы по степени акустической близости, чтобы сократить общий инвентарь единиц). Аналогичный подход применяется и при разработке систем распознавания речи, построенных на контекстно зависимых монофонах (трифонах). В случае разработки мультимедийных справочных экспертных систем принципиальное значение будет иметь наличие в речевом материале элементов, способствующих проявлению акцентной/диалектной вариативности речи на заданном языке, а также междикторской регионально не обусловленной вариативности.

В связи с изложенным выше, особое значение приобретает подготовка текстового материала в задачах, где набор необходимых элементов речевой базы заранее определен, особенно в тех случаях, когда ресурс для обработки и структурирования речевого материала ограничен. Помимо фонетической представительности, использование специального текстового

материала фиксированного объема обеспечивает компактность, что в дальнейшем позволяет существенно сократить время на его обработку. Считается, что при использовании больших объемов текстового материала полнота покрытия единиц достигается и без специального подбора, однако, во-первых, при таком подходе неизбежно возникает избыточность базы (которая может в дальнейшем потребовать пост-обработки материала для исключения повторяющихся элементов) и, во-вторых, отдельные редкие фонетические сочетания, возникающие на стыках слов, вполне могут так ни разу и не встретиться. Именно поэтому даже в сегодняшних синтезаторах, основанных на технологии Unit Selection и имеющих базы данных объемом более 10 часов речи, изредка встречаются «пропуски» или не вполне адекватные замены — по причине отсутствия необходимых элементов в исходной акустической базе.

С целью упрощения и автоматизации процесса формирования текстов с заданной фонетической представительностью, оценки степени полноты покрытия элементов в заданном тексте, а также сокращения избыточности текстового корпуса за счет удаления из него повторяющихся элементов, был разработан фонетический анализатор текстового материала, описание которого приводится ниже.

Описание анализатора частотности фонетических единиц в текстовом материале TextAnalyser

TextAnalyser позволяет решать следующие задачи:

- анализ статистики встречаемости в тексте речевых единиц различных уровней: фонем, аллофонов, слогов, звуковых последовательностей, слов;
- генерация слов (последовательностей слов), содержащих фонетические единицы с заданными параметрами;
- оценка степени фонетической информативности слов, входящих в состав текстового материала.

В состав программной системы входят следующие основные части:

- словарь словоформ русского языка, объёмом 180 000 словоформ [1]
- автоматический фонетический транскриптор русской речи по тексту, разработанный для синтезатора русской речи [2].
- модуль статистической обработки транскрипционного материала, осуществляющий сегментацию затранскрибированного материала на фонетические единицы различных типов — аллофоны, двучленные и трехчленные звуковые последовательности, слоги (выделенные по заданным правилам), — и сравнение полученной статистики с опорной статистикой или «нормой» (см. описание ниже), а также получение статистики встречаемости в тексте произвольно заданных типов звуковых последовательностей;

поиск в словаре и вывод слов или словосочетаний, содержащих заданные аллофоны, звуковые последовательности или слоги.

«Норма» встречаемости в тексте единиц различных типов определяется на основе статистического анализа фонетически представительного текстового корпуса. В текущей версии анализатора для расчета опорной статистики используется словарь объемом в 460 965 слов (2 500 288 фонемоупотреблений).

Поскольку основной произносительной единицей принято считать слог, для оценки слоговой представительности текстов в программе предусмотрено пять альтернативных вариантов слогоделения с последующим вычислением статистики встречаемости слогов:

1) Деление на открытые слоги (слоги всегда оканчиваются на гласный).
2) Деление на закрытые слоги (слоги оканчиваются на согласный, если за гласным следуют два или более согласных).
3) Выделение в тексте последовательностей типа «согласный + гласный» (при этом остальные элементы не учитываются).
4) Деление на слоги по правилу Р. И. Аванесова [3].
5) Деление на слоги по правилу Л. В. Щербы [4].

При этом слоговая статистика может сниматься как в «фонемном» (т. е. без учета ударности/безударности и редукции гласных), так и в «аллофонном» представлении, с полноценным учетом всех позиционных и комбинаторных аллофонов.

Кроме слоговой статистики, программа позволяет вычислять статистику встречаемости звуков и звуковых последовательностей в следующих вариантах:

- монофоны (статистика для каждого аллофона без учета контекста);
- дифоны (статистика для последовательностей из двух аллофонов, т. е. для каждого аллофона с раздельным учетом правого и левого контекстов);
- трифоны (статистика для последовательностей из трёх аллофонов, т. е. для каждого аллофона с учетом левого и правого контекстов).

Все перечисленные виды статистической оценки могут использоваться для произвольно выбранного текста.

Кроме того, может быть получена сравнительная оценка встречаемости в тексте фонетических единиц относительно аналогичной статистики в опорном тексте, принятом за «норму» (который также может быть произвольно задан пользователем). Это позволяет определить, каких элементов не хватает в тестовом материале, и при необходимости дополнить текст, используя приведенные примеры слов и словосочетаний с требуемыми фонетическими единицами. Слова, в которых встречаются искомые фонетические единицы, выводятся в модифицированной орфографии (промежуточный уровень между нормативным орфографическим написанием и фонетической транскрипцией). Ударные гласные передаются прописными символами. Пример использования данной функции приведен на Рис. 1:

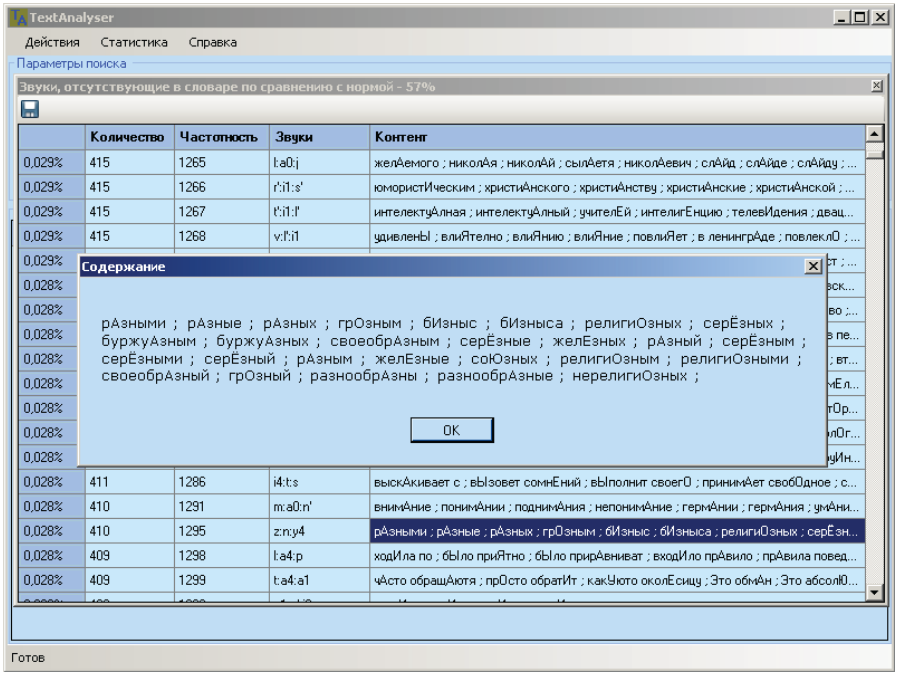


Рис. 1. Отображение типов слога (выделенных по правилам Р. И. Аванесова), отсутствующих в тексте, с примерами слов из опорного текстового корпуса, которые содержат заданный тип слога

Опция вывода статистики информативности слов позволяет оценить, насколько часто повторяются в словах текстового материала заданные фонетические единицы (аллофоны, дифоны, трифоны, слоги). Если какой-либо элемент заданного типа встречается в тексте дважды или чаще, то каждому слову, его включающему, сопоставляются те слова, в которых данный компонент дублируется.

Кроме того, для входящих в слово компонентов указывается их ранг по частотности в опорном словаре-норме (соответственно, на Рис. 2 это «39», «69» и «133» для слогов слова «доро́ги» при слогоделении «по Аванесову»):

Имеется возможность сортировать наборы элементов по минимальному порогу встречаемости в тексте (например, более двух раз, более трех раз и т. п.).

Кроме расчета статистики фонетических единиц данное приложение предусматривает также следующие возможности:

- сбор статистики встречаемости слов в тексте (частота встречаемости каждой словоформы);
- учет (исключение) повторяющихся словоформ при выведении текстовых последовательностей (слов) с заданным сочетанием фонем (аллофонов);

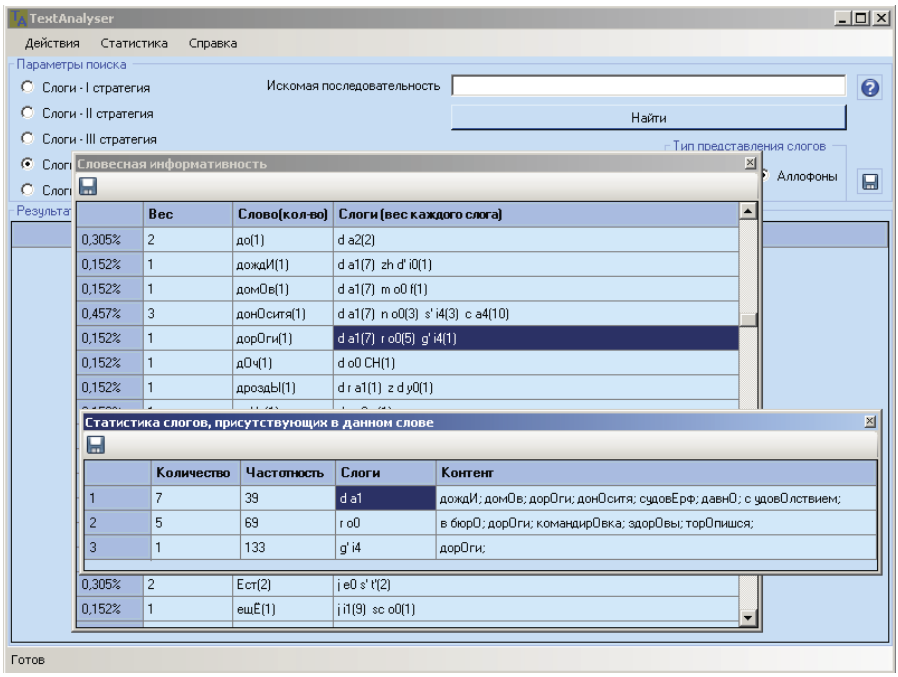


Рис. 2. Оценка информативности слов по выбранной единице анализа (слоги)

Некоторые статистические данные, полученные с помощью TextAnalyser

Сегодня исследование статистических характеристик речевых единиц является крайне востребованным и актуальным уже не столько в теоретических целях, сколько для эффективного решения конкретных прикладных задач, касающихся обработки и моделирования речевого сигнала (в частности, в вероятностных n-gram алгоритмах автоматического распознавания речи и др.). Поэтому в данном разделе мы приводим некоторые статистические данные, полученные с помощью программы анализатора статистик TextAnalyser, и частично сопоставляем их с ранее публиковавшимися сведениями.

Для вычисления статистических характеристик русского языка был сформирован текстовый корпус, включающий в себя примерно в равных пропорциях тексты из классической русской литературы (фрагменты произведений Ф. М. Достоевского, А. П. Чехова, Н. В. Гоголя, М. Ю. Лермонтова), современную русскую прозу (фрагменты произведений В. Г. Распутина, О. Михайлова, И. С. Шмелева, В. А. Солоухина, В. Н. Крупина, Ч. Айтматова и др.) и публицистику (опубликованные новостные репортажи, интервью, текстовые

расшифровки дискуссий и публичных лекций). Общий объем текстового материала составил более 460 тысяч (460 965) словоформ, более 1 млн. слогов, более 2,5 млн. (2 500 288) фонемоупотреблений. Учитывались все аллофоны фонем русского языка, включая предударные и заударные гласные, а также позиционные и комбинаторные аллофоны согласных не только внутри, но и на стыках слов.

На вход программы был подан текстовый материал в орфографической форме, после чего он был подвергнут автоматическому членению на синтагмы и транскрибированию в модуле транскриптора системы синтеза речи. На базе данного текстового корпуса была получена статистика реализации в русских текстах всех типов фонетических единиц, предусмотренных для выделения в TextAnalyser.

В таблице 1 приведены данные относительно встречаемости в тексте аллофонов русских фонем:

Таблица 1. Статистика аллофонов русских фонем в большом текстовом корпусе

Аллофон	Количество	Ранг	Аллофон	Количество	Ранг
a1	136 247	1	**r'**	35 575	26
a4	134 417	2	**z**	33 011	27
i4	129 413	3	**ch**	29 413	28
i1	123 175	4	**b**	28 764	29
a0	107 527	5	**sh**	28 388	30
t	107 481	6	**u1**	28 340	31
j	104 064	7	**g**	27 903	32
n	93 916	8	**u4**	27 670	33
o0	90 216	9	**u0**	27 620	34
s	85 979	10	**d'**	27 504	35
v	77 550	11	**f**	25 831	36
r	75 458	12	**v'**	24 539	37
e0	73 723	13	**zh**	23 804	38
k	72 485	14	**m'**	22 893	39
n'	59 279	15	**y0**	22 154	40
p	57 808	16	**h**	21 715	41
m	57 748	17	**c**	20 400	42
l'	53 374	18	**y1**	17 383	43
l	52 090	19	**k'**	11 502	44
a2	50 346	20	**p'**	11 267	45
d	46 516	21	**sc**	9 819	46
i0	45 618	22	**b'**	8 809	47
t'	44 941	23	**o1**	8 734	48
y4	42 162	24	**z'**	6 587	49
s'	39 080	25	**g'**	4 385	50

Аллофон	Количество	Ранг	Аллофон	Количество	Ранг
f'	2368	51	**C**	86	56
H	1748	52	**e1**	12	57
h'	1076	53	**SC**	11	58
o4	222	54	**e4**	9	59
CH	133	55			

В данной системе транскрипции для аллофонов русских гласных используются символы <a> (а), <i> (и), <e> (е,э), <u> (у), <o> (о), <y> (ы); индекс «0» обозначает ударный гласный, «1» — предударный (или первый предударный аллофон фонемы / а /), «2» — второй предударный аллофон фонемы / а /, «4» — гласный в заударном слоге. Значок «'» обозначает мягкость согласного. Приведем несколько примеров слов, записанных в данной системе транскрипции:

- город: g o0 r a4 t
- карандаш: k a2 r a1 n d a0 sh
- зонтик: z o0 n' t' i4 k

Как видно из таблицы, наиболее частотными являются безударные аллофоны фонем /а/ и /i/. У фонем /u/ и /у/, не подвергающихся качественной редукции в безударном положении, частотность безударных аллофонов также выше, чем ударных.

Таблица содержит также аллофоны, возникающие исключительно на стыках слов в результате процессов ассимиляции. В частности, это озвонченные аллофоны непарных глухих фонем / h /, / ch /, / c / и / sc / — соответственно / H /, / CH /, / C / и / SC /. Они, как видно, встречаются довольно редко, однако их моделирование является необходимым в полноценных системах синтеза и распознавания речи.

В силу некоторых различий в правилах транскрибирования текстового материала полученные нами данные не могли быть полноценно сопоставлены с ранее опубликованными исследованиями [5, 6]. Для целей сопоставления информация о частотности аллофонов, полученная с помощью TextAnalyser, была сведена к фонемам и таким образом оказалось возможным провести сравнение наших данных со статистикой, представленной в [6]. Результаты приведены в таблице 2.

Данные ЦРТ, полученные на материале 2,5 млн. фонемоупотреблений, сопоставляются с данными Института математики (ИМ) Сибирского отделения АН СССР (1964 г., на материале 1,5 млн. фонемоупотреблений в текстах технической и математической тематики) и статистикой, полученной в Лаборатории экспериментальной фонетики ЛГУ (ЛЭФ) на материале 100 000 фонемоупотреблений. В сравнительной таблице приводятся только ранги фонем, поскольку анализировались текстовые выборки разного объема.

Принятая для представления «ленинградская» трактовка фонемного статуса аллофонов «ы» и «а» в безударной позиции на месте орфографического «о» объясняется исключительно практическим удобством и не связана с определенной теоретической позицией авторов статьи.

Таблица 2. Сопоставление статистики русских фонем в различных исследованиях

Фонема	Ранг			Фонема	Ранг		
	ЦРТ	ЛЭФ	ИМ		ЦРТ	ЛЭФ	ИМ
a	1	1	1	**r'**	22	21	22
i	2	2	2	**z**	23	22	23
t	3	3	4	**ch**	24	24	25
j	4	6	5	**b**	25	25	24
o	5	4	6	**sh**	26	26	32
n	6	5	3	**g**	27	27	28
s	7	7	7	**d'**	28	30	27
u	8	8	9	**f**	29	28	29
y	9	15	20	**v'**	30	31	30
v	10	11	10	**zh**	31	29	33
r	11	9	8	**h**	32	32	31
e	12	13	12	**m'**	33	33	26
k	13	10	11	**c**	34	34	34
n'	14	14	16	**k'**	35	36	36
p	15	16	14	**p'**	36	37	35
m	16	17	15	**sc**	37	35	37
l'	17	18	13	**b'**	38	38	39
l	18	12	18	**z'**	39	39	38
d	19	19	21	**g'**	40	40	40
t'	20	20	17	**f'**	41	41	41
s'	21	23	19	**h'**	42	42	42

Как видно из таблицы, несмотря на различия в объемах и стилистической принадлежности текстовых выборок, в целом данные всех трех источников очень близки. Совпадает состав первых наиболее частотных семи фонем — с разницей в рангах не более двух (a, i, t, j, o, n, s) и наиболее редких девяти (c, k', p', sc, b', z', g', f', h'). Некоторые расхождения с одним из источников присутствуют в рангах фонем / j /, / l / и / m' /. Примечательно, что в данных ЦРТ ранг / j / занимает как бы промежуточное положение по сравнению с рангом этой фонемы по ЛЭФ и ИМ, ранг / l / совпадает с данными ИМ при существенном различии в ранге этой фонемы у ЛЭФ, тогда как ранг / m' / совпадает с данными ЛЭФ при существенных различиях с данными ИМ.

Наиболее существенные различия наблюдаются в ранге фонемы / y /(«ы»): 9 ранг у ЦРТ по сравнению с 15 у ЛГУ и с 20 по данным ИМ. Т.е. разница между ЦРТ и ИМ составляет 11 рангов (!). Не имея достаточно подробных сведений о специфике обработки текста исследователями ЛЭФ и ИМ, мы можем лишь предположить, что причина данного различия может скрываться в характере транскрибирования начальнословных «и» после слов, оканчивающихся на твердый согласный («как **И** ты», «в **И**нтересах», «всех **И**х» и т. п.). В нашем

случае они всегда транскрибировались как аллофоны «ы» (если модуль автоматического разбиения на синтагмы не ставил между ними синтагматической границы). Если же причислять их к аллофонам фонемы «и», то ранг фонемы «ы» понизился бы до 14, что близко к данным ЛЭФ и ИМ.

Приведем еще одну таблицу статистических данных — а именно, бифонемных сочетаний. Информация о типах и частотности таких сочетаний необходима для обеспечения полноты покрытия единиц в системах дифонного синтеза речи.

Статистика построена на анализе уже упомянутого выше текстового корпуса. Всего в тексте было обнаружено более 2 тыс. типов таких последовательностей (2176). Из них 2012 последовательностей встретились более 3 раз.

Наиболее частотные 164 последовательности, составляющие 50 % всех бифонемных сочетаний, приведены в таблице 3.

Таблица 3. Статистика дифонов — последовательностей из двух аллофонов

Дифон	Ранг	Количество	Дифон	Ранг	Количество
j:i4	1	27 609	**a1:t**	26	10 820
n:a4	2	22 845	**j:a4**	27	10 701
a4:_	3	22 647	**n:y4**	28	10 567
s:t	4	22 511	**p:a2**	29	10 404
i4:j	5	19 963	**v:a1**	30	10 314
i4:_	6	18 551	**n:a0**	31	10 270
n':i1	7	18 527	**r:a0**	32	10 252
n':i4	8	18 382	**t:a1**	33	10 232
v:a4	9	17 839	**p:r**	34	9 954
r':i1	10	17 040	**v:o0**	35	9 954
a4:j	11	16 383	**s:k**	36	9 854
l':i4	12	15 512	**j:i1**	37	9 821
t:a4	13	15 484	**e0:t**	38	9 677
r:a1	14	15 072	**_:a1**	39	9 203
k:a1	15	14 522	**a1:s**	40	9 142
k:a4	16	14 223	**p:r'**	41	9 130
t:o0	17	13 433	**s':t'**	42	9 051
l:a4	18	13 081	**_:p**	43	9 032
t':i4	19	13 058	**t:a0**	44	8 919
p:a1	20	11 680	**i1:s**	45	8 896
y4:j	21	11 451	**a4:m**	46	8 832
a0:j	22	11 321	**i1:v**	47	8 690
n:a1	23	11 086	**j:a0**	48	8 655
j:u4	24	11 066	**e0:n'**	49	8 645
a4:v	25	11 044	**k':i4**	50	8 572

Дифон	Ранг	Количество	Дифон	Ранг	Количество
_:i1	51	8552	**t:r**	92	6065
a0:l	52	8426	**a4:s'**	93	6030
v:a0	53	8363	**i1:r**	94	6023
d:a0	54	8183	**d:a1**	95	5891
_:k	55	8141	**t:_**	96	5890
r:o0	56	8085	**k:o0**	97	5863
n:a2	57	8076	**n:o0**	98	5863
d':e0	58	7947	**f:s'**	99	5858
sh:t	59	7508	**i4:m**	100	5858
a1:v	60	7428	**v':i4**	101	5819
m':i4	61	7385	**a1:j**	102	5787
c:a4	62	7359	**i4:s**	103	5785
i4:n	63	7329	**j:_**	104	5782
a1:k	64	7311	**o0:m**	105	5782
ch:i4	65	7141	**l':e0**	106	5781
o0:n	66	7128	**e0:j**	107	5765
n':e0	67	7029	**a1:b**	108	5749
s':i1	68	6979	**ch:i1**	109	5677
_:n	69	6908	**r':e0**	110	5675
v':i1	70	6899	_:v	111	5605
i4:t	71	6898	a1:d	112	5593
v':e0	72	6830	_:sh	113	5576
k:a0	73	6782	a0:l'	114	5549
r:a4	74	6748	a0:t	115	5546
l:a1	75	6611	i1:n	116	5529
o0:j	76	6589	l:a0	117	5501
_:s	77	6517	a1:g	118	5459
r:a2	78	6516	a0:s	119	5411
a1:r	79	6507	p':i1	120	5384
u4:_	80	6465	d':i4	121	5308
zh:y4	81	6454	n':i0	122	5264
m':i1	82	6436	m':e0	123	5262
l':i1	83	6381	t:o1	124	5219
a0:t'	84	6375	t:v	125	5196
o0:r	85	6363	i1:n'	126	5119
t':i1	86	6299	i1:t	127	5106
i1:z	87	6266	a1:n'	128	5092
m:_	88	6216	a4:f	129	5077
sh:y4	89	6168	z:a1	130	5034
o0:t	90	6143	i4:v	131	5013
m:a1	91	6115	t':i0	132	4995

Дифон	Ранг	Количество	Дифон	Ранг	Количество
l':i0	133	4964	a0:_	149	4668
a4:s	134	4961	a2:s	150	4626
d':i1	135	4955	i1:r'	151	4582
a0:n'	136	4953	a4:n	152	4532
o0:l'	137	4945	a1:d'	153	4522
a0:n	138	4932	s:p	154	4479
y4:_	139	4921	e0:s	155	4471
_:j	140	4915	a1:l'	156	4434
i4:l'	141	4903	a0:m	157	4432
a1:z	142	4884	a0:k	158	4419
d:a4	143	4883	r':i4	159	4412
i1:k	144	4871	i1:p	160	4390
a4:p	145	4866	l':n	161	4384
a1:n	146	4811	e0:r	162	4380
o0:v	147	4787	j:e0	163	4379
i1:m	148	4686	b:y0	164	4376

Аналогичным образом была получена статистика трифонов (всего 35 073 типа, из них 24 342 — трифоны с частотой встречаемости в опорном тексте более 3 раз), последовательностей «согласный плюс гласный» (всего 410 типов) и слогов, выделенных по различным правилам. Любопытно, что число типов слога при слогоделении по Л. В. Щербе на несколько сот превышает число слогов, выделенных по правилу Р. И. Аванесова: соответственно 11 354 и 10 801, при этом в обоих случаях лишь чуть больше половины слогов (6164 и 5841) имеют частоту встречаемости в текстовом корпусе более трёх раз. В силу ограниченного объема данной публикации мы не имеем возможности приводить полученные статистические данные в полном объеме.

Перспективы развития TextAnalyser

Дальнейшая разработка программы предполагает автоматизацию процесса формирования текстового материала с заданной представительностью фонетических единиц (фонем, дифонов, трифонов, слогов). Это необходимо, в частности, при разработке систем автоматического синтеза речи. Так, используя данную программу, появится возможность получить текстовый материал с полным покрытием всех возможных в русском языке дифонов, или, например, обеспечить покрытие нескольких сот наиболее частотных типов слога. Разумеется, речь не идет о генерации оригинального связного текста — такое машине пока не под силу. Имеется в виду компоновка текста из подходящих слов, словосочетаний и предложений, содержащихся в опорном текстовом корпусе. Подбор их будет осуществляться программой таким образом, чтобы минимизировать избыточность (повторяемость) единиц в тексте. Критерием могут служить

не только фонетические (сегментные) свойства текста, но и коммуникативно-синтаксическая составляющая (по знакам препинания): например, можно будет задать долю вопросительных предложений, или предложений, содержащих неконечные синтагмы (по запятым). Очевидно, что эффективность подбора зависит от параметров опорного текстового корпуса: чем более полным и разнообразным он будет, тем более информативным и компактным будет получаемый на его основе текстовый материал.

References

1. *Avanesov R. I.* 1954. On Syllable Boundary and Syllable Structure in Russian Language [O Slogorazdele I Stroenii Sloga v Russkom Iazyke]. Voprosy Iazykoznaniia, 6.
2. *Bondarko L. V.* 1998. Modern Russian Phonetics [Fonetika Sovremennogo Russkogo Iazyka] : 199–200.
3. *Bondarko L. V., Zinder L. R., Shtern A. S.* 1977. Some Statistical Characteristics of Spoken Russian [Nekotorye Statisticheskie Kharakteristiki Russkoi Rechi]. Slukh I Rech' v Norme I Patologii, 2 : 3–16.
4. *Elkina V. N., Iudina L. S.* 1964. Russian Speech Syllables Statistics [Statistika Slogov Russkoi Rechi]. Vychislitel'nye Sistemy, 10 : 58–62.
5. *Khomitsevich O. G., Rybin S. V., Talanov A. O., Oparin I. V.* 2008. Automatic Estimation of Accentuation in Unknown Words in the Speech Synthesis System [Avtomaticheskoe Opredelenie Mesta Udareniia v Neznakomykh Slovakh v Sisteme Sinteza Rechi]. Materialy XXXVI Mezhdunarodnoi Filologicheskoi Konferentsii (Proc. of the XXXVI International Conference on Phylology).
6. *Vol'skaia N., Koval' A., Koval' S., Oparin I., Pogareva E., Skrelin P., Smirnova N., Talanov A.* 2005. New Generation Russian Text-to-speech Synthesizer [Sintezator Russkoi Rechi po Tekstu Novogo Pokoleniia]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2005" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2005").

ОСОБЕННОСТИ ПОДГОТОВКИ ТЕРМИНОВ ДЛЯ РУССКО-АНГЛИЙСКОГО ТЕЗАУРУСА ПО КОМПЬЮТЕРНОЙ ЛИНГВИСТИКЕ

Е. Г. Соколова (minegot@rambler.ru)

С. Ю. Семенова (sonya_sem@mail.ru)

Российский государственный гуманитарный университет, Москва, Россия

И. С. Кононенко (irina_k@cn.ru)

Ю. А. Загорулько (zagor@iis.nsk.su)

Институт систем информатики имени А. П. Ершова СО РАН, Новосибирск, Россия

О. Ф. Кривнова (okrivnova@mail.ru)

Московский государственный университет им. М. В. Ломоносова, Москва, Россия

В. П. Захаров (vz1311@yandex.ru)

Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

Описывается начальная стадия разработки русско-английского терминологического тезауруса по компьютерной лингвистике. В задачи этого этапа входит обоснованный выбор потенциальных источников терминов с учетом двуязычной специфики разрабатываемого ресурса, выделение и подбор терминов для базового словника, изучение особенностей представления терминов, их толкований и отношений между ними.

Ключевые слова: тезаурус, термины, компьютерная лингвистика, двуязычие.

SELECTION AND PREPARATION OF TERMS FOR THE RUSSIAN-ENGLISH THESAURUS OF COMPUTATIONAL LINGUISTICS

E. G. Sokolova (minegot@rambler.ru)

S. Iu. Semenova (sonya_sem@mail.ru)

RSUH, Moscow, Russian Federation

I. S. Kononenko (irina_k@cn.ru)

Iu. A. Zagorul'ko (zagor@iis.nsk.su)

A. P. Ershov Institute of Informatics Systems SB RAS, Novosibirsk, Russian Federation

O. F. Krivnova (okrivnova@mail.ru)

MSU, Moscow, Russian Federation

V. P. Zakharov (vz1311@yandex.ru)

Saint-Petersburg State University, St. Petersburg, Russian Federation

The initial phase of the development of Russian-English thesaurus on terminology in the fi eld of computational linguistics is described. One of the fi rst tasks is the choice of candidate sources of terms allowing for the bilingual nature of the electronic resource. Other problems to be solved are those of terminology extraction and selection of basic term list as well as the study of peculiarities of representation of terms and relations between them. The diversity of the fi eld of computational linguistics, its interdisciplinary nature and the lack of Russian terminological sources and term defi nitions due to certain lagging of the fi eld in Russia as compared to the English-speaking countries — all these factors explain the kind of decisions made at this stage. One of them concerns the use of the Russian-language corpus of papers presented at the International Conference "Dialogue" (2000–2010). This corpus proved to be a helpful source of terms in real use. Besides, dictionaries as well as indices and glossaries of textbooks and manuals have been examined in order to derive defi nitions. As an additional source of terms for the Russian part of the thesaurus the English-language terminological sources have been utilized and their terms and defi nitions translated into Russian. This is especially important for the terms in some empirical and technologically advanced subfi elds, such as speech technologies.

Key words: thesaurus, terms, computational linguistic, bilingualism.

1. Введение

В статье описывается начальная стадия разработки электронного терминологического тезауруса по компьютерной лингвистике (КЛ). Главной целью этого этапа является выбор терминов и их толкований для дальнейшего включения в тезаурус. Основные трудности в подборе и содержательном определении терминов для русско-английского тезауруса по КЛ связаны с особенностями самой науки и состоянием ее развития в России. Важно отметить, в частности:

1. еще не преодоленное до конца отставание русскоязычной КЛ (РКЛ) от англоязычной КЛ;
2. неоднородность предметной области (ПО) КЛ;
3. неравномерность развития отдельных направлений КЛ;
4. междисциплинарный характер КЛ.

Определенные трудности обусловлены также нашей установкой на отражение терминологии РКЛ в составе оригинальных русскоязычных работ, а не обзоров и таких учебников, которые во многом являются пересказом англоязычных источников[1]. Следствием пп. 1, 3 является отсутствие русскоязычных учебных и лексикографических источников, достаточно полно отражающих структуру современной КЛ в отличие от англоязычных источников, где она представлена детально и отчетливо. До сих пор термины РКЛ входили лишь в состав словарей и глоссариев по лингвистике и смежным направлениям. Следствием пп. 2, 3 является наличие источников только по отдельным разделам смежных направлений и КЛ, например, по искусственному интеллекту (ИИ), информационному поиску (ИП) и почти полное отсутствие русскоязычных терминов по другим разделам КЛ, например, по оценке систем (system evaluation). Следствием пп. 2, 4 являются ситуации, когда один и тот же термин в смежных науках имеет различные толкования, например, «синтаксический анализ» в ИИ и в КЛ. Место КЛ среди смежных наук и состав КЛ обсуждаются в р. 2.

Учитывая вышеперечисленные особенности КЛ, мы стремились в начальной версии тезауруса найти источники «живых» терминов РКЛ и их толкований и именно их зафиксировать в словарных статьях терминов. Список основных лексикографических источников терминов — тезаурусов и толковых словарей смежных наук, — а также принятая в нашем проекте структура описания термина, представлены в статье [1] данного сборника.

В соответствии с направленностью тезауруса на РКЛ, в условиях недостаточности основных лексикографических источников рассматривались также дополнительные источники в виде предметных указателей и глоссариев русскоязычных учебников и монографий и самих их текстов, а также

[1] Это не противоречит тому факту, что локально мы опираемся именно на англоязычные источники, стараясь дополнить картину КЛ в отдельных ее частях.

коллекция научных текстов [Прил., рус. 1][2], из которой были получены статистические показатели встречаемости терминов. Параметры и методика исследования массива научных статей конференции «Диалог» описываются в р. 3.

Мы используем в качестве источника терминов массив текстов «Диалога», поскольку он содержит «живые» термины, которые реально употребляются в РКЛ, а также предметные указатели, хотя они имеют те же недостатки, что и сами источники — локально являются пересказом англоязычных источников, что может приводить к выделению термина, реально в РКЛ не используемого. См. р. 4 и 5, в которых обсуждаются дополнительные лексикографические источники.

Для английской части тезауруса рассматривались только лексикографические источники и учебники, коллекции текстов не исследовались. В р. 6 описывается тематика терминов начальной версии тезауруса.

2. РКЛ как предметная область

В РКЛ для обозначения всего направления в основном используется термин *Прикладная лингвистика.* В последние два десятилетия все более употребительным становится термин *Компьютерная лингвистика*[3]. Эти термины и термины, представляющие близкие направления и разделы КЛ, имеют следующие частотные характеристики в [Прил., рус. 1]:

автоматическая обработка текста — 155
автоматическая обработка (естеств.) языка –7+8
искусственный интеллект — 265 + ИИ (108)
когнитивная лингвистика — 58
компьютерная лингвистика — 900[4]
корпусная лингвистика — 159
лингвистическая технология — 11
прикладная лингвистика (языкознание — 9) — 120
речевая технология — 74
синтез речи — чуть более 300.

В связи с отмеченными особенностями 2, 3, 4 научной области КЛ трудно дать определение. В русскоязычных источниках даются экстенсиональные

---

[2] Ссылка на Приложение в конце статьи, в котором перечислены наиболее употребляемые составителями статей источники как терминов, так и толкований. Таким образом, библиографические ссылки в статье отсылают к одному из двух взаимодополняющих списков: Литература и Приложение.

[3] Включенный, в частности, А. С. Нариньяни в название конференции «Диалог» с момента ее возрождения после перестройки в 1995 г.

[4] Большинство этих вхождений относится к библиографическим ссылкам на статьи в сборниках Диалога.

определения, например, в [2]: «Раздел лингвистики, задачей которого является исследование проблем, связанных с машинной обработкой текста: организацией естественно-языкового интерфейса, машинным переводом и реферированием, статистическим анализом словарей и текстов на ЭВМ, автоматическим распознаванием речи». При этом КЛ рассматривается как ветвь разных наук, в приведенном определении — лингвистики, в других источниках — ИИ [Прил., рус. 6]) и прикладной лингвистики [Прил., рус. 7]. РКЛ также пересекается с социо-, квантитативной и другими «лингвистиками». В англоязычных источниках КЛ подчиняется когнитивной науке (cognitive science), в частности, в меморандуме Х. Ускорайта [3]. Он справедливо разделяет КЛ на два направления: «теоретическая КЛ» (*Theoretical CL*) и «прикладная КЛ» (*Applied CL)*. Теоретическая КЛ базируется на теоретической лингвистике и пересекается с психолингвистикой и когнитивной психологией. Относительно прикладной КЛ указывается, что эта область также иногда обозначается терминами «language engineering» и «(human) language technology» и направлена на достижение практических результатов в моделировании ЕЯ.

Представление о том, что входит в ПО КЛ, также изменялось в истории этого направления в течение чуть более 60 лет. Современный состав ПО КЛ можно обрисовать по англоязычным источникам, в частности — [Прил., англ. 7, 8, 10]. Согласно обзору [Прил., англ. 10], в состав КЛ входит анализ/понимание vs. генерация текстов, распознавание и синтез речи, а также диалог и дискурс, мультимодальность, математические методы, лингвистические ресурсы и оценка систем[5]. Учебники [Прил., англ. 7 и 8] выделяют эти же направления, но добавляют к ним прикладную тематику: машинный перевод (МП), ИП, автоматическое реферирование, вопросно-ответные системы, извлечение знаний, автоматическое индексирование, взаимодействие с компьютером на ЕЯ, интеллектуальный поиск в текстах и некоторые другие.

С распространением эмпирических компьютерных технологий в конце 80–90 гг. зарождается потребность разделения «широкой» КЛ и «узкой», собственно технологической, области скорее ИИ, чем КЛ, для которой письменные тексты и звучащая речь являются только одним из видов данных. Особенно это характерно для направлений, связанных с обработкой речи. Специфика этого направления обсуждается в р. 5.

Отличие КЛ от традиционной лингвистики, которая является основой, базой для КЛ, состоит в том, что предметом КЛ (и, в конечном счете, ее объектом) является информация, а не языковая форма. Это верно как для теоретических направлений (анализ/понимание текста — генерация текста),

[5] Здесь опять возникает неопределенность. По [3] этот список относится скорее к «теоретической КЛ», которая «…deals with formal theories about the linguistic knowledge that a human needs for generating and understanding language… and implement(s) them as computer programmes.». При этом обзор называется «The state of the art in Human Language Technology» — термин, который сам Х. Ускорайт вместе с термином «language engineering» относит к прикладной КЛ.

так и для прикладных разделов КЛ: МП, ИП, речевые технологии (РТ) и др. При этом КЛ опирается преимущественно на «новейшие» разделы теоретической лингвистики — семантику, теорию дискурса и прагматику, быстрое развитие которых в последние десятилетия было вызвано в том числе и самой КЛ.

В начальной версии тезауруса мы старались собрать наиболее частотные термины, которые реально встречаются в РКЛ. В связи с этим в качестве основного источника терминов РКЛ взята коллекция русскоязычных текстов, представленная в р. 3.

3. Коллекция текстов как источник терминов

Складывающаяся терминологическая система отражается в учебниках, справочниках, электронных ресурсах и материалах конференций. Учитывая недостаток современной справочной русскоязычной литературы по КЛ, было принято решение создать базовый словник русскоязычных терминов по КЛ, используя материалы ежегодной международной конференции «Диалог». Собранная коллекция документов содержит тексты докладов, представленных на конференции «Диалог» в 2000–2010 гг., и имеет следующие характеристики: число документов — 1 193, объем — 4 610 694 словоупотреблений, суммарный размер — 27,5 Мб.

На этапе создания словника применялась словарная технология КЛАН [7], которая позволяет на базе коллекции текстовых документов создать список использованных в ней слов и словосочетаний — кандидатов в термины ПО, — причем каждый термин снабжен следующими статистическими показателями: частота встречаемости в коллекции и частота по документам. В процессе автоматической обработки этой коллекции было реализовано первоначальное наполнение словника с использованием технологии обучения по массиву текстов на базе лингвистических моделей: универсальный морфологический анализ, сборка именных словосочетаний на основе 20 синтаксических шаблонов, предсказание незнакомых слов. В результате был получен исходный словник объемом 79 678 слов и 512 783 словокомплекса (СК).

На этапе фильтрации полученный список терминов-кандидатов был полуавтоматически отсортирован для выявления наиболее важных (статистически значимых) в данной ПО слов и СК, имеющих терминологический характер:

- удалены термины с частотой встречаемости 1–3;
- удалены термины, встретившиеся только в одном тексте;
- удалены или перенесены в стоп-словарь служебные слова;
- отфильтрована нетерминологичная лексика специальных лексико-семантических разрядов (топонимы, имена персон и организаций);
- удалены ошибочные предсказания, в том числе: ошибки в предсказании части речи или морфологического класса, ошибки при определении лемм, некорректности, основанные на ошибках в тексте.

При удалении неверных гипотез были автоматически удалены построенные на их основе словосочетания.

В результате этапа фильтрации объем словника сократился до 23 760 слов и 31 709 СК. Словарь отфильтрованных терминов-кандидатов был передан экспертам в данной ПО для проведения экспертной оценки. Работа экспертов поддерживается конкордансом, который позволяет получить все примеры употребления термина словаря вместе с его контекстами.

При автоматизированной подготовке базового словника были использованы возможности сортировки знаменательных слов по части речи, морфологическому типу и убыванию встречаемости, а также автоматической проверки вхождения слов в состав СК:

- терминообразующие лексические единицы разряда фамилий — перенесены в стоп-словарь (*Зализняк — словарь Зализняка*; *Мельчук — модель Мельчука*);
- нарицательные одушевленные существительные — выборочно удалены (*адъюнкт*, *коллекционер*);
- глаголы, наречия — отсортированы по части речи и в порядке убывания встречаемости, выборочно удалены общеупотребительные.

В результате такой обработки объем словника сократился до 13 865 слов и 27 458 СК.

Далее эксперты провели отсев терминов, не относящихся к КЛ. Итоговый объем словаря составил 6013 слов и 8489 СК. Была проведена разметка словника с помощью системы семантических признаков, которая соответствует делению КЛ на три направления: АОТ (автоматическая обработка текста), РТ (речевые технологии) и КорпЛ (корпусная лингвистика). Наиболее представительным оказался подсловник АОТ (1524 слова), из них в топ-список (встречаемость выше 20) отнесено 941 слово; топ-список многословных единиц АОТ составляет 1001 СК. Как и следовало ожидать, направление РТ представлено на «Диалоге» слабо: в топ-список вошли 105 терминов по РТ и прикладной фонетике. Эксперты могут работать с различными выборками из соответствующего итогового подсловника, сформированными на базе таких признаков как частота, часть речи, структура СК и т. п. Так, для базовой версии тезауруса было принято решение ограничиться топ-списком, составленным из наиболее важных терминов-существительных и именных групп.

На предварительном этапе эксперты существенно опирались не только на знания о предмете и направлениях КЛ, но и на общелингвистические представления о терминологичности и путях формирования терминологических словников. Так, наш подход, основанный, в том числе, на предварительном структурировании ПО, согласуется с общей методикой формирования словников на базе классификационных схем предметных областей, см., например, [8].

Что касается английской части словника, то для данной версии тезауруса, имеющей русско-английскую направленность, выбирались переводные эквиваленты из доступных англоязычных источников по КЛ.

4. Особенности русскоязычных терминоисточников по РКЛ

Термины КЛ из основных источников — толковых словарей и тезаурусов, относящихся к лингвистике и смежным с КЛ областям, — требуют проверки. Например, тезаурус ИНИОН [8] и ЛЭС [Прил, рус. 11] считают основным термином *автоматический перевод,* присвоив ему статус дескриптора, а *машинный перевод* — аскриптором к нему. Встречаемость в [Прил, рус. 1]: *машинный перевод* 318; *автоматический перевод* 58[6]. Более высокая частотность первого не объясняется ссылками на литературу, в частности, на название сборника «*Машинный перевод и прикладная лингвистика*» приходится лишь 28 вхождений термина *машинный перевод.* Интернет-энциклопедии [Прил, рус. 2,3] и учебники придерживаются этой же традиции, которую и мы не стали нарушать. На сайте Европейской ассоциации машинного перевода [10] также отмечается, что термин *machine translation* звучит архаично, но тем не менее сохраняется как основной общий термин для всей области.

Дополнительные источники терминов — предметные указатели и глоссарии к научным текстам, а также сами тексты — более субъективны. Указатели отражают текст конкретной книги, статьи, а не ПО. В описании термина в тезаурусе мы отмечаем, где он встретился, в тексте издания или в предметном указателе (глоссарии), считая, что включение термина в предметный указатель повышает его терминологический статус, хотя бывает, что достаточно значимый термин встречается только в тексте. Так, термин *structural transfer* упоминается в тексте учебника [Прил., англ. 8] (и определяется как *transformation of source language structures into equivalent target language forms*), но отсутствует в предметном указателе. Кроме того, если термин заимствован из английского, то он может органично выглядеть в тексте, например, «*По Р. Шенку скрипт — это некоторая общепринятая, общеизвестная последовательность причинных связей*», но становиться неадекватным, когда автор выносит его в предметный указатель. Термин «скрипт» входит в предметные указатели учебников [7] и [Прил, рус. 7], в обоих случаях являясь калькой английского термина. В [Прил, рус. 1] «скрипт» употребляется исключительно как термин информатики для обозначения определенного типа программ, например, в следующем контексте: «база данных жестов, включающая в себя файлы скриптов, управляющих виртуальным демонстратором». Таким образом, «скрипт» является дескриптором в информатике, а частотность его употребления в РКЛ со значением «сценарий» по [Прил, рус. 1] равна 0.

---

[6] Поиск в Интернете дает обратное соотношение: *машинный перевод* 640 000, *автоматический перевод* 1 960 000, которое объясняется тем, что если речь идет о МП с языка на язык (а не о переводе на другой тариф и т. п.), основную часть ответов составляет реклама он-лайн переводчиков, т. е. имеется в виду разновидность *полностью автоматического перевода* (*онлайн-перевод*).

5. Англоязычная литература как источник терминов

Учитывая скачок, совершенный в области РТ в течение последних нескольких лет, когда эта область окончательно сложилась как высокотехнологичное направление, имеющее огромный практический и коммерческий выход, а также тот факт, что это направление слабо представлено в [Прил, рус. 2], авторы избрали методику сбора терминов, обратную методике, принятой в разделах АОТ и КорпЛ, используя в качестве основных англоязычные источники. В некоторой степени этот подход применяется и в других разделах, таких как «направления КЛ».

В качестве основы для сбора терминологического материала по РТ были взяты предметные указатели нескольких современных и наиболее авторитетных англоязычных книжных источников обзорно-учебного профиля. Кроме того, активно использовались глоссарии, входящие в состав известных звуковых анализаторов Adobe Audition 1.5. 2004 и Speech Analyzer 1.5–2002 [Прил, англ. 2].

На данной терминологической базе был составлен англо-русский словник параллельных терминов по РТ, включающий более 700 парных терминов (англо-русских эквивалентов). На следующем этапе из собранного англо-русского словника была выделена базовая часть, которая далее была включена в состав первой редакции проектируемого тезауруса по направлению РТ. Выделение базовой части словника осуществлялось экспертами по данному направлению с учетом списка частотных терминов по РТ и прикладной фонетике, полученного в результате компьютерной обработки и анализа электронных материалов конференции «Диалог».

Направление РТ характеризует большой массив собственной терминологии, например, в подразделах «автоматический синтез речи», «автоматическое распознавание речи» и др. Но имеются и точки пересечения с АОТ (см. пример ниже). Имеются и общие проблемы, к числу которых относятся пробелы в русскоязычной терминологии, ведущие к необходимости перевода терминов, а также отсутствие сложившейся традиции в понимании и употреблении уже имеющихся терминов.

Рассмотрим в качестве примера термин *spoken language machine translation*. Задача автоматического перевода устной речи возникла на стыке МП и РТ. *Spoken Language Processing* обычно переводится как *Автоматическая обработка устного языка*, одной из задач которой является автоматический устный перевод (АУП) с его разновидностями, соответствующими АУП типа «Речь(L1) --> Текст(L2)» и АУП типа «Речь(L1) --> Речь(L2)». Вторая разновидность представлена английским термином *speech-to-speech translation*. В русскоязычной литературе такой традиции нет, как нет (или практически нет) и такого типа приложений. Поиск в Интернете дал в качестве эквивалента для *spoken language machine translation* вариант *автоматический перевод устной речи*, который встретился дважды: в рецензии на англоязычную книгу по МП и на сайте «Лингвистика в России» со ссылкой на Группу речевой информатики Санкт-Петербургского института информатики и автоматизации РАН (впрочем, в русскоязычной части сайта этого института термин найти не удалось).

Определенную трудность вызывает представление базовых терминов *Лингвистические технологии* и *Речевые технологии* и их дифференциация,

соответственно, с понятиями *АОТ* и *Автоматическая обработка звучащей/устной речи (АОЗР)*. В англоязычной литературе разграничение между *Speech Technology = РТ* и *Speech Processing = АОЗР* проводится нечетко. Последний термин покрывает все прикладные задачи, связанные с автоматической обработкой устной речи, и здесь можно выделить четыре основных направления: цифровая (параметрическая) обработка речевого сигнала (ЦОРС), автоматический синтез речи, автоматическое распознавание речи, создание речевых корпусов. Одна из возможных точек зрения: термин РТ равнозначен термину Автоматическая обработка звучащей/устной речи, так как ЦОРС — тоже речевая технология, направленная на создание автоматических звуковых анализаторов/редакторов речи. Однако некоторые авторы к РТ относят только синтез, распознавание и корпусные технологии, т. е. понимают РТ уже, чем АОЗР.

6. Тематические классы терминов начальной версии русско-английского тезауруса

В итоге для начальной версии базового словника тезауруса было выделено порядка 3 тысяч терминов, к настоящему моменту описаны и внесены в электронный ресурс около 1100 терминов: дескрипторов — около 700, аскрипторов — более 400, связей между терминами — около 2500, источников терминов и их определений — 126. Множество терминов распадается на пять основных терминологических областей:

1. «Направления КЛ» — термины, называющие отдельные направления КЛ. Мотив выбора этой группы — получение по возможности полной картины о возможном предметно-тематическом составе тезауруса. Термины этой группы включены экспертом независимо от частоты их встречаемости в [Прил рус. 1];
2. «РТ» — относительно самостоятельное и минимально пересекающееся с остальными направление КЛ;
3. «КорпЛ» — базовое направление для статистических методов ИИ и различных эмпирических подходов, которые проникают во все направления современной КЛ;
4. «ИП» — одно из основных прикладных направлений КЛ;
5. «МП» — важнейшее приложение КЛ, традиционно интегрирующее всю проблематику АОТ, а в последнее время тесно взаимодействующее с РТ в рамках задачи автоматического устного перевода;
6. Группа терминов «метаязык». К этой области относятся термины фонетического, морфологического, лексического, синтаксического и семантического уровней языка и представлений этих уровней. Здесь систематично рассматривались терминологические обозначения семантических отношений — основополагающая для КЛ группа терминов, используемых как в ресурсах, например, лексико-семантических базах, так и в моделях языка и описаниях для систем АОТ, например, *агенс, начальная точка* и т. д.

Заключение

В статье излагаются наблюдения над структурой ПО КЛ и терминологией КЛ, сделанные в процессе создания начальной версии русско-английского тезауруса по КЛ. Отмечены особенности самой ПО (междисциплинарность, неравномерность развития разных направлений КЛ и др.), показано их влияние на терминологию и создание терминологических словарей. Кроме того, сформулирована главная особенность КЛ по сравнению с общей лингвистикой — направленность на передаваемую информацию, а не на формы языка. Именно наличие объекта, отличного от традиционной лингвистики, выделяет КЛ в самостоятельную научную дисциплину. Описаны методика и процесс обработки корпуса текстов для выделения терминов РКЛ и частота встречаемости некоторых терминов. Проанализированы особенности терминов КЛ и их отбора из основных лексикографических и дополнительных источников.

Благодарности

Работа выполнена при поддержке Российского гуманитарного научного фонда (грант № 10-04-12108в).

Авторы также выражают благодарность за плодотворное сотрудничество студентам и аспирантам МГУ им. М. В. Ломоносова, принимавшим участие в исследовании.

References

1. *Artificial* Intellect Explanatory Dictionary [Tolkovyi Slovar' po Iskustvennomu Intellektu].1992, available at: http://www.raai.org/library/tolk/aivoc.html
2. *Computaional* Linguistics Knowledge Web-site, available at: http://uniserv.iis.nsk.su/cl
3. *Krongauz M. A.* 2001. Semantics.
4. *Linguistics.* Information and Search Thesaurus of INION RAS. [Iazykoznanie. Informatsionno-Poiskovyi Tezaurus INION RAN]. 2007.
5. *Pererva V. M.* 1976. On Principles and Problems of Terms Selection and Terms Dictionary Formation [O Printsipakh I Problemakh Otbora Terminov I Sostavleniia Slovnika Terminologicheskikh Slovarei]. Problematika Opredelenii Terminov v Slovariakh Raznykh Tipov : 190–204.
6. *Sidorova E. A. Сидорова Е. А.* Multipurpose Dictionary Subsystem of Object Vocabulary Extraction [MnogotselevaIa Slovarnaia Podsistema Izvlecheniia Predmetnoi Leksiki]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2008" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2008"), 7 (14) : 475–481.
7. *Sokolova E. G., Zagorul'ko Iu. A., Kononenko I. S.* 2009. The Experience of Knowledge and Web Resources Classification for the Computaional Linguistics Knowledge Web-site [Opyt Sistematizatsii Znanii I Internet-resursov dla Portala Znanii po Komp'iuternoi Lingvistike]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2009" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2009"), 8 (15) : 465–470.
8. *Uszkoreit H.* What is computational linguistics?, available at: http://www.coli.uni-saarland.de/~hansu/what_is_cl.html
9. *Website* EAMT (The European Association for Machine Translation), available at: http://www.eamt.org/
10. *Zagorul'ko Iu. A., Borovikova O. I., Kononenko I. S., Sokolova E. G.* 2011. Designing of Russian-English Computational Linguistics Thesaurus [Razrabotka Russko-Angliiskogo Tezaurusa po Komp'iuternoi Lingvistike]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2011" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2011").

Таблицы содержат источники, наиболее часто используемые в начальной версии тезауруса, в основном, по указанным в статье тематикам. Всего на начальном этапе создания тезауруса зарегистрировано 127 источников.

Русскоязычные источники

	Тип источника	Название источника, библиографическая ссылка, URL	Дается определение	Встреч. дескриптор	Встреч. аскриптор
1	коллекция текстов	Коллекция текстов Диалог 2000–2010	1	116	78
2	интернет-ресурс	Интернет- энциклопедия «Википедия» http://ru.wikipedia.org	61	8	8
3	интернет-ресурс	Интернет- энциклопедия «Кругосвет» http://www.krugosvet.ru	11		2
4	книга	Трахтеров А. Л. Английская фонетическая терминология. М., Изд-во литературы на иностранных языках, 1962.	22	14	8
5	коллекция текстов	Корпус текстов по корпусной лингвистике		18	
6	книга	Искусственный интеллект. Справочник в 3-х томах. — М.: Радио и связь, 1990.	7	4	2
7	учебник	Баранов А. Н. Введение в прикладную лингвистику. Учебное пособие. — М.: Эдиториал УРСС, 2001. — 360 с.	10	2	4
8	учебник	Кобозева И. М. Лингвистическая семантика: Учебник. Изд. 4-ое — М.: Книжный дом «ЛИБРОКОМ», 2009. — 352 с. (Новый лингвистический учебник).	10	3	1
9	учебник	Кодзасов С. В., Кривнова О. Ф. Общая фонетика. М, РГГУ, 2001.	20	27	13
10	учебник	Тестелец Я. Г. Введение в общий синтаксис М.:РГГУ, 2001.	18		4
11	энциклопедический словарь	Лингвистический энциклопедический словарь. / Под ред. В. Н. Ярцевой. М.: Советская энциклопедия, 1990. — 685 с. [3 изд. 2002.]	6	1	1
12	энциклопедия	Энциклопедия «Русский язык». Гл. ред. Ю. Н. Караулов. Научное издательство «Российская энциклопедия», М., 1997.	9	7	3
13	интернет-ресурс	Сайт кафедры перевода и переводоведения ТюмГУ http://tc.utmn.ru	7	4	4

Англоязычные источники

1	интернет-ресурс	Интернет- энциклопедия «Wikipedia» http://en.wikipedia.org	24	4	18
2	документация	Документация к комп.программе Speech Analyzer 1.5–2002 http://www.sil.org/computing/speechtools/	39	18	2
3	интернет-ресурс	Интернет- энциклопедия «Glottopedia» http://www.glottopedia.de/index.php/Main_Page	13	5	
4	книга	J. Harrington, S. Cassidy. Techniques in Speech Acoustics. Text, Language, Technology, vol.8. Kluwer Academic Publishers. Dordrecht/Boston/London, 1999.	14	7	2
5	книга	J. Holmes and W. Holmes. Speech Synthesis and Recognition. 2nd edition/ Taylor&Francis, London/New York, 2001.	9	8	1
6	книга	K. Johnson. Acoustic and Auditory Phonetics. Blackwell Publishers, Cambridge, 1997.	38	19	4
7	учебник	Jurafsky Danial, Martin James H. Speech and language Processing: An Introduction to Natural language Processing, Computational Linguistics and Speech Recognition. Prentice-Hall, 2000	13	12	5
8	обзор	The Oxford handbook of computational linguistics (Ruslan Mitkov ed.) N.Y.: Oxford university press, 2003	62	43	21
9	статья	Igor Mel'čuk. Actants in semantics and syntax I: actants in semantics //Linguistics, 2004, 42:1, p.1-66	3	1	7
10	обзор	Survey of the State of the Art in Human Language Technology (Ronald Cole, editor in chief) 1996 http://cslu.cse.ogi.edu/HLTsurvey/	3	1	6
11	интернет-ресурс	Glossary of linguistic terms http://www.sil.org/linguistics/GlossaryOflinguisticTerms/	5	4	3

ПАДЕЖ КАК ПРИЗНАК ИДЕНТИЧНОСТИ ПРИ ЭЛЛИПСИСЕ В РУССКОМ ЯЗЫКЕ

Я. Г. Тестелец (yakov_ts@mail.ru)

Институт лингвистики РГГУ, Москва, Россия

Ключевые слова: падеж, эллипсис, идентичность, признак идентичности.

CASE AS A CHARACTERISTIC OF IDENTITY UNDER ELLIPSIS IN RUSSIAN

Ia. G. Testelets (yakov_ts@mail.ru)

Russian State University for Humanities, Moscow, Russian Federation

In Russian, all elliptical operations except N'-ellipsis require the identity of case values in NPs of the antecedent and the elliptical gap, and identity of role or grammatical relation does not suffice when cases are different. Ellipsis follows the six-case model, the peripheral cases, like partitive or 'the case of expected object', pattern with their more standard counterparts. With direct and indirect object NPs, however, case values may be different in the antecedent and the gap, e. g. with recipients, addressees, and the genitive of negation.

Key words: case, ellipsis, identity, identity characteristic.

В докладе будет проанализировано одно из грамматических требований к эллипсису в русском языке: совпадение граммемы падежа в антецеденте и в пробеле. Под эллипсисом понимается невыражение элементов синтаксической структуры при антецеденте, создающем контекстную избыточность. Наша цель — показать, что совпадение по падежу является одним из факторов эллипсиса в русском языке, и определить те случаи, когда падежные различия не играют роли.[1]

Мы будем ориентироваться на материал письменной речи — вопрос о том, действуют ли грамматические ограничения на эллипсис в русской разговорной речи, и если да, то какие именно, требует отдельного рассмотрения.

[1] Автор признателен анонимному рецензенту «Диалога» за ценные замечания, которые по возможности были учтены.

Исследователи отмечают, что информация, восстановимая из контекста, в разговорной речи обычно не выражается; при этом не упоминаются никакие препятствующие грамматические факторы (Земская и др. 1981; Земская 2004; Николаева 2000; Сиротинина (ред.) 2003; Кибрик, Подлесская (ред.) 2009).

Требование, чтобы граммемы некоторых морфологических категорий в антецеденте и пробеле совпадали, — часть более общей проблемы идентичности при эллипсисе (Merchant 2008; Wilder 1994; 1997 и др.). Например, в русском языке совпадения по роду (1) не требуется, но желательно совпадение по времени (2); согласно Е. В. Падучевой (1974: 166), в последнем случае имеет место жесткий запрет:

(1) *Саша пришел. — А Лена ~~пришла~~?*

(2) ?*Я дежурил вчера, а ты ~~дежуришь~~ завтра*.

Сразу отметим, что реакции носителей на нарушение требований грамматической идентичности антецедента и пробела при эллипсисе менее четки, уверенны и последовательны, чем реакции, например, на ошибки в управлении, поэтому не всегда легко получить надежный негативный материал. Знак * ниже означает, что носители литературного русского языка оценивают некоторую цепочку как скорее неприемлемую в качестве предложения в письменной речи.

По-видимому, условие совпадения падежа эллиптированной ИГ и ее выраженного антецедента в русском языке до сих пор подробно не обсуждалось. Е. В. Падучева отмечает, что «различия по способу оформления на поверхностно-синтаксическом уровне одной и той же синтаксической связи (например, связи глагола с его вторым дополнением) могут быть препятствием для сокращения» (Падучева 1974: 166); приводимые там же примеры показывают невозможность эллипсиса при различии предложно-падежной формы сокращаемой ИГ и ее антецедента:

(3) **Угол A больше ~~угла B~~ или равен углу B*.

(4) **Окружность f вписана ~~в правильный многоугольник~~, а окружность b описана около правильного многоугольника*.

Е. В. Падучева, однако, замечает, что различие вин. и род. падежей прямого дополнения, вызванное сентенциальным отрицанием («родительный при отрицании»), не учитывается как признак идентичности при эллипсисе (там же):

(5) *Тебе выдадут ~~пропуск~~, а мне не выдадут пропуска*.

Требование идентичности падежа не выполняется по крайней мере для одной эллиптической конструкции. Эллипсис ИГ с сохранением одного препозитивного определения — N'-эллипсис в генеративной грамматике или,

в терминологии Е. В. Падучевой, «эллипсис с сохранением представителя» (1974: 178), не чувствителен не только к падежу, но и к различиям между членами предложения:[2]

(6) Первую мировую войну он провел в окопах, во время второй ~~мировой войны~~ был журналистом.

(7) Правая рука сильнее левой ~~руки~~.

(8) Следует решить если не все ~~задачи~~, то большинство задач.

Граммемы категории падежа обычно обозначают функцию ИГ в составе той единицы — предложения или словосочетания, в которую она входит. Понятийный аппарат для описания падежных функций выбирается в зависимости от теоретических установок лингвиста: в терминах членов предложения, семантических ролей или составляющих; последнее характерно для генеративной грамматики, где функция сводится к структуре. Чтобы выделить фактор падежа при эллипсисе в чистом виде, рассмотрим три класса случаев, когда различия в падежном управлении не выражают очевидных различий в ролевой семантике и синтаксической структуре. Первый класс включает семантически не мотивированные или слабо мотивированные различия в фиксированном падежном управлении; второй класс — когда различие падежей в антецеденте и опущенной ИГ связано со свободной вариативностью падежного управления; третий класс — когда падежное различие вызвано тем, что или только опущенная, или только антецедентная ИГ получает форму род. п. при отрицании.

В тех случаях, когда выбор разных падежей при разных управляющих словах кажется семантически немотивированным или эта мотивация связана с трудно уловимыми различиями в значении, эллипсис так же невозможен, как и при явном различии ролей. Рассмотрим три случая различия падежа при совпадении роли — пациенса (или темы), экспериенцера и стимула.

Пациенс или тема: вин. или твор.:

(9) а. **Руками не размахивай, ~~руки~~ положи на стол.*
б. **Он долго вертел ключом в замке, а потом вытащил ~~ключ~~.*
в. **Он взял в руки ~~флаг~~ и взмахнул флагом.*
г. *В первом полугодии не учат иностранный язык. — *А во втором занимаются ~~иностранным языком~~?*

Экспериенцер: им. или дат.:

(10) а. **Я скучаю, и ~~мне~~ тоскливо.*
б. **В такую погоду детям скучно, поэтому ~~дети~~ грустят по дому.*

[2] К. И. Казенин (2007), анализируя эллипсис в составе русских ИГ, предполагает для случаев «зависания» одиночных определений при эллипсисе регулярную субстантивацию.

в. **Скучают ~~отдыхающие~~ и тоскливо в такую погоду отдыхающим.*
г. **Он почувствовал, что ~~ему~~ скучно.*
д. ??*Мне скучно. — И ~~ты~~ грустишь?*

Стимул: вин. или твор.:

(11) а. **Такого общества мы не уважаем и ~~таким обществом~~ брезгуем.*
б. ?*Молодым певцом заинтересовались режиссеры и ~~молодого певца~~ заметили фирмы звукозаписи.*
в. **Все пренебрегают их указаниями, и я тоже игнорирую ~~их указания~~.*

Приведенный материал показывает, что совпадения семантических ролей — пациенса, экспериенцера или стимула, — недостаточно для того, чтобы эллипсис был возможен: требуется также совпадение в падеже. Это, впрочем, не означает, что совпадение падежей является достаточным условием. Кроме того, роль — довольно грубая семантическая характеристика, и поэтому мы не можем быть уверены, что затруднение эллипсиса в (9–11) не связано с более тонкими семантическими различиями в антецеденте и пробеле.

Есть, однако, контексты, в которых падежные различия игнорируется: так, различие вин. и дат. падежей дополнения, по-видимому, при эллипсисе не учитывается:

(12) *Им обычно выдают барахло, но на этот раз ~~их~~ обеспечили хорошей аппаратурой.*

(13) *Мне поручено сказать, что мы горячо поддерживаем ~~Центральный Комитет Коммунистической партии Российской Федерации во главе с Геннадием Андреевичем Зюгановым~~ и доверяем Центральному Комитету Коммунистической партии Российской Федерации во главе с Геннадием Андреевичем Зюгановым и надеемся на их твердость и упорство в решении наших задач.* (Интернет).

(14) *А через неделю их всех сделали помощниками, а начальниками назначили побывавших заграницей, их же Каганович специально вызвал и предложил ~~им~~ не обижаться.* [Л. К. Бронтман. Дневники и письма (1932–1942)]; здесь и далее в квадратных скобках ссылки на Национальный Корпус русского языка.

(15) *Если вам позвонили и ~~вас~~ пригласили на собеседование, значит, вы справились с задачей.* [Елена Голованова. Дорогу молодым львам (2002) // «Домовой», 2002.06.04]

(16) *В том, что я решил помочь ему и «подключить» ~~его~~ к одному из компьютеров, чтобы усилить таким образом интеллектуальную мощь его*

мозга. [Геннадий Максимович. Фраза из дневника // Техника — молодежи, 1977]

(17) *Я обратил внимание, что это очень счастливая семья, ибо они любят ~~друг друга~~, поддерживают ~~друг друга~~ и помогают друг другу.* [Эльдар Рязанов. Подведенные итоги (2000)]

(18) *Нам это льстило и ~~нас~~ возвышало над будничным течением жизни*. [Нина Воронель. Без прикрас. Воспоминания (1975–2003)]

В спонтанных примерах из НКРЯ типа (13–18) ролевые различия между парами дополнений невелики, так как одушевленные ИГ в вин. п. скорее выступают в ролях адресата или реципиента, а не пациенса. Однако если взять примеры с ИГ-пациенсом, эллипсис окажется лишь ненамного более затрудненным:

(19) а. *Я помогла ~~ей~~ и поддержала её морально*.
б. ??*Я помогла ~~ей~~ и втащила её в автобус*.

Различие вин. и дат. падежа экспериенцера имеет значение для эллипсиса:

(20) а. **Его знобит, и ~~ему~~ нехорошо*.
б. **Ему нехорошо и ~~его~~ знобит*.

Возможен эллипсис ИГ, управляемой предлогом, с сохранением предлога; при этом падеж и роль обычно совпадают:

(21) *Находясь внутри здания, помните, что враг может поразить вас не только внутри ~~строения~~, но и снаружи строения*. [Ю. Хацкевич. Как выжить в «стреляющем» городе (2004) // «Солдат удачи», 2004.03.10]

(22) *Тем не менее, это оружие массового уничтожения перемалывало человеческие жизни не только до ~~разоблачения~~, но и после разоблачения*. [Семен Резник. «Протоколы сионских мудрецов» шагают во второе столетие (2003) // «Вестник США», 2003.10.01]

(23) *«OZ Рейсинг»: и до ~~сварки~~, и после сварки к эксплуатации не пригоден*. [Андрей Бойко. Танцоры «Диско» (2004) // «За рулем», 2004.03.15]

Эллипсис, однако, возможен только с некоторыми, главным образом, непервообразными предлогами, имеющими самостоятельное ударение. Первообразные предлоги эллипсиса управляемых ими ИГ не допускают:

(24) **в ~~шкафу~~ и на шкафу*

(25) **У Аксиньино мы вышли к ~~шоссе~~ и пошли по шоссе.*

Есть примеры, когда эллиптируемая ИГ и ее антецедент получают от своих предлогов разное падежное управление:

(26) *И право, я повторяю, должно быть выведено из политики и оказаться не только вне ~~политики~~, но и над политикой*. [Право на самоопределение и будущий миропорядок (2003) // «Неприкосновенный запас», 2003.07.16]

(27) *Людей казнили, а после я подбирал письма читателей и печатал статьи за ~~их казнь~~ и против их казни*. [Ю. О. Домбровский. Обезьяна приходит за своим черепом. Пролог (1943–1958)]

(28) *Кто-то вернется до пятнадцатого числа, кто-то вернется после ~~пятнадцатого числа~~*.

Под «идентичностью» падежа при эллипсисе подразумевается идентичность граммемы падежа, а не его формы. Это видно из поведения омонимичных падежных форм (29–31) и несклоняемых существительных (32–33):

(29) а. **Они не хвалят ~~власти~~* (род. ед.), *а льстят власти* (дат. ед.).
б. **Власти* (род. ед.) *они не хвалят, а льстят ~~власти~~* (дат. ед.)

(30) ??*Этой новости* (род.) *он или не услышал, или ~~этой новости~~* (дат.) *не поверил*.

(31) **Он поддался ~~страху~~* (дат.) *и натерпелся страху* (род. 2).

(32) **Патимат* (дат.) *стало жарко и ~~Патимат~~* (ном.) *почувствовала духоту*.

(33) ??*Кофе он запасался* (твор.) *и пил ~~кофе~~* (вин.) *в больших количествах*.

Сокращение левого элемента иногда возможно, если у разных падежей совпадает форма — примеры (34а–36а), как кажется, несколько более приемлемы, чем (34б–36б):

(34) а. ?*число, большее ~~семи~~* (род. п.) *или равное семи* (дат. п.)
б. ??*угол, больший ~~угла А~~* (род. п.) *или равный углу А* (дат. п.)

(35) а. ?*Он не избегал ~~этих людей~~* (род. п.), *а, наоборот, привлекал этих людей* (вин. п.).
б. **Они не избегали ~~этих разговоров~~* (род. п.), *а, наоборот, поддерживали эти разговоры* (вин. п.).

(36) а. ?*Он или не услышал ~~этой новости~~* (род. п.), *или не поверил этой новости* (дат. п.).

б. **Он или не услышал ~~этих новостей~~* (род. п.), *или не поверил этим новостям* (дат. п.).

Установленный выше факт, а именно то, что эллипсис «ориентируется» на падежные граммемы, а не на форму падежа, приводит к интересной проблеме — какой именно набор падежных граммем используется в правилах эллипсиса в русском языке?

Парадигма падежа является не наблюдаемым фактом, а абстрактной моделью и существует в разных версиях. Общепринято, что в русском языке целесообразно выделять как минимум шесть основных падежей, однако традиционно спорным остается выделение нескольких периферийных падежей — «второго родительного», или партитивного (*воды, сахару,*...), и «второго предложного», или локативного (*в шкафу, в лесу*, ...). Кроме того, в работах А. А. Зализняка (1967, 1973) обсуждалась возможность выделения еще по крайней мере трех периферийных падежей — «ждательного», «включительного» и «временно́го». Интересно поэтому увидеть, в пользу какого из вариантов падежной парадигмы «голосует» русский эллипсис.

Несамостоятельный, т. е. не имеющий собственных форм, «ждательный» падеж может выделяться ввиду особого управления нескольких интенсиональных[3] глаголов, в частности, глагола *ждать* (Зализняк 1967: 49–50; Зализняк 1973: 72; Ицкович 1982: 29–30). В терминах шестипадежной системы падежом дополнения при глаголе *ждать* выступает «у одушевленных обычно винительный, у неодушевленных обычно родительный; у некоторых неодушевленных употребляются оба... Фразы типа *я жду результат, я жду конец* воспринимаются по меньшей мере как стилистически неряшливые, а такие фразы, как *я жду матери*, *я жду учительницы* находятся, по-видимому, за гранью грамматической правильности. Такое же управление, как *ждать*, имеют производные глаголы *прождать, подождать, поджидать, ожидать*; в разговорной речи к этому типу управления тяготеют также глаголы *бояться, слушаться*... в меньшей степени глаголы *остерегаться, опасаться*» (Зализняк 1967: 49).

Неодушевленные существительные, допускающие оба падежа, — это, главным образом, названия конкретных предметов, которые могут в том или ином смысле «появиться» в поле зрения наблюдателя (*автобус, поезд, машина, письмо*,...) и некоторых отрезков времени (*весна, война*, ...). В итоге можно определить «ждательный» падеж на трех классах имен:

---

[3] Глагол, обладающий валентностью на дополнение, называется *интенсиональным*, если экстенсионал (истинностное значение) простого предложения, в которое он входит, не полностью зависит от экстенсионалов его частей, в том числе прямого дополнения. Например, истинность предложения *Мальчик искал единорога*, содержащего интенсиональный глагол *искать*, не требует существования единорогов, — в отличие от предложения с экстенсиональным глаголом *Мальчик кормил единорога*. Ниже используются некоторые наблюдения над управлением русских интенсиональных глаголов, сделанные нами вместе с группой соавторов (Борщев и др. 2008), а также неопубликованное корпусное исследование, проведенное А. Б. Летучим, с результатами которого автор имел возможность ознакомиться.

Табл. 1

	Одушевленные (варьирование редко)	**Неодуш. с варьированием**	**Неодуш. без варьирования**
«Ждательный»: *Он ждет*	*мать, сестру, Машу, добрую фею/ доброй феи*	*машину/машины, весну/весны, матч/матча*	*конца, результата, возвращения*
Вин.: *Он видит*	*мать, сестру, Машу*	*машину, весну, матч*	*конец, результат, возвращение*
Род.: *Таково свойство*	*матери, сестры, Маши*	*машины, весны, матча*	*конца, результата, возвращения*

Как отмечает А. А. Зализняк, не исключено, что в ходе эволюции системы «к управлению «ждательным» падежом перейдут, помимо названных выше, также такие глаголы, как *избегать, чуждаться, ослушаться* и т. п.» (Зализняк 1967: 49). В итоге А. А. Зализняк отклоняет идею особого «ждательного» падежа, так как смысловое различие между род. и вин. в тех случаях, где они чередуются, еще не полностью стерто и может иногда выражать противопоставление по определенности (там же: 50). Тем более, на наш взгляд, следует отказаться от идеи «искательного», «просительного» или «желательного» падежей, так как с именами, которые допускают чередование род. и вин. п. при глаголах *искать, просить, хотеть, желать, требовать* налицо гораздо более четкое семантическое различие, чем с глаголами *ждать, ожидать, дожидаться*: род. п. выражает неопределенность или партитивность (Ицкович 1982: 29–34; Борщев и др. 2008: 160–161).

Поскольку семантическое различие вин. и род. п. при глаголе *ждать* в современном языке по сравнению с остальными перечисленными глаголами наименее выражено и, следовательно, «ждательный» остается наиболее перспективным кандидатом в несамостоятельные падежи, рассмотрим эллиптические свойства его дополнений в вин. и род. п.

Если формы вин. и род. п. различаются, требуется, чтобы антецедент эллипсиса был в том падеже, который совпадает со «ждательной» формой данного имени:

(37) *Эту сотрудницу мы давно знаем и ~~эту сотрудницу~~ с нетерпением ждем.*

(38) **Этой сотрудницы нам не хватало и ~~эту сотрудницу~~ мы с нетерпением ждем.*

(39) ?*Следующий вторник мы как раз имеем в виду и ~~следующего вторника~~ с нетерпением ждем.*

Если сама «ждательная» форма является антецедентом, вин. или род. падеж в эллиптическом пробеле должен с ней совпадать:

(40) а. *Мы встречу с президентом ждем и ~~встречу с президентом~~ репетируем.*
б. **Мы встречи с президентом ждем и ~~встречу с президентом~~ репетируем.*

(41) а. *Мы встречи с президентом ждем и ~~встречи с президентом~~ в то же время опасаемся.*
б. **Мы встречу с президентом ждем и ~~встречи с президентом~~ в то же время опасаемся.*

Эти факты говорят против выделения особого «ждательного» падежа. Во-первых, антецедент может быть в «ждательном», а мишень эллипсиса — в «обычном» род. или вин. падеже (40а–41а), что было бы мало вероятно, если бы «ждательный» с точки зрения эллипсиса представлял собой отдельный падеж. Во-вторых, «ждательная» форма, омонимичная с вин., требует, чтобы совпадающая с ней ИГ была в вин., а омонимичная с род., требует род. п. же от совпадающей формы, из чего можно заключить, что «ждательный» падеж в грамматике эллипсиса никак не проявляется, а проявляются то род., то вин. п., в соответствии с вариантами управления.

Еще один несамостоятельный «включительный» падеж (*пошел в армию, в солдаты…*) (Зализняк 1967: 50–51), предположительно совпадающий для части имен с вин., а для части — с им. п., а также «слабо дифференцированный», по А. А. Зализняку (1973: 76), второй предложный падеж (*в саду, шкафу…*), как кажется, не могут быть проверены на эллиптический тест, так как антецедент и мишень эллипсиса должны находиться в той же конструкции (*выдвинули ~~в академики~~ и выбрали в академики*; *была оставлена ~~на снегу~~ и валялась на снегу*). Примеры типа (42–43) не показательны, так как они, скорее всего, неприемлемы вследствие существенного различия ролевой семантики, и несовпадение в падеже не имеет значения:

(42) а. **Детьми они играли ~~в армию~~, и, повзрослев, пошли в армию.*
б. **Детьми они играли ~~в моряков~~, и, повзрослев, пошли в моряки.*

(43) **Вчера в снеге экологи обнаружили фенол, а сегодня ~~экологи — в снегу обнаружили~~ бутылку водки.*

«Факультативный», по А. А. Зализняку, второй родительный (партитивный) падеж (*дать чаю, воды…*) (Зализняк 1973: 80–84), выступая в качестве антецедента, как кажется, допускает эллипсис ИГ в обычном вин. п., что говорит против выделения его в качестве отдельного падежа, отличного от вин.:

(44) *Сначала налил в чашку чаю, потом вылил ~~чай~~* (**чаю*).

(45) *Снегу сверху набросали и ~~снег~~* (**снегу*) *утоптали.*

Совпадение с им. п. менее приемлемо:

(46) ?*Снегу утром нападало сантиметров пять, потом ~~снег~~ перестал.*

Можно заключить, таким образом, что эллипсис «ориентируется» на минимальную шестипадежную модель русской именной парадигмы.

Второй случай, когда фактор падежа выявляется в «чистом» виде, — антецедент с вариативным управлением, а опущенная ИГ — с не вариативным, причем падеж опущенной ИГ соответствует либо не соответствует падежу антецедента, в зависимости от выбора последнего. Обнаруживается, что эллипсис требует совпадения падежа:

Вин. и род. п. дополнения при интенсиональном глаголе:

(47) а. *Люди на берегу ждали наш корабль/нашего корабля.*
б. *Люди на берегу знали ~~наш корабль~~ и ждали наш корабль* (**нашего корабля*).

Варианты с предложно-падежным управлением:

Совпадение требуется и в том случае, когда ИГ-антецедент, или эллиптированная ИГ, или они обе управляются предлогом:

(48) а. *Об этом можно узнать у Маши/от Маши.*
б. *Об этом можно спросить ~~у Маши~~ и узнать у Маши/*от Маши.*

(49) а. *Встреча началась скандалом/со скандала.*
б. *Встреча и началась скандалом (*со скандала), и закончилась ~~скандалом~~.*

Таким образом, вариативное управление подтверждает общий принцип, согласно которому падеж ИГ в пробеле и антецеденте должен совпадать.

Третий случай, когда фактор падежа при эллипсисе можно предположительно отделить от ролевой семантики и структуры, — управление род. п. объекта и субъекта при сентенциальном отрицании (род. п. при отрицании). В этой конструкции ИГ в пробеле, ее антецедент и управляющее слово различаются только признаком полярности и падежом, выбор которого зависит от этого признака.

Рассмотрим отдельно случаи, когда род. п. оформляет актант, соответствующий подлежащему непереходного глагола (субъектный род. п. при отрицании) и когда он оформляет актант, соответствующий прямому дополнению (объектный род. п. при отрицании).

При бытийном глаголе-сказуемом эллипсис возможен в обе стороны:

(50) а. *У него нет денег, а у меня ~~деньги~~ есть.*
б. *У него есть деньги, а у меня ~~денег~~ нет.*

(51) а. *У него не было денег, а у меня ~~деньги~~ были.*
б. *У него были деньги, а у меня ~~денег~~ не было.*

При знаменательных глаголах-сказуемых субъект в род. п. чаще всего мало приемлем и как антецедент, и как мишень эллипсиса ИГ в им. п., причем суждения носителей о таких предложениях довольно сильно варьируют. В целом, как кажется, случай «антецедент в род. п. — мишень в им. п.» несколько более приемлем, чем противоположный случай «антецедент в им. п. — мишень в род. п.»:

(52) а. ?*Писем в понедельник не пришло, а во вторник ~~письма~~ пришли.*
б. **Письма в понедельник пришли, а во вторник ~~писем~~ не пришло.*

(53) а. *— Новых данных не поступало. ? — А ~~новые данные~~ ожидались?*
б. *— Новые данные не поступали. * — А ~~новых данных~~ ожидалось?*

При эллипсисе отличие объектного род. п. при отрицании от обычного вин. п. прямого дополнения, как кажется, не играет роли. Поскольку форма глагола, в отличие от субъектного род. п. при отрицании, не меняется, рассмотреть случаи, когда род. п. является мишенью эллипсиса, чаще всего невозможно: в этом случае при чередовании падежей нельзя показать, что в пробеле именно род., а не вин. п.:

(54) *Я написал ему ответ, а ты не написала ему ~~ответ~~ //~~ответа~~.*

В тех немногих случаях, когда вин. п. прямого дополнения затруднен, напр., с местоимением *это* (Mustajoki, Heino 1991), эллипсис с антецедентом — ИГ в вин. п. допускается:

(55) а. *Я этого (?это) не писал.*
б. *Она написала это, а я ~~этого~~ не писал.*

ИГ в род. п. может выступать в качестве антецедента ИГ в вин. п.:

(56) а. *Пети мы пока не звали, но ~~Петю~~ позовем.*
б. *Петю мы пока не звали, но ~~Петю~~ позовем.*

(57) *Мы вещей не паковали, но ~~вещи~~ уже приготовили.*

Можно заключить, что эллипсис в русском языке требует совпадения граммем падежа в антецеденте и пробеле в терминах шестипадежной системы, но это требование ослабляется для винительного, родительного и дательного падежей беспредложных дополнений.

References

1. *Borshchev V. B., Paducheva E. V., Parti B. Kh., Testelets Ia. G., Ianovich I. S.* 2008. Genetive in Russian Language, Reference and Semantic Types [Roditel'nyi Padezh v Russkom Iazyke, Referentnost' I Semanticheskie Tipy]. Ob"ektnyi Genitiv pri Otritsanii v Russkom Iazyke. Issledovaniia po Teorii Grammatiki, 5 : 148–175.
2. *Itskovich V. A.* 2010. Sketches of Syntactic Norm [Ocherki Sintaksichekoi Normy].
3. *Kazenin K. I.* On Some Ellipsis Limitations in Russian [O Nekotorykh Ogranicheniiakh na Ellipsis v Russkom Iazyke]. Voprosy Iazykoznaniia, 3 : 92–107.

4. *Kibrik A. A., Podlesskaia V. I.* 2009. Stories about Dreams. Russian Spoken Discourse Corpus Research [Rasskazy o Snovideniiakh. Korpusnoe Issledovanie Ustnogo Russkogo Diskursa].
5. *Merchant J.* 2008. An Asymmetry in Voice Mismatches in VP-ellipsis and Pseudogapping. Linguistic Inquiry, 39 : 169–179.
6. *Mustajoki A., Heino H.* 1991. Case Selection for the Direct Object in Russian Negative Clauses. Part II: Report on a Statistical Analysis. Slavica Helsingiensia, 9.
7. *Nikolaeva T. M.* 2000. From Sound to Text [Ot Zvuka k Tekstu].
8. *Paducheva E. V.* 2007. On Syntax Semantics [O Semantike Sintaksisa].
9. *Sirotinina O. B.* 2004. Spoken Russian in the Russian Literary Language Functional Styles System [Russkaia Razgovornaia Rech' v Sisteme Funktsional'nykh Stilei Sovremennogo Russkogo Iazyka].
10. *Wilder Ch.* 1994. Coordination, ATB and ellipsis. Minimalism and Kayne's Antisymmetry Hypothesis. Groningen Arbeiten zur Germanistischen Linguistik, 37 : 291–329.
11. *Wilder Ch.* 1997. Some Properties of Ellipsis in Coordination. Studies on Universal Grammar and Typological Variation. Linguistik Aktuell/Linguistics Today, 13 : 59–107.
12. *Zalizniak A. A.* 1973. Russian Noun Inflexion [Russkoe Imennoe Slovoizmenenie].
13. *Zalizniak A. A.* 1973. On the Understood Meaning of the Term "Case" in Linguistic Descriptions.I [O Ponimanii Termina "Padezh" v Lingvisticheskikh Opisaniiakh. I] : 53–87.
14. *Zemskaia E. A., Kitaigorodskaia M. V., Shiriaev E. N.* 1981. Spoken Russian. General Questions. Inflexion. Syntaxis [Russkaia Razgovornaia Rech'. Obshchie Voprosy. Slovoobrazovanie. Sintaksis].
15. *Zemskaia E. A.* 2004. Spoken Russian: Linguistic Analysis and Learning Problems [Russkaia Razgovornaia Rech': Lingvisticheskii Analiz I Problemy Obucheniia].

О ДИНАМИЧЕСКОЙ СЕМАНТИКЕ СЛОВА *МИФ*

В. М. Труб (trub44@ukr.net)

Институт украинского языка НАН Украины, Киев, Украина

В статье показано, что различные значения существительного миф могут рассматриваться как результаты сдвига фокуса внимания на разные компоненты его исходного толкования: само содержание мифа, опровержение содержащейся в нём информации, позитивную аксиологическую оценку персонажа или зафиксированного в нём явления.

Ключевые слова: миф, семантика, динамическая семантика, толкование.

ON THE DYNAMIC SEMANTICS OF THE WORD ‘MYTH’

V. M. Trub (trub44@ukr.net)

The Institute of Ukrainian language of the National Academy of Ukraine, Kyiv, Ukraine

One of the main problems of modern semantic research is polysemy. The way to explain polysemy consists in the representation of meaning of a polysemantic word as a result of semantic transitions. The derivative meanings can be explained as a result of transferring attention on one of the components of the initial meaning and suppressing another meaning. We illustrate this principle by the Russian polysemantic word миф ‘myth’. It is shown that different meanings of this noun can be interpreted as a result of focusing attention on the different components of its initial defi nition such as the content of the myth, the disproof of this content, the positive axiological evaluation of different aspects of this content.

Key words: myth, polysemy, dynamic semantics,interpretation.

Как известно, в современных семантических исследованиях едва ли не центральной проблемой лексики и грамматики признаётся многозначность. При этом «львиная доля многозначности — это не омонимия, а полисемия… Способ объяснить многозначность состоит в том, чтобы представить значение многозначного слова как полученное в результате семантических переходов, сдвигов» [Падучева 2005: 395]. Ещё раньше аналогичный подход был убедительно продемонстрирован на русском материале при описании дискурсивных слов, реализованном в рамках проекта Д. Пайара и его коллег — ср. [Путеводитель 1993]. Вторичные, производные значения могут быть, в частности, представлены как результат сдвига фокуса внимания на один из компонентов исходного значения и подавление другого. Одним из примеров слов с такой «многоликой» семантикой является часто употребляемое в современных контекстах существительное *миф*, которое и станет предметом рассмотрения в данной работе, существенно опирающейся на материал, почерпнутый в Национальном корпусе русского языка.

1. На прототипическом уровне *миф* может быть в первом приближении охарактеризован как отражённое в сознании многих поколений того или иного этноса (например, греческого) и эпизодически зафиксированное в литературных памятниках повествование, отражающее фантастическое мировосприятие ранних поколений этноса, их верования, вымышленные представления о происхождении Земли, Вселенной, природных явлениях, эсхатологии и т. д.

В прототипическом употреблении существительное *миф*, как и обозначение любого текста — информационного объекта (сказки, легенды, рассказа и т. д.) имеет валентность содержания, нормативно представимую предложно-падежной формой *о (об) + N (предл.)*. При этом можно условно выделить два близких значения данного существительного, различия между которыми связаны с разными вариантами оформления валентности содержания:

1) *миф о X* ≈ ‘Информационный объект — отражённое в общественном сознании многих поколений определённого этноса значимое для него фантастическое повествование, главным героем (протагонистом) которого является вымышленный персонаж или существо X’ — ср. *миф о Геракле, миф об Артемиде, миф об Аполлоне и Дафне:*

(1) *Взять, например,* ***миф об Антее****…;* ***Миф об Энее*** *был древний…; И мне припомнился* ***миф об Арахнее****…; Мне кажется, меня лучше поймут, если я напомню* ***миф об Антее****, чем если я дам марксистское объяснение…; Этот* ***миф об Аттисе*** *мы заимствуем из христианской апологии Ариобия.*

При этом в фокусе внимания оказывается отсылка к самому содержанию повествования, косвенным обозначением которого служит наименование, заполняющее данную валентность.

Другой вариант оформления валентности содержания предусматривает её заполнение в виде именной группы (ИГ), образующей «прямое» представление [Падучева 2004] одной или более пропозиций (Р) — мифологем,

составляющих миф. В этом случае существительному *миф* сопоставляется другое, несколько модифицированное толкование:

2) *миф о Р* ≈ ‘Информационный объект — отражённое в общественном сознании многих поколений определённого этноса значимое для него фантастическое повествование с сюжетом Р’.

При этом краткое раскрытие содержания мифа может достигаться:

а) посредством заполнения валентностей существительного — вершины ИГ, фигурирующей в позиции валентности содержания:

(2) *В частности высказывалось мнение, будто даже древнегреческий* ***миф об «эстафете власти»*** *на небесах* ***от Урана к Кроносу-Сатурну и затем к Зевсу-Юпитеру*** *заимствован из аналогичного древнейшего шумерского предания.*

б) посредством заполнения валентности содержания всей сложной ИГ, включающей в свой состав причастный оборот или придаточное относительное:

(3) *Так, в живописи часто являлся сюжетом* ***миф об Аполлоне и Дафне, превращённой в лавровое дерево****, или* ***миф о сёстрах Феронта, обратившихся в плакучие ивы****; В нём нашёл отражение* ***миф о природных духах, погибающих при смене сезона (рождённое зимой из снега существо при наступлении лета тает, превращаясь в облачко)****; Все знают* ***миф об Икаре, который взлетел к самому солнцу и погиб потому, что жар растопил воск, скреплявший перья****; …Это* ***миф об Орфее, который ведёт из царства мёртвых к новой жизни любимую Эвридику, но не исполняет назначенного самоотречения****…; Ведь* ***миф об Атланте, который пришёл в Европу, чтобы насадить в ней цивилизацию****, был куском поэтических преданий, озарявших утро флорентийской истории; Не мог же* ***миф об архитекторе Дедале, который якобы построил для критского царя Миноса легендарный замок Лабиринт****, возникнуть на пустом месте.*

2. Другое (переносное) значение существительного *миф* реализуется в случаях, когда оно приобретает синтаксическую функцию предиката, образующего главную рему предложения или фразового комплекса. При этом в качестве темы выступает «прямое» выражение пропозиции Р — информации, которая навязывается общественному сознанию как достоверная. Тем самым подразумевается, что данная информация уже актуализована в сознании говорящего и адресата. Таким образом, общая структура фразы или комплекса имеет вид *Р — миф*:

(4) *А то, что* ***учебно-издательский бизнес приносит какие-то баснословные барыши, — миф****; От бритья волос густоты не* ***прибавится****. Это* ***миф****; Утверждение, что* ***чем меньше конкурентов, тем лучше — миф****; Глобальное потепление — миф?; То, что* ***сейчас мы реформируем советскую систему образования, — это миф****, и от него пора отказаться. Говорят, что* ***рестораны потеряют доходы, если будет ограничено курение. Это миф****.*

В этих случаях предложение или фразовый комплекс выполняет функцию «разоблачения» — однозначной негативной оценки истинности (ОИ) информации, выраженной пропозицией Р. Таким образом, факт недостоверности информации, образующей содержание мифа, в рассматриваемых структурах приобретает статус единственного ассертивного компонента. В двух исходных значениях слово *миф* указывает на повествование о том, что происходит с вымышленными персонажами в некотором воображаемом мире. А данное его значение указывает на неправомерность приписывания того или иного свойства сущностям реального, современного мира, к которому принадлежат и говорящий, и слушающий.

Аналогичным образом интерпретируются предложения, у которых на поверхностном уровне связочная часть *Это миф* инвертирована относительно опровергаемой пропозитивной части:

(5) *Так что **это миф,** что **моей рукой написано** — больше для красного словца; **Это миф,** о том, что **правые настаивают на сугубо западном пути развития России**.*

Важно подчеркнуть, что в синтаксической позиции предиката значение негативной ОИ реализует именно существительное *миф*. Близкое по значению, но не тождественное существительное *сказка* (ср. [Пропп 1986]), также употреблённое в предикативной функции в единственном числе, приобретает значение положительной аксиологической ОИ:

(6) *Сочи — это мечта! Сочи — **это сказка**! Сочи — это город, из которого не хочется уезжать!; А побывать весной в Одессе — **это сказка**!; **Это сказка**, санаторий: запасы крупы, тюки с консервированной тушёнкой, вода; «Котлеты — **это сказка**, это радость для всего дома», — говорил он тогда, запустив свои маленькие сухие руки в огромную миску с фаршем.*

В других случаях актуальное членение предложения преобразуется: в его тематической части так или иначе утверждается о ложности некоторой распространённой информации, в связи с чем она именуется как *миф*, а в рематической части располагается «прямое» выражение этой информации:

(7) *Интересно, как во всё это встраивается распространённый в нашем обществе **миф** о том, что **Москва — это другая страна, в которой живут другие люди**; Есть устойчивый **миф** о том, что **чеченский террор на долгие годы — как на Ближнем Востоке**; С одной стороны, существует **миф** о том, что **Россия находится в эпицентре главных мировых событий, что всё так или иначе связано с ней**; А кому незнаком расхожий **миф** о том, что **безумие сопутствует гениальности**?; Это **миф** о том, что **Россия (преемник Союза) — Великое Государство;** При этом среди пациентов-язвенников упорно живёт **миф о чудесной «водкотерапии»;***

Иногда могут указываться и причины, послужившие основанием для опровержения соответствующей информации:

(8) *В эпоху социалистического реализма возник* ***миф*** *о том, что якобы* ***по вызову Сталина они приезжали в Москву для консультаций…****, однако* ***дневники Вернадского свидетельствуют, что никуда он [Сахаров] из Борового не уезжал и настроен был определённо против Курчатова***.

В ряде случаев говорящий преследует цель разоблачения сразу нескольких мифов, которые последовательно нумеруются:

(9) *Мне представляется, что попытки выполнить строгий отбор основаны на нескольких заблуждениях, слишком часто приводящих к неверному выбору.* ***Миф1. Прошлый положительный опыт гарантирует будущий успех.*** *…****Миф 2. Специализация равнозначна эффективности****…* ***Миф 3. Хорошо можно делать только одно дело.***

3. В рассмотренных выше примерах функцию негативной ОИ выполняет само предложение. Поэтому в качестве субъекта опровержения в них естественно выступает сам говорящий. В то же время часто может идти речь о фактах опровержений, ранее осуществлённых и другими лицами, которых говорящий не оспаривает. В таких случаях в предложении обычно фигурируют разнообразные предикаты, обозначающие акты опровержения некоторого ложного информационного объекта — ср. *развенчивать, развеивать, разрушать, подвергать сомнению дискредитировать, быть несостоятельным*… В этих условиях информация, демонстрация ложности которой описывается данными глагольными формами, представляется в виде прямого обозначения пропозиции, заполняющей содержательную валентность существительного *миф*, которое фигурирует в позиции актанта ложного информационного объекта при предикатах опровержения. При этом содержащееся в значении слова *миф* указание на недостоверность соответствующей информации плеонастически дублирует тот же смысл, который образует ассерцию предикатов опровержения:

(10) ***Миф*** *о том, что* ***военные необщительны, был развенчан****;* ***Не состоятелен*** *и* ***миф*** *о том, что* ***в Москве тратят намного больше денег, чем в провинции****; Одновременно* ***испаряется*** *усиленно внедряемый конфессионально-политический* ***миф*** *о том, что* ***в России существует всего четыре «традиционные» религии — православие, ислам, буддизм, иудаизм****; На полях Подмосковья было нанесено первое крупное поражение немецко-фашистской армии во второй мировой войне,* ***развеян миф о её непобедимости****; Психологи, протестировавшие интуитивные способности более 15 тыс. человек,* ***пришли к выводу,*** *что* ***женская интуиция — это миф****; Татьяна с блеском и азартом* ***развенчивала миф*** *о том, что* ***женщина с высшим образованием, мать — почти обязательно человек душевный и тактичный****; Я хотел бы* ***разрушить*** *ещё один* ***миф*** *о том, что* ***Федеральная программа***

выборочно кому-то даёт, кому-то не даёт; Миф *о том, что* ***Сампрас на травяных кортах непобедим, оказался развеян*** *в одночасье.*

Иногда в предложении может актуализироваться информация, ложность которой считается общеизвестной и потому квалифицируемой как миф:

(11) ***Миф*** *о том, что* ***красивая женщина непременно глупа,*** *выдуман из зависти и явно в женском кругу; Сталинграда ещё не было,* ***миф о войне как о прогулке*** *ещё не развеялся.*

В то же время важно указать на принципиальную несочетаемость данных предикатов со словом *миф* в его исходном значении — ср. (12) **развеять (развенчать, разрушить, подвергнуть сомнению…) миф о Геракле*. Ведь недостоверность подобных повествований и так считается общеизвестной. Поэтому выражения типа (12) обречены на коммуникативный провал.

Полезно также сопоставить слова *миф* (в рассматриваемом значении) и *иллюзия*. Подобно *мифу иллюзия* может быть *развеяна, разрушена*. Однако, в отличие от *мифа*, иллюзия не может быть *развенчана*. Ведь *развенчание* всегда преследует цель не только продемонстрировать несоответствие чьего-либо мнения действительности, но также и то, что это заблуждение является следствием обмана, т. е. нацелено и на уличение лжеца. Между тем *иллюзия* является следствием **самообмана.**

Ср. также допустимость фразы (13а) *Хочу развеять (разрушить) твою иллюзию* и аномальность (13б) **Хочу развеять (разрушить) твой миф*. Ведь носителем иллюзии может быть и одно лицо, тогда как миф всегда «адресован» общественному, коллективному сознанию.

4. Существительное *миф* часто может фигурировать в позиции результативной валентности глаголов, указывающих на кардинальное изменение природы той или иной сущности, часто приводящее и к изменению её таксономии. Если при этом в качестве пациенса соответствующего преобразования выступает некоторое положение дел, реальность существования которого по крайней мере в определённый период времени не вызывает сомнения, то превращение его в *миф* (т. е. в нереальность) указывает на прекращение его существования:

(14) *Однако, когда я начал работать,* ***бесплатная медицина стала мифом,*** *и мечту в полной мере реализовать не удалось; Трезвый ум обманул реалиста: пришли большевики, и* ***его «счастье» превратилось в миф…***

В то же время эта же модель (когда в качестве пациенса преобразования рассматривается общеизвестный факт или факты, т. е. сведения, входящие в общий фонд знаний говорящего и слушающего) может использоваться для отражения неверного мнения других лиц. В этих случаях имеется в виду, что другие лица ставят под сомнение (приравнивают к мифу) факты, достоверность которых является общеизвестной:

(15) *Она [правнучка Хрущёва] до сих пор уверена, что* ***Хрущёв*** *не* ***стучал ботинком по трибуне ООН****, просто* ***это*** *уже* ***стало мифом****, от которого никто не хочет отказаться, в том числе и ООН; Да с каких это пор* ***жертвы коллективизации, гладомора 30-х годов, Соловки и многое другое стало мифом****?*

Такой же тип интерпретации сохраняется и тогда, когда автор просто эксплицирует, что цитирует чужое мнение, согласно которому некоторое общеизвестное явление квалифицируется как миф:

(16) *В обществе до сих пор существует мнение, что* ***мобильные вирусы — это миф****, а антивирусные компании продают воздух.*

Существительное *мифами* (во множественном числе и творительном падеже), также как и *слухами*, часто может фигурировать в позиции результативной валентности предиката *обрастать*, употребляемого в переносном значении нецелевой каузации. В качестве же первого (объектного) актанта данного предиката выступает обозначение какой-либо реальной ситуации:

(17) а) *«Холодная война» сделала недоступной для современников* ***истинную историю битвы за космос****, которая* ***обросла мифами****, как древняя Атлантида;*
б) *В последнее время* ***тема ввозных пошлин обросла мифами****;*
в) *— Эндрю, спрашиваю я, — с годами* ***история корабля-призрака обросла мифами****, о ней написаны романы, сняты художественные фильмы;*
*г) …****Страшное происшествие уже обросло мифами****. То говорят, что врача не пустили три квартиры (не пускала одна, не выпускали три).*

В подобных случаях формы *мифами*, в отличие от примеров, рассмотренных выше, указывают не на нереальность некоторой ситуации, а на саму информацию о нереальной ситуации. При этом сама ложная информация (то, что является содержанием соответствующего мифа) остаётся не раскрытой, хотя и может факультативно эксплицироваться в пределах более широкого контекста. Так, в примере (17г) эта информация раскрывается во втором предложении: «*…врача не пустили три квартиры (не пускала одна, не выпускали три*». А первый (объектный) актант *обрастать*, указывающий на событие, которое мифологизировалось, соотносится с косвенным обозначением тематической валентности (= заглавия) существительного *миф* в его исходном значении. Таким образом, ситуация, на которую указывает первый актант глагола *обрастать*, является и причиной появления мифов, и одновременно темой, которой они посвящены. В примерах (17*а* — 17*г*) таким темам соответствуют фрагменты: «*История битвы за космос, …тема ввозных пошлин, …История корабля-призрака, …страшное происшествие*».

Понятие «миф» (в рассматриваемом значении) полезно сопоставить со слухами, которыми тоже может *обрастать* то или иное событие. В отличие от мифов как информации заведомо недостоверной, слухи представляют собой информацию с не гарантированной достоверностью. Поэтому можно *проверять слухи*,

т.е. устанавливать их истинность или ложность, но не **проверять миф*, т.е. устанавливать достоверность информации, о которой заведомо известно, что она недостоверна.

С точки зрения выражения (18) *Слухи о Р сильно(несколько) преувеличены* ситуации Р сопоставляется условное «числовое» значение на определённой шкале, которая образует как бы «родовой термин» относительно ситуаций типа Р. При этом утверждается, что в качестве реального положения дел выступает другая ситуация Q, которой на этой условной шкале соответствует намного (или несколько) меньшее значение, чем то, которое сопоставлено ситуации Р. Тем самым на глубинном уровне значение (18) содержит противопоставительное отрицание ‘не Р, а Q’:

(18) а) *Когда говорят, что у населения скопилось огромное количество средств, замечу,* ***эти слухи сильно преувеличены*** ≈ ‘У населения скопилось не огромное количество средств, а небольшое их количество’.

Здесь полезно упомянуть и известное изречение Марка Твена (19) *Слухи о моей смерти сильно преувеличены*. С одной стороны, слухи — это информация с негарантированной достоверностью. В то же время раз сам Марк Твен пишет о слухах о своей смерти, то, следовательно, он жив, т.е. эти слухи являются ложными. С другой стороны, между жизнью и смертью (по крайней мере с точки зрении языка) соотношение не шкальное, а дихотомическое — человек или жив, или мёртв. Между тем выражение *преувеличенные слухи* апеллирует к значению на шкале. Это противоречие и объясняет каламбурный эффект данного высказывания.

5. Следующий тип семантической интерпретации существительного *миф* предусматривает, что в коммуникативный фокус попадает позитивное отношение общества к тому, что образует содержание мифа независимо от степени его достоверности. Тем самым актуализируется компонент исходных толкований, указывающий на **значимость** содержания соответствующего повествования для **всего** общественного сознания. Такой миф обычно характеризуется как *красивый, светлый, великий* — ср. (20) *Если мы хотим что-то изменить в родной стране, то нужен если не миф, то правда, «красивая, как миф»*.

Можно выделить по крайней мере три подвида данного значения:

а) высокая позитивная оценка персонифицированного героя — лица, имя которого чаще фигурирует в позиции 1-го актанта глаголов превращения или создания:

(21) ***Френк Синатра*** *стал* ***мифом*** *Америки;* ***Он*** *[Высоцкий] сотворил из себя* ***миф****, понятный миллионам и понятый миллионами; Но зато мы имеем великий* ***миф о Нижинском***.

б) косвенное указание на явление, вызвавшее к себе повышенный интерес, длительное и пристальное внимание общества:

(22) *…Почему «Чайка» — комедия и что такое «Вишнёвый сад» (не пьеса, а* ***то, что вырвавшись из неё,*** *уже стало* ***мифом****)?; …****Их «амур труа»*** *стало*

***мифом** столетия. О нём написаны исследования; Что, **спектакль «Принцесса Турандот» с Юлией Борисовой** — действительно **миф**?*

в) раскрытие содержания мифа через «развёртывание» валентности содержания — «прямое» представление объекта высокой оценки — того, что составляет предмет всеобщей любви, гордости, почитания, позитивного имиджа:

(23) *Можно сказать, что Розанов создал **миф о Пушкине как потерянном рае нашей литературы**; Он создал **миф о сильном мужчине**, но сам таковым не был; Судьба Диснея блестяще иллюстрирует вечно живой американский **миф о чистильщике сапог, ставшем миллионером**; Возможно, я разрушила **миф о французской галантности**?; Тихонов — «последний герой» советской эпохи, создавший красивый — в прямом и переносном смысле — **миф о сильном, но не ожесточившемся человеке, ставший олицетворением мужского обаяния**…; У каждого индуса своя судьба, у каждого племени свой язык, … разнородную, многоверующую нацию объединяет лишь **миф о едином отце государства Махатме Ганди**…*

Как видим, в подобных ситуациях факультативно реализуется «географическая» валентность существительного *миф*, функция которой состоит в указании на народ, в общественном сознании которого хранится значимая для него информация. Эта валентность может быть представлена формой *N род.* — ср. *миф **Америки*** или прилагательным (***американский** миф*). Такие же сведения могут быть уточнены и в рамках более широкого контекста (ср. *Тихонов — последний герой **советской** эпохи; У каждого **индуса** своя судьба…*).

Таким образом, различные значения существительного *миф* могут рассматриваться как результаты сдвига фокуса внимания на разные компоненты его исходного толкования: само содержание мифа, опровержение говорящим или другими лицами содержащейся в нём информации, позитивную аксиологическую оценку персонажа или явления, зафиксированного в мифе.

References

1. *Paducheva E. V.* 2004. Dynamic Models in Lexical Semantic [Dinamicheskie Modeli v Semantike Leksiki].
2. *Paducheva E. V.* Modality as a Script [Modal'nost' kak Stsenarii].Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2005" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2005") : 395–400.
3. *Propp V. Ia.* 1986. The Historical Roots of the Wonder Tale [Istoricheskie Korni Volshebnoi Skazki].
4. *Russian* Discursive Words Guide [Putevoditel' po Diskursivnym Slovam Russkogo Iazyka]. 1993.

К ПРОБЛЕМЕ НЕЕДИНСТВЕННОСТИ ЛИНГВИСТИЧЕСКОГО ОПИСАНИЯ: СОЮЗ *ХОТЯ* И ОБМАНУТОЕ ОЖИДАНИЕ*

Е. В. Урысон (uryson@gmail.com)

Институт русскго языка им. В. В. Виноградова РАН, Москва, Россия

В литературе предлагались разные описания семантики уступительного союза *хотя*, адекватные для примеров типа *Хотя шел дождь, мы гуляли*. В докладе анализируется новый материал, ср. *Хотя я иногда беру такси, обычно я езжу на метро*. Демонстрируется, что основа семантики союза *хотя* — это «обманутое ожидание».

Ключевые слова: союз, обманутое ожидание, уступительный союз, семантика союза.

CONCESSIVE CONJUNCTION *KHOTIA* 'THOUGH' AND "CANCELLED EXPECTATION"

E. V. Uryson (uryson@gmail.com)

V. V. Vinogradov's Russian Language Institute, Russian Academy of Sciences, Moscow, Russian Federation

The paper is focused on the semantics of the concessive conjunction khotja 'though'; cf. (1) Khotja pogoda byla ochen' plokhaja (P), oni kazhdyj den' kupalis' (Q) 'Though the weather was very bad (P), they bathed every day (Q)'. Different defi nitions of the meaning 'khotja' are found in the literature. In typology it is generally agreed that its main components are implication and negation: Though P, Q = 'usually if P, then not-Q; in this case P and Q'. In traditional Russian grammar it is commonly supposed that the basic semantic component of khotja 'though' is "cancelled expectation": situation P in the subordinate clause induces the expectation not-Q, and this expectation fails in the main clause. Both defi nitions are adequate for examples like (1). Some new material enables to choose

* Работа выполнена при финансовой поддержке РГНФ (грант 10-04-00273а), гранта НШ-4019.2010.6 и Программы ОИФН РАН «Генезис и взаимодействие социальных, культурных и языковых общностей».

between them. I analyze sentences like (2) Khotja bol'shinstvo potukhshikh vulkanov — eto gory konusoobraznoj formy (P), ne vsjakaja takaja gora — byvshij vulkan (Q) 'Though most of the extinct volcanoes are conical mountains (P), not every conical mountain is an extinct volcano (Q)'; (3) Khotja Fjodor zarabatyvaet bol'she Ivana (P), on tozhe ne mozhet soderzhat' semju (Q) 'Though Fjodor earns more than Ivan (P), he cannot provide for his family either (Q)'. I show that the main semantic component of 'khotja' is "cancelled expectation". In cases like (1) "cancelled expectation" is due to our knowledge of usual casual relations between situations (fi xed in frames or scenarios). In cases like (2)–(3) "cancelled expectation" is not due to any frames or scenarios but, rather, to some properties of human consciousness. For instance, (2) presupposes a stable association between volcanoes and conical mountains. This association is the basis of our "expectation" and the consessive conjunction khotja marks that this expectation fails. (3) presupposes a comparison of two people, and in such a context situation P induces our "expectation", which fails in the main clause (cf. Grice's informativity postulate). Thus, in these cases "cancelled expectation" is due to our standard line of reasoning. The concessive conjunction khotja is a means for marking that reasoning of this type is wrong. Some linguistic problems of description of the concessive conjunction khotja are discussed.

Key words: conjunction, concessive conjunction, conjunction semantics, cancelled expectation.

1. Объект исследования и постановка задачи

Объект предлагаемого исследования — уступительный союз *хотя*. В своем центральном значении он представлен в примерах типа:

(1) *Хотя на улице было сыро (P), Ваню повели гулять (Q).*

(2) *Хотя в квартире все спали (P), мальчики говорили в полный голос (Q).*

(3) *Он не снимал пиджака (Q), хотя было жарко (P).*

Предварительно семантику уступительного союза *хотя* можно описать так:

(4) *Хотя P, Q* [*Хотя на улице было много детей (P), Ваню повели гулять (Q)*] ≈
'имеет место ситуация P;
обычно ситуация типа P препятствует тому, чтобы имела место ситуация типа Q;
в данном случае ситуация Q имеет место'.

Так, в (1): сырая погода — это обычно препятствие для детской прогулки. В примере (2): ситуация 'все в квартире спят' обычно препятствует тому, чтобы имела место ситуация 'люди говорят в полный голос'. В примере (3): ситуация

'жарко' обычно препятствует существованию ситуации 'мужчина не снимает пиджака'.

Этот «каркас» значения союза *хотя* и других слов с аналогичным уступительным значением в той или иной форме выделен в целом ряде работ по русистике [Богомолова 1955; Гречишникова 1971; Печенкина 1976; Перфильева 1985; Теремова 1986; Апресян В. 1999].

Оговорим, что союз *хотя* употребляется и в других, достаточно разных контекстах.

Так, словари выделяют еще одно значение союза *хотя*, представленное в примерах типа

(5) *Она умная, хотя очень злая*.

(6) *Пьера поразила скромность маленького, хотя и чистенького домика* (Л. Толстой).

(7) *Я тоже артист, хотя плохой* (И. С. Тургенев).

В подобных контекстах союз *хотя* квалифицируется как сочинительный, близкий союзу *но*.

В следующих примерах представлена еще одна лексема слова *хотя*. Ср.

(8) *Дед жил тогда с нами на даче (Q). Хотя ты был тогда маленьким и, конечно, этого не помнишь (P)*.

(9) *Петровы переехали и своего нового адреса не оставили (Q). Хотя спросите у соседей напротив, может быть, они больше знают (P)*.

(10) *Стены домика были зелеными, хотя это их позже перекрасили, сначала дом был голубой*.

Данная лексема *хотя* обладает ярко выраженной синтаксической спецификой. Предложение P, вводимое ею, может располагаться только после предложения Q, причем, как правило, отделяется от него достаточно большой паузой (на письме предложения P и Q часто разделяются точкой). Перед нами, очевидно, фразовая частица. Существенно, что примеры (8)–(10) сильно отличаются от (1)–(3) семантически: в данном случае ни ситуация P не может интерпретироваться как препятствие для ситуации Q, ни, наоборот, ситуация Q не может быть препятствием для P. Наконец, в некоторых контекстах союз *хотя* выступает как неточный конверсив лексемы *хотя*, представленной в (1)–(3). Ср.

(11) *Петя теперь очень много занимается, хотя учителя им по-прежнему недовольны* [ср. *Хотя Петя теперь очень много занимается, учителя им по-прежнему недовольны*].

(12) *Вовсю палило солнце, хотя было не так душно, как в Москве* (Ф. Незнанский) [ср. *Хотя вовсю палило солнце, было не так душно, как в Москве*].

Союз *хотя* в контекстах типа (5)–(12) нами здесь не рассматривается.

Что касается центральных примеров типа (1)–(3), то они, казалось бы, не представляют особых трудностей для семантического анализа: подобные высказывания вида *Хотя P, Q* обладают вполне четкой и прозрачной семантической структурой. Тем не менее, союз *хотя* в подобных контекстах описывался в литературе достаточно разными способами. Задача данной работы — проанализировать эти описания с точки зрения проблемы неединственности лингвистического описания. Мы ограничимся работами по русистике — они содержат все те идеи, которые высказывались в европейской и американской науке.

2. Основные подходы к описанию семантики союза ХОТЯ

Приведенное выше толкование (4) уступительного союза *хотя* в своих существенных чертах восходит к работе [Богомолова 1955]. Однако в академической грамматике того времени [Грамматика русского языка 1954] уступительные предложения описываются по-другому — как выражающие «обманутое ожидание». Действительно, в высказывании *Хотя P, Q* ситуация ‘P’ индуцирует у нас ожидание ‘не-Q’. Однако описываемая реальность оказывается противоположной. Как пишет академическая грамматика, в предложении с придаточным уступительным «главное предложение содержит сообщение о факте, противоположном тому, чего можно было бы ожидать на основании того, о чем говорится в придаточном предложении» [Грамматика русского языка 1954: 337]. Аналогичным образом интерпретируется семантика союза *хотя* во многих более поздних работах. В частности, в работах [Крейдлин, Падучева 1974а; Падучева 2004: 47] союз *хотя* толкуется так:

(13) *Хотя P, Q* = ‘P; поэтому ожидалось, что не-Q; Q’.

Очевидно, что ожидание ‘не-Q’ создается не наличием ситуации P как таковой, а нашим знанием действительности. Так, применительно к (1), мы знаем, что в сырую погоду маленьких детей не выводят на улицу. Тем самым, высказывание с союзом *хотя* не просто описывает некоторое положение дел, но и апеллирует к общеизвестным закономерностям. При этом союз *хотя* указывает на то, что в данном конкретном случае закономерность нарушается. Эта особенность союза *хотя* обсуждается, в частности, в монографии [Ляпон 1986]: «высказывание *Мальчик с пальчик, хотя был мал, но был очень ловок и хитер* строится на априорной истине ‘если мал, значит не ловок, не хитер’ <…>, которая опровергается актуальной истиной ‘мал и в то же время ловок и хитер’, соответствующей реальному положению дел» [Ляпон 1986: 137]. Очевидно, что

под «актуальной истиной» понимается описываемое положение дел. «Априорная истина» — это фрагмент общего знания людей, он и индуцирует ожидание. Априорная истина естественно формулируется с помощью союза *если*, ср. ‘если мал, значит не ловок, не хитер’.

Отметим, что отсылка к общеизвестным закономерностям, содержащаяся в значении союза *хотя*, принципиально отличает этот союз от других подчинительных союзов, например, *когда, потому что, если*. Высказывания с этими союзами просто описывают некоторое положение дел.

В дальнейшем во многих работах, особенно типологических, уступительная семантика сводится к указанию на «априорную истину» и ее нарушение в описываемом случае. Априорная истина формулируется в виде импликации ‘обычно если P, то не-Q’. Значение союза *хотя* в его центральном значении представляется так (ср., например, [Koenig 1988; Храковский 2004]):

(14) а) *Хотя P, Q* [*Хотя на улице было сыро (P), Ваню повели гулять (Q)*] ≈
‘P;
обычно если P, то не-Q;
Q’.

Или, в более полном виде:

б) *Хотя P, Q* [*Хотя на улице было сыро (P), Ваню повели гулять (Q)*] ≈
‘имеет место ситуация P;
обычно если имеет место ситуация типа P, то не имеет место ситуация типа Q;
в данном случае ситуация Q имеет место’.

Идея препятствия, четко выраженная в описаниях [Богомолова 1955; Теремова 1986; Апресян В. 1999] и других работах, в этом толковании не выражена. Но эта идея и не отвергается — она как будто выражается компонентом ‘обычно если P, то не-Q’.

Как показано в работе [Урысон 2002], указание на препятствие, или иначе указание на каузальную зависимость между P и не-Q, необходимо в толковании союза *хотя*. Уточненное толкование союза *хотя* имеет следующий вид:

(15) *Хотя P, Q* [*Хотя было сыро (P), Ваню повели гулять (Q)*] =
‘(i) имеет место ситуация P;
(ii) обычно ситуация типа P влияет на положение дел; в результате если имеет место ситуация типа P, то не имеет место ситуация типа Q;
(iii) в данном случае имеет место ситуация Q’.

В предлагаемой работе мы не будем подробно останавливаться на обосновании необходимости этого уточнения. Нас будет интересовать идея обманутого ожидания в семантике союза *хотя*.

Толкование типа (14)–(15) привлекает своей простотой: уступительная семантика в целом сводится к импликации и отрицанию. Действительно, в общем, уступительное предложение вида *Хотя P, Q* представляется через условное предложение вида ‘Если P, не-Q’.

Заметим, что идея обманутого ожидания не формулируется в этом толковании прямо (в нем нет никакого компонента типа ‘ожидать’), но естественно вытекает из импликации, выражающей «априорную истину», — она и индуцирует определенное ожидание. Кажется очевидным, что в описании союза *хотя* не требуется никакого специального указания на ожидание: оно просто избыточно. Однако поставленный вопрос оказался далеко не таким простым.

Дело в том, что высказывание вида *Хотя P, Q* может описывать и принципиально иные случаи, когда ситуации P и Q связаны друг с другом не столько в «объективном» мире, сколько в сознании говорящего. Рассмотрим некоторые из них.

ЗАМЕЧАНИЕ. Ожидание чего-то — это, прежде всего, некоторое мнение о будущем, о том, что может произойти, и готовность к этому возможному событию [Апресян 2004]. Когда речь идет о семантической структуре высказывания, то «ожидание» естественно интерпретировать как готовность адресата к получению вполне определенной информации. Вводя в семантическую структуру высказывания компонент ‘ожидать’, необходимо представлять порядок подачи (и восприятия) информации, т. е. то, как высказывание разворачивается во времени.

Мы хорошо представляем, как компонент ‘ожидание’ встраивается в семантическую структуру высказывания с препозитивным придаточным *Хотя P, Q*: первая пропозиция — P — индуцирует ожидание ‘не-Q’, а союз *хотя* указывает на то, что оно «обманется». Сложнее обстоит дело в случае постпозиции придаточного: *Q, хотя P*. Ср. *Мы купались, хотя шел дождь*. Здесь «ожидание» сложным образом переплетается с реальным порядком подачи информации. По-видимому, «реальное ожидание», связанное с порядком развертывания высказывания во времени, учитывается отдельными, динамическими, правилами. Современная лингвистика ограничивается описанием «статической» структуры высказывания, без учета динамики его развертывания (одно из исключений — работа [Пекелис 2008]). В предлагаемой работе тоже не рассматриваются вопросы, связанные с линейностью высказывания.

3. Лексема «ХОТЯ логической ловушки»

В научном тексте союз *хотя* часто выступает в некоем специфическом значении. Ср.

(16) *Хотя большинство потухших вулканов — это горы конусообразной формы (P), не всякая такая гора — бывший вулкан (Q).*

(17) *Хотя всякий равносторонний треугольник является равнобедренным (P), не всякий равнобедреннный треугольник будет еще и равносторонним (Q).*

Буквальное применение толкования (15) к подобным примерам приводит к абсурду. Действительно, основной компонент этого толкования таков: 'обычно если P, то не-Q'. Этот компонент обычно отражает какую-то конкретную общеизвестную истину, которая в данном случае опровергается. Возьмем теперь пример (16) и попытаемся выявить его пресуппозицию. Если бы высказывание (16) было устроено как предыдущие, то его пресуппозиция имела бы следующий вид:

(18) 'Обычно если потухшие вулканы — это горы конусообразной формы, то всякая такая гора — бывший вулкан'.

Но очевидно, что (18) не только не является фрагментом нашего знания о мире, но и прямо противоречит законам логики. В чем же отличие высказываний типа (16)–(17) от примеров, обсуждавшихся выше?

Высказывание (16) не апеллирует ни к какому знанию о каузальных связях между ситуациями P и Q. Оно подразумевает, что в данном случае не применим естественный, «накатанный» ход нашей мысли, что нужно отвергнуть напрашивающийся вывод. Действительно, представим, что мы изучаем объекты типа O и что все известные нам объекты этого типа обладают свойством M. Пока мы не встретим никакого исключения, т.е. никакого объекта O, не обладающего этим свойством, мы будем готовы ожидать, что все объекты O обладают свойством M, — естественно напрашивается именно это обобщение. Такова естественная гипотеза об изучаемых объектах, наше естественное, вполне осознанное предположение об их устройстве. Союз *хотя* в (16) указывает на неправильность этого предположения. Однако в данном случае ожидание индуцировано не какой-либо «априорной истиной» об устройстве мира, а логикой познания действительности, стандартным ходом нашей мысли.

Ясно, что высказывание (17) устроено точно так же. Союз *хотя* предупреждает здесь о ловушке, которую готовит нам наезженный ход мысли: мы склонны ожидать, что если объекты Rs являются Rb, то и объекты Rb являются Rs. Иными словами, мы склонны считать, что множества Rs и Rb тождественны, упуская из виду, что одно множество может быть подмножеством другого. Наше ожидание неправильно и потому обманывается.

Очевидно, что в случаях (16)–(17) перед нами особая лексема союза *хотя*. В отличие от лексемы «*хотя* препятствия», она не указывает на нарушение обычного порядка вещей, а предупреждает о возможности логической ошибки. Назовем эту лексему «*хотя* логической ловушки».

Приведем еще некоторые примеры.

(18) *Хотя союз между частями сложного предложения может отсутствовать (P), стандартно компоненты предложения связаны друг с другом с помощью формального показателя такого типа (Q).*

Ситуация P, т.е. отсутствие формального показателя связи между предложениями, влияет не на «объективное» положение дел, а лишь на наше ожидание относительно устройства сложного предложения: если формального

показателя связи может не быть, то естественно ожидать, что он и неважен для данного класса случаев. Это предположение оказывается неправильным.

(19) *[Горные выработки позволили определить строение алмазных трубок Кимберли.] Хотя ближе к поверхности эти трубки имеют цилиндрическую форму (P), на большой глубине они становятся похожими на трещины (Q).*

Форма алмазных трубок у земной поверхности (ситуация P) влияет на наше ожидание: мы готовы считать, что данная форма присуща этим трубкам вообще, независимо от глубины их расположения. Это ожидание обманывается.

(20) *Через несколько мгновений приступ кончился, хотя обычно он продолжался часами* (В. Карцев) [обычная продолжительность приступа влияет на наши ожидания относительно его длительности и в данном случае].

Значение союза «*хотя* логической ловушки» можно представить так:

(21) *Хотя P, Q* [*Хотя большинство потухших вулканов — горы конусообразной формы (P), не всякая такая гора — бывший вулкан (Q)*] =
‘(i) неоднократно имеет место ситуация типа P;
(ii) это влияет на формирование ожидания относительно устройства мира;
(iii) в результате естественно было бы ожидать: ситуация типа P всегда имеет место;
(iv) ситуация типа P не всегда имеет место; в некоторых случаях имеет место ситуация типа Q’.

Несколько иное, хотя и похожее значение имеет союз *хотя* в следующем примере:

(22) *Хотя Федор зарабатывает больше Ивана (P), он тоже не может содержать семью (Q).*

Это высказывание уместно, например, когда два человека разговаривают о Федоре и Иване. Один из собеседников говорит, что Иван зарабатывает мало и не может содержать семью, а затем добавляет, что Федор зарабатывает больше. Второй собеседник из последней реплики может сделать вывод, что Федор может содержать семью. Эта импликация объясняется постулатом информативности Грайса. Ясно, что ситуация ‘Федор зарабатывает больше Ивана’ в данном случае лишь индуцирует определенное ожидание относительно жизни Федора, причем это ожидание наводится еще и общим контекстом — сравнением Федора с Иваном. Это ожидание неправильно — на это указывает союз *хотя*. Данный пример отличается от предыдущих тем, что неправильный вывод навязывается здесь постулатом информативности Грайса.

4. Лексема «ХОТЯ испорченного телефона»

В близком значении союз *хотя* широко употребляется в самой обычной речи. Однако данная лексема *хотя* предупреждает не об ошибочности первоначальной гипотезы, не о логической ошибке в прямом смысле слова и даже не о нарушении постулата информативности, а о возможности «испорченного телефона». Ситуации «испорченного телефона», т. е. ненамеренного искажения информации, подчиняются своим закономерностям. Эти закономерности подробно описаны в работе [Гловинская 1998]. Проиллюстрируем действие некоторых из них на нашем материале.

(23) *Хотя Петя иногда остается в гостях (P), обычно он приходит ночевать домой (Q).*

Допустим, что человек узнал от кого-то, что Петя иногда остается на ночь в гостях. При неоднократной передаче этой информации «из уст в уста» она может приобрести такой вид: Петя не приходит ночевать домой, Петя вообще не ночует дома (конкретное высказывание «превращается» в общее). Говорящий предупреждает адресата о возможности такого искажения информации с помощью союза *хотя*. Коротко говоря, в данном случае ситуация P влияет на наше ожидание относительно того, где Петя проводит ночь: узнав, что Петя иногда остается в гостях, мы готовы считать, что он вообще не ночует дома. Союз *хотя* — это сигнал того, что такое мнение неправильно.

(24) *Хотя Вася вчера был пьяным (P), он вообще-то не выпивает больше рюмки, и то лишь на дне рождения тестя (Q).*

Представим, что мы увидели мало знакомого нам человека пьяным, или услышали, что он вчера был пьяным. На этом основании мы можем предположить, что этот человек — алкоголик. В дальнейшем мы можем забыть, что это всего лишь наше предположение. И потом, особенно если эта информация неоднократно переходит от собеседника к собеседнику, она передается уже не как предположение, а как факт. Как и в предыдущем примере, союз *хотя* предупреждает, что это заключение ошибочно: хотя Вася вчера был пьяным, он не пьяница.

Еще один пример (предложен анонимным рецензентом):

(24) а) *Хотя я иногда беру такси (P), обычно я езжу на трамвае (Q).*

Данное высказывание отличается от предыдущих только тем, что оно — от 1-го лица. Говорящий заранее опровергает возможное мнение адресата, что он (говорящий) часто или обычно ездит на такси: то, что он иногда берет такси, не значит, что это — его обычный способ передвижения. Таким образом, союз *хотя* здесь также предупреждает о возможном неверном умозаключении адресата.

Применительно к таким разговорным контекстам будем говорить о лексеме «*хотя* испорченного телефона». Она свойственна диалогической речи.

Эта лексема безусловно не толкуема через более простые понятия. Дело в том, что она указывает на специфические ментальные операции, приводящие к искажению информации, которые, скорее всего, не осознаются говорящим, а потому не имеют нетерминологического обозначения в естественном языке. Отразить эту специфику данной лексемы *хотя* мы можем только с помощью символа, условного знака, но не обычного слова.

Декомпозиция лексемы «*хотя* испорченного телефона» могла бы выглядеть так:

(25) *Хотя P, Q* [*Хотя Петя иногда остается в гостях (P), обычно он приходит ночевать домой (Q)*] =
‘(i) имеет место ситуация типа P;
(ii) «действие механизма переработки и хранения информации»;
(iii) это влияет на формирование ожидания;
(iv) в результате естественно было бы ожидать: ситуация типа P всегда имеет место;
(v) ситуация типа P не всегда имеет место; в некоторых случаях имеет место ситуация типа Q’.

Обратим внимание на компонент (ii) этого выражения — он взят в обычные, двойные кавычки. Этот компонент представляет собой апелляцию к деятельности сознания, и, разумеется, не проще семантики союза *хотя*. Однако для обозначения этого весьма сложного смысла язык не располагает обычным словом. С похожими ситуациями мы уже сталкивались при описании целого ряда союзов — многие из них указывают на специфические ментальные операции, которые, однако, не обозначаются никакими знаменательными словами (по крайней мере, не терминами). Компонент (ii) представляет собой «долексемный» компонент, близкий кваркам Ю. Д. Апресяна. Более подробно эта проблематика обсуждается в работе [Урысон 2011].

5. Вместо заключения

Многообразие употреблений союза *хотя*, указывающего не на нарушение обычного хода вещей, а лишь на обманутое ожидание, не ограничивается рассмотренными случаями. В некоторых высказываниях союз *хотя* имеет еще более ослабленное значение. Ср.

(26) *Хотя с утра шел сильный дождь (P), к вечеру небо расчистилось (Q).*

(27) *Хотя с утра больному стало лучше (Q), к вечеру он умер (P).*

(Эти примеры являются модификациями примеров с союзом *но* из книги [Санников 1989]: *С утра шел сильный дождь, но к вечеру небо расчистилось; С утра больному стало лучше, но к вечеру он умер.*)

Разумеется, ситуация типа 'с утра шел сильный дождь' как-то влияет на положение дел. Однако результатом этого влияния не может быть продолжительность самого дождя или длительность плохой погоды. В данном случае ситуация типа 'с утра шел сильный дождь' влияет лишь на формирование нашего ожидания: мы готовы думать, что погода изменится еще очень нескоро. Аналогичным образом обстоит дело и в примере (27): состояние больного утром вряд ли влияет на ход его болезни, но оно внушает нам надежду, т. е. создает ожидание. Но в этом случае ожидание основано не на знании объективных закономерностей и даже не на «законах испорченного телефона», а на каких-то особенностях нашего сознания, нашего восприятия действительности, пронизанного эмоциями и оценками. В подобных высказываниях значение союза *хотя* весьма выхолощено и, по существу, указывает только на обманутое ожидание.

Не исключено, что «*хотя* логической ловушки» и «*хотя* испорченного телефона», а также употребление союза *хотя* в (26)–(27), как и в (22), можно трактовать как единую, весьма абстрактную лексему, основная функция которой — указание на «обманутое ожидание», индуцированное нашими особенностями восприятия информации. Толкование этой лексемы могло бы выглядеть так:

(28) *Хотя P, Q* =
'(i) имеет место ситуация типа P;
(ii) это влияет на формирование нашего ожидания;
(iii) в результате можно было бы ожидать: не имеет место ситуация типа Q;
(iv) имеет место ситуация Q'.

Однако легко видеть, что толкование (28) вполне обслуживает и контексты с лексемой «*хотя* препятствия». Ср. *Хотя на улице было сыро (P), Ваню повели гулять (Q)*: 'на улице было сыро; поэтому было бы естественно ожидать, что Ваню не поведут гулять; Ваню повели гулять'. Ясно, что идею обманутого ожидания выражают все рассмотренные лексемы союза *хотя*. Выражение (28) можно рассматривать как инвариант рассмотренных лексем союза *хотя*.

Различие между ними — в том, на чем основано это ожидание. В случае «*хотя* препятствия» основанием ожидания является наше знание о мире, о причинно-следственных связях между разными ситуациями. В других случаях ожидание основано не на знании, а скорее, на особенностях нашего сознания, на нашем неумении мыслить до конца логически, на том, что человеку свойственно ненамеренно, несознательно искажать информацию.

Итак, «обманутое ожидание» — это семантический мост между разными значениями союза *хотя*. На наш взгляд, естественно ввести данный компонент в толкование и центральной лексемы *хотя*. С учетом сказанного, ее толкование переписывается так:

(29) *Хотя P, Q* [*Хотя было сыро (P), Ваню повели гулять (Q)*] =
'(i) имеет место ситуация P;
(ii) обычно ситуация типа P влияет на положение дел; в результате если имеет место ситуация типа P, то не имеет место ситуация типа Q;

(iii) поэтому в данном случае ожидалось, что ситуация Q не имеет места;
(iv) ситуация Q имеет место'.

Одно замечание о предложенном способе описания разных контекстов с *хотя*. В принципе можно усматривать во всех этих контекстах единую лексему, а различие между ее употреблениями относить на счет экстралингвистического знания: ожидание, на которое указывает союз *хотя*, возникает на разных основаниях, по разным причинам. Не имея хорошего инструмента для представления такого знания, мы выбрали более традиционный подход описания семантики *хотя*.

Разные точки зрения на союз *хотя* — это следствие сложности описываемого объекта. Между тем, каждый подход открывает свою «долю истины». Привлечение нового материала, до сих пор как будто выпадавшего из поля зрения исследователей, показало, что в данном случае требуется не столько выбор одного описания из нескольких, сколько интегрирование этих описаний.

В заключение отметим, что по мысли Е. В. Падучевой «компонент «ожидание» входит в семантику большого числа самых разных слов», причем «претендует на роль строевого» [Падучева 2004: 47]. Анализ союза *хотя* подтверждает эту мысль. В терминологии Ю. Д. Апресяна [Апресян 2006], смысл 'ожидание' следует отнести к системообразующим.

References

1. *Apresian Iu. D.* 2004. Dictionary Paragraph 'ZHDAT' [Slovarnaia Stat'ia 'ZHDAT']. Novyi Ob"iasnitel'nyi Slovar' Sinonimov Russkogo Iazyka.
2. *Apresian Iu. D.* 2006. The Fundamentals of Systematic Lexicography [Osnovaniia Sistemnoi Leksicografii]. Iazykovaia Kartina Mira I Sistemnaia Leksikografiia.
3. *Apresian V.* 1999. The Meaning of Concession in the Language [Ustupitel'nost' v Iazyke I Slova so Znacheniem Ustupki]. Voprosy Iazykoznaniia, 5.
4. *Bogomolova A. V.* 1955. Concessive Constructions with the Conjunction 'KHOTIA (KHOT')' in Modern Literary Russian Language [Ustupitel'nye Konstruktsii s Soiuzom 'KHOTIA (KHOT')' v Sovremennom Russkom Literaturnom Iazyke].
5. *Glovinskaia M. Ia.* 1998. "Chinese Whispers" as the Main Mean of Another's Speech Transmission ["Isporchennyi Telefon" kak Osnovnoi Sposob Peredachi Chuzhoi Rechi]. Russkii Iazyk v ego Funktsionirovanii. Tezisy Dokladov Mezhdunarodnoi Konferentsii "3 Shmelevskie Chteniia" (Russian Language in its Funcioning. Proc. of International Conference "3 Shmelev Recital").
6. *Grechishnikova R. M.* 1971. Complex Sentence with Phraseologicalized Concession Expression Means in Modern Literary Russian Language [Slozhnoe Predlozhenie s Frazeologiziruiushchimisia Sredstvami Vyrazheniia Ustupitel'nykh Otnoshenii v Sovremennom Literaturnom Russkom Iazyke].
7. *Khrakovskii V. S.* 2004. Concessive Constructions: Semantics, Syntaxis, Typology [Ustupitel'nye Konstruktsii: Semantika, Sintaksis, Tipologiia]. Tipologiia Ustupitel'nykh Konstruktsii.

8. *Kreidlin G. E., Paducheva E. V.* 1974. The Meaning and Syntactic Characteristics of a Conjunction [Znachenie I Sintaksicheskie Svoistva Soiuza]. NTI, 2 (9).
9. *Koenig E.* 1988. Concessive Connectives and Concessive Sentences: Cross-linguistic Regularities and Pragmatic Principles. Explaining language Universals.
10. *Liapon M. V.* 1986. Semantic Structure of a Complex Sentence and the Text [Smyslovaia Struktura Slozhnogo Predlozheniia I Tekst].
11. *Paducheva E. V.* 2004. Dynamic Models in Lexical Semantics [Dinamicheskie Modeli v Semantike Leksiki].
12. *Pekelis O. E.* 2008. Coordination and Subordination: Communicative Approach [Sochinenie I Pochinenie: Kommunikativnyi Podkhod]. Russkii Iazyk v Nauchnom Osveshchenii, 2 (16).
13. *Pechenkina T. G.* 1976. Syntactic Category of Concession and Forms of its Expressions in Literary Russian Language of the 2nd half of the XIX c. [Sintaksicheskaia Kategoriia Ustupitel'nosti I Formy ee Vyrazheniia v Russkom LIteraturnom Iazyke Vtoroi Poloviny XIX veka].
14. *Perfil'eva N. P.* 1985. Syntactic Status of Concessive-Adversative Constructions [Sintaksicheskii Status Ustupitel'no-Protivitel'nykh Konstruktsii]. Strukturno-Funktsional'nyi Analiz Iazykovykh Edinits.
15. *Russian* Grammar [Grammatika Russkogo Iazyka], 2, 2. 1954.
16. *Teremova R. M.* 1986. Semantics of Concession and its Expression in Modern Russian Language [Semantika Ustupitel'nosti I ee Vyrazhenie v Sovremennom Russkom Iazyke].
17. *Uryson E. V.* 2011. Experiment of Conjunction Semantics Description: Language Data on Conscience Activity [Opyt Opisaniia Semantiki Soiuzov: Dannye Iazyka o Deiatel'nosti Soznaniia].

ЭНАНТИОСЕМИЯ В РУССКОЙ ФРАЗЕОЛОГИИ

М. М. Вознесенская (voznes-masha@yandex.ru)

Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Работа посвящена явлению энантиосемии в русской фразеологии. Рассматриваются источники и типы энантиосемии, связанные с различным переосмыслением внутренней формы (антифразис, импликации) или особенностями прагматической интерпретации ситуации (эмоционально-оценочное противопоставление).

Ключевые слова: энантиосемия, фразеология, внутренняя форма, типы энантиосемии.

ENANTIOSEMY IN RUSSIAN PHRASEOLOGY

M. M. Voznesenskaia (voznes-masha@yandex.ru)

V. V. Vinogradov Russian Language, Institute, Russian Academy of Sciences, Moscow, Russian Federation

The phenomenon of enantiosemy in Russian phraseology, its sources and types are considered. Enantiosemy is shown to be connected with different interpretation of an inner form (antiphrasis, implications of different kinds etc.). In some cases enantiosemy arises as a result of pragmatic interpretation of a corresponding situation (emotive connotations).

Key words: enantiosemy, phraseology, inner form, enantiosemy typology.

0. Энантиосемия определяется как появление у одной языковой единицы «противоположного значения (иногда вытесняющего первоначальное, иногда сосуществующего с ним)» [Шмелев 1977: 208]. Несмотря на то, что явление энантиосемии в русском языке считается малоизученным, оно привлекает к себе внимание ученых уже более ста лет, с момента опубликования в 1883–1884 гг. работы В. Шерцля «О словах с противоположными значениями (или о так называемой энантиосемии)» [Новиков 2001: 226–238]. Исследователями выделяется номинативная и эмоционально-оценочная энантиосемия, полная (ядерная)

и частичная (периферийная), эксплицитная (реализованная, конвенционализованная, ингерентная, языковая) и имплицитная (потенциальная, окказиональная, адгерентная, речевая) энантиосемия, диахроническая и синхронная, внутри и межъязыковая энантиосемия [Горелов 1986; Ермакова 2002; Кравцова 2006; Махмутова 2009; Новиков 2001; Соколов 1980; Цоллер 1998, 2000; Шмелев 2009; Шмелев 1977]. Представления о природе этого явления противоречивы — «энантиосемичны». Обычно появление энантиосемии объясняется как результат развития разных, доходящих до противоположности, значений древних синкретичных корней [Новиков 2001: 229–232]. Вместе с этим энантиосемия наблюдается и в некоторых новых словоупотреблениях, характерных, например, для современного жаргона [см. Ермакова 2002: 66–67]. Энантиосемия рассматривается и как крайне «неудобное», затрудняющее нормальную коммуникацию явление, и как процесс, непосредственно отражающий философскую суть любого предмета или явления — единство противоположностей. Не случайно, что Гегель не обошел вниманием энантиосемию, признавая за ней способность «доставлять радость мышлению» (цит. по [Горелов 1986]). В современном понимании энантиосемия признается системным, регулярным явлением, имеющим определенные когнитивные и коммуникативные источники [см. Ермакова 2002; Шмелев 2009].

1. Предмет рассмотрения в данной работе — энантиосемия в русской фразеологии. Обычно фразеологическая энантиосемия не является объектом отдельного исследования, ее рассматривают или вместе с лексической и словообразовательной энантиосемией, или фразеологический материал привлекается для иллюстраций некоторых общих положений наряду с лексическими и словообразовательными примерами. Однако представляется, что фразеологические единицы оказываются более «приспособленными» к развитию энантиосемии в силу специфической природы своей знаковой функции: являясь результатом вторичной номинации, основанной на различных переосмыслениях внутренней формы, они характеризуются такими особенностями значения, как ситуативность (полипризнаковость), диффузность и экспрессивная окрашенность. Естественно, что в небольшой статье невозможно исчерпывающе описать столь сложное и противоречивое явление. Нашей задачей является показать, что энантиосемия в русской фразеологии представлена шире, чем это принято думать, а также привлечь внимание к некоторым фактам, связанным с фразеологической энантиосемией, которые еще не попадали в круг рассмотрения, и продемонстрировать основные источники и, соответственно, типы энантиосемии, наиболее характерные для фразеологии.[1]

1.1. Одним из источников энантиосемии, широко представленным во фразеологии, является антифразис, под котором понимают стилистическую

[1] Фразеологические единицы, рассматриваемые в статье, брались из словарей, список которых приведен в конце статьи. Для экономии места в каждом конкретном случае источник не указывается. Также некоторые толкования даются в сокращенном виде.

фигуру отрицания, когда слово (словосочетание, предложение) употребляется в противоположном смысле. Возможны две разновидности антифразиса — ирония (завышение оценки с целью понижения) и мейозис (занижение оценки с целью повышения). Ироническое отрицание приводит к энантиосемии таких ФЕ как *хорошее дело*, *милое дело*[2] — ‘выражение одобрения’ и ‘выражение неодобрения, возмущения’, где первое, «положительное» значение (ср. 1а) непосредственно мотивировано внутренней формой выражения, а второе представляет собой его ироническое переосмысление (2а):

(1) *а.* — А куда? — К Черному морю. — *Хорошее дело*. — В санаторий. — *Хорошее дело.* (В. Шукшин. Печки-лавочки).

б. Увидев оперативников, он остановился как вкопанный и, прищурив глаза, произнес не предвещающим ничего хорошего тоном: — Ага, голубчики! Вот вы где! Хорошее дело — никто не знает, где у нас находятся Гуров и Крячко! Просто неуловимые ковбои какие-то! (Н. Леонов, А. Макеев. Гроссмейстер сыска).

То же ироническое переосмысление лежит в основе появления энантиосемичного значения у фразеологизма *увидеть небо в алмазах*. Первое значение — ‘испытать что-л. приятное’ (2а) и второе, основанное на ироническом отрицании — ‘испытать что-л. неприятное; иметь серьезные проблемы, неприятности’ (2б):[3]

(2) *а.* Кое-что тебе отказало, но уж «Зиппо»-то не откажет никогда. Щелкни зажигалкой, выкури сигарету, успокойся — и все у тебя пойдет как надо. Ты еще *увидишь небо в алмазах*. (Итоги).

б. — Ну, Платон Михайлович, теперь ты увидишь небо в алмазах, узнаешь, почем фунт лиха, хлебнешь горюшка! Я тебе напомню, сука, восьмидесятый год… (Ю. Дубов. Большая пайка).

Отрицательное — ироническое — переосмысление широко распространено во фразеологии при формировании фразеологического значения. Именно этот процесс приводит к фразеологизации свободных сочетаний типа *устроить веселую жизнь*, *от большого ума*, *сказать пару ласковых [слов]*, *с тем/таким же успехом*, *Очень нужно!* и др. Ср. примеры употреблений свободных (3а), (4а) и фразеологических (3б), (4б) сочетаний:

(3) *а.* <…> почему-то так получилось, что любовь всегда пассивна, а ненависть зато всегда активна и потому очень привлекательна, и говорят

---

[2] Интересно, что варианты этих идиом с диминутивными компонентами *хорошенькое дело*; *миленькое дело* употребляются только в «отрицательных» значениях.

[3] Подробнее о различиях между значениями этой идиомы см. в [Добровольский 2011].

еще, что ненависть — от природы, а любовь — от ума, *от большого ума* <…> (А. и Б. Стругацкие. Хищные вещи века).

б. Муж вел машину и более ничего не спрашивал, умея молчать. Что умел — то умел. Тут еще и почерк: он считал (от большого ума, конечно), что сдержанностью и таким вот нудным молчаньем он много выигрывает, день ко дню набирая психологические очки в их ссоре, болван. (В. Маканин. Предтеча).

(4) *а.* Поездку в Россию Берлиоз вспоминал как счастливый сон. Его удалось повторить через двадцать лет *с тем же успехом*: слава, деньги и любовь. (Московский комсомолец).

б. <…> борцы с экономической преступностью решили ударить по беззаконию произволом. Вольному, как говорится, воля, но с тем же успехом можно ударять чумой по холере или гололедом по травматизму. (Общая газета).

О фразеологизированности выражений типа *Этого еще не хватало (недоставало)! А ты думал! Эко важное дело! Эко важная птица! Велика беда! Велика важность! Эка беда (важность)! Будьте покойны!* (при угрозе и т. п.), *Не беспокойтесь! Извольте радоваться! Будьте уверены! Будь здоров!* упоминает Д. Н. Шмелев в [Шмелев 1958: 75]. При этом типе энантиосемии словосочетание (или предложение), употребляемое в противоположном смысле при произнесении обязательно сопровождается особой, экспрессивно-иронической, интонацией. При письменной фиксации для определения энантиосемичного значения необходим достаточно широкий контекст. Отметим, что в лексике энантиосемия, имеющая своим источником ироническое переосмысление (антифразис), относится к речевым явлениям, а во фразеологии — к языковым. Именно ироническое переосмысление исходного свободного сочетания является внутренней формой, мотивирующей фразеологическое значение. Таким образом, можно говорить об энантиосемии между буквальным и фразеологическим значением словосочетания.

1.2. Источником, приводящим к появлению разных, подчас противоположных, значений у фразеологических единиц может быть и различное осмысление ситуации (ее признаков, причин, следствий и т п.), лежащей в основе внутренней формы идиомы. Так, идиома *сводить с ума (кого-л.)* может указывать на разные причины, приводящие к состоянию, метафорически уподобляемому прекращению способности адекватно мыслить в результате психического заболевания, — что-л. очень неприятное, раздражающее или, наоборот, что-л. чрезвычайно приятное, вызывающее сильное чувство удовольствии или любви. Соответственно, идиома может иметь противоположные значения — ‘сильно раздражать кого-л.’ (5а), (5б) и ‘очень нравиться кому-л.’ (5в), а для понимания того, какое именно из значений реализуется в высказывании (впрочем, как всегда при энантиосемии) необходим широкий контекст:

(5) *а.* «Думаю, большинство людей устает сейчас от такого музыкального однообразия. Эта музыка *сводит с ума* и покупателей, не только торговых работников», — полагает Билл Гибсон. (Публицистика Интернета).

б. Ей мешали наушники, положенные в студии выступающим. Она вертела их на голове, нисколько не заботясь о сохранности прически. — Они меня сводят с ума! Я слышу там сорок семь разных голосов! Наконец она надела их задом наперед, отчего стала похожа на нахохленного воробья, и начала отвечать на вопросы. (Аргументы и факты)

в. — Она сложена как ты, госпожа, только тяжелее, мощнее тебя! И это сводит с ума мужчин, особенно таких, как этот мальчишка... (И. Ефремов. Таис Афинская).

В идиоме *распустить слюни* исходное выражение метонимически отсылает к образу младенца, и различные эмоциональные состояния маленького ребенка (плач, беспомощность, радость) служат внутренней формой разных значений фразеологизма: 1) ‘расстраиваться, жаловаться и т. п.’, 2) ‘быть неуверенным, растерянным, бездеятельным’, 3) ‘приходить в умиление, восторг’. Видно, что первое и третье значения энантиосемичны, ср. (6а) и (6б):

(6) *а.* Ох, только б не думать об этом, не я их — так они б меня убили. Жизнь — это борьба. Нечего *слюни распускать*. (Ю. Семенов. Непримиримость).

б. Но другая половина ее души так хотела верить в то, что все это — правда, что Олег действительно нежный, заботливый, чуткий сын, боготворящий свою мать, что он талантливый, целеустремленный, честный, порядочный мальчик! «Не распускай слюни, — постоянно одергивала себя Наталья Евгеньевна, — ему нельзя верить, ты ведь отлично понимаешь, что он такое». (А. Маринина. Не мешайте палачу).

В идиоме *выйти из окопов* внутренняя форма основывается на ситуации военного конфликта. Различные цели, с которыми совершается это действие — для того, чтобы сдаться или для того, чтобы перейти в атаку, начать наступление — метафорически переосмысляются и приводят к появлению двух противоположных значений фразеологической единицы (‘перестать активно действовать, конфликтовать’, и ‘начать активно действовать’). Ср. значения (7а), (7б) и (7в) соответственно:

(7) *а.* Ситуацию точно описал Александр Сокуров: «Я надеюсь, что таким образом мы сможем *выйти все из окопов*. «Небесная линия» — это премия, когда все *вышли из окопов*, сели за стол переговоров. И в эти окопы мы больше возвращаться не будем, потому что там находиться больше нельзя». (Публицистика Интернета).

б. Продолжавшаяся около года «газовая» война «Газпрома» и Беларуси завершена. Противники вышли из окопов и раскурили трубку мира. (Публицистика Интернета).

в. Используя военную терминологию, полковник Чавес заявил: «Мы уже вышли из окопов и идем в атаку. Социализм — это единственный путь для спасения планеты!». (Публицистика Интернета).

Различное осмысление ситуации, лежащей в основе внутренней формы фразеологизма, хорошо демонстрирует сравнение «полных» и «кратких» форм фразеологических единиц. Полные формы обычно представляют собой пословицы или устойчивые сочетания, а краткие — это собственно идиомы или поговорки, формально совпадающие с частью полной формы, типа *бог правду видит [да не скоро скажет]*. Краткая форма может быть как результатом сокращения полной («сгущение мысли» по Потебне), так и служить основой для расширения. (Подробнее об этих процессах как отражении тенденций к эксплицитности и имплицитности во фразеологии см. [Мокиенко 1989: 96–156]). Полная форма обычно представляет собой двухчастную структуру, во второй части которой излагаются некоторые пояснения *девичий стыд до порога [переступила и забыла], собака на сене [лежит, сама не ест и скотине не дает],* следствия *голод не тетка [пирожка не поднесет]*, причины *не все коту масленица [будет и великий пост], везет как [субботнему] утопленнику [баню топить не надо],* связанные с ситуацией, обозначенной в первой части. Краткие варианты подобных выражений имеют то же значение, хотя в ряде случаев мотивировка уже не осознается носителями языка, как, например, в поговорке об утопленнике. Особый тип представляют собой пословицы, во второй части которых указываются неожиданные следствия, пояснения и т. п., «аннулирующие» все, сказанное в первой части *ума палата [да разума маловато], кто старое помянет — тому глаз вон [а кто забудет — тому два], шито-крыто[а узелок-то тут], рыбак рыбака видит издалека [поэтому стороной и обходит], пьяному море по колено [а лужа — по уши], бедность — не порок [а вдвое хуже], в здоровом теле здоровый дух [– редкость], против лома нет приема [окромя другого лома], век живи, век учись [дураком помрешь]* и т. п. Естественно, что при отсутствии эксплицитно выраженной «парадоксальной» мотивации за оставшейся частью закрепляется значение, основанное на «нормальных» причинно-следственных связях, т. е. тем самым краткая форма становится энантиосемична своему «полному» варианту.[4]

1.3. К появлению энантиосемичных значений у фразеологизма может приводить и переосмысление его внутренней формы, основанное на ее буквальном

[4] Ср. замечание Г. Л. Пермякова о различии между развертыванием (без изменения смысла) и дополнением (с изменением смысла) исходного фразеологического выражения [Пермяков 1988:57].

понимании. Так происходит с библейским выражением *нищие духом*. Существует множество объяснений значения этого выражения, восходящего к Евангелию от Матфея (Мф 5, 3) см., например, [Аверинцев 2004]. Несмотря на различные, подчас противоположные друг другу истолкования, ясно, что «библейское» значение (что бы под ним ни подразумевали — или смиренных, лишенных гордости люди, либо людей, осознанно отказавшихся от богатства) противоположно новому значению этого выражения, обозначающему людей, лишенных духовных интересов:

(8) — И все-таки, знаете ли, богатство в руках богатого человека — это лучше, чем богатство в руках нищего. — А *нищие духом* на «мерседесах»? Разве они думают о живом, о реках? (Московский комсомолец).

1.4. Отдельно можно говорить об эмоционально-оценочной энантиосемии в сфере фразеологии.[5] Обычно к этому типу энантиосемии относят явления двух типов. Во-первых, это возможность фразеологизмов выражать различные, вплоть до противоположных, эмоциональные состояния, реакции, оценки. В первую очередь речь идет о фразеологизмах-междометиях, типа *Бог ты мой! Вот это да! Полный п…ц!* и т. п. Во-вторых, сами явления, обозначаемые фразеологизмами, могут по-разному оцениваться (одобряться — не одобряться) говорящим, что тоже приводит к эмоционально-оценочной энантиосемии («эмотивной полисемии» по определению В. Н. Телии). Например, значения идиом типа *под крылышком*, *как у себя дома* может меняться в зависимости от отношения говорящего:

(9) а. После всех этих передряг не могло быть и речи о том, чтобы ехать на Банную. Только в Клуб. Только в Клуб! В наш ресторанный зал, обшитый коричневым деревом! В атмосферу прельстительных запахов! За мой столик под крахмальной скатертью! *Под крылышко* к Сашеньке… хотя нет, сегодня нечетный день. Значит, *под крылышко* к Аленушке! (А. и Б. Стругацкие. Хромая судьба).

б. Кстати, вспоминая тургеневских женщин, я как-то не представляю ни одну из них с ребенком на руках. Что-то не могу представить. И потом, эти длительные поездки-побеги в Париж под крылышко Полины Виардо. Издали, когда они не мешают, видимо, как-то легче воспевать отечественных женщин. У Тургенева, по-моему, был такой метод: влюбился, укрылся, написал. (Ф. Искандер. Сандро из Чегема.).

Конечно, эта разная эмоционально-оценочная окрашенность фразеологизмов не входит в значения идиом, а объясняется особенностями прагматической интерпретации ситуации.

[5] Подробнее об эмоционально-оценочной энантиосемии фразеологизмов см. [Цоллер 2000].

2. Фразеологическая энантиосемия может быть основана на энантиосемии слов-компонентов фразеологического выражения. Так в идиомах *вертится в голове* ('постоянно присутствует в мыслях, припоминается' и 'никак не вспоминается'), *вертится на языке* ('очень хочется сказать, спросить и т. п.' и 'никак не вспоминается'), энантиосемичные значения связаны с противоположными значениями глагола *вертеться* — 'постоянно присутствовать где-л.' и 'поворачиваться из стороны в сторону, меняя положение' (Ср. *Он постоянно вертится около взрослых* и *Не вертись, стой на месте*). С другой стороны, энантиосемия фразеологических выражений может иметь своим следствием развитие энантиосемии у слов-компонентов. Так, говорят о «зачатках энантиосемии» [Шмелев 2009: 188] у слова *заслуги* под влиянием идиомы *получить по заслугам*, которая может значить как 'наградить кого-л. в соответствии с его достижениями', так и 'наказать кого-л. за его неблаговидные поступки'[6].

3. Таким образом, энантиосемия, связанная со спецификой фразеологического знака, присутствует на всех этапах «жизни» фразеологических единиц — сопровождает их появление (ироническое переосмысление буквального значения словосочетаний, приводящее к их фразеологизации), участвует в формировании противоположных значений у одной идиомы (при различных осмыслениях ситуации, лежащей в основе фразеологического образа), характеризует отношения между генетически родственными фразеологическими выражениями (пословицами и поговорками).

References

1. *Averintsev S. S.* 2004. Proceedings. Translations: Gospel of Matthew. Gospel of Mark. Gospel of Luke. Book of Job. Psalms of David [Sobranie Sochinenii. Perevody: Evangelie ot Matfeia. Evangelie ot marka. Evangelie ot Luki. Kniga Iova. Psalmy Davidovy].
2. *Gorelov I. N.* 1986. Enantiosemy as the Conflict of Contradictory Language Development Tendencies [Enantiosemiia kak Stolknovenie Protivorechivykh Tendentsii Iazykovogo Razvitiia]. Voprosy Iazykoznaniia, 4 : 86–96.
3. *Dobrovol'skii D. O.* 2011. Conversion and Actantial Derivation in Phraseology [Konversiia I Aktantnaia Derivatsiia vo Frazeologii]. Slovo I Iazyk. Sbornik Statei k 80-letiiu Akademika Iu.D. Apresiana.
4. *Ermakova O. P.* 2002. Does Enantiosemy Exist in Russian Language as a Regular Phenomenon? Remembering General Etymology of Beginning and End [Sushchestvuet li v Russkom Iazyke Enantiosemiia kak Reguliarnoi Iavlenie?

[6] В современном русском языке более широко представлено употребление этой идиомы в значении 'наказания'. Некоторые исследователи рассматривают это значение как единственное у идиомы, а выражение *получить по заслугам* со значением 'поощрения' считают «коллокацией, коррелирующей с выражением *воздать по заслугам*» [Добровольский 2011].

Vspominaia Obshchuiu Etimologiiu Nachala I Kontsa]. Logicheskii Analiz Iazyka. Semantika Nachala I Kontsa : 61–68.
5. *Kravtsova V. Iu.* 2006. Enantiosemy of Lexical and Phraseological Unities: Language and Speech [Enantosemiia Leksicheskikh I Frazeologicheskikh Edinits: Iazyk I Rech'].
6. *Makhmutova L. R.* 2009. Basic Types of Russian Enantiosemy [Osnovnye Tipy Enantiosemii v Russkom Iazyke].
7. *Mokienko V. M.* 1989. Slavic Phraseology [Slavianskaia Frazeologiia].
8. *Novikov L. A.* 2001. Proceedings, I. Linguistic Meanings Problems [Problemy Iazykovogo Znacheniia].
9. *Permiakov G. L.* 1988. Fundamentals of Structural Paremiology [Osnovy Strukturnoi Paremiologii].
10. *Shmelev A. D.* 2009. "Insignificant " and "Unexpressed" Negotion (Cognitive and Communicative Origins of the Enantiosemy) ["Neznachashchee" I "Nevyrazhennoe" Otritsanie (Kognitivnye I Kommunikativnye Istochniki Enantiosemii)]. Logicheskii Analiz Iazyka. Assertsiia I Negatsiia : 173–202.
11. *Sokolov O. M.* 1980. Enantiosemy in the Circle of Adjacent Phenomena [Enantiosemiia v Krugu Smezhnykh Iavlenii]. Filologicheskie Nauki, 6 : 36–42
12. *Shmelev D. N.* 1977. Modern Russian Language. Vocabulary [Sovremennyi Russkii Iazyk. Leksika].
13. *Shmelev D. N.* 1958. Expressive-Ironic Statement of Negation and Negative Estimation in Modern Russian Language [Ekspressivno-Ironicheskoe Vyrazhenie Otritsaniia I Otritsatel'noi Otsenki v Sovremennom Russkom Iazyke]. Voprosy Iazykoznaniia, 6 : 63–75.
14. *Tsoller V. N.* 1998. Emotional-Evaluative Enantiosemy in Russian Language [Emotsional'no-Otsenochnaia Enantiosemiia v Russkom Iazyke]. Filologicheskie Nauki, 4 : 76–83.
15. *Tsoller V. N.* 2000. Emotional-Evaluative Phraseologisms Enantiosemy [Emotsional'no-Otsenochnaia Enantiosemiia Frazeologizmov]. Filologicheskie Nauki, 4 : 56–64

Dictionaries

16. *Baranov A. N., Voznesenskaia M. M., Dobrovol'skii D.O., Kiseleva K. L., Kozerenko A. D.* 2009. Phraseology Explanatory Dictionary of Russian Language.
17. *Birikh A. K., Mokienko V. M., Stepanova L. I.* 2007. Russian Phraseology. Historical-Etymological Dictionary.
18. *Dal' V.* 1978. Explanatory Dictionary of the Live Great Russian Language.
19. *Dictionary-Thesaurus* of Modern Russian Idiomatics. 2007.
20. *Explanatory* Dictionary of Russian Language with Etymological Information Included. 2007.
21. *Lubenskaia S. I.* 1997. Russian-English Phraseology Dictionary.
22. *Russian* Folk Proverbs. Collection of V. Dal'. 2 vv. 1984.
23. *Phraseology* Dictionary of Russian Language. 1986.
24. *Zhukov V. P.* 1993. Dictionary of Russian Proverbs and Folksays.

РВАТЬ ЗУБЫ И *МЫТЬ ДЕНЬГИ*: ОБ ОДНОМ ТИПЕ УПОТРЕБЛЕНИЯ ПРОСТЫХ ИМПЕРФЕКТИВОВ В РУССКОМ ЯЗЫКЕ

А. А. Зализняк (anna.zalizniak@gmail.com)
Институт языкознания РАН, Москва, Россия

И. Л. Микаэлян (irina.mikaelian@yandex.ru)
Penn State University, США

В статье обсуждается использование в небрежной, фамильярной или профессиональной речи простого имперфектива вместо вторичного (ср. *мыть деньги* вместо *отмывать*) и сопровождающий такое словоупотребление стилистический эффект, сочетающий элементы «сниженности» и «герметичности» (жаргонности).

Ключевые слова: имперфект, простой имперфект, фамильярная речь, небрежная речь, профессиональная речь.

ONE USE OF SIMPLE IMPERFECTIVES IN RUSSIAN

A. A. Zaliznak (anna.zalizniak@gmail.com)
Institute of linguistics, Russian Academy of Sciences, Moscow, Russian Federation

I. L. Mikaelian (irina.mikaelian@yandex.ru)
Penn State University, USA

The article discusses a specifi c stylistic effect produced by the use of simple (non-prefi xed) imperfective verbs when, under special conditions, they substitute for prefi xed imperfective verbs, as in myt' den'gi used instead of otmyvat' den'gi ("to launder money"), varit' truby, instead of svarivat' ("to weld pipes"), etc. There are three main features that characterize this effect: (1) The verb and its complement constitute a more or less idiomatic expression. (2) The simple imperfective belongs to a lower or substandard register. Thus žeč' gorški is a low-register use with respect of obžigat' gorški ("to fi re pots"); myt' den'gi is cruder than otmyvat' den'gi, and pisat' pulju is even lower than raspisyvat' pulju ("to play a game of preference").

(3) Simple imperfectives signal hermetic or jargon-like expressions: they are used in the speech of professional communities or communities of interest, cf. varit' truby instead of svarivat', gruzit' fi lm instead of zagružat' ("to download a fi lm)." Such pairs of imperfective verbs are not very numerous, but they constitute an open list, which proves that the Russian verb system possesses a mechanism that generates this kind of "quasi" non prefixed verbs.

Key words: imperfective, simple imperfective, jargon, professional community, idiomatic expression.

Предметом настоящей статьи является один специфический тип употребления бесприставочных глаголов несов. вида. Речь идет о том, что в небрежной, фамильярной или профессиональной речи может использоваться бесприставочный имперфектив — в том значении, в котором в стандартном литературном языке и/или нейтральном стиле речи выступает вторичный приставочный имперфектив. Таким образом, простой имперфектив употребляется как бы «вместо» вторичного, и при этом возникает вполне определенный стилистический эффект, ср.: *рвать* зубы вместо *вырывать*, *рвать мосты* вместо *взрывать*; *варить трубы* вместо *сваривать, мыть деньги* вместо *отмывать, мешать краски* вместо *смешивать*, *писать на пленку* вместо *записывать* и т. п. Некоторые из перечисленных глаголов уже упоминались в наших работах в связи с проблемой видовой коррелятивности (см. [Зализняк, Шмелев 2000: 50, Mikaelian, Shmelev, Zalizniak 2007, Зализняк, Микаэлян, Шмелев 2010, Зализняк, Микаэлян 2010]), однако данный феномен до сих пор не был объектом самостоятельного исследования. Здесь нас будет интересовать в первую очередь собственное, процессное (или узуальное) значение рассматриваемых имперфективов (независимо от наличия или отсутствия событийного значения, обеспечивающего им возможность вступать в видовую корреляцию).

Сознательно несколько упрощая реальную картину, напомним, что в русском языке имеется два основных морфологических типа видовых пар — суффиксальные и префиксальные. Добавление имперфективирующего суффикса образует морфологический имперфектив и, соответственно, суффиксальную видовую пару, для глаголов сов. вида с «полнозначными» приставками (*переписать — переписывать*); в префиксальную видовую пару входят глаголы сов. вида с десемантизированными, или «чистовидовыми», приставками (*построить — строить, написать — писать*).[1]

Однако, с одной стороны, между полнозначным и десемантизированным вхождением приставки имеется довольно большая зона промежуточных случаев; с другой стороны, в русском языке существуют позиции обязательной имперфективации, и их заполнение обеспечивается мощным механизмом, порождающим вторичные имперфективы практически от любого глагола сов. вида, в том числе от уже имеющего бесприставочный коррелят. Поэтому для многих

[1] «Чистовидовые» приставки здесь понимаются в том смысле, в каком они определены в [Зализняк, Шмелев 2000: 81]. См. также недавнюю статью [Janda, Nesset 2010].

глаголов оказываются применимы одновременно обе стратегии, и в таких случаях возникают «видовые тройки», ср. *намазать — мазать/намазывать*.

Между этими двумя имперфективами — простым и вторичным — могут быть различные отношения: от полной синонимии (и тогда это свободное варьирование в пределах одной и той же пары; таких немного, напр. *гибнуть-погибать*) — до их дополнительного распределения по контекстам, означающего, что мы имеем дело не с видовой тройкой, а с двумя парами, как в случае имперфективных коррелятов глагола *сшить*, который образует две отдельные видовые пары в двух своих различных значениях: *сшить1* <два куска материи> ‘соединить при помощи шитья’– *сшивать*; *сшить2* <платье> ‘изготовить при помощи шитья’ — *шить*.

Чаще всего между двумя глаголами несов. вида имеется неполная синонимия. Именно в таких случаях может возникать обсуждаемый феномен «замены» вторичного имперфектива простым (*варить трубы* вместо *сваривать*), при которой возникает обсуждаемый стилистический эффект. Существенно, что этот эффект проявляется в разной степени, так, *мазать масло на хлеб* по сравнению с *намазывать*, *капать в нос* по сравнению с *закапывать* или *густеть* по сравнению с *загустевать* (о краске), *плести* и *сплетать венок* и многие другие аналогичные пары имперфективных глаголов различаются лишь в сторону большей разговорности за счет некоторой упрощенности в обозначении ситуации простым имперфективом по сравнению с вторичным, и, наоборот, приставочный имперфектив может восприниматься как более эксплицитный и техничный. В таких случаях мы имеем дело с тем, что в статье [Зализняк, Микаэлян 2010] названо «образцовыми тройками», т. е. такими, в которых первичный и приставочный имперфективы сосуществуют в языковой практике носителей языка на равных правах.

Однако имеются случаи, когда стилистический эффект проявляется особо ярко и не ограничивается просто разговорностью. Нашу задачу мы видим в том, чтобы выявить те глаголы, в которых возникает специфический стилистический эффект, описать, в чем именно он состоит, и попытаться понять, за счет чего он возникает (в частности, почему в одних случаях есть, а в других его нет).

Характер обсуждаемого эффекта употребления простых имперфективов «вместо» вторичных, определяется следующими признаками:

1) сниженность[2]: *жечь горшки* более «низменно», чем *обжигать*; *мыть деньги* более вульгарно, чем *отмывать*, и даже *писать пулю* — несколько «снижено» по сравнению с *расписывать*;

2) герметичность, или жаргонность: они встречаются в речи профессиональных сообществ (ср. *варить трубы*) или просто людей, причастных к некоторой сфере деятельности (ср. *мыть деньги*, *писать пулю*);

3) простые имперфективы в обсуждаемом значении выступают во фразеологически связанных сочетаниях (или просто в очень ограниченном числе контекстов — дополнений).

---

[2] Что приблизительно соответствует помете «разговорно-сниженное» в словаре [Апресян 2004].

Перечисленные признаки не являются полностью независимыми. Их соотношение можно предварительно определить следующим образом: фразеологическая связанность является, по-видимому, обязательным условием возникновения интересующего нас эффекта, а сам эффект состоит из компонентов «сниженность» и «герметичность», которые в разных глаголах обнаруживают себя в различном соотношении.

В большинстве глаголов присутствуют в равной мере оба (*жечь горшки, рвать зубы, варить трубы, мыть деньги, писать на пленку, качать фильм*), в некоторых преобладает эффект сниженности (*жечь ведьм, писаться в очередь, шкурить помидоры*), в других — эффект герметичности (*бить масло, мыть золото, крепить балку, лить колокола, вязать паруса*). Эффекты сниженности и герметичности представляют собой содержательно очень близкие явления, которые все же имеет смысл различать.

Эффект сниженности возникает, по-видимому, в тех случаях, когда внимание переносится с более абстрактного ***результата***, представление о котором включено в значение приставочного имперфектива, на более конкретный, чувственно воспринимаемый ***процесс***, «материальная грубость» которого усугубляется именно оторванностью от результата, который придавал бы ему некий внеположенный самому процессу, и тем самым более «высокий» смысл: горшки *обжигают* для придания им их функциональных свойств, ведьм *сжигают* в целях борьбы со злом, письма *сжигают* для того, чтобы порвать с прошлым или уничтожить какие-то свидетельства, и т.д. Да и сам процесс оказывается при этой подмене представлен как более примитивный: *обжигать горшки* — это достаточно сложный технологический процесс; обозначение его при помощи глагола *жечь* его упрощает, и тем самым «снижает». В *жечь ведьм* переключение внимания на материальный процесс привносит дополнительный элемент кровожадности, отсутствующий в *сжигать* (как жгут листья или хворост, наблюдая, как они горят)[3], в *рвать зуб* — элемент грубости, и т.д.

Относительно источника возникновения эффекта герметичности можно высказать следующее предположение. Этот эффект наблюдается в случаях, когда вторичный имперфектив имеет узкоспециальное и, часто, техническое значение и в этом значении используется только с определенным дополнением. В таком случае, даже если приставка не является чистовидовой, т. е. вносит некий существенный дополнительный смысловой компонент в значение глагола, ее отсечение не ведет к потере этого смыслового компонента, поскольку бесприставочный имперфектив, благодаря фиксированному дополнению, сужающему его значение, получает возможность соотноситься с той же ситуацией, принадлежащей определенной профессиональной или предметной сфере и известной «посвященным». Но этот специфический компонент значения остается в бесприставочном имперфективе невыраженным, имплицитным, что создает эффект герметичного, или жаргонного, значения, функционирующего по принципу “sapienti sat”. Фактически у этого морфологически как бы простого, но на самом деле семантически сложного (потому что приставка

[3] Ср. начало романа Р. Брэдбери «451 по Фаренгейту»: *Жечь было наслаждением. Какое-то особое наслаждение видеть, как огонь пожирает вещи, как они чернеют и меняются.*

отсечена, а ее семантический вклад остался), т. е. «квазибесприставочного» имперфектива возникает новое, фразеологически связанное значение, «стандартным» носителем которого является приставочный имперфектив, ср. [Апресян 1995: 106].

Оба обсуждаемых эффекта (сниженность и герметичность) демонстрирует, например, сочетание *варить трубы*. В литературном языке глагол *сварить* относится к тому типу что *сшить*, с четким дополнительным распределением имперфективных коррелятов по значениям: *сварить* <суп> (= ‘изготовить’) *vs. варить*; *сварить* <рельсы, трубы > (= ‘соединить сваркой’) — *сваривать* (ср. пример (1)). Однако в профессиональной речи сварщиков, а также в разговорной речи людей самых разных профессий наряду с литературным *сваривать* в этом контексте употребляется простой имперфектив *варить* (ср. примеры (2), (3); в примере (4) языковая игра построена на том, что мальчик не владеет специальным значением глагола *варить*).

(1) *Я играла сварщицу, а Блинов — замечательный актер — моего возлюбленного. Я* ***сваривала*** *какую-то трубу, а он стоял позади.* [Лидия Смирнова. Моя любовь (1997)] [пример из НКРЯ]

(2) *Требуются бригады сварщиков по внутренним трубопроводам умеющие* ***варить трубы*** *под давлением*.

(3) *Нельзя* ***сваривать трубы****, кромки которых покрыты ржавчиной, маслом, краской или грязью <…>*

(4) *Мальчик с папой идут по дороге. Мальчик видит — дядя* ***варит трубы****. Мальчик спрашивает у папы: — А что это дядя делает? — Это дядя* ***варит трубы****. — А с чем их едят?* [Коллекция анекдотов: дети (1970–2000)] [пример из НКРЯ]

Обратим внимание на то, что глагол *варить* в таком употреблении отличается от *варить* <суп> не только отсылкой к разным процессам (‘плавить металл’, в отличие от ‘готовить пищу кипячением’), но еще и по своей внутренней форме: Это *варить*, равное по смыслу *сваривать*, т. е. ‘соединять, плавя металл’, но из него как бы удален префикс — и тем самым идея ‘соединять’ оказывается выраженной имплицитно.

Типичный пример собственно герметичного эффекта представлен значением глагола *бить*, которое реализуется в сочетании *бить масло* (наряду с *сбивать масло*). Рассмотрим его подробнее.

(5) *На месте кожзавода позже был построен маслозавод. Сюда съезжались* ***бить масло*** *со всего района*.

Сбивать и *бить* до какой-то степени взаимозаменимы, хотя и не всегда. *Бить масло* встречается преимущественно на сайтах энциклопедий и технических справочников, а также в разного рода повествованиях о деревенской жизни; *сбивать масло* — преимущественно в рекламных текстах туристических фирм.

(6) *Во всех домах хозяева заняты делом: топят печку торфом, доят коров,* ***сбивают масло****, прядут пряжу, пекут пироги по старым рецептам.* [Об Ирландии: реклама турфирмы]

(7) *Расспрашиваем жителей, чем они тут заняты. Узнаем, что пасут и доят коров, делают сыр,* ***бьют масло****.* [обозначение процесса взято как бы со слов этих людей]

(8) *Мария и Георгий — из Смоленска, шесть лет прожили при Оптиной пустыни, а потом перебрались во Фроловское. Они ухаживают за стадом, доят коров,* ***бьют масло****, изготавливают сметану.*

Бить масло — это конвенциализованный способ наименования технологического процесса *сбивания масла*, включающего этап, на котором крупинки, образовавшиеся в результате определенного типа механического воздействия на жирные сливки или сметану, *сбиваются в масло*. Тем самым значение приставки *с-* здесь — соединительное (а не чисто результативное, как мог бы счесть носитель языка, не знакомый с технологией производства сливочного масла). Точнее, в *сбить масло*, как и в *связать человека* мы имеем дело со значением, производным от значения соединения[4]; оно является промежуточным звеном на пути семантической деривации (ср. *связать* <два бревна друг с другом> и *связать* <человека>), который приводит к «чисто» результативному значению, (*связать кофту*).

Приведем теперь предварительный список бесприставочных глаголов несов. вида, которые в сочетании с определенным дополнением обнаруживают обсуждаемый эффект — в том или иной его варианте. Для каждого такого глагола мы указываем тот вторичный имперфектив, «вместо» которого он употребляется, указываем в скобках количество вхождений, найденных поисковой системой Яндекс[5], приводим некоторые примеры (из найденных Яндексом)[6] делаем некоторые комментарии.

- *бить масло* (670) вм. *сбивать* (4100)
  См. примеры (5)–(8).
- *варить трубы* (1800) вм. *сваривать* (1300)
  См. примеры (1)–(4).
- *грузить программу* (550) вм. *загружать* (3600)

---

[4] Подробнее об этом значении приставки *с-* см. [Зализняк, Шмелев 2001: 240].

[5] Мы приводим точные (в пределах 100) или округленные (если более 100) цифры, указывающие на количество вхождений, найденных путем поиска целиком словосочетания в инфинитиве (если в другой форме, то это специально оговаривается). Цифры, полученные таким образом, отражают реальную картину весьма приблизительно, но тем не менее они могут быть использованы для того, чтобы составить представление о сравнительной частотности обсуждаемых форм, для чего они и приводятся.

[6] В случае если частотность приставочного имперфектива превосходит частотность простого в несколько раз, и при этом приставочный имперфектив является нейтральным словом общелитературного языка, примеры его употребления не приводятся.

(9) *А зачем **грузить программу** — Гугл Планета Земля? Можно открыть напрямую — карты Гугл, выбрать Рославль, поставить в настройках галку — фотографии, и всё видно без лишних программ.*

(10) *Всё работает нормально, по крайней мете под WinXP. Надо только **грузить программу** со второго сайта.*

(11) *Основной задачей сайта является предоставление возможности **загружать программу** телепередач.*

Глагол *загружать* в этом техническом значении возник, очевидно, как калька с английского *upload/download*; от него вторично образовано *грузить* именно как термин профессионального жаргона.

- *жать масло* (450) вм. *отжимать* (450)

(12) *из семян тунгов (их несколько видов) чаще всего **жмут масло**, которого в них много.*

(13) *Он изобрел и запатентовал гидравлический пресс, при помощи которого из какао бобов можно было **отжимать масло**.*

- *жечь ведьм* (1400) вм. *сжигать* (1800)

(14) *В любые времена, в любой стране дай плебеям в руки христианский крест — пойдут **жечь ведьм** и еретиков*

(15) *…эпизод из фильма «Эльвира, повелительница тьмы», в котором жители одного городка Новой Англии выкапывают закон, доселе разрешающий **сжигать ведьм** на костре.*

Здесь присутствует лишь эффект сниженности. *Жечь ведьм* — употребляется в эмоционально окрашенных и полемических контекстах, *сжигать* — в более спокойно-академических (ср. примеры (14) и (15)). Эффект эмоциональной окрашенности простого имперфектива может проявляться и в менее экзотических контекстах, ср.:

(16) *он начал лихорадочно **жечь документы** и загранпаспорта*

(17) *Тот факт, что по закону **сжигать документы** нельзя, не останавливает многих желающих избавиться от надоевших бумаг.*

- *жечь горшки* (15) вм. *обжигать горшки* (2000)

(18) *Вера, по предсмертному завету отца, стала **жечь горшки**, чем спасла от голода всю семью и во время войны и сразу после нее.*

(19) *Хорошо, что налоговики не указывают мне, как шить сапоги,* ***жечь горшки****, доить корову и т. д.*

В словосочетании *жечь горшки* опущена полнозначная приставка (*об-*), и это — термин профессионального жаргона (т. е. помимо сниженности есть еще и герметичность).

- *качать фильм* (из Интернета) (38 тыс.) вм. *скачивать* (43 тыс.)

(20) *Где-то видел что можно настроить торрент-клиент так, чтобы* ***качать*** *и смотреть фильм одновременно без проблем.*

(21) *Зачем* ***качать фильм****, который можно и по ТВ посмотреть?*

(22) *просьба прежде чем начинать* ***скачивать фильм*** *по частям, проверьте все ссылки на работоспособность.*

(23) *Зачем* ***скачивать фильм****, если его можно смотреть онлайн?*

- *лить колокола* (1350) вм. *отливать колокола* (1350)

(24) *В 1557 году он* ***льет колокол*** *на Соловки совместно с Кузьмой Михайловым*

(25) *В этом году он вместе с литейщиком Гаспаром Нади* ***отливает колокол*** *и поднимает его на городскую башню Аринго*

- *мешать краски* (660) вм. *смешивать* (22 тыс.)

(26) *Можно читать молитву, Можно жать на курок, Можно* ***мешать краски*** *палитры, Можно слякоть дорог.*

(27) *Папа не будет дзен — он слишком вложился в сына, А сын хочет* ***мешать краски*** *и ночью писать картины.*

- *морозить продукты* (300) вм. *замораживать* (5500)

(28) *В некоторых случаях достаточно бывает отрегулировать работу этого механизма, чтобы холодильник вновь начал прекрасно* ***морозить продукты****.*

(29) *Я не большая любительница* ***морозить продукты****, поэтому я решила, что буду готовить целую курицу!*

- *морозить цены* (300) вм. *замораживать* (6600)

(30) *Если продолжим* ***морозить цены*** *на продукты, тем самым мы будем бить по своим крестьянам*.

(31) *«Они жалуются на свои проблемы, мы — на свои, все в рабочем режиме», — сказал источник в министерстве, добавив, что* ***«морозить» цены*** *пока никого не просили и, возможно, и не попросят*.

Ср. также примеры, показывающие, что переносное значение глагола *морозить* используется не только с дополнением *цены*.

(32) *Китай* ***морозит свинину*** *вместе с ценами*

(33) *Онищенко* ***морозит птицу*** (= ‘не пускает в продажу’)

- *мыть деньги* (550) вм. *отмывать* (21 тыс.), иногда в кавычках (маркирующих стилистическую окрашенность данного употребления):

(34) *Менеджер Сбербанка помогал* ***«мыть» деньги***

(35) *не будет там никакого стадиона, просто будут* ***мыть деньги****, как моют на строительстве моста уже 20 лет*

(36) *а еще через Чечню удобно* ***мыть деньги*** *— сначала из бюджета на военные действия, а теперь — на восстановление*.

(37) *Итальянская мафия хотела* ***отмывать деньги*** *на последствиях землетрясения*.

- *писать пулю* (1000) вм. *расписывать* (600)

(38) *Офицеры сели* ***писать пулю****, но им нужен четвертый* [начало анекдота]

(39) *если вы сутки напролет способны* ***расписывать пулю*** *под пиво, то это отнюдь не означает, что те же сутки вы можете проводить в пути*.

- *писать на пленку* (= ‘делать аудиозапись’) (700) вм. *записывать* (4500)

Сейчас говорят также *писать на цифру*; «на цифру» скорее именно *пишут*, а не *записывают*.

(40) *в режиме WEB, камера* ***писать на пленку*** *не будет и звук берется тоже не от нее, а от отдельного Вашего микрофона*

- *писаться в очередь* (4200) вм. *записываться* (15 тыс.)

(41) *Уже хотя бы за то, что в истории российских продаж это была первая модель Audi, на которую покупатели в массовом порядке начали* ***писаться в очередь*** *и ждали ее выхода на рынок*

- *рвать мосты* (850) вм. *взрывать* (7500)

(42) *…с призывом к населению громить тылы немецких армий,* ***рвать мосты****, развинчивать рельсы, поджигать леса, уйти в партизаны, все время беспокоить немцев-угнетателей.*

- *рвать зубы; зуб* (5400; 2600) вм. *вырывать* (3700; 3600) [7]

(43) *А я знаю как без боли* ***рвать зубы****. И коренные тоже.*

(44) *Шли столетия, врачи продолжали* ***вырывать зубы****, а их щипцы-зубодеры все больше совершенствовались.*

(45) *Что написать другу, которому сейчас будут* ***рвать зуб****, чтоб он прочитал и забыл о боли?*

(46) *есть ли вероятность того, что когда будут* ***вырывать зуб****, он сломается?*

Рвать зубы — это скорее разговорное, а не профессиональное выражение. Обратим внимание на то, что здесь основная часть значения заключена в опущенной приставке: ‘извлечь, вынуть наружу’, значение глагольной основы здесь менее существенно, от нее сохраняется лишь идея «грубой силы». Ср. сходное распределение смысловой нагрузки между приставкой и глагольной основой в *отмывать деньги* (см. пример (41)), где важно именно значение пристаки *от-*: ‘удаление, избавление от криминального происхождения’.

- *шкурить помидор* (6)[8] (*баклажан*, *банан*) вм. *ошкуривать* (110)

(47) ***Помидоры шкурим****, мелко кромсаем и не спеша тушим вместе с луком, чесноком и перцем.*

(48) *Остывший* ***баклажан шкурим*** *и даем стечь лишней жидкости.*

(49) ***Ошкуриваем помидоры****, режем их тончайшими полукружьями.*

---

[7] Для многих глаголов из нашего списка предпочтительным, или даже единственно возможным, является множественный объект, что до определенной степени коррелирует с семантикой итеративности. Обсуждение этого их интересного свойства выходит за пределы данной статьи.

[8] В силу экзотичности этих глаголов поиск осуществлялся более сложным способом.

Имеется, кроме того, ряд входящих в идиоматические сочетания простых имперфективов, для которых соответствующий вторичный имперфектив в составе данной идиомы малоупотребителен. Во всех этих парах частотность простого имперфектива в десятки и сотни раз превосходит частотность приставочного. Вторичный имперфектив при этом сам по себе обычно является стилистически нейтральным литературным глаголом.

- *грести деньги* (15 тыс.) вм. *загребать* (3800)

(50) *кормят отвратно, **денег гребут** не хило, так еще и обслуживающий персонал походу не знает своих обязанностей!*

(51) *Больше всего **денег загребают**, конечно же, адвокаты, менеджеры, арт-директора и прочий руководящий разработкой люд*

- *качать мышцы* (28 тыс.) вм. *накачивать* (5600)

(52) *мне 17 лет, хочу накачать большие мышцы, как правильно **качать мышцы**, какие делать упражнения, и как правильно питаться, хочу накачать все тело.*

(53) *как надо делать упражнения, чтобы сжигать жир, а не **накачивать мышцы**?*

- *копать картошку* (4900) вм. *выкапывать* (800)[9]

(54) *Редко какая девушка поместит на аватарку фотку, где она в поте лица занимается генеральной уборкой или **копает картошку** на дачном участке.*

(55) *Поэтому осенью огороднику не нужно долго **выкапывать картошку**, а достаточно просто выдернуть чулок с урожаем за оставленный сверху край.*

(56) *Глупо **выкапывать картошку**, едва посадив ее, только потому, что «очень кушать хочется».*

Выкапывать картошку — свободное сочетание, результирующее значение которого композиционально. Между тем, *копать картошку*, — сочетание идиоматическое: *копают* не любую картошку, а только *выращенную*, т. е. в *копать* <картошку> включено некоторое специальное знание (отсутствующее в *выкапывать*). Отметим, что между парами *копать яму* (24 тыс. ответов на Яндексе) и *выкапывать яму* (800 ответов), подобного противопоставления нет: в данном контексте эти глаголы различаются лишь большей эксплицитностью вторичного имперфектива по сравнению с первичным.

[9] Цифры приводятся для формы 3 л.ед. ч.

- *крепить плинтус* (720), вм. *прикреплять* (32)

(57) *В этой статье речь пойдет о плинтусáх, а именно о том, как* ***крепить плинтус****.*

(58) *Как лучше всего* ***прикреплять плинтус****? Нам тут посоветовали приклеить его на акриловый (кажись) клей, а потом еще прибить гвоздями.*

Прикреплять плинтус сказать можно, но это не будет «правильным», идиоматичным обозначением определенного элемента строительного дела (при том что в других значениях именно *прикреплять*, а не *крепить* является нормальным имперфективом).

- *мыть золото* (11 тыс.) вместо литературного *намывать* (900)

(59) *Вскоре Лев Иванович нашёл эти богатые золотом песчаные россыпи и предложил* ***намывать золото*** *в реках.*

(60) ***Мыть золото*** *в Подмосковье — это реально?*

(61) *Куда же еще поехать учиться* ***мыть золото****, как не в Калифорнию!*

(62) *Артель золотоискателей* ***моет золото*** *в сибирской тайге.*

(63) (67) *Обычный старатель* ***намывает*** *по три грамма* ***золота*** *в день.*

Глагол *мыть* в контексте дополнения *золото* имеет практически исключительно процессное значение и, в отличие от глагола *намывать*, который в основном употребляется именно с квантифицированным объектом, ср. (66) и (67).

- *шить дело* (28 тыс.), вм. *пришивать* (64)[10]

(64) *В Челябинске партизанам будут* ***шить дело*** *за «экстремизм»*

(65) *Не собираюсь вам* ***пришивать дело****, Милн, но если в моих руках будет достаточно улик, тогда берегитесь!*

Идиома *шить дело <кому-то>*, имеющая свою историю возникновения (которую здесь нет возможности обсуждать) производна от глагола сов. вида *пришить <дело>*. Вторичный имперфектив *пришивать* употребляется в этом значении крайне редко.

Список простых имперфективов, обнаруживающих, в определенных условиях, эффект сниженности и/или герметичности, может быть продолжен. Хотя

[10] Цифры приводятся для формы 3 л.ед. ч.

таких глаголов не очень много, но очевидно, что это системное явление, т. е. в русском языке действует некий механизм, который такие квазибесприставочные имперфективы порождает.

References

1. *Apresian Iu. D.* 1995. Redundant Aspectual Paradigm Interpretation in the Explanatory Dictionary [Traktovka Izbytochnykh Aspektual'nykh Paradigm v Tolkovom Slovare]. Proceedings, II : 102–113.
2. *Apresian Iu. D.* 2004. New Russian Explanatory Synonym Dictionary [Novyi Ob"iasnitel'nyi Slovar' Sinonimov Russkogo Iazyka].
3. *Janda L., Nesset T.* 2010. Taking Apart Russian RAZ- . Slavic and East European Journal, 45 (3) : 477–502.
4. *Mikaelian I., Shmelev A., Zalizniak A.* 2007. Imperfectivization in Russian. Meaning-Text Theory 2007, Proceedings of the 3rd International Conference on Meaning-Text Theory : 315–323.
5. *Zalizniak A. A., Mikaelian I.* On the Place of Aspect Triplets in the Russian Aspectual System [O Meste Vidovykh Troek v Russkoi Aspektual'noi Sisteme]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2010" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2010") : 130–135.
6. *Zalizniak A. A., Mikaelian I., Shmelev A. D.* 2010. Aspectual Correlation in Russian Language: in Defense of Aspectual Pair [Vidovaia Korreliativnost' v Russkom Iazyke: v Zashchitu Vidovoi Pary]. Voprosy Iazykoznaniia, 1 : 3–23.
7. *Zalizniak A. A., Shmenelv A. D.* 2001. Conveni, convici, convixi. [Glagol'nye Prefiksy i Prefiksal'nye Glagoly]. Moskovskii Lingvisticheskii Zhurnal, 5 : 233–252.
8. *Zalizniak A. A., Shmenelv A. D.* 2000. Introduction to the Russian Aspectology [Vvedenie v Russkuiu Aspektologiiu].

ЭКСКЛАМАТИВЫ В РУССКОМ ЯЗЫКЕ: КОРПУСНОЕ ИССЛЕДОВАНИЕ

Н. А. Зевахина (natalia.zevakhina@gmail.com)

Московский Государственный Университет
имени М. В. Ломоносова, Москва, Россия

В статье исследуются контексты экскламативного употребления (т. е. выражения экскламации) вопросительных местоимений в Национальном Корпусе Русского Языка (www.ruscorpora.ru). Под экскламацией понимается удивление говорящего по отношению к наблюдаемому им положению дел в мире, которое нарушает его ожидания. Результаты исследования показывают, что вопросительные местоимения можно разделить на те, которые имеют независимое экскламативное употребление, и те, которые требуют дополнительного контекста, т. е. встречаются в специальных синтаксических конструкциях. Кроме того, рассматриваются экскламативные употребления вопросительных местоимений в сентенциальных актантах. Делаются выводы о лексико-грамматических свойствах матричных предикатов, допускающих такие контексты.

Ключевые слова: экскламатив, экскламация, местоимение, вопросительное местоимение.

EXCLAMATIVES IN RUSSIAN: A CORPUS STUDY

N. A. Zevakhina (natalia.zevakhina@gmail.com)

Lomonosov Moscow State University, Moscow,
Russian Federation

The paper investigates the exclamative use of wh-words in the National Corpus of Russian Language. Exclamation is a verbally uttered speaker-oriented emotion occurring when the state of affairs in the real world violates the speaker's expectations. The paper shows that, basically, Russian wh-words as exclamatives can be split into four groups according to four criteria: (i) independent exclamative use; (ii) 'anaphoric' use (i. e., with reference to phrases within a particular discourse); occurrence in special syntactic constructions (iii) with the particle tol'ko, (iv) with the particle vot. The first group of wh-exclamatives satisfies all of the criteria, the second group complies with (ii)-(iv), etc. We also discuss the exclamative use of Russian wh-words in sentential arguments and lexico-grammatical properties of matrix predicates allowing for such contexts that rather exhibit a continuum than clear-cut classes.

Key words: exclamative, exclamation, pronoun, wh-word.

1. Введение

Данная работа посвящена изучению эксклamативов в русском языке на материале Русского Национального Корпуса, далее НКРЯ (www.ruscorpora.ru). Следуя работам Michaelis and Lambrecht 1996, Michaelis 2001, Portner and Zanuttini (2000, 2003), Ono (2002, 2006), Kalinina in press, под эксклamативами мы будем понимать морфологические показатели и синтаксические конструкции, которые выражают эксклamацию, то есть удивление говорящего по отношению к наблюдаемому им положению дел в мире, которое нарушает его ожидания. Другими словами, цель эксклamации — не столько передача информации адресату о некоторой ситуации, имеющей место в действительности, сколько эмоциональная реакция (удивление) говорящего на эту ситуацию. Предложения, содержащие эксклamативы, мы будем называть эксклamативными предложениями.

Мы оставляем за рамками рассмотрения восклицательную интонацию, которая характерна для выражения эксклamации, и сосредоточимся на анализе синтаксических и прагматико-семантических свойств ряда конструкций, с помощью которых выражается эксклamация.

Прототипическим примером эксклamатива, на наш взгляд, можно считать японский показатель *nante*[1]:

(1) японский [Ono 2006]

kono	*hana-wa*	*nante*	*utukusii*	*no da roo.*
этот	цветок-Top	Exc	красивый	Fin Foc Mood

'Какой красивый цветок!'

Во многих языках, в том числе и в русском, отдельного показателя эксклamатива нет. В Kalinina in press утверждалось, что эксклamация также зачастую выражается синтаксически зависимыми конструкциями: инфинитивными оборотами (2), номинализациями (3) и *что*-предложениями (4).

(2) *Находить такие книжки да еще приносить их в дом!* [Валентина Осеева. Динка (1959)][2]

(3) аварский[3] [Kalinina in press]

dur	*ħaliqa-ł-i*	*šši-b!*
2Sg.Gen	подлый-Nml	что-n

'Какой ты подлец!'

---

[1] Показатель *nante* имеет два алломорфа *nanto* и *nantoyuu*, дистрибуция которых зависит от их синтаксического окружения. Подробнее смотреть Ono (2002, 2006).

[2] Здесь и далее, если специально не оговорено, примеры взяты из НКРЯ.

[3] Аварский язык относится к аваро-андийской группе нахско-дагестанской семьи языков.

(4) немецкий [Buscha 1976]
[Ich wundere mich,] Daß du immer noch Witz-e machen kann-st.
я удивлять я.Acc что ты всегда еще шутка-Pl делать мочь-Pres.2Sg
‘Удивительно, ты все еще шутишь (об этом)!’
(букв. ‘[Я удивляюсь тому,] что ты все еще шутишь (об этом)!’)

В данной работе мы остановимся на еще одном широко распространенном способе выражения экскламации, который отмечается в Michaelis 2001 и заключается в использовании вопросительных местоимений в экскламативных клаузах. Такое употребление вопросительных местоимений встречается, например, в русском (5), английском (6), итальянском (7), а за пределами европейской семьи языков — например, в турецком (8) и в адыгейском (9).

(5) ***Какая*** *радость бывает встретиться!* [митрополит Антоний (Блум). О встрече (1969)][4]

(6) английский [Portner and Zanuttini 2003]
What a delicious dinner you’ve made!
‘Какой вкусный обед ты приготовил(а)!’

(7) итальянский [ibid.]

Che libri,	*a tua sorella,*	*le hanno*	*regalato!*
какой книги	Dat твоя сестра	ей Aux.Pres.3Pl	дать.Pass.Part

‘Какие книги они подарили твоей сестре!’

(8) турецкий [Michaelis 2001]

Neler	*bulduk,*	*(neler)!*
что.Pl	найти.Pst.1Pl	что.Pl

‘Что мы нашли!’ (букв. ‘Что мы нашли, что!’)

(9) адыгейский [Kalinina in press]

wə-ne-xe-r	*səd-ew*	*dax-a!*
2Sg.Poss-глаз-Pl-Abs	что-Adv	красивый-Q

‘Какие красивые у тебя глаза!’

Таким образом, цель данной работы — провести корпусное исследование употребления русских вопросительных местоимений (*где, зачем, как, какой, когда, куда, почему, сколько*) в качестве экскламативов. Мы оставляем за рамками рассмотрения конструкцию *что за*, синтаксические и семантические особенности которой подробно описаны в Рахилина 1990, Подлесская 2007 и Кузнецова 2009.

[4] Здесь и далее полужирным шрифтом в русских примерах мы будем выделять вопросительные местоимения.

2. Вопросительные местоимения в эксклamативных клаузах

Согласно Rett 2008, в английском языке только местоимения *what* ‘что/какой’[5] и *how* ‘как’ употребляются в эксклamативных клаузах, в то время как остальные вопросительные местоимения, как *who* ‘кто’, *which* ‘который/какой’, *when* ‘когда’, *where* ‘где/куда’, *why* ‘почему’, не грамматичны в подобных контекстах. Дж. Ретт объясняет данное явление, вводя ограничение по степени: в эксклamативных клаузах употребляются только такие вопросительные местоимения, которые представимы в виде степеней некоторого признака, причем степень удивления говорящего соответствует высшей степени данного признака.

Как можно заметить, местоимения со значениями ‘что/какой’ и ‘как’, во-первых, характеризуют некоторый признак (и этим отличаются от остальных вышеперечисленных местоимений), а во-вторых, этот признак качественный. Однако нетрудно убедиться, что значение ‘сколько’ также можно представить в виде степеней, но не на качественной, а на количественной шкале. И действительно, данное значение употребляется в качестве эксклamатива и выражается с помощью конструкций *how many, how much*[6].

Стоит отметить, что это ограничение также позволяет объяснить неграмматичность некоторых вопросительных местоимений в качестве эксклamативов в других языках: *qui* ‘кто’, *què* ‘что’, *on* ‘где’, *quan* ‘когда’, *per què* ‘почему’ в каталанском; *perche* ‘почему’ в итальянском; *aš’tew* ‘почему’ в адыгейском. Однако, согласно Portner and Zanuttini 2003, в итальянских эксклamативных предложениях грамматичны вопросительные местоимения *dove* ‘где’, *quando* ‘когда’, *chi* ‘кто’.

Что касается русского языка, то еще А. М. Пешковский в 1914 году (Пешковский 2001) отметил, что в русских эксклamативных клаузах грамматичны вопросительные местоимения *кто, что, какой, куда, как* и др. Мы провели небольшое исследование на материале НКРЯ, проанализировав около 350 примеров. Как и следовало ожидать, в эксклamативных предложениях употребляются вопросительные местоимения *какой*, *как* (10)–(11):

(10) ***Какие** выпуклые характеры! **Какой** грандиозный сюжет, **как** всё в нём сцеплено и взаимосвязано!* [Владимир Войнович. Иванькиада, или рассказ о вселении писателя Войновича в новую квартиру (1976)]

(11) *О **каком** же добре речь!* [Юрий Трифонов. Предварительные итоги (1970)][7]

Также встречается ожидаемое местоимение *сколько*:

---

[5] Именная группа после этого местоимения должна быть неопределенной, что выражается в ед. ч. с помощью неопределенного артикля, а во мн. ч. существительным без артикля.

[6] Данные конструкции встречаются в эксклamативных, но не в вопросительных предложениях.

[7] Также грамматичны употребления местоимения *какой* в прочих косвенных падежах.

(12) ***Сколько*** *трагедий на этой почве!* [С. П. Мельгунов. «Красный террор» в России (1924)]

Что касается других местоимений (*кто, что, куда, где, когда, зачем, почему*), многие найденные примеры выражают значение риторических вопросов:

(13) ***Кто*** *ожидал, что эта туча доберется и до нас грешных!* [М. Ю. Лермонтов. Вадим (1833)]

(14) ***Кому*** *под суд попасть хочется!* [М. Е. Салтыков-Щедрин. Игрушечного дела людишки (1886)][8]

(15) ***Что*** *может случиться за тридцать шесть часов!* [Татьяна Окуневская. Татьянин день (1998)][9]

(16) ***Что*** *они потеряли и что получили взамен!* [М. А. Алданов. Пещера (1932)][10]

(17) ***Куда*** *черт понес!* [В. Я. Шишков. Угрюм-река. Ч. 1–4 (1913–1932)]

(18) *Ох, господи!* ***Куда*** *же он побег, проклятый! — с тревогой заголосили солдатки.* [Валентина Осеева. Динка прощается с детством (1969)][11]

(19) *Сам знаешь: где нашему брату достать денег?* ***Где*** *их взять!* [Д. В. Григорович. Пахатник и бархатник (1860)]

(20) ***Когда*** *я их приучу к порядку! — проговорила генеральша и дернула за сонетку.* [А. Ф. Писемский. Тысяча душ (1858)]

(21) *— Хватит!* ***Зачем*** *стреляться!* ***Зачем*** *увеличивать детскую смертность!* [Александр Проханов. Господин Гексоген (2001)]

(22) *— И* ***почему*** *мне в последнее время в голову лезут дурацкие мысли!* [Татьяна Тронина. Никогда не говори «навсегда» (2004)]

---

[8] Также грамматичны употребления местоимения *кто* в других косвенных падежах.

[9] Аварский язык относится к аваро-андийской группе нахско-дагестанской семьи языков.

[10] Интересно также отметить, что местоимение что в контексте риторического вопроса может приобретать значение цели (‘зачем’): ***Что*** *ему ваши поклоны!* [М. Н. Загоскин. Москва и москвичи (1842-1850)]

[11] К сожалению, употребление местоимения *куда* в эксклamативных клаузах не было найдено в нейтральном контексте, в котором бы отсутствовали какие бы то ни было частицы (*же, бы* и др.). Подробнее о них см. ниже. С другой стороны, ***Куда*** *черт понес!* является скорее оформившимся устойчивым выражением, значение которого в высокой степени идиоматично.

Однако, на наш взгляд, местоимения *кто, что, куда, где* могут выражать эксклamативное значение (примеры (23)–(26) мои — *Н.З.*)[12]:

(23) ***Кого*** *я вчера встретил!*

(24) ***Что*** *Маша нарисовала!*

(25) ***Куда*** *я летом ездил!*

(26) ***Где*** *я вчера побывал!*

Заметим, однако, что данные примеры отличаются от примеров (10)–(12) тем, что после их произнесения говорящий должен упомянуть, кого же он вчера встретил, что именно нарисовала Маша и т. д. Другими словами, употребления таких местоимений в эксклamативных клаузах являются анафорическими, требуют расширенного контекста.

Местоимения *когда, зачем* и *почему* не употребляются в такой анафорической функции, что иллюстрируется ниже (примеры (27)–(29) мои — *Н.З.*):

(27) ?? ***Когда*** *я ездил на море!*

(28) ?? ***Зачем*** *Маша ездила в Москву!*

(29) ?? ***Почему*** *Маша нарисовала эту картину!*

Однако из всего вышесказанного не следует, что вопросительные местоимения, помимо *какой, как* и *сколько*, не употребляются в эксклamативных предложениях. Характерной особенностью русских вопросительных местоимений *кто, что, куда, где* (и, по всей видимости, отличием их от вопросительных местоимений в других языках) является их употребление в контексте сказуемого под отрицанием и с (видимо, факультативной) частицей *только*[13]:

(30) ***Где*** *не бывал я, по* ***каким*** *дорогам не ходил!* [И. С. Тургенев. Рудин (1856)][14]

---

[12] Заметим, что местоимения *какой*, *как* и *сколько* также могут употребляться в данной функции.

[13] Конструкции *какой только* и *как только* со сказуемым под отрицанием также грамматичны и, как другие подобные им конструкции, выражают количественные признаки. Конструкция *сколько только* не грамматична, видимо, в силу фонетических причин – повторения последовательности олько, однако частица *только*, переставленная на синтаксическую позицию перед сказуемым, синтаксически и семантически относится к глаголу, а не к местоимению *сколько*: ***Сколько*** *людей всю жизнь только готовятся к жизни!* [Николай Сладков. Зарубки на памяти (1970-1996) // «Звезда», 2000]

[14] При добавлении частицы *только* предложение остается грамматичным: ***Где*** *только не бывал я, по* ***каким*** *только дорогам не ходил!*

(31) ***Куда*** *только его, военного переводчика, не носила судьба!* [И. Грекова. Фазан (1984)]

(32) *И* ***кто*** *только не сложил тут свою вольную голову!* [Ю. О. Домбровский. Хранитель древностей, часть 1 (1964)]

(33) ***Что*** *только не писали о Тесле журналисты!* [Наталья Галкина. Вилла Рено // «Нева», 2003]

В случае местоимения *когда* частица *только* не является факультативной, достаточно сравнить (27) и (34):

(34) ***Когда*** *только я не ездил на море!*[15]

Семантически все перечисленные конструкции являются экскламативными в принятом здесь смысле, поскольку выражают высшую степень на шкале того или иного количественного признака: *где только не* значит ‘везде’ (или ‘во многих местах’); *куда только не* значит ‘всюду’ (или ‘во многие места’); *кто только не* значит ‘каждый’ (или ‘многие’); *что только не* — ‘всё’ (или ‘многое’); *когда только не* значит ‘всегда’ (или ‘часто’).

Что касается местоимений *зачем, почему*, они употребляются в таком контексте, где глагол не обязательно находится под отрицанием. Частица *только*, вероятно, также является факультативной в подобного рода контекстах. Значения, выражаемые такими предложениями, близки к значениям риторических вопросов: на них не предполагается или затруднительно дать ответ, говорящий выражает свое удивление и даже непонимание, какова причина того или иного действия или состояния, имеющего место в действительности (34), или какова его цель (35).

(34) ***Почему*** *только он меня с собой не взял!* [Межа и кровь (2003) // «Криминальная хроника», 2003.07.24]

(35) *И* ***зачем*** *только я на том уроке занимался посторонним делом!* [Валерий Медведев. Баранкин, будь человеком! (1957)]

Упомянутая выше анафорическая функция присуща экскламативным конструкциям, состоящим из вопросительных местоимений и указательной частицы *вот*. В примерах (36)–(42) такие экскламативные конструкции отсылают к предшествующему или последующему контексту:

(36) *Так вот* ***как*** *выглядит наш предок! Вот* ***какое*** *у него было обезьянье лицо, звериные, острые скулы, полусогнутое, очевидно, волосатое тело.* [Ю. О. Домбровский. Обезьяна приходит за своим черепом, часть 1 (1943–1958)]

[15] Пример мой – *Н. З.* В НКРЯ подобных примеров найдено не было.

(37) *Вот **сколько** полезного получил от государства Ланкин!* [Анатолий Азольский. Лопушок // «Новый Мир», № 8, 1998]

(38) *Мне тогда вспомнилось наше посещение знаменитых Кунгурских пещер на Урале — вот **где** простор для фантазии неравнодушных людей!* [И. А. Архипова. Музыка жизни (1996)]

(39) *Вот **куда** бы нужно было съездить! К Мариетте!* [Ю. О. Домбровский. Факультет ненужных вещей, часть 5 (1978)]

(40) *Ужасной была обратная поездка в Москву по вызову Ярославского. Вот **когда** я была на волосок от самоубийства!* [Е. С. Гинзбург. Крутой маршрут (1990)]

(41) *Так вот **зачем** он кататься-то звал! он хотел меня за городом-то на тот свет отправить…* [Н. Г. Чернышевский. Что делать? (1863)]

(42) *Так вот **почему** у кофе был такой странный вкус! Ты отравила меня, подлая!* [В. Н. Комаров. Тайны пространства и времени (1995–2000)]

Заметим также, что данная конструкция имеет, помимо анафорического, и дейктическое значение, т. е. может отсылать к ситуации, имеющей место в действительности:

(43) *В кабинете, над диваном, висел портрет хозяина в уланском мундире, писанный масляными красками. — А, вот **где** ты, обезьяна бесхвостая!* [И. С. Тургенев. Конец Чертопханова (1872)]

Таким образом, можно справедливо утверждать, что в русском языке, во-первых, некоторые вопросительные местоимения (*какой, как, сколько*) имеют независимое употребление в качестве эксклamативов; во-вторых, есть группа местоимений (*что, кто, куда, где*), эксклamативное употребление которых бывает только «анафорическим»; и, в-третьих, существует группа вопросительных местоимений (*когда, зачем, почему*), которые имеют эксклamативное употребление только в специальных эксклamативных конструкциях (например, с указательной частицей *вот*).

3. Междометия в эксклamативных клаузах

Как было уже сказано во Введении, эксклamативы выражают эмоциональную реакцию (а именно удивление) говорящего на то действие или состояние, которое имеет место в действительности и которое нарушает его ожидания. Такая эмоциональная установка эксплицитно выражается с помощью различного рода междометий, что, согласно Michaelis 2001, является одним из широко распространенных признаков эксклamативов в языках мира. Характерно это

и для русского языка. В НКРЯ засвидетельствованы многие междометия в экскламативных клаузах:

(44) *О-о,* ***какая*** *могла быть публикация,* ***какой*** *реферат!* [Юрий Давыдов. Синие тюльпаны (1988–1989)]

(45) *Э-э-эх,* ***как*** *кипело вагонное депо!* [Виктор Астафьев. Обертон (1995–1996)]

(46) *Ах,* ***сколько*** *воспоминаний окружило мою бедную голову!* [Вениамин Смехов. Театр моей памяти (2001)]

(47) *Ой,* ***кто*** *только не прошёл тогда перед судом!* [Ю. О. Домбровский. Факультет ненужных вещей, часть 2 (1978)]

(48) *Господи, о* ***чем*** *только мы не спорим!* [Юлий Даниэль. Письма из заключения (1966–1970)]

(49) *Ох, и* ***зачем*** *только она вспомнила своего Илью Муромца!* [Елена Ильина. Четвертая высота (1945)]

4. Синтаксическая зависимость эксклamативных клауз

Мы уже упомянули во Введении, что эксклamативные клаузы могут оформляться синтаксически зависимыми структурами, которые, однако, употребляются как независимые предложения (см. примеры (2)–(4)). Проблема, получившая наиболее сильный резонанс в литературе, обсуждающей типологические данные, заключается в том, могут ли эксклamативные клаузы быть сентенциальными актантами фактивных предикатов.

Согласно Portner and Zanuttini 2003, в английском эксклamативные клаузы могут быть сентенциальными актантами фактивного предиката *know* ‘знать’ (49а) и эмотивных конструкций (например, *it is amazing* ‘удивительно’ в (49б)). Однако они не могут выступать в качестве сентенциальных актантов, во-первых, предикатов *think* ‘думать’, *wonder* ‘интересоваться’, *regret* ‘сожалеть’ (49а), а во-вторых, предикатов под отрицанием в форме настоящего времени и с подлежащим в форме первого лица (49в):

(49) английский
- а. *Mary knows/* thinks/* wonders/ *regrets how very cute he is.*
  ‘Мэри знает / думает / интересуется/сожалеет, как он мил.’
- б. *It is amazing how very cute he is.*
  ‘Удивительно, как он мил!’
- в. **I don’t know/realize how very cute he is.*
  ‘Я не знаю/понимаю, как он мил.’

Японский похож на английский тем, что эксклamативные клаузы могут выступать в качестве сентенциальных актантов эмотивных предикатов со значениями ‘удивляться’ и ‘изумляться’ (50б), и отличается от английского тем, что глагол мыслительной деятельности со значением ‘знать’ (а также ‘помнить’) не грамматичен в подобного рода контекстах (50б), а глагол со значением ‘думать’, напротив, грамматичен (50а).

(50) японский

а. *John-wa Mary-ga nante takusan-no hon-o yonda noda-roo-ka-to omotteiru*.
Джон-Top Мери-Nom Exc много-Gen книга-Acc читать
Foc-Mood-Q-Comp думать.быть
‘Джон думает, сколько книг Мери прочитала.’

б. *John-wa Mary-ga nante takusan-no gakusee-ni okotta noda-roo-to odoroiteiru/**sitteiru.
Джон-Top Мери-Nom Exc много-Gen студент-Dat сердитый Foc-Mood-Comp удивляться.быть знать.быть
‘Джон удивился/*знает, на скольких студентов Мери сердится.’

В каталанском эксклamативные клаузы могут быть сентенциальными актантами перцептивных глаголов в форме повелительного наклонения (51), но не эмотивных глаголов (52):

(51)

Mira *quin* *home* *tan graciós* *que surt* *per* *la* *tele!*
смотреть-Imp какой мужчина так забавный что выходить-3Sg Prep Det телевизор
‘Посмотри, какой забавный человек выступает по телевизору!’

(52)

??*És* *increïble* *que alt* *que és!*
быть.Pres.3Sg невероятно что высокий что быть.Pres.3Sg
‘Невероятно, какой он высокий!’

В русском языке, как справедливо отмечается в Падучева 1996, «восклицательные предложения [к которым относятся в том числе и эксламативные предложения — *Н.З.*] синтаксически неподчинимы как грамматический класс». Исключение составляет то, что Е. В. Падучева называет «конструкцией с косвенным восклицанием» (там же):

(53) а. ***Какой** он осторожный!*
б. *Это удивительно, **какой** он осторожный!*

Другими словами, пример (53б) демонстрирует грамматичность эксклamативной клаузы в качестве сентенциального актанта эмотивного предиката. В НКРЯ были найдены примеры употребления перцептивных глаголов (например, *посмотреть, слышать, видеть*) в формах повелительного (54)

и сослагательного наклонения (55)–(56); некоторых глаголов мыслительной деятельности (*представлять, вспомнить, понять, подумать*) в формах повелительного (57), изъявительного наклонения настоящего времени первого и второго лица (58)–(59), прошедшего времени с местоимением первого лица (60) и инфинитива (61); примеры употребления глагола *знать* в формах сослагательного наклонения (62) и глаголов речи (например, *рассказать*) в форме инфинитива (63) и повелительного наклонения (64).

(54) *Посмотри,* ***какая*** *красавица!* [Сати Спивакова. Не всё (2002)]

(55) *А слышали бы они,* ***какая*** *складная, красивая, чудесная речь звучит у него внутри!* [И. Грекова. Фазан (1984)]

(56) *Вы бы видели,* ***как*** *наша Аня работала!* [Алексей Варламов. Купавна // Новый Мир, № 11–12, 2000]

(57) *Вспомните,* ***какие*** *замечательные, действительно интеллектуальные издания выходили миллионными тиражами!* [Владимир Плотников. СМИ без цензуры — диктатура халтуры (2003) // «Советская Россия», 2003.08.19]

(58) *Представляешь,* ***какая*** *чушь!* [Сати Спивакова. Не всё (2002)]

(59) *Представляю,* ***как*** *замирал зал от неожиданности!* [Григорий Горин. Иронические мемуары (1990–1998)]

(60) *И вот тут я понял,* ***как*** *же это прекрасно — новая хорошая машина!* [Андрей Колесников. Бублики Мондео (2002) // «Автопилот», 2002.01.15]

(61) *А* ***каких*** *гигантских успехов достигает наука за двадцать лет, страшно подумать!* [Юрий Трифонов. Предварительные итоги (1970)]

(62) *И* ***как*** *я теперь, Наташка моя, счастлива, если бы ты знала!* [Невеста (2000) // «Туризм и образование», 2000.06.15]

(63) *А вот рассказать бы ей,* ***как*** *мне самому бывает боязно!* [Андрей Волос. Недвижимость (2000) // Новый Мир, № 1–2, 2001]

(64) *Расскажи,* ***какой*** *у тебя хороший обычай тут заведен!* [Терский берег (1895–1896)]

Конечно, приведенные примеры не покрывают все возможные употребления эксклamативных клауз в качестве сентенциальных актантов. Однако можно сделать следующий вывод: фазовые предикаты (*начать, закончить* и др.), предикаты манипуляции (*просить, приказать* и др.), предикаты желания

(*хотеть, надеяться* и др.), эмотивные глаголы (например, *радоваться, сожалеть*) не употребляются с экскламативными конструкциями[16]. Перцептивные глаголы грамматичны в экскламативных контекстах в форме повелительного, сослагательного наклонения и инфинитива. Что касается глаголов мыслительной деятельности и глаголов речи, видимо, каждый из них проявляет свои индивидуальные свойства. Достаточно попарно сопоставить, к примеру, *представлять* и *знать*, *рассказать* и *утверждать*. Наконец, формы изъявительного наклонения неграмматичны с большинством глаголов, за исключением некоторых предикатов (например, *понять*).

5. Заключение

В данной работе мы рассмотрели примеры употребления в НКРЯ русских вопросительных местоимений в качестве экскламативов и в экскламативных конструкциях. Мы попытались показать, что вопросительные местоимения в функции такого рода представляют собой разнородный класс. Первую группу составляют местоимения *какой, как* и *сколько*, встречающиеся во всех экскламативных контекстах (в независимом употреблении, в анафорической функции, в конструкциях с частицами *только* и *вот*). Вторую группу формируют местоимения *кто, что, куда* и *где*, грамматичные в экскламативных контекстах в анафорической функции и в конструкциях с частицами *только* и *вот*. Третью группу образует местоимение *когда,* употребляемое в конструкциях с частицами *только* и *вот*. Наконец, *зачем* и *почему* представляют собой четвертую группу и употребляются в конструкции с частицей *вот*. Таким образом, со смелостью можно утверждать, что все вопросительные местоимения грамматичны в тех или иных экскламативных контекстах. Если же говорить в терминах Дж. Ретт, ограничение по степени объясняет независимое употребление в качестве экскламативов первой группы местоимений (*как, какой, сколько*) и неграмматичность независимого употребления остальных местоимений. Однако остается непонятным, например, почему в русском возможны особые употребления местоимений других групп или почему в итальянском возможны употребления местоимений со значениями ‘где’, ‘когда’, ‘кто’. Каковы эти употребления — независимые или, может быть, «анафорические»?

Что касается употребления экскламативных клауз, содержащих вопросительные местоимения в качестве сентенциальных актантов, то можно наблюдать вариативность между и внутри различных классов матричных предикатов. Имеются два вида континуумов. На одном полюсе первого, лексико-семантического, континуума находятся перцептивные глаголы, каждый из которых допускает употребление экскламативных клауз, на другом — предикаты речи и мыслительной деятельности, среди которых имеются глаголы, встречающиеся в контекстах экскламативных клауз, и глаголы, которые в них не употребляются. Второй континуум является грамматическим: контексты повелительного, сослагательного наклонения и инфинитива допускают употребления экскламативных

[16] По всей видимости, это типологически верно.

конструкций, в то время как контексты изъявительного наклонения, как правило, их блокируют. К сожалению, какие бы то ни было типологические обобщения делать еще рано, но, по всей видимости, эмотивных предикаты и перцептивные глаголы в форме повелительного наклонения благоприятствуют употреблению эксклamативных клауз в качестве сентенциальных актантов.

Список сокращений

1 — первое лицо, 3 — третье лицо, Abs — Абсолютив, Acc — Аккузатив, Adv — Адверсатив, Aux — вспомогательный глагол, Comp — подчинительный союз, Cop — глагол-связка, Dat — Датив, Det — определенный/неопределенный артикль, Exc — эксклamатив, Foc — фокус, Fin — показатель финитности, Imp — повелительное наклонение, Mood — показатель наклонения, n — показатель нейтрального класса, Nml — номинализация, Nom — Номинатив, Pass. Part — страдательное причастие прошедшего времени, Pl — мн. число, Poss — поссессив, Prep — предлог, Pst — прошедшее время, Q — вопросительная частица, Res — результатив, Sg — ед. число, Top — топик.

References

1. *Buscha A.* 1976. Isolierte Nebensätze im Dialogischen Text. Deutsch als Fremdsprache, 13 : 274–279.
2. *Kalinina E.* Exclamative Clauses in the Languages of the North Caucasus and the Problem of Finiteness. Tense, Aspect, Modality and Finiteness in Daghestanian Languages. Diversitas Linguarum, Sprachtypologie und Universalienforschung.
3. *Kuznetsova E. S.* 2009. The Syntax of Construction ‘CHTO ZA KH’ in Russian Language [Sintaksis Konstsruktsii ‘CHTO ZA KH’ v Russkom Iazyke].
4. *Michaelis L., Lambrecht K.* 1996. The Exclamative Sentence Type in English. Conceptual Structure, Discourse and Language : 375–389.
5. *Michaelis L.* 2001. Exclamative Constructions. Language Typology and Language Universals. An International Handbook : 1038–1050.
6. *Ono H.* 2002. Exclamatory Sentences in Japanese: A Preliminary Study. The proceedings of the third Tokyo Conference on Psycholinguistics : 305–326.
7. *Ono H.* 2006. An Investigation of Exclamatives in English and Japanese: Syntax and Sentence Processing.
8. *Paducheva E. V.* 1996. Semantics Researches [Semanticheskie Issledovaniia] : 308.
9. *Peshkovskii A. M.* 2001. Russian Syntax in Scientific Reporting [Russkoi Sintaksis v Nauchnom Osveshchenii].
10. *Podlesskaia V. I.* 2007. Polysemanticism of the Construction ‘CHTO plus ZA plus NOUN GROUP’ in the light of National Corpus Data [Mnogoznachnost’ Konstruktsii ‘CHTO plius ZA plius NOUN GROUP’ v svete Dannykh Natsional’nogo Korpusa Russkogo Iazyka]. Komp’iuternaia Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii “Dialog 2007” (Computational

Linguistics and Intelligent Technologies: Proceedings of the International Conference “Dialog 2007”).
11. *Portner P., Zanuttini R.* 2000. The Force of Negation in wh-Exclamatives and Interrogatives. Studies in Negation and Polarity: Syntactic and Semantic Perspectives : 201–239.
12. *Portner P., Zanuttini R.* 2008. Nominal Exclamatives in English. Ellipsis and Nonsentential Speech.
13. *Rakhilina E. V.* 1990. Semantics or Syntax? (Answering wh-questions in Russian). Slawistische Beiträge, 268.
14. *Rett J.* 2008. Degree Modification in Natural Language.

ТИПЫ СКРЭМБЛИНГА В СЛАВЯНСКИХ ЯЗЫКАХ

А. В. Циммерлинг (meinmat@yahoo.com)

РГГУ, Москва, Россия

Статья содержит классификацию разновидностей скрэмблинга т. е. отношения нефиксированного порядка категорий предложения в славянских языках. Принимается точка зрения о том, что все виды скрэмблинга возникают в результате факультативных синтаксических перемещений. Для систем порядка слов и формальных грамматик, предназначенных для распознавания структур со скрэмблингом, релевантны как свойства конечных областей перемещения, так и свойства исходных областей. В славянских языках представлены все четыре возможных типа скрэмблинга полноударных элементов предложения. Впервые выделены диагностические признаки двух типов скрэмблинга клитик.

Ключевые слова: скрэмблинг, классификация, категория предложение, клитика.

SCRAMBLING TYPES IN THE SLAVIC LANGUAGES[1]

A. V. Zimmerling (meinmat@yahoo.com)

Russian State University for the Humanities, Moscow, Russian Federation

The paper discusses the types of scrambling in the Slavic languages and in Universal Grammar. It is argued that all kinds of scrambling may be explained as instances of optional movement. Scrambling types are classifi ed on the basis of fi nal and initial movement domains in the clausal complex where sentence categories move. Slavic languages have all four theoretically possible scrambling types of non-clitic elements and both types available for clitic elements. The diagnostic features of clitic scrambling are described for the fi rst time.

Key words: scrambling classification, sentence category, clitic.

[1] This paper was supported by the project RGNF 11-04-00282а «Типология морфосинтаксических параметров» (“Typology of morphosyntactic parameters”). I am grateful to the anonymous reviewer of ‘Dialogue 2011’ for the criticism and valuable suggestions. The responsibility for the final formulations is my own.

1. Basic terminology and framework

In this section I am giving working definitions of the basic terms and specifying the framework of my paper. The term 'scrambling' is used a characteristics of languages generating well-formed sentences which can be linearized in two or more ways. Linearization is defined as an operation preserving syntactic structure i. e. a hierarchy of syntactic positions but changing the linear order of sentence categories manifested by spelled-out (non-zero) elements. The scrambling condition is defined in a scrambling language L_{Sc} for any two sentence categories x and y if their relative order may be inverted in the linear variants of the same sentence structure with a fixed number of positions and a fixed number of non-zero categories filling these positions: *x…y ~ y….x*. Linguistically interesting cases pertain to scrambling of sentence categories of the same type and/or the same phrase level: a) scrambling of verbal arguments; b) scrambling of adjuncts; c) scrambling of modifiers; d) scrambling of verbal heads; e) scrambling of phrasal constituents. In this paper I mostly discuss argument scrambling: the term 'argument' below is used both for internal arguments ('objects' of traditional grammar) and external arguments ('subjects' of traditional grammar).

The framework of this paper is the theory of formal grammars and its applications to natural language processing; I am particularly interested in formal grammars capable of generating languages with partly unordered sentence trees, cf. [Stabler 1997], [Michaelis & Gärtner 2007], [Rambow 1994]. The mode of representation of sentence trees as dependency trees vs constituency trees does not affect generative capacity of a language and does not play a crucial role for my argumentation; however, in this paper I opt for a constituency notation. Natural languages and their word order systems are treated in this paper mainly as instantiation of formal languages and their grammars: the data from natural languages are considered relevant for checking and revising formal grammars and parsing procedures. A word order system is defined as a set of language-specific constraints on word order or as set of type-specific word order constraints shared by similar languages. I am assuming that meta-linguistic knowledge about well-formed and ill-formed expressions can always be retrieved and am adopting the criterion of intuitive adequacy. The judgments on well-formedness or ill-formedness of the test sentences are based on normative grammars, representative descriptions and opinions of the native speakers.

Formal grammars capable of generating scrambling languages may be either context-sensitive or tree-adjoining/mildly context-sensitive. Stablerian Minimalist Grammars [Stabler 1997], [Michaelis & Gärtner 2007] and Chomsky's Minimalist Program [Chomsky 2005] pattern with the last class. In the Minimalist-type grammars scrambling may be licensed due to two reasons: a) the pair of sentence categories *…x, y… ~ …y, x…* remains unordered if the grammar has a special *scrambling operator,* so neither order results from a reordering mechanism; b) the order *…y, x..* is derived from the order *…x, y…* by a unidirectional mechanism called *movement*. In section 2 I am briefly discussing the pros and contras of the movement vs non-movement approaches to scrambling and adopting the movement approach. I am assuming that the direction of movement can

be established in all pairs *…x, y…* → *…y, x…* and that each instance of movement has some functional motivation. At the same, all kinds of unverifiable stipulations concerning the amount of movement and scrambling patterns licensed on the level of Universal Grammar (UG) are rejected. I am assuming that word order systems of natural languages do not violate UG but the proportion of language-specific and universal features is irrelevant for my analysis. Furthermore, I am not aiming at describing cross-linguistic variation or singling out language types in this paper: the data from Slavic languages are used merely as an illustration of formal models represented in natural languages and a motivation for revising these models.

2. Free word order, scrambling and movement

The term 'free word order' is metaphoric since all world's languages are restrictive: no language seems to allow for all possible linear orders or sentence categories in 100 % of sentences and it is reasonable to think that linearization constraints are salient for all word order systems. Meanwhile, there is a general agreement that free word order is a condition when sentence categories may be linearized in two or more different ways, at least in some well-formed sentences of a given language. This condition is known as scrambling of predicate arguments and/or other sentence categories. The term 'scrambling' is sometimes used just as a synonym for 'free word order' but may also convey a more formal meaning and be linked with hypotheses on mechanisms triggering free word order. It has become customary to classify natural languages into a class of languages with a fixed order of lexical sentence categories and a class of scrambling languages. For instance, an English sentence like *Pete ate a tomato* does not have a linear variant **A tomato ate Pete*, since this language blocks for OVS orders[2]. The class of scrambling languages can be defined in a twofold way — either as a) languages displaying a number of diagnostic movement patterns responsible for the alternations like SVO > VSO, SVO > OSV, SVO > OVS, SVO > SOV; or b) languages completely lacking any fixed order of diagnostic sentence categories, say S and O or S, O and V, cf. [Kosta 2006]. Both approaches to scrambling share the assumption that the same numeration, i. e. tree structure with a given number of nodes filled by identical elements, may be linearized differently.

A movement approach to scrambling languages capitalizes the idea that there is a unidirectional relation between different linear variants of the same numeration, one of the variants being the source of the other (s), cf. the presumably base-generated order in Rus. […] *Петя съел помидор* and the derived order [*Помидор*$_i$

[2] A sentence like *A tomato ate Pete* will be proven well-formed if we assume that carnivorous vegetables exist but again the sentence *A tomato ate Pete* won't get a linear variant *Pete ate a tomato* used in the same bizarre meaning "A human has been eaten by a vegetable". Consequently, the ungrammaticality of the SVO > OVS alternation in English does not depend on ontological assumptions about carnivorous vegetables and human vegetarians.

] *Петя съел* t_i: the symbol *t* marks the initial placement of the moved category before the reordering, and the brackets […] mark the target position of the movement. A non-movement approach to scrambling denies the idea of a fixed order of sentence categories in a scrambling language and treats all linear variants as representing the same level of derivation, cf. Rus. *Петя съел помидор* (SVO) ~ *Петя помидор съел* (SOV) ~ *Помидор Петя съел* (OSV) ~ *Помидор съел Петя* (OVS) ~ *Съел Петя помидор* (*VSO*) ~ *Съел помидор Петя* (VOS). The domain where categories scramble may be called scrambling domain. In the standard case illustrated by the Russian examples above, argument scrambling is bounded with a single clause, while all scrambled arguments S, O..U..W belong to one and the same verbal head v°:

(i) Local Scrambling: $[_S\{_{\text{SCRAMBLING DOMAIN}}$ …S…v°…O…}].

Scrambling of the type (i) is called 'local' or 'bounded'; it does not pose big problems for linguistic theory with either non-movement or movement analysis, since all positions available for a scrambled category are located in one and the same domain. Meanwhile, there is undeniable evidence that world's languages have unbounded scrambling, where the permuting arguments may belong to different verbal heads v^1, v^2.. v^n. This has been proven in [Rambow 1994] for Modern German, where unbounded argument scrambling takes place in complement clauses (CPs) in the domain between the complementizer (Comp) and the verbal complex, cf. (ii). Note that the verbal heads themselves are placed in German in a rigid order, so that the scrambling domain is smaller than the complement clause:

(ii) Unbounded Scrambling in German:
Ger. $[_{CP}$ Comp $\{_{\text{SCRAMBLING DOMAIN}}$ A^1 +B^2 + C^3} $[_{VP}$ [v^3, [v^2, [v^1]]] AUX].

Many formal grammars and semi-formal models of language representation including Chomsky's Minimalist Program [Chomsky 1993], [Chomsky 2005] and Stablerian Minimalist Grammars [Stabler 1997] generate ordered trees. Grammars of this type are mildly context-sensitive [Michaelis & Gärtner 2007] and can be adjusted for parsing scrambling languages: in this case their formalism must be extended by a special Scrambling operator in addition to standard Merge and Move operators responsible for merging and moving of sub-trees [Perekrestenko 2008][3]. At first glance this technical detail speaks in support for a non-movement analysis of scrambling, at least in a generative framework sharing the basic assumptions of the Minimalist Program. However, a reasonable linguistic interpretation of unbounded scrambling in (ii) is only possible under movement analysis: otherwise the question how an element of an already ordered subtree shows up in a higher clause remains unexplained. Since I am aiming at a unified account of all scrambling types, I am adopting movement analysis for all theoretically possible types of scrambling:

[3] However, the parsing problem for Minimalist Grammar extended with Unbounded Scrambling operators remains unresolved as shown in [Perekrestenko 2008].

for the reasons of space I am using a simplified notation of target positions and scrambling domains.

The distinction of local vs unbounded scrambling is consistent and useful both for formal grammars and for data-oriented linguistic research. Under movement analysis, the scrambling type (local vs unbounded) is established in the end positions scrambled elements assume after the movement has taken place, not in their initial positions before the reordering. Unfortunately, there is a different terminological tradition in generative linguistics, where scrambling is frequently understood as a characteristics of the initial domains. For instance, J. Baylin [Baylin 2004] sorts out 'short' scrambling when an element moves to a target position in the same clause, and 'long-distance scrambling' when an element is extracted (raised) into a higher clause. This distinction makes sense only if initial positions of the moved sentence material are relevant: it is clear that the terms are misleading and extraction won't entail scrambling in the final domain if the moved element takes just one position in the higher clause. The puzzle is explained by the fact that in the standard case, under local scrambling, where the scrambled elements remain in the same clause, the initial and the final movement domains match or coincide. This proportion does not hold for other scrambling types and it would be better to reserve the specific term 'scrambling' only for the pair 'local vs unbounded scrambling' and replace it by the general term 'movement' in the pair 'short vs long-distance scrambling'. Unless this is done, the term 'scrambling' remains ambiguous but one may try to tackle the problem from the other side and check which theoretically possible combinations of local vs unbounded scrambling & short vs long-distance scrambling are attested. If such combinations really exist and represent productive scrambling types used by the native speakers, this would confirm that a multidimensional analysis of linear alternations both in terms of final vs initial movement domains is on the right track.

This paper summarizes the data of Slavic languages — a group of languages known for a wide variety of movement patterns, cf. [Ковтунова 1976], [Kosta 2006], [Baylin 2004], [Циммерлинг 2008], [Franks 2009] The analysis has shown that almost all combinations of scrambling types are available for sentence categories represented by non-clitic words, while the number of scrambling types available for clitics is more reduced. Unless the opposite is explicitly stated, the scrambling types attested for non-clitic words are treated to be Pan-Slavic: the general prediction is that other Slavic languages likely have well-formed sentences within the same scrambling type but no prediction that an exact equivalent of a well-formed sentence with scrambling will be equally well-formed in other Slavic languages is made.

2.1. Local short scrambling and local long-distance scrambling of non-clitic elements

Let us agree that *Local* scrambling indicates that permuting elements belong to the same verbal head, *unbounded scrambling* indicates that the permuting elements belong to different verbal heads. With *short* scrambling, the moved element remains

in the same clause. With *long distance* scrambling, the moved element is extracted to a higher domain. The combinations 'Short & Local Scrambling', 'Long-Distance & Local Scrambling', 'Long-Distance & Unbounded Scrambling' are common, the combination 'Short & Unbounded Scrambling' is rare. All cases where an element is extracted out of non-finite clauses (IPs) count as long-distance scrambling, along the same lines as extraction out of finite clauses (TPs). Almost all combinations of Local/Unbounded Scrambling with Short/Long-Distance Scrambling were found. The Scrambling condition was tested on sentences perceived as completely grammatical or acceptable by the native speakers and on authentic examples from extinct languages. A minor part of the test sentences with scrambling proven to be well-formed does not sound quite natural in a oral discourse or are generally avoided in written texts on stylistic reasons. This is not an obstacle for my analysis since my aim was to check syntactic parameters enabling or blocking for scrambling and not to find linear orders that could be used in a maximal number of different contexts. I am assuming that movement of sentence categories triggering the scrambling condition always has some communicative motivation but do not prove this point formally here. The term 'non-clitic sentence category' in the following refers to phrases, not phrasal heads.

Fig. 1. Scrambling of non-clitic elements in the Slavic languages

	A. Local Scrambling	**B. Unbounded scrambling**
1. Short scrambling	+	(+)
2. Long Distance Scrambling	+	+

A1. Short & Local Scrambling.

This option is standard: the moved element is not extracted to a higher clause, no unprojective crossing of constituents arises:

(1) a. Rus. Профессор Иванов посетил нашу лабораторию в июне (S+V+O+Adv$_{Temp}$)
Professor$_{Nom.Sg.M.}$ Ivanov$_{Nom.Sg.M}$ visit$_{3Sg.M.Pst.}$ our$_{Acc.Sg.F.}$ laboratory$_{Acc.Sg.F.}$ in june$_{Loc.Sg.}$
"Professor I. visited our laboratory in June"

b. ⇒ [Нашу лабораторию]$_i$ профессор Иванов посетил t$_i$ в июне (O+V+S+Adv$_{Temp}$).
our$_{Acc.Sg.F.}$ laboratory$_{Acc.Sg.F}$ visit$_{3Sg.M.Pst.}$ Professor$_{Nom.Sg.M.}$ Ivanov$_{Nom.Sg.M}$ in june$_{Loc.Sg.}$
"the same".

A similar relation can be shown for adjuncts, cf. Czech examples in (2).

(2) a. … že Maria profesora$_i$ [v jeho$_i$ bytě] už několikrát navštívila4.
That Maria$_{Nom.Sg.F.}$ professor$_{AccSg.M.}$

[4] The examples in (4) are from [Kosta 2006].

in his$_{\text{Gen.Sg.M.}}$ flat$_{\text{Loc.Sg.}}$ already several.time visit$_{\text{3Sg.F.Pst.}}$
"…that Mary has already several times visited the professor$_i$ [in his$_i$ flat]".

b. ⇒ чеш. … že [v jeho$_i$ bytě] Maria profesora$_i$ t$_i$ už několikrát navštívila.
That in his$_{\text{Gen.Sg.M.}}$ flat$_{\text{Loc.Sg.}}$ Maria$_{\text{Nom.Sg.F.}}$ professor$_{\text{AccSg.M.}}$ already several.time visit$_{\text{3Sg.F.Pst.}}$
"the same".

c. ⇒ чеш. že Maria [v jeho$_i$ bytě] profesora$_i$ t$_i$ už několikrát navštívila.
That Maria$_{\text{Nom.Sg.F.}}$ in his$_{\text{Gen.Sg.M.}}$ flat$_{\text{Loc.Sg.}}$ professor$_{\text{AccSg.M.}}$ already several.time visit$_{\text{3Sg.F.Pst.}}$
"the same".

A2. Long-Distance & Local Scrambling.

The scrambling condition is found in the initial domain but not in the final domain. This is possible if the extracted element has just one target position in the higher domain.

(3) a. Rus. Мы бы хотели, чтобы министерство назначило профессора И. куратором нашей лаборатории
We$_{\text{Nom.Pl}}$.Cond.Pclwant$_{\text{1Pl.Cond}}$thatministry$_{\text{Nom.Sg.N.}}$appoint$_{\text{3Sg.N.Cond.}}$professor$_{\text{Acc.Sg.M.}}$ I. curator$_{\text{Instr.Sg.M.}}$ our$_{\text{Gen.Sg.F.}}$ laboratory$_{\text{Gen.Sg.F.}}$
"We would like that the ministry appointed professor I. curator of our laboratory".

b. ⇒ [[Профессора И.]$_i$, [мы бы хотели, [чтобы министерство назначило t$_i$ куратором нашей лаборатории]]].
Lit. 'Professor I $_i$, we would like [that the ministry appointed t $_i$ curator of our laboratory]'
"the same".

Cf. also Bulgarian example with extraction out an NP containing an embedded relative clause:

(4) a. Bulg. Ще=бъдат [две тоалетните, [като всеки от състезателите ще=може да ползва [която пожелае]].
Fut.Pcl. be$_{\text{3Pl.Fut}}$ two toilet$_{\text{Nom.Pl.Def.}}$ which any from sportsman$_{\text{3Pl.Def.}}$ Fut.Pcl can$_{\text{3Sg.Pres.}}$ Comp use$_{\text{3Sg.Pres.}}$ who want$_{\text{3Sg.Pres.}}$
"There will be two toilet rooms [which can be used by any of the sportsmen [who wants]]".

b. ⇒ [[Тоалетните] $_i$ ще бъдат [две t$_i$, като всеки от състезателите ще=може да ползва [която пожелае]]].
toilet$_{\text{Nom.Pl.Def.}}$ Fut.Pcl. be$_{\text{3Pl.Fut}}$ two which any from sportsman$_{\text{3Pl.Def.}}$ Fut.Pcl

can$_{3Sg.Pres.}$ Comp use$_{3Sg.Pres.}$ who want$_{3Sg.Pres.}$
"the same".

B2. Long-Distance & Unbounded Scrambling.

Sentences with three scrambled NPs A[1], B[2], C[3] linked with three hierarchically arranged verbal heads are rare. Sentences with two scrambled NPs A^m, B^n, linked with two hierarchically arranged verbal heads v^m, v^n are wide-spread. One of the common cases of long-distance unbounded scrambling is triggered by non-projective embedding of a constituent or its element into a higher clause. Let A° B° C° D° E be the basic word order, A° B° C° D ° be lexical heads and each next head be a dependent of the preceding one. It gives a projective structure (5), where blocks DE, CDE, BCDE, ABCDE are embedded constituents:

(5) [A ° [B° [C° [D °E]]]].

(5') Rus. *Арбитры[1] не имели права[1]* [$_{IP}$ *фиксировать[2]* [*победу[2] «Триумфа»*]].
Referee$_{Nom.Pl.}$ not have$_{3Pl.Pst.}$ right$_{Gen.Sg.}$ fix$_{Inf}$ win$_{Acc.Sg.}$ 'Triumph'$_{Gen.Sg}$.
'The referees[1] had no right[1] to fix[2] the win[2] of 'Triumph".

Moving the blocks DE, CDE and embedding the heads A°, B° into lower constituents one can get orders like [CDE]$_i$ A°B° t$_i$, [[DE]$_j$ C° t$_j$]$_i$ A°B° t$_i$, [[DE]$_j$… A°$_k$ … C° t$_j$]$_i$ t$_k$ B° t$_i$, …A°$_k$ …[[DE]$_j$ C° t$_j$]$_i$ t$_k$ B° t$_i$, where t$_{i, j, k}$ — traces of the moved heads or blocks. An illustration is provided in fig. 2.

Fig. 2. Long-Distance Unbounded Scrambling in Russian

	Pattern	**Illustration**
Basic word order	[A ° [B° [C° [D °E]]]]	(6a) рус. *Арбитры[1] не имели права[1]* [$_{IP}$ *фиксировать[2]* [*победу[2] «Триумфа»*]].
Derived orders	[CDE]$_i$ A°B° t$_i$,	(6b) Ю
	[[DE]$_j$ C° t$_j$]$_i$ A°B° t$_i$,	(6c) Ю
	[[DE]$_j$… A°$_k$ … C° t$_j$]$_i$ t$_k$ B° t$_i$	(6d) Ю
	…A°$_k$ …[DE]$_j$ C° t$_j$]$_i$ t$_k$ B° t$_i$	(6e) Ю

2.2. Unbounded Short Scrambling and Unbounded Long-Distance Scrambling of non-clitic elements

B1. Short & Unbounded Scrambling.

If the initial domain does not contain embedded structures, Short Unbounded Argument Scrambling may only arise due non-projective crossing of groups not involved in an immediate dominance relation, cf. (7). Such examples are rare.

(7) $[_{x}$AB$]\ldots[_{y}$ CD$]$ Ю $[_{x}$A $[_{y}$ C$_{x}\ldots$ B$]$ $\ldots{}_{y}$ D$]$.

Sentences with disjoint constituents and embedding are slightly more acceptable than examples with non-projective crossing. Cf. Russian data (8a-c).

(8) a. [$_{x}$Жители° столицы] [любят° [$_{y}$ пивную продукцию° Клина]].
Resident$_{\text{Nom.Pl.}}$ capital$_{\text{Gen.Sg.}}$ love beer$_{\text{AdjAcc.Sg.F.}}$ production$_{\text{Acc.Sg.F.}}$ Klin$_{\text{GenSg.}}$
“The residents of (our) capital love the beer production <from the city of> Klin”

b. $^{??}[_{y}$ Клина$]_{i}$ [$_{x}$жители° столицы] [любят° [$_{y}$ пивную продукцию° t$_{i}$].
Klin$_{\text{GenSg.}}$ resident$_{\text{Nom.Pl.}}$ capital$_{\text{Gen.Sg.}}$ love beer$_{\text{AdjAcc.Sg.F.}}$ production$_{\text{Acc.Sg.F.}}$

c. *$[_{y}$ Клина$]_{i}$ [$_{x}$столицы$]_{j}$ [любят° [$_{y}$пивную продукцию° t$_{i}$] [$_{x}$ жители° t$_{j}$].
Klin$_{\text{GenSg.}}$ capital$_{\text{Gen.Sg.}}$ love beer$_{\text{AdjAcc.Sg.F.}}$ production$_{\text{Acc.Sg.F}}$ resident$_{\text{Nom.Pl..}}$

If one cancels the requirement that the scrambled elements must represent one and the same sentence category or the requirement that they must be hierarchically independent, Short Unbounded Scrambling may be in found in other constructions, especially in constructions with second-position clitics splitting the initial constituent, as in the Old Russian examples (9a-b).

(9) a. Old. Rus. а и-Суждальской {$_{\text{Scrambling}}$ =*ти* (1) **земле** (2)} Новагорода не рядити (ГВНП, №. 1, 1264 г.).
And from Suzhdal$_{\text{Adj.Gen.Sg.F.}}$ you$_{\text{2Dat.Sg.}}$ land$_{\text{Gen.Sg.F.}}$ Novgoroda$_{\text{Gen.Sg.M}}$ not rule$_{\text{Inf}}$
“And from Suzdal’s **land** (2), *you* (1) should not rule Novgorod”.

b. а и-земле (1) {$_{\text{Scrambling}}$ =*ти* (2) суждальской (2)} Новагорода не рядити}.
And from **land**$_{\text{Gen.Sg.F.}}$ *you*$_{\text{2Dat.Sg.}}$ Suzhdal$_{\text{Adj.Gen.Sg.F.}}$ Novgoroda$_{\text{Gen.Sg.M}}$ not rule$_{\text{Inf}}$

c. *а [*и-Суждальской земле*] =*ти*.

The Old Russian pronoun *ти* in (9) is a fixed position pronominal clitic that must be placed after the first stressed word form, cf. the ill-formedness of (9c), while the NP *земле* lacks a fixed position in a clause. But since the optional movement of just one category in the pair (…x, y…) ~ (…y,x…) is a sufficient condition of scrambling and the NP *земле* (x) may end up both the right and to the left from the clitic *ти* (y), nothing prevents from recognizing Short Unbounded Scrambling here. The scrambling domain in (9) is short — it includes only the clitic position and the position of the subsequent non-clitic element — while the clitic and the NP are linked with predicate heads of a different level[5]. One might theorize that clitics do not scramble with

[5] The NP *земле* in (9) is the complement of the PP [$_{\text{pp}}$ *и(з) Суждальской земли*] which is dependant of the infinitival head *рядити* ‘to ordain’, while the dative clitic *ти* ‘you’ is the modal subject and may be viewed as an argument of the (zero) auxiliary head. Note that the infinitive *рядити* in (9) is not the head of an embedded clause but part of the main predicate.

non-clitic elements but this stipulation lacks an independent verification since clitics do scramble with each other which is demonstrated in the next section.

3. Clitic classes and Scrambling

The term 'clitic' has many uses, cf. [Zwicky 1997], [Sadock 1995], [Зализняк 2008, 8], [Циммерлинг 2009]. Let us define [syntactic] clitics as prosodically deficient sentence categories linearized by syntactic mechanisms. In a Chomskyan framework, syntactic clitics may be analyzed either as heads (X°), cf. [Franks 2008], or as the so called left-branching elements i. e. reduced phrases (XP/X°), cf. [Bošković 2002]: the choice of the interpretation in the context of our paper is irrelevant. There are fixed position clitics and floating free clitics. Clitics can also be clusterizing i. e. capable of making up clitic clusters arranged in a rigid order or non-clusterizing i. e. not imposing any restrictions on contact position of two or more clitics. Fixed position clitics that do not move and do not make up clusters are of no interest for scrambling theories. Floating free non-clusterizing clitics scramble in the same way as non-clitic categories. Finally, if one accepts scrambling of clitic and non-clitic arguments in example (9) above, this type of scrambling patterns with scrambling of non-clitic elements: even if the clitic has a fixed position in a clause, as *mu* in (9), its relative placement respective to a non-clitic category still may be different, cf. variants (9a) and (9b).

Clusterizing clitics exhibit non-trivial features. Cross-linguistically, clusterization of clitics always takes place in some canonical syntactic position and may be blocked in other positions[6]. That means that clusterizing clitics are a subclass of fixed position clitics. At the same time, clusterizing clitics move, the whole clitic cluster may shift its location in a clause or be split in certain contexts; that means that some or all clusterizing clitics may occasionally end up outside their canonical position of clusterization [Зализняк 1993], [Циммерлинг 2009]. All Slavic languages except for Modern Russian, Modern Ukrainian and Modern Belorussian have clusterizing clitics [Dimitrova-Vulchanova 1999]. No Slavic language has phrase-level clusterizing clitics (in NPs or other non-predicative phrases[7]), cf. [Ćavar, Wilder 1999], [Циммерлинг 2011].

---

[6] This point is proved formally in [Циммерлинг 2011] and [Kosta, Zimmerling 2011]. The crucial fact is that in many world's languages one and the same clitic may clusterize on the clause level and be non-clusterizing on the phrase level. This is attested in Slavic languages where pronominal dative clitics, cf. Bulg. *ми* 1Dat.Sg. «me» clusterize as verbal arguments but do not clusterize as possessive markers on the DP-level. The same duality is characteristic of Ossetic dative-genetive pronouns: they clusterize only as verbal arguments on the clause level but not as possessive markers on the NP/DP level. This indicates that at least in languages of the Slavic/Ossetic type clusterizing capacity of a clitic is not an inherent lexical feature but a characteristics of the syntactic configuration.

[7] The anonymous reviewer objects that the data of Modern Bulgarian might falsify my formulation. The checking of this claim is linked with the discussion about the so called Possessor Raising out of Bulgarian DPs: [Schürcks & Wunderlich 2004] argue that Bulgarian allows Raising of possessive datives out of DPs, while [Cinque & Krapova 2011] argue against a Raising analysis. Whatever view of Bulgarian DP is taken, the only candidates for the role of a clitic cluster in DPs are combinations of the definite article and the possessive pronoun,

In most cases Slavic languages put clitic clusters/ single clusterizing clitics after the first spelled-out constituent /first phonetic word[8] or after the complementizer: main clauses vs subordinate clauses, finite clauses vs non-finite clause apply the same set of clusterizing clitics. These facts lead to the following generalization:

Slavic clusterizing clitics are clause-level second-position clitics (2P clitics).

The generalization (iii) holds for the following Slavic languages: Serbo-Croatian, Slovene, Czech, Slovak, Burgenland Croatian, Vojvodina Rusinsky, Old Novgorod Russian, Bulgarian. Bulgarian (and Macedonian) word order systems have a constraint on contact realization of clusterizing clitics and verbal forms. It has become customary to divide Slavic word order systems with clusterizing clitics into systems with clause-level 2P clitics and into systems with clause-level Verb-Adjacent clitics, cf. [Franks & King 2000], [Franks 2009]. This practice is justified but no analysis of the Bulgarian word order system can ignore the fact that this language retains a constraint on the number of groups preceding pronominal and auxiliary clitics. Cf. examples with a compound verbal form consisting of an l-participle and a BE-auxiliary in the past tense in (10): the compound form takes one position as shown in (10a) but a combination of a compound form with another constituent before the clitic *я* is excluded in whatever order as shown in (10b) and (10c):

(10) a. Bulg. #[$_{VP}$ Купил бих]=*я* книгата.
[bought$_{PerfPart.Sg.M.}$ Be.Aux$_{1Sg.Cond}$] she$_{Acc.Sg.}$ book-the$_{Acc.Sg.F.Def.}$
'I would rather buy this book', lit. '[bought would$_{1Sg}$] = *it* the book',

b. *[$_{DP}$Книгата] [$_{VP}$ купил бих]=*я*,
book-the$_{Acc.Sg.F.Def.}$ [bought$_{PerfPart.Sg.M.}$ Be.Aux$_{1Sg.Cond.}$] she$_{Acc.Sg.}$

c. *[$_{VP}$ купил бих] [$_{DP}$книгата]=*я*.

This gives a ground to state that the principle of 2P placement is not violated in Bulgarian, whatever the reason may be. Therefore, Bulgarian clusterizing pronouns and auxiliaries should be treated both as 2P clitics and as Verb-Adjacent clitics — cf. the ungrammatical order (10c) where the constraint of on clitic-and-verb adjacency is violated.

---

cf. Bulg. ужасни-те (1) *си* (2) грешки lit. awful$_{Pl.}$ Det.Pl. (1) Refl.Poss (2) mistake$_{Pl}$. (3) 'one's awful mistakes' ~ грешки-те (1) *си* (2) ужасни (3) mistake$_{Pl}$. Det.Pl. (1) Refl.Poss (2) awful$_{Pl.}$ (3) 'the same'. The enclitic definite article is attached to the first stressed word of DP, while the dative possessive pronoun is cliticized to the first element containing a definiteness morpheme. Hence, the Bulg. definite article is merged pre-syntactictally on the morphological level, while Bulg. dative possessives are merged in syntax. Consequently, no clitic cluster arises: [$_{DP}$ [ужасни-*те* (1)] = *си* (2) °грешки] ~ [$_{DP}$ [°грешки-*те* (1)] = *си* (2) ужасни].

[8] The exact formula of the first spelled-out constituent/first phonetic word variation is irrelevant for a scrambling analysis: in [Kosta, Zimmerling 2011], we address this issue in detail. Cf. also a general discussion in [Anderson 1995] and a case study of the 2nd position phenomena in Czech in [Avgustinova, Oliva 1997].

3.1. Clitic clusters and clusterizing clitics

Clitic clusters are by definition contact strings of clitics excluding permutation of elements and insertion of non-clitic words [Зализняк 1993: 289]. That means that if a°, b° and c° are clusterizing clitics and the fixed order of clitics is [$_{\text{Clitic Phrase}}$ a°, b°, c°], no other order like *[$_{\text{Clitic Phrase}}$ b°$_i$, a°, t$_i$ c°], *[$_{\text{Clitic Phrase}}$ c°, a°, b° t$_i$] should be possible in the canonical position of clusterisation. This amounts to saying that clusterizing clitics do not have short scrambling in sentences without cluster splitting. With cluster splitting orders as …X°= [$_{\text{Clitic Phrase}}$ c°] …Y° [$_{\text{Clitic Phrase}}$ a°, b°]… where the clitic c° is placed earlier than clitics a°, b° preceding it in the cluster may arise, if parts of the cluster are attached to different sentence categories. However, such cases are difficult to recognize as scrambling, since the clitic(s) leaving the clusterization position (or not reaching it) almost invariably end up in a position adjacent to a verbal head [Циммерлинг 2011]. I am assuming here that this a special pattern of clitic movement that should be treated separately both from Short Scrambling and from Long-Distance Scrambling. A Clitic Template generating clitic clusters is illustrated by Old Novgorod data in fig. 3 below.

Fig. 3. Old Novgorod Russian clitic template

A					**B**		**C**
Particles					Pronouns		Present tense indicative BE-auxiliary
1	2	3	4	5	6	7	8 = AUX1
Affirm	Quest	Cause	Evid	Opt	Dative 1–2 p. (incl. Dat.Refl)	Accusative 1–3 p. (incl. Acc. Refl)	1–2 p. Sg.Du.Pl.
Že	Li	Bo	Ti	By	Mi, ti, si, ny, vy, na, va	M'a, t'a, s'a, ny, vy, na, va, i, ju, je, ě, ja	Jesm', jesi, jesme, jeste, jesvě, jesta

Cluster splitting is illustrated by the Old Russian example (11) where the alternative particle *ли* which precedes the auxiliary clitic *еси* in the cluster ends up outside the clusterization position (2P) and is attached to the verbal head *слышалъ* '(you) heard'. On reasons specified above I do not treat such cases as scrambling and analyze them in terms of communicatively driven clitic movement: the initial topical PP *а у королева мужа* 'and from the king's man' has effect only on the surface position of *ли* but not on the surface position of *еси*.

(11) O. Russ. [$^{\text{BARRIER}}_{\text{a}}$ {$_{\text{TopicP}}$ **А** [$_{\text{PP}}$ **оу королева**}]=*еси*$^{\text{b}}$ мужа] слышалъ=*ли*$^{\text{a}}$ о томъ чстномъ крстѣ? (Ипат., under 1152 AC, list 166 rev.).
And from king's.$_{\text{GEN.SG.}}$ BE.$_{\text{AUX.PRES.2SG.}}$ man.$_{\text{GEN.SG.}}$ hear.$_{\text{PRF.2SG.M.}}$ Q about that.$_{\text{LOC.SG.M.}}$ worthy.$_{\text{LOC.SG.M.}}$ cross.$_{\text{LOC.SG.M.}}$

'Haven't you heard about that worthy cross from the king's man?'

A puzzling fact is that clusterizing clitics that lack options for short scrambling do allow extraction into a higher clause: the parameter responsible for extraction is known as Clitic Climbing. Most though not all Slavic languages have Clitic Climbing of argument and reflexive pronouns out of embedded non-finite clauses, while the so called Clitic Templates[9] generating clitic clusters have slots for the clitics raised from embedded clauses [Franks & King 2000], [Kosta, Zimmerling 2011]. Clitic Climbing is a prerequisite of Clitic Scrambling but not its sufficient condition. Three different scenarios are possible:

a) If the extraction is obligatory, no scrambling relation arises.
b) If the extraction is optional and the extracted clitic has one and only one available target position in a higher clause, Clitic Climbing leads to a condition resembling or identical with Local Unbounded Scrambling.
c) If the extraction is optional and the extracted clitic has multiple (more than one) target positions in a higher clause, Long-Distance Unbounded Scrambling arises.

Different Slavic languages show all these scenarios, as shown in fig. 4.

Fig. 4. Scrambling of clusterizing clitics in the Slavic languages

	A. Local Scrambling	**B. Unbounded Scrambling**
1. Short Scrambling	–	–
2. Long-Distance Scrambling	Clitic Climbing (+)	Clitic Scrambling +

In Slavic languages only argument and reflexive clitics climb into higher clauses. I am unaware of any examples of auxiliary and particle clitic climbing.

3.2. Clitic Climbing and Optional Movement

Let us examine Clitic Climbing first. In all Slavic languages except for Bulgarian and Macedonian clitic clusters have slots for clitic pronouns syntactically belonging to heads of embedded clauses. That means that e. g. a reflexive clitic dependent on an infinitival head must/may raise to a higher clause if the cluster has a slot for this category of clitics. In (12) clitics a, b, d belong syntactically to the head $v^{1\circ}$, located in the main clause (TP), while the clitic c^2 clusterizes with a^1, b^1, d^1 but belongs syntactically to the head $v^{2\circ}$, located in the embedded infinitival clause (IP)[10].

[9] The term 'Clitic Template' used in the Western tradition, cf. [Anderson 1995], [Franks & King 2000], [Browne 2008], [Kosta, Zimmerling 2011] corresponds to the term 'Ranking Rule' (Rus. *правило рангов клитик*) coined by Andrej Zaliznjak [Зализняк 1993] and adopted in [Циммерлинг 2008a], [Зализняк 2008], [Циммерлинг 2011].

[10] The tag CliticP in (12) indicates that the clitic cluster a^1 b^1 c_i^2 d^1 is a phrase (clitic group). The tag TP (Tense Phrase) stands for a finite clause, its boundaries being marked with brackets.

(12) $[_{TP}\ldots [_{CliticP}\ a^1\ b^1\ c_i^{\,2} d^1\ ]\ v^{1\circ}]\ [_{IP}\ v^{2\circ}\ t_i]\ ]$.

The pattern (11) is illustrated by the Rusinsky example (13), where clitics *=ше* and *=му* belong syntactically to the infinitive *поклонїц*, while the clitic *=би* belongs to the head of the main clause, the verb *пошол*.

(13) Rusin. *же=би*^{1}*=*{*=**ше**^{2}**=му**^{2}*}$_i$ и я пошол$^{\circ 1}$ [$_{IP}$ поклонїц$^{\circ 2}$ t$_i$]. (Mat. 2.8).[11]
That Cond.Pcl Refl.Pcl. him$_{3Dat.Sg.M.}$ and I$_{1Nom.Sg.}$ go$_{1Sg.Pst.}$ bow$_{Inf.}$
Lit. 'that=*Pcl*^{1}=*{**=REFL**^{2}**=to-him**^{2}*}* and I went$^{\circ 1}$ to bow.low.$^{\circ 2}$'.

The structure (12) conforms to the definition of Local Long-Distance Scrambling: scrambling in the initial domain, no scrambling in the final movement domain. But since Clitic Climbing is obligatory in Rusinsky, the example (13) does not exhibit scrambling. The linear variants (where the clitics do not climb (14a) or do not reach the clusterization position in the main clause (14b) are ill-formed.

(14) a. Rusin. **же=**би**$_i$ и я пошол поклонїц=**ше=му.***

b. Rusin. **же=би и я пошол=**ше=му** поклонїц.*

Clitic Climbing is obligatory in the Croatian variety of Serbo-Croatian [Ćavar, Wilder 1999: 447] and in most other literary Slavic languages. Nevertheless, Slavic idioms with optional Clitic Climbing exist. Zaliznjak [Зализняк 1993: 295–296] discusses Old Novgorod Russian usage of the XIV-XV centuries, where the reflexive clitic *ся* normally did not climb. Sentences with the climbing of *ся* are however attested, cf. the authentic example (15a). The standard option is shown in (15b).

(15) a. Old Novg. а холоп и роба не оучноу$^{\tau}$ = *ся*$_i$ [$_{IP}$ тяга$^{\tau}$ t$_i$] (a XV century copy from a 1396 letter)[12].
And servant$_{NomSg.M.}$ and bondmaid$_{Nom.Sg.F}$. not start$_{3Pl.Pres.}$ Refl. litigate$_{Inf.}$
«And (if) a servant and a bondmaid do not start litigating ».

b. а холоп и роба не оучноут [$_{IP}$ тягать=*ся*].
And servant$_{NomSg.M.}$ and bondmaid$_{Nom.Sg.F}$. not start$_{3Pl.Pres.}$ litigate$_{Inf.}$ Refl.

The tag IP stands for a non-finite clause headed by an infinitive or participle. The finite verbal head of TP is marked in (12) as v^1, the non-finite verbal head of IP is marked as v^2. The uppercase indexes $a^1\ b^1\ c_i^{\,2} d^1$ indicate to which of the two verbal heads— v^1 or v^2 — each clitic belong. The lowercase index $c_i^{\,2}$ indicates that the clitic c syntactically belonging to the head v^2, has been raised into the main clause by Clitic Climbing. The symbol t_i marks the initial placement of this clitic before Clitic Climbing took place.

[11] The examples are from [Browne 2008].

[12] The example is from [Зализняк 1993: 296].

3.3. Long-Distance Unbounded Clitic Scrambling

This type of Clitic Scrambling requires a combination of two non-trivial parametric settings — 1) Clitic Climbing should be optional, not obligatory; 2) clusterizing clitics extracted from an embedded clause should have more than one target position in a higher domain. Previous accounts of Clitic Climbing took for granted that this combination is excluded and Clitic Scrambling was ignored, but F. Marušić [Marušić 2007] found it in Modern Slovene. According to him, each verbal head mediating between the main clause verb and the head of the embedded infinitival clause may attract the extacted clitics in Slovene. In (16a–f) it is the pronominal clitic *=jo* «her».

(16) a.Slov. [$_S$ { $_{SCRAMBLING}$ On=***jo***^{2} $_i$ **=*je***^{1} hotel$^{1°}$ [$_{IP}$ nehati° [$_{IP}$ hoteti° [$_{IP}$ videvati° t $_i$ vsak dan]]]}].
He$_{3Nom.Sg.M}$ her$_{3Acc.Sg.F}$. BE.$_{Aux.3Sg.Pres}$. want$_{3Sg.Pst.}$ not.want$_{Inf.}$ want$_{Inf.}$ see$_{Inf.}$ every day
"He wanted to stop wanting to see her every day".
Lit. 'he=**her=BE.AUX** wanted to stop to want to see her every day'.

b. [$_S$ { $_{SCRAMBLING}$ On=**je**^{1}#=**jo**^{2} hotel° [$_{IP}$ nehati° [$_{IP}$ hoteti° [$_{IP}$ videvati° t $_i$ vsak dan]]]}].

c. [$_S$ { $_{SCRAMBLING}$ On=**je**^{1} hotel°# **=jo**^{2} [$_{IP}$ nehati° [$_{IP}$ hoteti° [$_{IP}$ videvati° t $_i$ vsak dan]]] }].

d. [$_S$ { $_{SCRAMBLING}$ On=**je**^{1} hotel° [$_{IP}$ nehati#**=jo**^{2} [$_{IP}$ hoteti° [$_{IP}$ videvati° t $_i$ vsak dan]]] }].

e. [$_S$ { $_{SCRAMBLING}$ On=**je**^{1} hotel° [$_{IP}$ nehati° [$_{IP}$ hoteti°#**=jo**^{2} [$_{IP}$ videvati° t $_i$ vsak dan]]] }].

f.[$_S$ { $_{SCRAMBLING}$ On**=je**^{1} hotel° [$_{IP}$ nehati° [$_{IP}$ hoteti° [$_{IP}$ videvati°=**jo**^{2} vsak dan]]] }].

Marušić himself does not use the term 'Scrambling' for the examples (16a-f) but his Slovene data clearly demonstrate Long-Distance Unbounded Clitic Scrambling: the clusterizing clitics in (16a-f) initially belong to different verbal heads but scramble in the final domain i. e. S. Other Slavic languages lack Long-Distance Unbounded Clitic Scrambling. Slovene data prove that it is a possible but not typical linearization strategy for clusterizing clitics, while the same scrambling type is more common for Slavic non-clitic elements.

4. Conclusion

The account of a scrambling theory outlined here demonstrates that scrambling in pairs of sentence categories (x, y) may be effectively triggered by optional movement of one of these categories. Two pairs of parameters — local/unbounded scrambling and short/long-distance scrambling give rise to four scrambling types all of which are attested in Slavic languages. Local vs Unbounded Scrambling are opposed by the final

movement domains, Short vs Long-Distance Scrambling — by the initial movement domains. The combination of Short and Long-Distance Scrambling is rare but theoretically not excluded since the final movement domain with Long-Distance Scrambling may be smaller than a single clause. Clusterizing clitics have more reduced scrambling possibilities than non-clitic sentence categories. They do not have Short Scrambling but may under certain conditions have Long-Distance Scrambling. The movement domains for elements of this class must be checked in positions where the raised clitics clusterize with other clitics, not in positions where they are base-generated. The movement pattern known as Clitic Climbing requires or allows for a clitic generated in an embedded clause to raise and reach its canonical position in a higher clause. If the raised clitic has exactly one position in a higher clause, Local Long-Distance Scrambling arises. If the raised clitic has two or more available positions in a higher clause / clauses, Unbounded Long-Distance Scrambling arises.

References

1. *Anderson Stephen P.* 1995. Toward an Optimal Account of Second-Position Phenomena. Optimality Theory: Phonology, Syntax, and Acquisition : 302–333.
2. *Avgustinova T., Karel O.* 1997. On the Nature of the Wackernagel Position in Czech. Formale Slavistik : 25–47.
3. *Bailyn J. F.* 2004. Generalized Inversion. Natural Language and Linguistic Theory, 22 : 1–49.
4. *Bošković Željko.* 2002. Clitics as Nonbranching Elements and the Linear Correspondence Axiom. Linguistic Inquiry, 33 (2) : 329–40.
5. *Browne Wayles.* 2008. Porjadok Klitikox u Vojvodjaskim Rusinskim. Shvetlosts, 3 : 351–362.
6. *Chomsky Noam.* 1993. A Minimalist Program for Linguistic Theory. The View from Building, 20.
7. *Chomsky Noam.* 2005. Three Factors in Language Design. Linguistic Inquiry, 36 : 1–22.
8. *Ćavar Damir, Chris Wilder.* 1999. Clitic Third in Croatian. Clitics in the languages of Europe (Eurotype 20–5) : 429–467.
9. *Dimitrova-Vulchanova Mila.* 1999. Clitics in the Slavic languages. Clitics in the languages of Europe (Eurotype 20–5) : 83–121.
10. *Franks Steven.* 2008. Clitic Placement, Prosody and the Bulgarian Verbal Complex. Journal of Slavic linguistics, 16 (1): 91–137.
11. *Franks Steven.* 2009. Clitics in Slavic. The Slavic Languages. An International Handbook of their Structure, their History and their Investigation, I : 725–738.
12. *Franks Steven, Tracy King.* 2000. A Handbook of Slavic Clitics.
13. *Gärtner Hans Martin, Jens Michaelis.* 2007. Some Remarks on the Locality Conditions and Minimalist Grammars. Interfaces + Recursion = Language? Chomsky's Minimalism and the View from Syntax and Semantics : 162–195.
14. *Kosta Peter.* 2006. On Free Word Order Phenomena in Czech as Compared to German. Zeitschrift fuer Slawistik, 51 (3) : 306–320.

15. *Kosta Peter, Zimmerling Anton*. 2011. Slavic Clitic Systems in a Typological Perspective. Studies in Generative Grammar (SGG): The syntax of DP/NP in Slavic.
16. *Kovtunova I. I.* 1976. Modern Russian Language. Word Order and Actual Sentence Segmentation [Sovremennyi Russkii Iazyk. Poriadok Slov I Aktual'noe Chlenenie Predlozheniia].
17. *Marušic Frank*. 2007. Positioning Slovenian clitics. SLS 2 conference.
18. *Perekrestenko Aleksander*. 2008. Minimalist Grammars with Unbounded Scrambling and Non-discriminating Barriers are NP-hard. LATA 08 conference.
19. *Progovac Liliana*. 1996. Clitics in Serbian/Croatian: Comp as the second position. Approaching second: Second position clitics and related phenomena : 411–28.
20. *Rambow Owen*. 1994. Formal and Computational Aspects of Natural Language Syntax.
21. *Sadock Jerrold M*. 1995. A Multi-hierarchy View of Clitics. CLS 31 (2) : 258–79.
22. *Stabler Edward P.* 1997. Derivational Minimalism. Logical Aspects of Computational Linguistics : 68–95.
23. *Zimmerling A. V.* 2008. Local and Global Rules in Syntax [Lokal'nye I Global'nye Pravila v Sintaksise]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2008" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2008"), 7 (14) : 551–562.
24. *Zalizniak A. A.* 1993. On Writings on Birch Bark Language Studues [K Izucheniiu Iazyka Berestianykh Gramot]. Novgorodskie Gramoty na Bereste is Raskopok 1984–1989 : 191–319.
25. *Zalizniak A. A.* 2008. Old Russian Enclitics [Drevnerusskie Enklitiki].
26. *Zimmerling A. V.* 2008. Word Order in Slavic, Germanic and Romanic Languages [Poriadok Slov v Slavianskikh, Germanskikh i Romanskikh Iazykakh]. Ot Imen k Faktam (Slaviano-Germanskie Issledovaniia), 3 :165–239.
27. *Zimmerling A. V.* 2009. Clitics in the Old Russian Language Space [Klitiki v Prostranstve Drevnerusskogo Iazyka]. Russkii Iazyk v Nauchnom Osveshchenii, 1 (17) : 259–277.
28. *Zimmerling A. V.* 2011. Word Order Systems in the Slavic Languages [Sistemy Poriadka Slov v Slavianskikh Iazykakh] : 9–67.
29. *Zwicky Arnold M.* 1977. On Clitic.

Section III.
Other areas of the “Dialogue”

В данном разделе публикуются доклады, которые несколько выходят за рамки основных направлений Диалога, но были тем не менее включены в программу конференции по решению Редсовета.

МНОЖЕСТВО ПОВЕСТВОВАТЕЛЕЙ ВАРЛАМА ШАЛАМОВА

М. Ю. Михеев (m-miheev@rambler.ru)

НИВЦ МГУ, Москва, Россия

Рассмотрены разные виды авторской точки зрения в 135 прозаических текстах «Колымских рассказов» (КР) Шаламова; предложены признаки для описания нетривиальных случаев того, что может быть названо *Я*-повествованием *и Он*-повествованием; при этом учитывается не только то, от чьего лица ведется повествование, но и названо ли это лицо каким-либо именем кем-то «со стороны» (или им самим), или же данное лицо осталось безымянным (б/и). Оказывается, что рассказы в начальных циклах заключают в себе несколько типов повествования — ведущегося как бы на разных уровнях, но затем, к концу КР, это множество ипостасей автора убывает, все более приближая текст к традиционно автобиографическому.

Ключевые слова: авторская речь, повествователь, автор, автобиография.

MULTIPLE NARRATORS IN VARLAM SHALAMOV'S TEXTS

M. Iu. Mikheev (m-miheev@rambler.ru)

Lomonosov Moscow State University, Moscow, Russian Federation

I examine the author's point of view in 135 prosaic texts taken from the Kolyma Tales (KT) by Varlam Shalamov. I consider certain characteristics of non-trivial cases that might be called I-narration and He-narration (first- and third-person narration) considering not just the narrator's perspective alone, but also whether that person is called by another name by others or if the person remains nameless. The result: the stories in the primary cycle contain a few types of narration at different levels, but by the end of the KT, the multiple incarnations of the author start to decrease and the text gradually approaches traditional autobiography.

Key words: Shalamov, authorial speech, narrator, author, autobiography.

> …«Субъективен» дискурс, в котором явно или неявно маркировано присутствие «я» (или отсылка к нему)…. И обратно, объективность повествования определяется как полное отсутствие отсылок к рассказчику.…
> (Женетт)[1]

В нарратологии принято различать *диегетического* повествователя (т. е. принадлежащего, или включенного в сам мир текста) и — *экстра-диегетического*, описывающего события как бы откуда-то извне (Шмид). Первый представляет читателю всё с более субъективной точки зрения, чем второй. В текстах Шаламова мы встречаемся и с тем, и с другим типами повествования. Однако для описания их различия, на мой взгляд, лучше вернуться к более простому, как будто оставленному наукой различению, так называемого — Ich-Erzaehlung, или *Я*-повествования, и того, что хотелось бы назвать — *Er*-Erzaehlung, или *Он*-повествованием.

В прозе Шаламова выделяется такой «автор», который упоминает или прямо описывает среди персонажей — в их числе — и самого себя, т. е. «я» (1-е лицо единственного числа), зачастую называя его каким-то именем, но не обязательно совпадающим с его собственным (именем реального автора), или даже вовсе никак его не называя, оставляя это «я» безымянным. И тот и другой случаи (только лишь с упоминанием или еще и с описанием «я») — это Ich-Erzaehlung, или *Я*-повествование. Их надо отличать, во-вторых (или даже: *во-первых*, по значимости), от такого способа изложения, в котором повествователь формально вовсе устраняется, рассказывая исключительно только то, «как было

[1] Четыре точки подряд здесь и далее обозначают прерывание цитаты не на точке исходной фразы.

дело» — об окружающем, персонажах, происходящих с ними событиях, но себя при этом не называет и не описывает, никаким боком не «вставляя» в текст. Последний способ и хочется назвать — *Er*-Erzaehlung, или *Он*-повествование. (Иное дело, что оба способа могут порождать в сознании читателя образ как всеведуще-вездесущего, так и сколь угодно «ненадежного» повествователя, даже, к примеру, неантропоморфного — в результате вполне может оказаться, что речь ведется от имени какого-нибудь пса, кота, лошади, белки и т. п.)

Иначе говоря, внешний (*экстра-диегетический*, «объективный», при Er-Erzaehlung) повествователь ведет рассказ, не обсуждая и не упоминая того, кто всё рассказывает[2], а повествователь внутренний (*диегетический*, или субъективный, при Ich-Erzaehlung) еще и сам материально присутствует в тексте, что-то читателю сообщая о самом себе (и о первичном, вторичном или каком угодно рассказчике, если есть еще и таковой, когда они различаются). Эта информация просачивается либо из метатекста, автокомментариев, входя как *эго*-текст, либо вводится кем-то из действующих лиц путем обращения к тому, кто в речи от автора поименован местоимением 1-го л. ед. числа. — С этой стороны Ich-Erzaehlung выглядит сложнее, чем Er-Erzaehlung, повествовать от «я» сложнее, чем если об этом «я» ничего не говорить[3].

Но с другой стороны, наоборот, более простым оказывается как раз Ich-Erzaehlung. Ведь надо различить, с одной стороны, «я» соответствующее реальному человеку, т. е. тому Ego, кто написал данный текст (автору), а с другой стороны, «я» им воображаемое и навязываемое читателю –повествователя, от лица которого текст написан, «подставного» автора, который с тем или иным успехом подменяет в наших читательских глазах реального. В первом случае мы имеем дело с автобиографическим, дневниковым или мемуарным текстом[4], во втором — с текстом фикциональным: скажем, с автором «Лолиты» В. Набоковым (1) или неким *Гумбертом Г.* (2). В литературе «я» или «он» (за которым кто-то стоит: автор или кто-то еще) почти всегда оказывается более сложной величиной, чем «я» очерков, дневников или воспоминаний[5].

---

[2] Остается в наличии только *повествующее* «я», но не *повествуемое*, или чисто *изображающее* начало, но не *изображаемое* (первое согласно Шмиду, второе — Бахтину).

[3] Последний способ — самая простая, по выражению Бютора, даже «примитивная и фундаментальная форма повествования», когда авторская инстанция вообще отсутствует, ее и духа нет в тексте, текст как бы рассказывает сам себя, без каких-либо нарративных *прибамбасов*. Это тип классического, «безличностно-авторского повествования», в котором отсутствует субъект высказывания и создается иллюзия «отсутствия текста», «наличного бытия самого объекта изображения» (Апанович 1997 со ссылкой на Гей 1989).

[4] О разновидностях такового — Михеев 2007. При этом «Повествование от первого лица удовлетворяет законное любопытство читателя и умеряет столь же законные угрызения автора. Кроме того, оно обладает хотя бы видимостью пережитого, достоверностью, которая вызывает уважение читателя и умеряет его недоверие» (Саррот 1956).

[5] Есть и обратная точка зрения — что эго-текст строится еще сложнее, чем художественный (Савкина 2010).

В первом сборнике прозы Шаламова «Колымские рассказы» (1954–1962) насчитывается в общей сложности 33 рассказа, во втором (1956–1965), «Левый берег», — 25, в третьем (1955–1965), «Артист лопаты», — 28. Эти тексты писались в течение около 12 лет[6]. Сюда же примыкают 30 текстов цикла «Воскрешение лиственницы» (1965–1967), а также 21 текст цикла «Перчатка, или КР-2» (1962–1973). Цикл «Очерки преступного мира» стоит несколько в стороне, выпадая из КР в целом[7]. Поэтому будем говорить здесь только о пяти *основных* циклах — собственно «Колымских рассказах» (КР-1), «Левом береге» (ЛБ), «Артисте лопаты» (АЛ), «Воскресении лиственницы» (ВЛ) и «Перчатке, или КР-2» (КР-2)[8].

---

[6] В скобках везде указаны годы написания (а не издания) произведений. Ни одного из своих рассказов Колымских циклов при жизни Шаламов как будто так и не видел в печатном виде (напечатанными в журнале или в книге), а свои ранние прозаические тексты, опубликованные в 30-е годы, оценивал весьма низко: «В свое время я много потратил труда на остросюжетные рассказы — часть их была напечатана, а 150 рассказов — пропало. # Очень важно не переписывать рассказ много раз. Первый вариант — как в стихах — самый искренний» (Шаламов 1965).
Здесь и далее знак # обозначает пропуск абзацного отступа при воспроизведении цитаты. В случаях, где комментарий отходит от обсуждаемого сюжета несколько в сторону: переходит в метатекст или становится репликой возражения или вопросом самому себе («читательскими» комментариями), т.е. напрямую не связан с общей темой, он заключен в квадратные скобки — и в тексте цитат, и в основном тексте, и в тексте примечаний, как здесь: [Что здесь, в приведенной цитате, в последнем замечании, после знака абзаца? Нетерпимость ли к собственному, уже преодоленному, оставленному в прошлом, творчеству? Доведенный до максимума *перфекционизм*? Непереносимость фальши, причину которой автор видит теперь в том, что слишком долго сидел над своими ранними рассказами?]
Более подробно к обсуждению того же он обращался еще 10-ю годами ранее: «С год тому назад вздумалось мне перечесть те несколько рассказов, которые были напечатаны в 36-м, 37-м году («Октябрь», «Литературный современник», «Вокруг света»). Перечел с крайним отвращением, не со стыдом, а просто унизительно было как-то думать, что они и принимались в журнал, и прессу имели хорошую. Не так, не это, не об этом надо было писать» (Шаламов 1965).

[7] Это соответствует уже устоявшейся в шаламоведении традиции — сошлюсь на работу Сухих 2001, где сказано, что первые три цикла КР связаны между собой наиболее тесно, «образуя трилогию с пунктирным метасюжетом», причем в третьем из них («Артист лопаты») разрыв между очерками и новеллами увеличивается: «новеллистика перестает мимикрировать под документ, обнаруживая свою литературность», а в последнем («Перчатка, или КР-2») «продуманная композиция первых книг исчезает совсем. Большая часть вошедших в сборник текстов — очерки-портреты колымских заключенных, начальников, врачей или физиологические срезы лагерной жизни, опирающиеся уже не столько на воображение и память, сколько на память о своих прежних текстах (не нужно забывать, что все двадцать лет рассказы Шаламова не публикуются, и автор лишен возможности посмотреть на них со стороны, почти лишен обратной читательской реакции, лишен ощущения творческого пути). «Перчатка» — книга большой усталости. По структуре она аналогична не КР-1, а «Очеркам преступного мира». Отдельные тексты («Тачка I», «Подполковник медицинской службы», «Уроки любви»), кажется, не дописаны, да и весь сборник остался незавершенным.... (…) Последняя работа Шаламова — уже чистые воспоминания о Колыме, с непреломленным авторским «я»».

[8] Далее пять перечисленных циклов «Колымских рассказов» (КР) обозначаю сокращенно по первым буквам.

Из 84-х рассказов трех первых циклов более половины написано от безымянного «я» («я» + б/и) и около трети — рассказы об «он» + имярек: конкретно о ком-то из героев, без формальных следов авторского присутствия[9]. В 51 рассказе двух последних циклов от «я» + б/и уже 42 рассказа, т. е. более 4/5, и только 1 рассказ — об «он» + имярек.

Таким образом, из 135 текстов во всех пяти циклах написаны от «я» + имярек всего лишь 7 рассказов (5 %). И только в одном случае этому «я» соответствует действительная фамилия автора — *Шаламов* (очерк «Мой процесс» 1960). В другой раз к «автору» обращаются по имени и отчеству — *Василий Петрович* («Надгробное слово» 1960); еще дважды его называют фамилией *Андреев* («Заговор юристов» 1962, «Инженер Киселев» 1965), однако во многих рассказах с виду тот же самый *Андреев* фигурирует и как «он»! Еще дважды «я» назван фамилией *Крист* («Человек с парохода» 1962); даже с именем и отчеством — *Роберт Иванович Крист* («Геологи» 1965). Наконец, в седьмой раз, повествователь выведен все-таки под действительными именем и отчеством — *Варлам Тихонович* («Галина Павловна Зыбалова» 1970–71). Но при этом многие, подавляющее большинство рассказов от «я» + б/и несут на себе те или иные следы, указывая биографические обстоятельства, однозначно относимые к Шаламову.

Есть, например, один рассказ, где повествование ведется от «я» с такими указанными в тексте характеристиками, как рост и вес. В самом начале рассказа «Домино» (1959) говорится, что «я» весил 48 килограммов — при росте 180 сантиметров, т. е. за вычетом веса костей, на нем *мяса осталось 16 килограмм, ровно пуд всего* (как мы знаем из других текстов Шаламова, из-за «алиментарной дистрофии» в лагере он столько и весил).

Или в рассказе «Богданов» (1965 АЛ, «я» + б/и) бригадир, именем которого озаглавлен рассказ, прямо на глазах повествователя сжигает письма, пришедшие тому за долгое время от жены, — разрывая их и бросая в печь. Внутренняя речь повествователя: «Я два года не переписывался с женой, не мог связаться, не знал о ее судьбе, о судьбе моей полуторагодовалой дочери». — Перед нами детали явно автобиографические: у автора во время заключения оставалась в Москве жена с маленькой дочкой, он страстно ждал писем, писем долго не было. Но неясно, был ли действительно сам вопиющий случай издевательства, это демонстративное сожжение писем? — скорее всего, это уже авторская фантазия, одна из деталей, на которых рассказ построен. Биографические детали могут совпадать с текстовыми или же расходиться, уводя читателя в область художественного вымысла.

Попутно еще вопрос: какая именно форма оказывается ближе к автобиографическому, а какая — к художественному повествованию? 1) *Я*-повествование (Ich-Erzaehlung) + б/и, 2) с названным в тексте именем («я» + имярек); 3) *Он*-повествование (Er-Erzaehlung) с именами героев и без авторского «я» («он» + имярек)? К тому же: 3а) с именем героя реальным — или все-таки 3б) придуманным? И еще: 2а) с именем повествователя, совпадающим с авторским, или

[9] Без того, что Набоков называет авторским представителем.

2б) нет? Какой из вариантов (1–3) ближе к *эго*-тексту[10]? — Решения этой формальной стороны дела в тексте бывают весьма различны.

Про повествователя в КР, в отличие от персонажа, верно подмечено, что «это иная фигура, чем центральный персонаж, тот — объект, а этот — субъект повествования, поводырь читателя по колымскому аду. Он знает больше, чем его герои. И главное, он понимает больше. Он близок к тем немногим героям «Колымских рассказов», кто возвышался до понимания времени» (Лейдерман 1992)[11].

В рассказе «Тифозный карантин» (1959, КР-1) на месте авторского «я» — герой, «он», названный *Андреевым Павлом Ивановичем*. — По замечанию В. В. Есипова (сделанному в переписке), тут ясно, что это сам Шаламов: «случай в карантине взят из его жизни, и все детали, что ему 31 год (действие происходит в 1938 г.), то что он <назван по профессии> дубильщиком, <читает на память> стихи и т. д. Что бы произошло, если бы Шаламов писал везде «я» — как он обманывал, ловчил, чтобы не отправили на прииски, на смерть? (…) Может быть, было бы достовернее… Но, видимо, есть какие-то пределы откровенности — и этические, и эстетические. (…) будь рассказ напечатан — писателя заклеймили бы позором».

Действительно, такие неблаговидные детали, например, как сбор *липких и вкусных мясных остатков* с тарелок, оставшихся на столах у начальников, наверно, приличнее было с точки зрения литературной конвенции поручить своему повествовательному заместителю. Но ведь Шаламов, надо это признать, не боится подобных упреков со стороны читателя, привыкшего к классически построенному нарративу. Он будто не принимает «нормальных» соображений в расчет и как раз строит свой текст на нарушении нормы, на эпатаже читателя.

Повествование об «он» ведется также в рассказе «Академик» (1961, ЛБ), но только там — о журналисте с фамилией *Голубев*, пережившем концлагерь и берущем интервью у академика, который теперь (в настоящее время рассказа) резко изменил взгляды на кибернетику. Раньше, 30 лет назад, он клеймил ее как лженауку [в начале 30-х такой науки еще не знали, хотя наверно под дисциплиной, которую мог публично осуждать тогда персонаж, имеется в виду *математическая логика*], а сейчас — перестроился [«вместе с партией»]

---

[10] Как считает Е.Волкова, опираясь в этом на М.Бахтина (Волкова 1997), в Шаламовском «творчестве доминирует рассказчик (Ich-Erzaehlung) или его ипостаси (Андреев, Крист, Голубев), — все они, если следовать Бахтину, “изображенные образы, имеющие своего автора, носителя чисто изображающего начала. Мы можем говорить о чистом авторе в отличие от автора изображенного, показанного, входящего в произведение как часть его”…(Бахтин 1979)». — По ее мнению, у Шаламова “авторское «я» не персонифицировано, а выступает как «чистое изображающее начало», которому ведомы душевные движения и действия героя”. На мой взгляд, следовало бы все-таки формально развести, с одной стороны, ипостась автора «невидимого», со спрятанным «я» («я» + б/и), — а с другой, такого, у которого это «я» хоть как-то обозначено внутри текста, активно себя в нем проявляет («я» + имярек).

[11] Здесь же точно сказано об «осевой линии» для КР в целом — «между традиционной новеллой и риторическими жанрами».

и выступает уже от имени самой кибернетики (автор награждает антигероя *бегающими черными глазами*, с иронией называя — *популярным академиком*).

Под журналистом в рассказе проглядывает как будто сам Шаламов (он действительно работал журналистом во время своего первого возвращения из лагеря, в 1932–1936 гг., и *мог*, в принципе, встречаться с подобным человеком[12]). Журналист Голубев, так же как и вышедший из лагерей Шаламов, плохо слышит; *после болезни* он не переносит лифта: *ни подъема, ни спуска, особенно спуска с его коварной невесомостью* — и потому вынужден подниматься на 6-й этаж пешком (у Шаламова после заключения развилась болезнь Меньера, в результате чего поездки на транспорте и нахождение в лифте были для него крайне мучительны); Голубев боится произносить звук «у» — потому что от этого у него вылетает зубной протез [возможно и эта деталь взята из собственных ощущений автора].

Вот ударная фраза, которой заканчивается этот рассказ:

> Плечевые суставы Голубева были разорваны на допросах в тридцать восьмом году. ##[13]

Однако в текстах Шаламова много раз повторено, что пытки на допросах стали применяться с июля 1938, а он сам прошел следствие до этого, в 1937-м. (Но такое вполне могло быть с автором позже, в следственной тюрьме в Магадане, когда он был арестован уже по лагерному «делу юристов», в конце 1938-го, или в 1943-м, когда ему в очередной раз добавили срок[14].)

В другом замечательном — и тоже остросюжетном — рассказе («Кусок мяса» 1964, ЛБ: опять-таки от «он» по фамилии *Голубев*), с совершенно классическим *саспенсом*[15], переживания главного героя также описаны «авторскими» глазами: в лагерной больнице он приносит в жертву часть своего тела, дав вырезать аппендицит, симулируя его острый приступ, — чтобы только не попасть в командировку, на угольный или золотой забой [было ли подобное в биографии самого Шаламова? Скорее всего, это опять-таки реальность *домысленная*]. Как много раз

---

[12] За его подписью в это время увидели свет больше 80 статей, заметок и очерков (Клайн). «За четыре года он написал более ста произведений малой прозы, основная часть которых, к сожалению, была сожжена семьей после его второго ареста в 1937 году» (там же). Кстати, один из ранних рассказов, «Пава и древо» (Литературный современник. 1937 №3), о вологодской кружевнице, был удостоен литературной премии (что произошло, когда автор находился уже в тюрьме).

[13] Знак ## здесь и далее обозначает — конец цитируемого произведения (рассказа, очерка, новеллы…).

[14] В книге «Комментарии» Г. Г. Красухина, близко общавшегося с Шаламовым с 1966-го по 1971–72 гг., [именно по поводу данного рассказа] сказано: «На допросах ему вывернули руки, порвали сухожилия, отчего ему трудно было попадать рукою в рукав. Чекисты били его по ушам, повредив барабанные перепонки, — он стал плохо слышать».

[15] Саспе́нс (англ. *suspense* — неопределённость, беспокойство, тревога ожидания, приостановка; от лат. *suspendere* — подвешивать) – состояние тревожного ожидания, беспокойства; обычно в кинематографе (Википедия).

повторено в КР, начальство «не допускало пятьдесят восьмую статью <политических> ни к каким работам, кроме кайла и тачки, пилы и топора»[16]. Но далее в рассказе Голубев чуть было не попадает под нож уже вторично — представляя из себя совершенно беззащитный *кусок мяса* — для уголовника, который в припадке *блатарской* откровенности выкладывает начистоту, что собирался было сначала зарезать его соседа по койке, а потом решил, что нет, лучше его самого, но — вот беда, только что получил письмо от «своих», что надо срочно бросать больницу и возвращаться на прииск, где «*суки наших режут*». — Перед читателем снова фигура героя, за которым угадывается автор, каким последний в самом деле *мог бы* быть, если бы обстоятельства сложились определенным образом, если совпало бы одновременно множество сгущенных в единое художественное целое *перипетий*[17].

У писателя была следующая, несколько раз повторенная, установка. Имея в виду лагерь, Шаламов писал:

> Рассказывать об этой жизни нельзя от первого лица («Память», 1970-е[18]).

С одной стороны, в циклах КР, конечно, достаточно много рассказов, написанных от автора как бы совершенно безлично («Бизнесмен» 1962, «Калигула» 1962, «Утка» 1963,…) — только о ком-то из героев, людей или животных, представленных читателю как «он», с намеренно «атрофированным» авторством, будто ex machina. С другой стороны, гораздо больше все-таки рассказов, написанных от авторского «я», 1-го л. ед.числа, хотя большей частью без произнесения имени собственного; в исключительных случаях фамилия или имя-отчество «высвечиваются» как бы нехотя, только при обращении к автору кого-то из персонажей. Конкретные художественные обличия автобиографического рассказчика различны, при этом почти все рассказы — как бы о самом себе. Авторское «я» рассыпается, дробясь как бы на множество «частичек зеркала»:

---

[16] О Шаламове (возможно, несправедливо, в полемическом увлечении): «Он был человеком, люто ненавидевшим всякий физический труд. Не только подневольный, принудительный, лагерный — всякий. Это было его органичным свойством» (Лесняк 1998, с. 209).

[17] Перипети́я (греч. περιπέτεια, «внезапный поворот») — внезапная перемена в жизни, неожиданное осложнение, трудно преодолимое обстоятельство, один из существенных элементов драматургии, обозначающий всякий неожиданный поворот в развитии сюжета и усложняющий фабулу (Википедия).
Ф.Апанович (1997) даже склонен усматривать в рассказах колымских циклов Шаламова триединство автора, повествователя и героя. Позднее (2002) он предложит считать Голубева — наиболее неуловимой (из трех), а именно духовной ипостасью автора (при отождествлении *голубь* = Дух Святой). Эта ипостась противостоит, с одной стороны, образу Бога Отца, персонифицированному в *Андрееве*, с другой, собственно божественному началу, Христу, представленному в образе *Роберта Криста* или в других фигурах повествователя, принимающих точку зрения героев-жертв.

[18] Более широкий контекст: «Как заставить понять, что мышление, чувства, действия человека просты и грубы, что его психология чрезвычайно проста, что его словарь сужен, а чувства его притуплены? Рассказывать об этой жизни нельзя от первого лица. Ибо это будет рассказ, который никого не заинтересует, — настолько беден и ограничен будет душевный мир героя.»

«Шаламов не даёт нам никакого права искать в этих персонажах автобиографические черты: несомненно, они на самом деле есть, но автобиографизм здесь не значим эстетически. Наоборот, даже «Я» — это один из персонажей, уравненный со всеми, такими же, как он, заключёнными, «врагами народа». Все они — разные ипостаси одного человеческого типа» (Лейдерман).

Можно подумать, что такое «я» способно предстать для него в произвольной форме, в форме любого из тюремных или лагерных товарищей автора, впитав в себя произвольный, ставший известным ему опыт. Описываемые события в самом деле могли бы произойти с тем и с другим, и с третьим.

Неужели безразлично, с кем? — Так, может быть, такое множественно-размытое «я» и есть искомый субъект автора-рассказчика в *анти-романе*, стилистические принципы которого, как и сам термин, Шаламов, скорее всего, заимствовал у французов (после встречи зарубежных писателей в Ленинграде в 1963 году)[19]?

Натали Саррот называла такую инстанцию — «анонимным «я» (…), чаще всего лишь отражением самого автора». Она отмечает хитроумный прием, употребленный Фолкнером в «Шуме и ярости», когда «два разных персонажа названы одним и тем же именем. Это имя, которым автор дразнит читателя, поводя им от персонажа к персонажу, словно куском сахара перед носом собаки, вынуждает читателя всё время быть начеку». — Вот и Шаламов, скорее всего, независимо от Фолкнера, проделывает с читателем КР нечто подобное, только именуя по-разному уже не персонажа, а себя, повествователя.

---

[19] Благодарю за указание на этот источник коллегу Францишека Апановича. Он поясняет (в электронной переписке): «С понятием „антироман“ Шаламов встретился после ленинградской встречи европейских писателей в 1963 году, с участием, между прочим, представителей французского „нуво романа“, на которую было немало откликов, в том числе: *Роман*, *человек*, *общество*. *На встрече писателей Европы в Ленинграде.* (Иностранная литература 1963, № 11, с. 204–246); Т. Балашова, Споры о „новом романе“ (Вопросы литературы 1963, № 12). Сомневаюсь в том, читал ли Шаламов кого-либо из представителей „нового романа“, хотя бы потому, что их романы не были переведены тогда на русский язык, но за дискуссией вокруг этой встречи следил, многое в их положениях одобрял, свои открытия считал близкими им, но не тождественными».
И в другом месте: «Шаламов ничего от французов не мог заимствовать, так как большинство рассказов первого цикла возникло еще до ленинградской встречи, многие писались в пятидесятые годы, когда о французском „новом романе“ он еще, скорее всего, не слышал. (…) „Колымские рассказы“ — именно попытка создать новый роман, анти-роман. Это не просто сборник рассказов, а одно целое, но сцепленное на основе иных принципов, чем единая, последовательная и выдуманная фабула в традиционном романе. (…) Шаламов отказывается от всяких условностей и канонов романного жанра, в частности, от единства и последовательного развития действия, от четко выстроенных характеров и психологического анализа, от единства хронотопа, или единства образа мира, от аукториального повествователя (абстрактного, возвышающегося над изображенным миром, всеведущего и всемогущего, или иначе олимпийского) — казалось бы, от всякой упорядоченности. (…) <Такая> раздробленность участвует в создании шаламовского образа раздробленного мира, который распался, как распался и человек в этом мире».

Как Шаламов выразился однажды (в пересказе Ирины Сиротинской):

«Я сам себя собрал из осколков»[20].

С другой стороны, в раннем письме О. С. Неклюдовой, по поводу доработки ею написанного, но так и не опубликованного романа «Ветер срывает вывески», Шаламов высказывал такие наставления — [будущей жене и] самому себе:

«Я не знаю, как пишутся книги от первого лица, но думаю, что это связано с гораздо большей затратой сердца и нервов. Это — гораздо ответственней, чем во всяком другом случае, как ни отделяй себя от героя» (24.7.1956).

То есть тут следует понять, что для него автобиографическое «я» представлялось наиболее сложной позицией при повествовании — устроенным даже сложнее, чем он-повествование?

Интересно, что в более позднем Шаламовском цикле рассказов, написанном в течение двух лет («Воскрешение лиственницы» 1965–1966) статистика несколько другая, чем в первых трех. Тут уже миниатюр от безымянного «я», становится значительно больше — 26 из 30, а вот рассказов с *Он*-повествованием, как ни странно, всего лишь три[21].

Это кажется удивительным, если считать Er-Erzaehlung более **простой** авторской позицией. — Значит, и тут не всё так просто? Говорит ли это о меньшей «отделанности» вещей этого сборника? — Вряд ли, но о меньшем времени, затраченном на их отделку, все-таки да, безусловно. Т.е. Он-повествование в художественном тексте оказывается все же сложнее, чем Я-повествование в исповеди-дневнике или мемуаре, но проще, нежели Я-повествование в тексте фикциональном.

Шаламовские рассказы иногда включают в себя «вложенных» рассказчиков[22], неся в себе несколько типов повествования, как бы на разных уровнях.

---

[20] В свою очередь, со ссылкой на Федота Сучкова: «Шаламов сказал: „Как ты можешь мной восхищаться? Я же совсем не то, за что ты меня принимаешь! Я же состою из осколков, на которые раздробила меня Колымская лагерная республика…“» — Сиротинская поправляет: «Слова мои воспроизведены Федотом Федотовичем не совсем точно. В. Т. сказал: „Я сам себя собрал из осколков“» (Сиротинская 2006).

[21] В двух случаях «он» — это Крист («Облава» 1965 и «Смытая фотография» 1966) и один раз — Шелгунов («Боль» 1967). Один же рассказ — вообще без «я» и без «он» («Графит» 1967). В последнем, шестом цикле, как заметила Е. Волкова, повествование от «я» доминирует: «Бросающаяся в глаза особенность последнего цикла, „Перчатка, или КР-2“, — доминирование знаменитого Ich-Erzaehlung, рассказчика от имени „я“, иногда конкретизированного как Варлам Тихонович Шаламов».

[22] Более подробно не буду касаться здесь весьма интересного для изучения вопроса о типологии *рассказчиков* (в отличие от *повествователей*). Впрочем, как и вопроса об использовании воровских выражений у Шаламова — при активном отторжении самой по себе *блатарской* идеологии. Сколько раз, например, в КР нам встречается слово *параша* — в значении ‘активно распространяющийся слух’ (более десятка) или слово *фраер* для обозначения простого человека, в отличие от вора (около полусотни). Или почему для Шаламова так характерна ирония – взять хотя бы многократные повторы

Кажется логичным разделить все его тексты — в соответствии с тем, от чьего лица они написаны, — по трем следующим признакам:

а) каким личным местоимением сопровождается авторская или та, которую принято называть «авторской», речь — вообще никаким, *я*, или *он*[23];
б) каким конкретным именем собственным названа «авторская» ипостась: вообще никаким — или любым Иваном, Петром, Павлом (при этом, может быть, именем самого автора, в данном случае, *Шаламов* или *Варлам Тихонович*); и наконец,
в) какие характеристики ему, этому авторскому «я» или «он», приписаны в тексте, т. е. какие действия он совершает.

По последнему признаку, как было сказано, во множестве рассказов нам отчетливо видна автобиографичность (по месту, времени, конкретным поступкам и совпадающим деталям биографии), хотя возникает и некоторый «зазор», давая воображаемый образ, «сгущение» событий, авторскую проекцию, домысленную логически…

Понятно, что и по второму признаку, т. е. по произносимому в тексте имени, будь то имя самого автора или его повествователя, или при обращении к кому-то из них — кого-то из героев: *Роберт Иванович, Христ, Андреев, Василий Петрович* или *Голубев*… — читатель всегда может (и видимо неизбежно всякий раз достраивает) некий собирательный портрет — зе́ка, каким он хотел

выражений «*чистый воздух*» (в кавычках) для противопоставления лагеря — тюрьме; сочетания «*друзья народа*» (для обозначения воров) — противостоящего «врагам народа» (т. е. политзаключенным), а также особая интонация ернической, парадоксальной издевки, с которой, например, о побережье Охотского моря говорится следующим образом: *…Арман, Ола, поселки, в которых останавливались если не Колумб, то Эрик Рыжий*. Или: *Обычай — это многовековая лагерная традиция еще со времени Овидия Назона, который, как известно, был начальником Гулага в Древнем Риме*.… («Иван Богданов» 1970–71); где это повтор мотива, встречавшегося ранее, с его усложнением. Ср: *…и Овидий Назон был начальником ГУЛАГа* («У стремени» 1967).
[Что изначально имелось в виду под этим иронико-саркастическим замечанием? — Только то, что поэта Овидия император Август за какой-то его проступок сослал на север империи, в Крым или Румынию?]
Отнесем сюда же, к крайне интересным для дальнейшего исследования, и вопрос о практически полном отсутствии в рассказах «гетероглоссии» (согласно Е.Михайлик 1997 [и Л. Токер]): «Персонажи „Колымских рассказов“ не говорят языком автора — это автор говорит их — и своим — общим — языком»; а также пристрастие к «черной, или гиньольной иронии», «юмору висельников»… Ср.: «Говорил он <Шаламов> трудно, подыскивая слова, прерывая речь междометиями. В его бытовой речи многое оставалось от лагерного бытия» (Лесняк 1989, с. 212).

[23] А *ты/вы* или *мы* — уже как разновидности: 1) подстановки читателя/ей на свое (авторское) место; или 2) включения себя в более крупную группу описываемых в рассказе персонажей; 2а) — группу вместе с читателем. Из рассказа «Тачка II» (1972): *Докатив тачку до своего забоя, ты просто бросаешь ее. Тебе готова другая тачка на рабочем трапе.… ты хватаешься за ручки.…*

себя показать, в той или иной ипостаси. Таким же показывает себя и любой другой автор, скажем, Солженицын — в *Иване Денисовиче* или в *Нержине*[24].

Да и по первому, наиболее формальному, наименее содержательному признаку, местоимению, каким обозначит себя автор перед читателем: вообще никаким, «он», «я» (или, может, каким-то другим, еще более замысловатым) — вероятен целый спектр возможностей.

Когда нет ни местоимения, ни имени, ни узнаваемых фактов и свойств, касающихся собственно авторской инстанции[25], речь идет о *человеке вообще* («В бане» 1955), а текст исчерпывается, как правило, некими общими утверждениями с фиксацией фактов как бы с птичьего полета или формулировкой обобщенных законов (в масштабах лагеря, тюрьмы, страны в целом), с неизбежно возникающим оттенком доктринальности, нравоучительности, морализаторства — из-за непонятно откуда берущегося «всезнания» автора-хрониста[26] («Комбеды» 1959, ЛБ, «Зеленый прокурор» 1959 АЛ, «Графит» 1967, ВЛ)[27]. Подобным же образом выглядят весьма значительные части и во множестве других рассказов (в частности, «Красный крест» 1959, КР-1) — с отдельными обращениями к конкретности, как бы иллюстрирующими примерами, от «я» или об «он», и — *сентиментальными включениями* (Волкова 2007). Такой рассказ может быть лирическим монологом-миниатюрой, развернутым воспоминанием («По снегу» 1956, «Стланик» 1960, «Тропа» 1967), иллюстрацией какого-то положения, «сентенции»[28], стихотворением в прозе или даже геро-

---

[24] Сопоставление двух поэтик, Шаламова и Солженицына, — весьма актуальная тема, на сегодня почти не исследованная (только с одной из сторон эта тема представлена в работе *Есипов* 2007).
Поэтика Солженицына-антимодерниста затронута в статье Ричарда Темпеста 2010, где замечено (о таких «несимпатичных» автору героях «Красного колеса», как Ленин и Парвус):
«в тех больших отрезках текста, что написаны в форме свободной непрямой речи, рассказчик часто иронизирует над воззрениями конкретных (несимпатичных) героев; „выдает“ их, так сказать, манерой их речи. Тем самым отдельные персонажи и явления, которым противоречит голос рассказчика, постоянно подвергаются внутритекстовому ниспровержению, иногда саботажу, даже если их голоса членораздельны, осмысленны, сведущи и интеллектуально убедительны. (…) имплицитный автор интеллектуально вездесущ!»
Вместе с тем в последних «узлах» романа Солженицына Темпест фиксирует преобладание *диегесиса* над *мимесисом* (значение голоса рассказчика становится максимальным по сравнению с голосами героев). Тот же самый процесс, мне кажется, можно наблюдать и в прозаических текстах Шаламова.

[25] Хотя, может быть, и остаются такие их реликты, как безличное авторское «мы» без самого местоимения, только лишь с глаголом в 1 л. мн. ч.: *Оговоримся сразу…* («Тачка I» [год написания или публикации этого текста не удается установить], КР-2).

[26] Иначе можно назвать это «обобщенным риторико-публицистическом» планом повествования (*Волкова* 1998, гл. 7).

[27] Так устроены целиком и «Очерки преступного мира»: в них встречаем только отдельные вкрапления Я-повествования, в рассказе «Жульническая кровь» (1959).

[28] «…весь рассказ как бы нанизывается на некую уже найденную и неопровержимую формулу, иллюстрацией к которой он служит» (Шкловский).

ической балладой, некоей легендой («Сентенция» 1965, «Водопад» 1966, «Воскрешение лиственницы» 1966[29], «Золотая медаль» 1966[30], «Последний бой майора Пугачева» 1959[31]).

С другой стороны, если в тексте представлено 1 л. ед. ч., а имя, факты биографии и прочее совпадает с авторским «я» (Шаламов), — перед нами очерк, автобиографическая проза («Мой процесс»). Но имярек субъекта, ведущего повествование и называющего себя в рассказах «я», чаще всего никак не обозначен — в половине всех рассказов трех первых циклов и в подавляющем большинстве двух последних. В позднейших циклах рассказ движется в сторону всё большей автобиографичности.

Называемое при «я» имя и ранее по большей части **не** совпадало с авторским — это были *Крист* или *Андреев,* хотя по дополнительным деталям читатель и там догадывался, что имеется в виду автор или на его место подставùм обобщенный человек — такой же, как автор, или вообще любой, кто мог бы оказаться на его месте, в тех обстоятельствах.

Выбору автором того или иного способа повествования, вообще говоря, не всегда возможно подобрать вполне однозначное объяснение. Если сравнить рассказ «Тифозный карантин» (1959, «он» + *Андреев*), с более поздним рассказом «Вечная мерзлота» (1970, «я» + б/и), то в первом случае мы имеем классически художественное повествование, а в последнем «острие рассказа как бы повернуто внутрь, к самому автору, невольно ставшему причиной гибели другого человека. Это — суд над самим собой» (Шкловский).

В самом начале этого рассказа сообщаются факты из биографии автора — окончив фельдшерские курсы на Колыме, он *принимает фельдшерский участок на Адыгалахе*… По поводу смены повествовательной дистанции в последнем рассказе В. В. Есипов (которому я благодарен за обсуждение) констатирует, что «„я“ (Шаламов-фельдшер) вынуждает человека к самоубийству из-за угрозы отправки на прииски. Здесь автор предельно откровенен — и это тоже реальный случай. Последняя фраза — *я понял, что мне уже поздно учиться и медицине, и жизни.* — говорит, что Шаламова тоже поглотила «вечная мерзлота» человеческого равнодушия. Он сам признает это, и мы его понимаем и прощаем. Будь здесь какой-нибудь Андреев — такое же признание было бы фальшиво. А «я» — беспощадно и правдиво. (…) И подтверждает главное открытие Шаламова — «новые закономерности в поведении человека» (в лагере)».

То есть здесь мы, читатели, как бы приглашаемся стать участниками суда автора над самим собой. Одновременно это можно прочесть и как приглашение — к суду над самими собой. А готовы ли мы к этому?

---

[29] Здесь «я» уходит на периферию повествования: *Я тоже ставил ветку лиственницы в банку с водой: ветка засохла, стала безжизненной, хрупкой и ломкой — жизнь ушла из нее.*

[30] А тут «я» мерцает, *всплывая* как бы только в отдельных местах повествования.

[31] Елена Михайлик назвала ее — стилизацией под военную кинобалладу (*Михайлик* 1997а). Там нет Ich-Erzaehlung.

Наиболее выразительная деталь, по-моему, создающая здесь повтор вместе с пуантой всего рассказа, помимо приведенной финальной сентенции, — мотив *банного тазика*, на особенности изготовления которого сначала специально было задержано внимание читателя (нам объяснено, что тазики искусно делают на Колыме из отслуживших *консервных банок.* Ранее, из других рассказов мы уже знаем, что *трехлитровая консервная банка* могла служить котелком и в ней же вываривали на огне белье, чтобы избавить от вшей… — «Дождь» 1958, «Сухим пайком» 1959 «я» + б/и). Потом этот самый тазик привлекает взгляд повествователя как единственное возвышение, с которого смог *повеситься* бедняга Леонов в конюшне — при слишком низком ее своде для нахождения какого-либо иного возвышения.

Но субъект авторской (а вслед за ним и читательской) *эмпатии* вполне может скрываться и за 3-м лицом ед. ч. Так, в новелле «Крест» (1959, АЛ) главный герой — безымянный «он». Здесь, судя по многим совпадающим деталям, *слепой священник* — скорее всего, это отец Шаламова Тихон Николаевич; его младший сын, о котором сказано, что он *за участие в подпольном митинге был арестован и выслан, и след его затерялся*, — сам автор (только реально он посажен в Бутырскую тюрьму и отправлен в Вишерские лагеря не за подпольный *митинг*, а за распространение подпольного *письма* Ленина к съезду).

Вот и в рассказе «Почерк» (1964, АЛ) дело идет об «он», названном *Роберт Иванович Крист*. Хотя в обоих случаях рассказ о тех событиях, которые, как известно из биографии, происходили или могли происходить с его родным отцом и с ним самим — в частности, у Шаламова поначалу действительно был хороший почерк[32], который *мог* использовать для переписывания документов лагерный следователь, — но автору почему-то удобнее смотреть на себя со стороны. [Различие формальных повествовательных инстанций все-таки весьма существенно?]

Герой третьего рассказа («Геологи» 1965, ЛБ), названный, как и во втором случае, *Крист*, предстает уже одновременно — и как «он», и как «я»! Вернее, на протяжении всего рассказа (в нем всего 4 страницы) главное действующее лицо упоминается в 3-м лице, но на полстраничке в середине рассказа — вдруг является еще и в 1-м, пять раз подряд. А в последней фразе отрывка повествование вновь возвращается к употреблению 3-го лица применительно к Кристу. Тем самым равенство: Крист = автор, то ли осознанно, то ли бессознательно (может быть, по недосмотру?) прочерчивается. Скорее, все-таки сознательно, поскольку аналогично этому и в более раннем рассказе «Человек с парохода» (1962, уже в позднем цикле КР-2), прямо на первой полустранице соседствуют разные обозначения одного и того же лица: 1) обращение персонажа-доктора к «я» героя-повествователя — *Крист*, 2) его, этого повествователя, само-называние, объединительно с доктором: *мы*, 3) его же обозначение

[32] Так, в архиве Лесняка хранится небольшой самодельный альбом, куда аккуратным, четким, красивым почерком Шаламов по памяти вписывал для главврача лагерной больницы Нины Савоевой любимые стихи поэтов (Шкловский).

мои глаза (т. е. глаза Криста), но далее 4) уже отдельно, как бы со стороны повествователя — *он, Христ*.

Как представляется, в целом игра со множеством разнородных повествовательных инстанций — особенно в начальных циклах КР — важное приобретение Шаламовской «новой» прозы. Жаль, что оно не было развито и не получило продолжения во всех его текстах.

References

1. *Apanovich.* 1997. On Semantic Functions of Intertextual Relations in "The Kolyma Tales" of Varlam Shalamov [O Semanticheskikh Funktsiiakh Intertekstual'nykh Sviazei v "Kolymskikh Rasskazakh"Varlama Shalamova]. IV Mezhdunarodnye Shalamovskie Chteniia (Proc. of the IV International Shalamov Recital): 52.
2. *Apanovich.* 2002. Descent to Hell (the Image of the Trinity in "The Kolyma Tales") [Soshestvie v ad (Obraz Troitsy v "Kolymskikh Rasskazakh")]. Shalamovskii Sbornik, 3, available at: http://shalamov.ru/research/89/
3. *Bakhtin M.* 1979. On the Question of the Methodology of Aesthetics in Written Works [Estetika Slovesnogo Tvorchestva] : 288.
4. *Biutor M.* 2000. The Use of Personal Pronouns in the Novel [Upotreblenie Lichnykh Mestoimenii v Romane]. Roman kak Issledovanie : 71–77.
5. *Esipov V. V.* 2007. Shalamov and Solzhenitsyn: One-on-One in the Hystory Space [Shalamov I Solzhenitsyn: Odin na Odin v Istoricheskom Prostranstve]. Varlam Shalamov I ego Sovremenniki, available at: http://shalamov.ru/research/102/
6. *Gei N. K.* 1989. Pushkin's Prose. Poetics of Narration. [Proza Pushkina. Poetika Povestvovaniia] : 70–76.
7. *Genett G.* 1967. The Boundaries of Narrativeness. Figures II [Granitsy Povestvovatel'nosti. Figury II].
8. *Gorelik G. E.* 2009. From the Notes of a Historian Experimentator. 1984 [Iz Zapisok Istorika-Eksperimentatora. 1984]. Sovetskaia Zhizn' L'va Landau Glazami Ochevidtsev : 158–171.
9. *Clein L.* 2002. Technique Mastering (On Early Prose of Shalamov) [Ovladenie Tekhnikoi (O Rannei Proze Shalamova)]. Shalamovskii Sbornik, 3, available at: http://shalamov.ru/research/91/
10. *Krasukhin G. G.* Comments [Kommentarii], available at: http://www.e-reading.org.ua/bookreader.php/137737/Krasuhin_-_Kommentariii._Ne_tol%27ko_literaturnye_nravy.html
11. *Leiderman N.* 1992. "…V Metel'nyi, Ledeniashchii Vek". Ural, 3, available at: http://shalamov.ru/research/159/
12. *Lesniak B. N.* 1998. "I've Come to You!"["Ia k Vam Prishel!"]. Arkhivy Pamiati, 2 :209, available at: http://www.sakharov-center.ru/asfcd/auth/auth_book-3cff.html?id=85208&aid=91)
13. *Mikhailik.* 1997. In the Literary and Historical Context [V Kontekste Literatury I Istorii]. Shalamovskii Sbornik, 2. available at: http://shalamov.ru/research/50/

14. *Mikhailik.* 1997. Other Shore. “The Last Battle of Mayor Pugachev”: The Problem of the Context [Drugoi Bereg. “Poslednii Boi Mayora Pugacheva”: Problema Konteksta]. Novoe Literaturnoe Obozrenie, 28, available at: http://shalamov.ru/research/109/
15. *Mikheev M. Iu.* 2007. Diary as an Ego-Text [Dnevnik kak Ego-Tekst].
16. *Modern* Citation Dictionary [Slovar’ Sovremennykh Tsitat]..2002 : 389–390
17. *Sarrot N.* 1956. The Era of Suspects [Era Podozrenii], available at: litext.narod.ru/texts/sarrot.doc
18. *Savkhina I.* 2010. “Rewriting Oneself”: Autobiographic Narrations of Immigrants to Finland (1992-2000-s): Svetlana’s Case [“Perepisyvaia Sebia”: Avtobiograficheskie Narrativy Immigrantov v Finliandiiu: Sluchai Svetlany]. Mezhdunarodnaia Konferentsiia “Marginalii 2010: Granitsy Kul’tury I Teksta” (Proc. of International Conference "Marginalias 2010: Confines of Culture and Text"), available at: http://uni-persona.srcc.msu.ru/site/conf/marginalii-2010/thesis.htm
19. *Sirotinskaia I. P.* 2006. No Memoires, but there are Memoirists [Net Memuarov, Est’ Memuaristy]. Moi Drug Shalamov, available at: http://shalamov.ru/memory/37/5.html
20. *Sukhikh I.* 2001. To Live After Kolyma (1954-1973. “The Kolyma Tales”of Shalamov) [Zhit’ Posle Kolymy (1954-1973. “Kolymskie Rasskazy”)]. Znamia, 6 : 198–207, available at: http://shalamov.ru/research/42/
21. *Tempest R.* 2010. Allexander Solahenitsyn – (Anti)Modernist. Novoe Literaturnoe Obozrenie, 103 : 259
22. *Volkova.* 1998. The Integrity and Variability of Books-Cycles [Tsel’nost’ I Variatovnost’ Knig-Tsiklov]. Shalamovskii Sbornik, 2: 130–157.
23. *Volkova.* 2007. “TheKolyma Tales” of Varlam Shalamov in Terms of Neorhetorical and AntiRhetorical Meanings Producing of Iu.M. Lotman [Teksty “Kolymskikh Rasskazov” Varlama Shalamova v Rakurse Neoritoricheskikh I Antiritoricheskikh Smysloporozhdenii Iu.M.Lotmana]. K Stoletiiu so Dna Rozhdeniia Varlama Shalamova. Mezhdunarodnaia Konferentsiia (Proc. of International Conference on the 100 Shalamov´s Anniversary) : 30.
24. *Shmid V.* 2003. Narratology [Narratologiia] : 80–81
25. *Shalamov V.* 1955. A Letter to A. Z. Dobrovol’skii 12.3.1955 [Pis’mo A. Z. Dobrovol’skomu 12.3.1955].
26. *Shalamov V.* 1965. From the Letter to Ia.D. Grodzenskii 17.5.1965 [Iz Pis’ma Ia.D. Grodzenskomu 17.5.1965]. Znamia, 1993 (5) : 142.
27. *Shalamov V.* 1970. Memory. 1970-s [Pamiat’. 1970-e], available at: http://shalamov.ru/library/25/1.html
28. *Shklovskii E. A.* 1991. Varlam Shalamov, available at: http://shalamov.ru/

МЕЖЪЯЗЫКОВЫЕ КАЛАМБУРЫ В РУССКИХ АНЕКДОТАХ*

Е. Я. Шмелева (eshkind@mail.ru)

А. Д. Шмелев (shmelev.alexei@gmail.com)

Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

В статье речь идет об одном из приемов словесного юмора, используемом в современных русских анекдотах, а именно — разным типам межъязыкового каламбура, в основу которого кладется межъязыковая омонимия или паронимия,– столкновения двух сходно звучащих выражений, принадлежащих разным языкам.

Ключевые слова: анекдот, каламбур, юмор, омонимия, паронимия.

INTERLINGUAL PUNS IN RUSSIAN JOKES

E. Ia. Shmeleva (eshkind@mail.ru)

A. D. Shmelev (shmelev.alexei@gmail.com)

Vinogradov, Institute of Russian Language, Russian Academy of Sciences, Moscow, Russian Federation

The paper deals with a certain mechanism of verbal humor used in Russian jokes, namely, interlingual puns, that is, the contrast between two linguistic expressions of different languages that sound alike. The interlingual puns, entailing the interplay between languages, are based on interlingual homonymy or paronymy and are the products of a transaction between languages. The paper describes various types of interlingual puns.

Key words: pun, joke, humor, omonymy, paronymy.

* Данная статья продолжает серию докладов на конференции «Диалог», начатую в 1998 г. (см. [Шмелева, Шмелев 1998]).

1. Вступительные замечания

Когнитивные теории юмора предполагают, что комический эффект возникает в том случае, когда какая-то ситуация или какое-то языковое выражение допускает двоякое понимание, причем юмористический текст устроен таким образом, что вначале на ум неизбежно приходит одно понимание ситуации или языкового выражения, а затем дается некое указание, свидетельствующее, что имеется в виду другое возможное понимание [Raskin 1985: 31–36; Attardo 1994: 47]. Именно в этом эффекте обманутого ожидания, затем озарения, догадки и заключается пуанта анекдотов и других юмористических текстов. Соответственно, различают референциальный и словесный юмор [Attardo 1994: 95]: в первом случае комизм связан с неоднозначностью ситуации, во втором — с неоднозначностью языкового выражения.

В соответствии со сказанным словесный юмор базируется на столкновении двух пониманий некоторого языкового выражения: причем поначалу строение речи должно подсказывать «маскирующее» понимание, а «замаскированное» понимание становится явным для адресата речи лишь вследствие внезапной догадки или озарения. Тем самым комический эффект достигается лишь при условии, что адресата речи удалось направить по ложному пути: важна внезапность догадки или озарения, поэтому «замаскированное» понимание не должно стать явным прежде времени.

Способы замаскировать второе понимание могут быть различны (так, в статье [Plungian, Rakhilina 2008] описано использование для этой цели свойств синтаксических конструкций). В настоящей статье речь пойдет об одном из приемов словесного юмора, используемом в современных русских анекдотах, а именно — столкновению двух сходно звучащих выражений, принадлежащих разным языкам, межъязыковом каламбуре, в основу которого кладется межъязыковая омонимия или паронимия. При этом, как правило, слушатель первоначально воспринимает произнесенное выражение в рамках одного из языковых кодов, хотя часто ему с самого начала дается подсказка, указывающая на возможность использования иного кода.

Далее мы рассмотрим основные типы использования в русских анекдотах межъязыкового каламбура. Напомним, что, как мы уже указывали в предшествующих публикациях (в частности, [Шмелева, Шмелев 2002]), под «русскими анекдотами» мы понимаем анекдоты, которые рассказываются по-русски в русской языковой среде, каким бы ни был генезис этих анекдотов.

2. Русский язык как средство «маскировки»

Один из самых распространенных типов анекдотов, построенных на межъязыковых каламбурах, составляют анекдоты, в которых русское языковое выражение скрывает за собою возможность понимать его как сходно звучащее иноязычное выражение. Поскольку анекдот рассказывается в русской среде и рассказчик и употребившие данное выражение герои, как

правило, говорят по-русски, для слушателя естественно интерпретировать это выражение как русское. При этом чаще всего делается «подсказка» относительно возможности иной интерпретации, напр. указывается, что герой анекдота — иностранец или инородец (далее оказывается, что он интерпретирует русское выражение как элемент своего родного языка). Понимание таких анекдотов требует от слушателей некоторой лингвистической грамотности, знания того, что означает соответствующее выражение в иностранном языке. В противном случае рассказчику приходится сопровождать анекдот пояснениями, которые снижают комический эффект или вообще уничтожают его (поскольку утрачивается внезапность «проникновения в замаскированный смысл»). Напомним приведенный нами в статье в сборнике «Диалог 2008» (см. [Шмелева, Шмелев 2008]) анекдот о хохлушке, которая в ответ на единогласное решение феминистского конгресса о том, что женщины должны «дарить ласки» мужчинам не чаще, чем три раза в неделю, спрашивает: «Що три раза у нэдилю, то я согласная. А як будэ у будни?». Этот анекдот останется непонятен, если не знать, что слово *неділя* в украинском языке (как и аналогичные слова в большинстве других славянских языков, в частности в церковнославянском) означает ‘воскресенье’[1].

Для понимания следующего анекдота важно знать, что на идише *тухес* означает ‘задница’:

(1) Написал Чуковский «Муху-Цокотуху». Приходит к Ленину. «Владимир Ильич! Я стихотворение написал. Хотел бы его опубликовать». — «Ну, читайте». — «Муха, Муха, цокотуха, / Позолоченное брюхо, / Муха по полю пошла, / Муха денежку нашла. / Пошла Муха на базар / И купила самовар…» — «Стоп,стоп. Това’ищ Чуковский! Почему на база’,а не в коопе’атив? Это политическая ошибка. Пе’епишите стихотво’ение!» Приходит к Сталину. «Товарищ Сталин, я написал стихотворение. Разрешите опубликовать?» — «Читайтэ». — «Муха, Муха-Цокотуха, Позолоченное брюхо! Муха по полю пошла, Муха денежку нашла…» — «Постойтэ, постойтэ. Дэнэжку нашла, пропаганда капиталистичэских отношэний? Запрэтить!» Умер Сталин, к власти пришел Хрущев. Чуковский идет к нему с той же просьбой. Начинает читать: «Муха, Муха-Цокотуха, Позолоченное брюхо! Муха по полю пошла…» — «Постойте, постойте. Если все по полям ходить будут, как же кукуруза будет расти? Запретить!» Сняли Хрущева, к власти пришел Брежнев. Чуковский идет к нему с новой редакцией стихотворения: «Муха, Муха-Цокотуха, Позолоченное брюхо…» — «Постойте, постойте. Это у кого позолоченное брюхо, если у меня всего пол-брюха позолочено? Запретить!» Умер Брежнев, к власти пришел Андропов. Чуковский идет к нему: — «Юрий Владимирович, Никак не могу опубликовать

[1] Ср. исторический анекдот, состоявший в том, что в России в XVIII в. некий слуга, подавал жене французского посла после приёма её салоп и произнёс: «Ваш салоп». Он тут же получил пощечину за оскорбление дамы, поскольку фраза была воспринята как *vache salope* (что, понятно, по-французски означает нечто вроде ‘грязная корова’).

стихотворение,помогите». — «Читайте». — «Муха, Муха-Цокотуха…» — «Что вы сказали? ЦК — тухес?»

Заметим, что этот же анекдот иногда рассказывался и в другом варианте, не включавшем эксплицитного произнесения слова *тухес*, но предполагавшем умение догадаться, что имеется в виду. В этом варианте, опубликованном нами в статье [Шмелева, Шмелев 2009] (там же были приведены различные продолжения этого анекдота), Андропов произносит: «Какая такая ЦэКатуха? Что вы там про ЦК сказали?!» В любом случае вероятной мотивировкой возможности каламбурного понимания были слухи о еврейском происхождении Андропова.

К анекдотам данного типа относятся некоторые анекдоты на тему межъязыковой коммуникации, которая сама по себе предполагает возможность межъязыкового каламбура. Напр.:

(2) Встретились в океане две подводные лодки: американская и русская. Американский капитан выходит на связь: “I am Captain Smith.” — «Капитан Фокин». В ответ тишина. Через час снова : “I am Captain Smith.” — «Капитан Фокин». Снова тишина. Еще через час : “I am Captain Smith.” — «Капитан Фокин». “What? Still fucking!?!”[2]

В следующем анекдоте каламбурное осмысление эксплицитно задается указанием на латинский язык, но поддерживается упоминанием того, что участники диалога — студенты-медики:

(3) Два студента-медика пьют пиво у ларька. Один говорит другому: «Что-то пиво сегодня пенистое!» Второй отвечает: «Да задолбал уже этой латынью, сказал бы по-русски».

Некоторая усложненность этого примера состоит в том, что каламбурное осмысление касается только части слова, причем латинское осмысление должно быть использовано в качестве производящей основы для качественного прилагательного.

В некоторых случаях интерпретация выражения как иноязычного может казаться ничем не мотивированной, как в следующем анекдоте:

(4) «Штирлиц, закройте окно, дует». — “Do it yourself, Bormann!”

Дело в том, что, поскольку есть целая серия анекдотов про Штирлица, основанных на эффекте каламбура, можно обойтись без «подсказки», что дальше может последовать переосмысление слова *дует*. Однако выбор английского языка оказывается несколько неожиданным: действие анекдота,

[2] Ср. анекдот о первом русском посольстве в Англии, в которое вошли бояре Лонгинов, Тихонов, Путятин, Фокин и Неверов. Их представляют королеве: “Your Majesty! Long enough, thick enough, put it in, fuckin' and never off!“

по-видимому, происходит в Германии, поэтому немецкий язык здесь был бы гораздо уместнее.

Это же касается и ряда других каламбурных анекдотов о Штирлице, напр.:

(5) Штирлиц приготовился к бою. Но пришла герла.

Несколько особняком стоят прибаутки, подобные следующей:

(6) К хорошему программисту на Новый Год приходит Дед Мороз, к плохому Дед Лайн, а к слишком умному Дед Лок.

Во-первых, прибаутки такого рода не являются анекдотами в собственном смысле слова, их произнесение в норме не предваряется метатекстовым вводом *Слышали анекдот? / Знаете анекдот? / Хотите, расскажу анекдот?* Во-вторых, выражения *дедлайн* (deadline)и *дедлок* (deadlock), строго говоря, не являются иноязычными: первое из них активно используется теми, кому надо сдавать работу к какому-либо сроку (далеко не только программистами); второе действительно преимущественно используется компьютерщиками для обозначения ситуации в многозадачной среде, при которой несколько процессов находятся в состоянии бесконечного ожидания ресурсов, захваченных самими этими процессами. Поэтому каламбур, используемый в этой прибаутке скорее должен быть отнесен к разделу 5, в котором речь идет о явлениях, смежных с межъязыковыми каламбурами.

3. «Замаскированный» русский язык

Иной, почти противоположный тип анекдотов, основанных на межъязыковом каламбуре, представлен анекдотами, в которых в качестве персонажа выступает русский, попавший в иноязычную среду или вынужденный вести беседу на иностранном языке и воспринимающий услышанные им иностранные слова и выражения как русские. Слушатель анекдота, напротив того, с самого начала понимает выражение, служащее базой каламбура, как иноязычное и для него неожиданной оказывается возможность восприятия его как русского. Как и в предыдущем случае, понимание анекдота требует от слушателей некоторой лингвистической грамотности, знания того, что означает соответствующее слово в иностранном языке, чтобы избежать пояснений, снижающих или уничтожающих комический эффект (впрочем, обычно эти знания самые примитивные). Так, например, анекдот о новом русском, который на отдыхе заграницей в ответ на вопрос лифтера “Down?” хватает его за грудки и говорит: «Это еще кто тут даун!», — требует понимания того, что *down* по-английски значит ‘вниз’.

Из анекдотов данного типа одним из самых известных является анекдот о разговоре в авиакассе, существующий в ряде вариантов. Самый простой:

(7) Ирландец просит кассира в Москве: “Two tickets to Dublin.” — Куда, блин, «туда, блин»?[3]

Несколько иную разновидность данного типа анекдотов представляющий собою следующий анекдот:

(8) Приезжает русский в Эстонию. Пограничник спрашивает его: “Occupation?” “No, -успокаивает его русский, — just visiting”.

Здесь русский понимает, что вопрос задан по-английски, но воспринимает английское слово как тождественное по значению с его русским аналогом («ложным другом» переводчика).

4. Межъязыковой каламбур: два иностранных языка

Среди анекдотов, рассказываемых в русской среде, есть и такие, персонажи которых — иностранцы, говорящие на разных языках. Понимание межъязыкового каламбура в таких анекдотах требует от русских слушателей понимания сходных по звучанию выражений двух разных иностранных языков. Поэтому в таких анекдотах, как правило, обыгрываются самые простые немецкие, английские или французские слова. Напр.:

(9) Немцы после второй мировой войны в Париже стараются не говорить по-немецки, поэтому делают заказ по-английски: “Two martinis, please”. “Dry?” — переспрашивает официант. „Nein, zwei,“ — отвечают немцы [немцы интерпретируют английское слово dry ‘сухой’ как немецкое drei ‘три’].

(10) Двенадцать англичан насилуют немку. Она кричит: найн, найн!!! Трое англичан одеваются и уходят [англичане понимают немецкое nein ‘нет’ как его английский пароним nine ‘девять’].

Многие из таких анекдотов являются международными и рассказываются на разных языках. Как уже говорилось, возможность рассказывать их по-русски обусловлена тем, что иноязычные выражения, вступающие в них в отношения каламбурного столкновения, являются общеизвестными. В противном случае потребовался бы комментарий, который бы свел на нет комический эффект. В качестве примера можно было бы привести ряд анекдотов, которые рассказывались на идише или на немецком, а по-русски напечатаны в русском переводе книги

[3] Этот анекдот дал название подборке стихов Александра Левина «кудаблин-тудаблин», напечатанной в журнале «Знамя» (1999, № 11). Одно из стихотворений из этой подборки так и начинается: «Уходит на запад кудаблин-тудаблин, / спокоен, взволнован, упрям и расслаблен», — и кончается: «Уходит на запад кудаблин-тудаблин. / Вернется ль обратно? / Да нет, никогда, блин».

Der jüdische Witz [Ланцман 2006] (в скобках содержатся комментарии, без которых анекдот может быть непонятен русскоязычному читателю). Ограничимся одним:

(11) Экскурсия по Лувру. Экскурсовод никак не может оторваться от «Моны Лизы». Наконец он говорит:– Mettons-nous en marche (ну, в путь — последнее слово звучит похоже на немецкое am Arsch — на заднице)!
— Вот и хорошо,– с облегчением говорит фрау Поллак,– хоть посидим немного.

5. Смежные явления

К анекдотам, построенным на межъязыковых каламбурах примыкают анекдоты, в основу которых положено сходство между единицами разных кодов: знаками разных графических систем, выражениями литературного языка и вариантов языка, отклоняющихся от стандарта, фонетически правильной речью и речью с акцентом. Приведем ряд примеров с минимальными комментариями.

Пример на сходство знаков разных графических систем:

(12) Один дирижер, грузин, заболел, и оркестру на замену прислали русского дирижера. Придя на первую репетицию, он открывает партитуру и видит надпись на первой странице: «Тоналность — сол». Он взял, да и дописал для грамотности в конце слова «сол» мягкий знак. Прошло время, грузинский дирижер выздоровел. Приходит на репетицию, заглядывает в партитуру: «Ничего не понимаю! Быль тоналность — сол, а сталь — сол-бемол!»

Анекдот построен на внешнем сходстве знака «бемоль» нотной грамоты и мягкого знака русской азбуки.

Пример на сходство единиц литературного языка и вариантов языка, отклоняющихся от стандарта:

(13) Наркоман, идя мимо песочницы, раздавил машинку сидящего там малыша. Тот плачет: «Ты мою машинку поломал, отдавай теперь новую!» Наркоман дает ему шприц. «Нет, моя была с колесами». Тот дает ему колеса. «Не нужна мне твоя машинка, пойду я лучше на травке посижу». Наркоман: «Иди, иди. Я в твое время тоже на травке сидел».

Анекдот построен на внешнем совпадении слов наркоманского жаргона с общелитературными словами: *машинка* ‘шприц’, *колеса* ‘таблетки (наркотического действия)’, *травка* ‘марихуана’.

Пример на сходство фонетически правильной речью и речью с акцентом:

(14) Сидит мужик в горах на узкой тропе. Мимо идут два грузина и несут на палке убитого медведя. Мужик спрашиват: «Гризли?» А грузин: «Зачэм гризли, так застрелили».

Анекдот построен на внешнем совпадении произнесения глагола *грызли* с «грузинским» акцентом и названия медведя *гризли*. Фактически в случаях такого рода мы имеем дело с внутриязыковой паронимией, которая в речи с акцентом превращается в омонимию.

6. Заключительное замечание

Итак, мы видим, что русские анекдоты не замыкаются исключительно на русском языковом материале; понимание многих из них предполагает умение опознать межъязыковой каламбур. При этом типы межъязыковых каламбуров, используемых в русских анекдотах довольно разнообразны. В то же время понимание таких анекдотов требует языковой компетенции и умения быстро осознать необходимость смены языкового кода или же одновременного использования двух языковых кодов.

References

1. *Attardo S.* 1994. Linguistic Theories of Humor.
2. *Lantsman Z.* 2006. Jewish Wit [Evreiskoe Ostroumie].
3. *Plungian V. A., Rakhilina E. V.* 2008. La Construction des Anecdotes du Point de Vue de la Grammaire des Constructions. Questions de la Linguistique Slave : 235–248.
4. *Raskin V.* 1985. Semantic Mechanisms of Humor.
5. *Shmeleva E., Smelev A.* Jokes Telling as a Genre of Modern Spoken Russian [Rasskazyvanie Anekdota kak Zhanr Sovremennoi Russkoi Ustnoi Rechi]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 1998" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 1998") : 262–271.
6. *Shmeleva E. Ia., Smelev A. D.* 2002. Russian Joke. Text and Speech Genre [Russkii Anekdot. Tekst I Rechevoi Zhanr].
7. *Shmeleva E. Ia., Smelev A. D.* 2008. "We" or "Others": Imitation of Ucrainian Speech in the Russian Joke ["My" ili "Drugie": Imitatsiia Ukrainskoi Rechi v Russkom Anekdote]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2008" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2008"), 7 (14) : 581–584.
8. *Shmeleva E.Ia., Smelev A. D.* 2009. Variability, Continuity and Seriation of Jokes: The Problems of Database Creation [Variativnost', Prodolzhenie I Seriinost' Anekdotov: Problemy Postroeniia Bazy Dannykh]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2009" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2009"), 8 (15) : 548–553.

Abstracts

AUTOMATIC DETECTION OF NEAR-SYNONYMS IN NEWS CLUSTERS

Alekseev A. A. (a. a.alekseevv@gmail.com), **Loukachevitch N. V.** (louk_nat@mail.ru)
Lomonosov Moscow State University

The paper presents a method for extraction of alternative names of a concept or a named entity mentioned in a news cluster. The method is based on the structural organization of news clusters and exploits comparison of various contexts of words. Word contexts are used as basis for multiword expression extraction and detection of alternative names. As a result of cluster processing we obtain groups of near-synonyms, in which the central synonym of each group is determined.

PARALLEL CONSTRUCTION OF SLAVIC GRAMMATICAL RESOURCES

Avgustinova T. (avgustinova@coli.uni-saarland.de) DFKI GmbH & Saarland University

We present the idea of parallel construction of HPSG-based grammatical resources for Slavic languages using a common Slavic core module in combination with language specific extensions and corpus-based grammar elaboration at all stages of the project.

SEMANTIC RELATIONS IN PHRASEOLOGY

Baranov A. N. (baranov_anatoly@hotmail.com), **Dobrovol'skii D. O.** (dm-dbrv@yandex.ru)
Russian Language Institute, Russian Academy of Sciences

Traditionally, the following types of semantic relations in the lexical system are distinguished: synonymy, antonymy, polysemy, hyponymy, conversion, and causativity. In the field of phraseology, these phenomena display some specific properties. The focus of our paper is on revealing and discussing some of these properties. The starting point of the discussion is the category of semantic field. It provides the theoretical framework for considering semantic relationships between idioms. The semantic field is defined as a set of lexical units which are connected with each other by some salient semantic features. The totality of semantic fields along with the conceptual links between them constructs the thesaurus of a given language, which can be represented in the form of a semantic network.

The most important type of semantic relations within the semantic field is synonymy. Full synonymy is a rare phenomenon in phraseology, because the meaning of an idiom contains additional semantic features, namely the so-called image component. Idioms with identical actual meanings often reveal differences in their image components, and are perceived as near-synonyms, rather than full synonyms. Antonymy is not typical of phraseology because in most cases it is impossible to single out the central semantic feature that could be considered responsible for meaning contrasts. Although traditionally idioms were mainly regarded as monosemous units of the lexicon, the results of our recent research prove that idioms' polysemy is a quite typical phenomenon.

WHAT ARE SOCIOLINGUISTS AND LEXICOGRAPHERS LACKING IN A DIGITIZED WORLD?

Belikov V. I. (vibelikov@gmail.com) Russian Language Institute, Russian Academy of Sciences

It is a common belief that text corpora provide the best testing ground for solving any kind of linguistic problems. As far as grammar is concerned, this may be true, but if we focus on investigating the lexicon the results often appear to be rather superficial.

WWW contains some relatively homogeneous arrays of texts formed independently of linguists, in some cases emerging quite spontaneously. Text arrays with the most prominent social characteristics of their authors are regarded as independent Internet segments (digitized classical literature and 2010 teenager blogs are the most contrasting examples). Frequencies of the same lexical items differ greatly from one segment to another, and this statistics is very significant for sociolinguistics. The main problem in applying the method of segmental statistics is the lack of a suitable instrument for automatic data processing. Several case studies are presented, and the results of segmental statistics seem to be more indicative than those obtained from the Russian National Corpus.

ITALIAN CONSTRUCTIONS WITH SUPPORT VERB FARE IN COMPARISON WITH RUSSIAN

Benigni V. (benigni@uniroma3.it) Università Roma Tre, **Cotta Ramusino P.** (paola.cottaramusino@unimi.it) State University Milan,

The paper deals with Support Verb Constructions (SVC) in Italian that are formed by the verb *fare* 'to make' and its nominal object (V+NOBJ) in an interlinguistic perspective with the Russian SVC with the verb *delat'*.

The study has been carried out for Italian on ITWac (gathered by Baroni) and, for Russian, on the Russian Web Corpus (gathered by Serge Sharoff, University of Leeds), both are available as pre-loaded corpora within The Sketch Engine corpus query system (http://the.sketchengine.co.uk). About 280 types of SVC with a token frequency ≥200 resulted from the query in the Italian corpus. The Italian SVC have been classified into lexical-semantic patterns, on the basis of Nsubj and Nobj semantic features and the Support verb lexical-semantic meaning. Subsequently, the patterns have been grouped into the well-known actional classes of accomplishments, achievements, semelfactives, activities and states (Vendler 1967, Comrie 1976). The overall classification shows that most SVCs go hand in hand with the features of telicity (as regards verbs) and of concreteness and referentiality (as regards NOBJ), and in these classes (accomplishments, achievements) there is a partial parallelism with Russian, whereas fewer Russian SVCs can be found in the activity and states verb classes. Moreover, the presence of a high number of SVCs in the Russian corpus may be considered as a further evidence of the typological shift towards the analytic type that contemporary Russian is apparently undergoing (see e. g. the simplification of noun declension, the expansion of invariable words and the increasing number of bi-aspectual verbs).

E-MAIL VS. CHAT: THE INFLUENCE OF THE COMMUNICATION CHANNEL ON THE LANGUAGE

Berdichevskii A. (alexander.berdichevsky@if.uib.no) University of Bergen

Does the mere change of the communication channel, unaccompanied by any other changes in situational characteristics, affect the language? Quantitative analysis of two corpora of Russian texts that differ solely by the communication channel from which they originate (e-mail vs. chat) proves that it does.

MODERN RUSSIAN PUBLIC DISCOURSE: DO CHANGES IN INFORMATION TECHNOLOGY LEAD TO NEW DISCOURSE STRATEGIES, OR TO NEW WORLDVIEW?

Bergel'son M. B. (mirabergelson@gmail.com) Lomonosov Moscow State University

This study aims at looking into various formats of modern Russian-language internet communication in order to discover changes in sociocultural patterns and models of the discourse behavior that characterize values and norms of the contemporary Russian public life. Specific public discourse genres — high officials' internet blogs — are analyzed with a special emphasis on whether the public discourse represented in the modern electronic modes is different in the language used from that of the traditional official discourse. This analysis should allow to better understand ideas and beliefs prevailing in the Russian public opinion, to trace its changes and emerging linguistic patterns.

QUALITY ASSURANCE TOOLS IN THE OPENCORPORA PROJECT

Bocharov V. (bocharov@opencorpora.org), Mathlingvo, **Bichineva S.** (bichineva@opencorpora.org), **Granovskii D.** (grand@opencorpora.org), **Ostapuk N.** (nataxan90@gmail.com), **Stepanova M.** (mariarusia@gmail.com), OpenCorpora

OpenCorpora is a project that aims at creating an annotated corpus of Russian texts, which will be fully accessible to researchers, the annotation being crowd-sourced. The article deals with annotation quality assurance tools.

SOME LEXICAL «DISCOVERIES» ON THE MATERIAL OF RUSSIAN SPONTANEOUS SPEECH, A CORPUS STUDY

Bogdanova N. V. (nvbogdanova_2005@mail.ru), **Os'mak N. A.** (nataly.androsova@gmail.com) Faculty of Philology, Saint-Petersburg State University

The article presents results of the first attempt at lexicographical description of Russian spontaneous speech. Analysis is based on the material from the Corpus of Spoken Russian "One Speech Day". New linguistic units (words and phrases) not represented in dictionaries yet, new meanings and definitions or connotations of "old" words are described along with the trends of use in everyday speech. It is shown that a new area of lexicography, which could be called "speech lexicography", is emerging. Its overall principles have not been completely determined yet, although some of the directions can already be specified: 1) creation of a dictionary of common Russian colloquial speech, which should reflect linguistic units used in everyday speech; 2) creation of a dictionary of context-dependent expressive units; 3) creation of a dictionary of discursive units, and 4) collection of a corpus of aphetic and reduced units. The paper outlines controversial problems for each direction and provides linguistic examples.

A LARGE ELECTRONIC DICTIONARY AS A POLYTHEMATIC GUIDE AND SHAPER OF QUERIES TO THE WEB

Bol'shakov I. (bolshakov34@mail.ru) Independent Researcher, **Gel'bukh A.** (gelbukh@gelbukh.com) National Polytechnic Institute, Mexico

A large Russian electronic dictionary is presented. It contains both fundamental information on the Russian language (grammatical and combinatory properties of words, semantic and paronymic relations between words) and ample encyclopedic information on geographical objects, famous people, organizations, and artifacts. The dictionary includes technical terms and basic concepts of science, humanities, business, and economy. Among its applications is the possibility to form queries for Internet search engines on medicine, commerce, tourism, and other topics.

A CORPUS-BASED STUDY OF NOUN CRYPTOTYPES IN EIGLISH

Boriskina O. O. (olboriskina66@mail.ru) Voronezh State University

We develop a method of identifying noun cryptotypes in English, relying primarily on the Corpus of Contemporary American (COCA) and the results of typological studies. The study uses data-oriented and theory-oriented approaches to linguistic description. A cryptotype is referred to the principle of distribution of nouns among classes in accordance with a certain semantic feature and with reference to the typological principle of contrastive grammar. The class membership of a noun is evidentially revealed in syntax, particularly in collostructions which bear the classifying function of the noun class. The semantic, morphologic and syntactic criteria for identification of a noun class are discussed. The study of cryptotypes concerns the issues of grounding, recognition, and reasoning. An adequately formalized description of cryptotypes can be used in computational modeling and text processing.

PARAMETER OF NEARNESS IN THE METAPHORICAL SPACE

Borisova E. G. (egbor@mail.ru) The Moscow City Teachers' Training University,
Ovchinnikova T. E. (teomax@ya.ru) Moscow State Linguistic University

In this work a conception of using deictic means as indicators of spatial relations for function of modal (intensifying) particles is developed. The relation with the deictic function implies metaphorisation of spatial relations and transfer of their parameters on relations in the field of a discourse, the speaker's and listener's general knowledge and more delicate semantic relations connected with the degree of importance for Speaker and Listener. They are able to be metaphorised through the concepts "speaker's space", "speaker's and listener's common space".

Obviously, the modal particles VOT and VON are connected with index (spatial) particles. The modal meanings, different from index ones, are the approximation meanings, the intensifying meanings and a number of other ones.

We are guided by the opposition of the spatial index particles VOT and VON connecting "indication near subject — indication distant subject" with opposition.

As identification is an action quite widespread in metaphorical space, we often meet with the use of VOT for indicating an object or phenomena so that we can speak about metaphorical "sense space" (the thesaurus of conversation participants) or "speech space" (i.e. the semantic network of discussed events).

The situation with the indication VON being a sign (or an instruction) of searching the required object, is quite different. Naturally, the question on searching arises usually when the object is far. However, it is possible to search in rather near space.

Therefore the description of difference in use of particles VOT and VON in modal functions can be compared with concepts of nearness and distance, but taking into account the described distinctions in their semantics.

It turns out that the features of metaphorical use of VOT and VON are connected with their meanings in terms of indication, and not just with opposition on degree of distance, but with ways of indication. Therefore the opposition is metaphorised, too. We can say that particles specify different ways of searching objects in metaphorical spaces. We had sense spaces (the speaker's and listener's thesaurus), intercourse space, that is meaning of messages during the intercourse. In all cases the concept of nearness and distance (as derivatives of identification and searching operations) is not connected with the participants of the intercourse, but with distance between the actual representations realized at present intercourse, and new concepts, objects, properties involved for semantic, emotional and other problems.

TO FIND OUT OR TO BUY? PRODUCT REVIEW VS. WEB SHOP CLASSIFIER

Braslavskii P. (pb@yandex-team.ru), Yandex, **Kiselev Iu.** (yurikiselev@yandex-team.ru), Ural Federal University

We examine two categories of search results retrieved in response to product queries. This classification reflects the two main kinds of user intents — product reviews and online shops. We describe the training and test samples, classification features, and the classifier structure. Our findings demonstrate that this method has good quality and performance, suitable for real-world applications.

"FUNCTIONAL" STANDARD IN RUSSIAN AND ENGLISH DEGREE CONSTRUCTIONS

Bylinina E. G. (e. g.bylinina@uu.nl) Institute for Linguistics OTS, Utrecht University

We develop a notion of functional standard, which refers to the 'functional standard degree construction' (*John is a little bit too tall for this job*). The construction involves a 'purpose' proposition parameter that determines the set of degrees compatible with the purpose. The maximal degree belonging to this set serves as a standard in the construction. We argue against contextual and comparative analyses either explicitly or implicitly assumed in the literature. Instead, we propose that the purpose is an argument of (certain) gradable adjectives, and the whole construction is a positive construction. We try to pinpoint the difference between Russian and English functional standards.

THREE-WAY MOVIE REVIEW CLASSIFICATION

Chetverkin I. (ilia2010@yandex.ru) Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, **Loukachevitch N.** (louk_nat@mail.ru) Research Computing Center, Lomonosov Moscow State University

We consider a three-way classification approach for Russian movie reviews. All reviews are divided into groups: "thumbs up", "so-so" and "thumbs down". To solve this problem we use various sets of words together with such features as word weights, punctuation marks and polarity influencers that can affect the polarity of the subsequent words. We also estimate the maximum upper limit of automatic classification quality in this task.

VOICE EMOTION CLASSIFICATION: PROBLEMS AND SOLUTIONS

Davydov A. G. (davydov-a@speechpro.com), **Kiselev V. V.** (kiselev-v@speechpro.com), **Kochetkov D. S.** (kochetkov-d@speechpro.com), Speech technologies LLC, Belarus

An algorithm for automatic emotion recognition from the speaker's voice has been developed. A number of tests were performed using the widely known corpus of Emotional Speech — Berlin Database (Emo-DB). The classification efficiency for different acoustic features was estimated and a very small set of the most reliable characteristics was extracted in order to obtain a robust and quick emotion state classification. Using the SVM classifier with quadratic kernel and this feature set provides the recognition accuracy of approximately 96%. between «anger» and «neutral» emotional states. GMM classifier was less effective and demonstrates a classification error of up to 6%. A

brief comparison of this feature set and SVM kernel effectiveness was performed using the Munich openEAR toolkit. A recommended set of 384 features and linear-kernel SVM was used to solve the same problem. The classification efficiency of such algorithm reached 98%. This value is only ~2% higher than the respective value for the designed feature set and classifier. Under the several conditions, such as in the case of obtaining a decision support factor in the systems of real-time speech analytics the simplified classification scheme would be more preferable than a complex one.

SYNTAX PARSING FOR TEXTS WITH MISSPELLINGS IN DICTASCOPE SYNTAX

Erekhinskaia T. N. (te@dictum.ru), **Titova A. S.** (titova@dictum.ru), **Okat'ev V. V.** (oka@dictum.ru) Dictum Ltd., Nizhny Novgorod, Russia

The paper deals with syntax parsing of natural language texts with misspellings and misprints in DictaScope Syntax.

We propose a method for integration of a spellchecker and parser, which allows us on the one hand to correct typographical errors considering the context and on the other hand to increase the robustness of the parser.

We start by outlining various types of misprints and ways to correct them, taking account of the specific character of keyboard typing and typical mistakes.

To correct the misspellings and misprints we propose to use a modified Levenshtein algorithm, in which each pair of characters involved in calculation of the Levenshtein distance is assigned a specific weight from the interval. This accounts for keyboard typing, phonetically similar characters, similarity between Russian and Latin alphabet symbols, numbers and other symbols.

The paper states the need to take into account the lexical context of the words to be corrected in order to achieve the maximum accuracy of correction, which helps correct words used in an unusual context. As a result we get a number of correction options for the words. The final choice is made by the DictaScope parser

Basing on the modified Eisner algorithm, the parser builds a dependency tree for the sentence. The modification includes punctuation checking and some additional linguistic limitations. In our model several vertices of interpretations correspond to one word, and variants of spell correction could be processed in the same way as morphological interpretations.

The integration of misprint correction and syntactic analysis is illustrated by a simple case (correcting a single word) and a more complex case — splitting a word in two or merging two words into one.

The proposed method of integration of the parser and the spellchecker modules was implemented in the DictaScope Syntax system. This made it possible to considerably increase the stability of the parser and provided an opportunity to use it as a component of the opinion mining system for monitoring of blogs and forums.

EXPERIMENTAL ANALYSIS OF DISCOURSE: THE IMPACT OF A POTENTIAL REFERENTIAL CONFLICT ON THE CHOICE OF THE REFERRING EXPRESSION (ON THE MATERIAL OF RUSSIAN)

Fedorova O. V. (olga.fedorova@msu.ru), **Uspenskaia A. M.** (ania.quies@gmail.com) Lomonosov Moscow State University

The paper describes an experiment carried out in order to study the referential choice in the situation of potential referential conflict. The results showed that in the situation participants choose full NPs. The results confirmed that referential choice depends on the participants' working memory and made some additions to the model of referential choice.

TAGGING LEXICAL FUNCTIONS IN RUSSIAN TEXTS OF SYNTAGRUS

Frolova T. (tfrolova@iitp.ru), **Podlesskaia O.** (olga@iitp.ru) Laboratory of Computational Linguistics, A. A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences

The paper describes the process and the results of tagging with Lexical Functions the texts of SynTagRus (Syntactically tagged Russian corpus available at www.ruscorpora.ru). The work, begun in 2009, is still in progress. The lexical items which are identified as values and arguments of collocate Lexical Functions (LFs) are tagged in syntactically annotated Russian sentences. So far, about 4,300 sentences (5,500 LF collocations) have been supplied with LF annotation. Examples of possible linguistic and educational uses for the corpus with LF tagging are given.

CHARACTERISTICS OF STUDENT-PROFESSOR E-MAIL COMMUNICATION

Giliarova K. A. (hilaris@yandex.ru) Russian State University for the Humanities

We analyse student-professor e-mail interaction in Russian universities in terms of Field, Tenor and Mode [Halliday 1978]. According to their content, we classify all e-mail messages into three types: "container e-mails", "organizational e-mails" and "essential e-mails".

Even though the e-mail correspondence is a variety of the written communication mode, in organizational e-mails many speech-like features are present. They contain temporal and spatial deixis, anaphora and references to common ground. The word order is typical for colloquial speech, which makes organizational e-mails closer to phone calls. E-mails series resemble oral dialogues.

Both students and professors use different discourse styles: formal, informal, slang, etc. The mode of writing depends more on the authors' age and computer skills rather than on their social status. However, the differences in tenor between the e-mails of students and professors do exist. They are explained by the different perceptions of the norms of social communication and politeness. The analysis of opening and closing formulae also shows that there is no significant difference between the mode of writing e-mails by students and professors. Nevertheless, some specific traits can be found.

THE SEMI-TAGGED CORPORA METHOD EXEMPLIFIED WITH A STUDY OF OSSETIC NOMINALIZATION

Grashchenkov P. (pavel.gra@gmail.com), Institute of Oriental Studies, **Ionov M.** (max_ionov@mail.ru), Lomonosov Moscow State University, **Maliutina S.** (i-am-stupid@list.ru), Lomonosov Moscow State University

We propose the method of Semi-Tagged Corpora (STC) for grammar research in languages that are not expected to have corpora in the nearest future. We exemplify this method with an STC study of internal structure of nominalization in Ossetic. The research was implemented in three major steps: 1) a set of valid surface structures was established; 2) theoretical predictions were made; 3) the initial hypothesis was tested on the text corpora. The corpora were created in two steps. First we selected a significant amount of texts available for Ossetic and merged them in a single text collection. Then we supplied the collection with specific search tools. The initial hypothesis was confirmed that made our field results more accurate and allows a further elaboration of the syntactic structure that we proposed for Ossetic nominalizations.

MULTIMODAL CLUSTERS IN SPOKEN RUSSIAN

Grishina E. (rudi2007@yandex.ru), Russian Language Institute, Russian Academy of Sciences

The paper introduces the notion of multimodal cluster (MMC). MMC is a multicomponent spoken unit, which includes diads "meaning + gesture", "meaning + phonetic phenomenon" (double MMC) or triad "meaning + gesture + phonetic phenomenon" (triple MMC). All components of the same MMC are synchronized in the speech, gestural and phonetic components conveying the same idea as the semantic component (naturally, with available means). To put it another way, MMC is a combination of speech phenomena of different modi (semantic, visual, sound), which are closely connected in the spoken language, and roughly speaking mean the same, i. e. convey the same idea by their own means. The paper describes some examples of double and triple MMCs specific for the Spoken Russian.

A STUDY OF THE NEWS TEXT STRUCTURE AS A SEQUENCE OF CONNECTED SEGMENTS

Iagunova E. V. (iagounova.elena@gmail.com), **Pivovarova L. M.** (lidia.pivovarova@gmail.com) Saint-Petersburg State University

The main object of this study is connected segments (collocations, compound nominations, predicative constructions, multiword expressions, etc.) extracted from the text by different statistical measures and during experiments with native speakers.

This paper deals with news texts: i) 2010 news from lenta.ru (40000 texts, 9.5 million tokens); ii) a small highly homogeneous corpus that deals with some particular event: Schwarzenegger in Moscow (360 texts,110 thousand tokens) and The appointment of Sobyanin (660 texts,170 thousand tokens); iii) three individual texts about Schwarzenegger and two individual texts about Sobyanin. These texts are part of both the small homogeneous corpus and the large news corpus.
In this paper we use an open-source "Cosegment" system (http://donelaitis.vdu.lt/~vidas/tools.htm). The program cuts the text into strongly connected segments depending on the corpus. We study different types of context using overlapping corpora as the input of the system. We also compare result based on the whole corpus and on individual texts from this corpus. During the experiments with native speakers we ask 18 students to put a number from 0 to 5 between every two words in the text. 5 means that these two words are strongly connected, 0 that there is no connection at all. Then we use a cutoff 3.7 to divide a text into connected segments.
Our results are the following: i) Longer connected segments are found in the more homogeneous corpus; ii) Frequent connected segments in highly homogeneous corpora (as opposed to lenta.ru corpus) are mostly predicative constructions; iii) The computer processing data are very close to the native speakers' data; iv) Native speakers tend to extract longer segments; they also prefer predicative constructions to collocations.

ACCENT PLACEMENT PRINCIPLES IN RUSSIAN

Ianko T. E. (tanya_yanko@list.ru) Institute for Linguistics, Russian Academy of Sciences

The basic constituents of intonation structure are pitch accents. Pitch accents designate topic-focus distinctions, contrast, and discourse structure. The question arises as to what phonetic words the accents are placed on. This paper gives an account of various accent placement principles in modern Russian.

HOW DIFFERENT LANGUAGES CATEGORIZE EVERYDAY ITEMS

Iomdin B. (iomdin@ruslang.ru), Russian Language Institute, Russian Academy of Sciences, **Piperski A.** (apiperski@gmail.com), Lomonosov Moscow State University, **Russo M.** (rousseau@mail.ru), **Somin A.** (somin@tut.by)

Classifications of everyday items (category words for clothing, stationery, personal hygiene, beauty products etc.) are studied. A survey of 41 languages was performed. Several results are reported, in particular:
1. Speakers of some languages provide generic terms relatively easy, while for speakers of other languages it is often difficult to perform this task.
2. Some items (such as keys, ear plugs, umbrellas) are virtually unclassifiable in all languages.
3. All languages have covert classes without well-established names (such as personal hygiene or data storage), and people either resort to awkward official phrases like Russian предметы личной гигиены or highly colloquial occasional words like Russian мыльно-рыльное. For items belonging to such classes, high variation of category words was observed.
4. Classes existing in several languages often overlap and include different items. So, посуда in Russian corresponds to dishes, cookware and cutlery in English.
Possible areas of further research are discussed, including studies of language acquisition and bilingualism and comparisons with folk biology and folksonomies.

THE TALKING ETAP. USING THE ETAP PARSER IN RUSSIAN SPEECH SYNTHESIS

Iomdin L. L. (iomdin@iitp.ru), A. A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, **Lobanov B.** (lobanov@newman.bas-net.by), **Getsevich Iu.** (mix1122@gmail.com), United Institute of Informatics Problems, National Academy of Sciences of Belarus

The paper presents an attempt to create an experimental hybrid system of Russian speech synthesis, which makes use of surface-syntactic analysis of the text to be read. The syntactic structure of the sentence, a labeled dependency tree formed by the parser, provides better speech parameters as compared to the classical system of speech synthesis, which does not take explicit account of the

information on how words are related in a sentence. The hybrid algorithm works as follows: the text to be read is sent to the parser of the ETAP-3 linguistic processor sentence by sentence; the ready syntactic structure of each sentence is treated by a number of specially designed rules that mark certain elements of the sentence as prosodically salient, specifying several element types like sentence head, last element of noun phrase etc. The Multiphone speech synthesis module uses this information to produce intrasentential pauses and emphasize certain words or word groups.

SOME METHODS FOR LANGUAGE MODEL PRUNING

Karpenko M. P. (m.karpenko@rambler-co.ru), **Protasov S. V.** (s.protasov@rambler-co.ru), Rambler

This paper describes a pruning system of statistical language models. We present a method for pruning the internal vocabulary which is made completely automatically based on users requests and texts drawn from the Internet. The spellchecker system is one of the components of a search engine, and uses this dictionary for language modeling. The described methods can significantly reduce the size of a language model, and open the possibility to improve spellchecker quality. Experimental results show an improvement in the efficiency of the spellchecker. In our tests the pruning method removed 48% of the language model without sacrificing the quality (in fact, the quality went up 2.7%). This reduction resulted in the speed increase by 87%. Pruning the model allows using a greater volume of query logs in the scenario when the amount of available RAM is fixed. This in turn can improve the quality of the spellchecker.

MEANING OF ESTIMATION IN SEMANTIC SHIFTS OF REBRANDING TYPE IN ADJECTIVES AND ADVERBS (ON THE MATERIAL OF THE DATABASE OF SEMANTIC SHIFTS IN RUSSIAN ADJECTIVES AND ADVERBS)

Karpova O. S. (o_k_@inbox.ru), RSUH, **Rakhilina E. V.** (rakhilina@gmail.com), Russian Language Institute, Russian Academy of Sciences, **Reznikova T. I.** (tanja.reznikova@gmail.com), VINITI, Russian Academy of Sciences, **Ryzhova D. A.** (daska1990R@yandex.ru), Lomonosov Moscow State University

The article is focused on the description of meaning of positive and negative estimation of re-branding type in qualitative adjectives and adverbs (for example, *bezumnyj* 'mad': *chelovek* 'man' */ plat'e* 'dress'; *blestyashchij* 'shining': *pugovica* 'button' */ obrazovanie* 'education'; *dikij* 'wild' *zver'* 'animal' */ pricheska* 'hairdo'; *zolotoj* 'golden': *slitok* 'bar' */ detstvo* 'childhood'; *uzhasnij* 'terrible': *zver'* 'animal' */ vkus* 'taste' etc.). The investigation is fulfilled on the material of the Database of semantic shifts in Russian adjectives and adverbs. The work contains analysis of different aspects of functioning of estimation meanings derived by the semantic shift "re-branding": semantic zones as sources of estimation meanings, mechanism of their generation, lexical combinatory. We also discuss the interaction of estimation meanings with other meanings of re-branding type: combinations of estimation meaning with meanings of intensity, quantity, size, and variety.

THE BASIS OF NATURAL HUMAN LANGUAGE AND ITS MAIN PARAMETERS

Kibrik A. E. (aekibrik@gmail.com) Lomonosov Moscow State University

The paper discusses some, although not all, basic properties of language. I discuss language and sign systems (symbolic signs, indexical signs, iconic signs), as well as the functions of language, including the primary (epistemic, cognitive, and communicative) and the secondary ones (the functions of: social solidarity, individuation, support of social comfort, getting in contact [phatic]; the aesthetic function, the fascination function, the emotional function, and the metalinguistic function). I also treat the main social registers of language (idiolect, subdialect, dialect, language, literary language) and the issues of language death, language change, and linguistic diversity.

ANTONYMS IN PHRASEOLOGY: FORMAL SIMILARITY AS A CONDITION OF SEMANTIC OPPOSITENESS

Kiseleva Ks. (xenkis@mail.ru) Russian Language Institute, Russian Academy of Sciences

The paper deals with semantic oppositeness on various levels of the phraseological system. The data come from an attempt to look at the Russian phraseology in one particular perspective, i. e. to investigate the role that the oppositeness plays between and within idioms. We propose that the semantic oppositeness between idioms can be formed by lexical, contextual and grammatical means. The paper argues that strict antonymy emerges when two idioms have similar structures and are based on the same image. The paper focuses, in particular, on the cases when this oppositeness is formed by negation. Some ways to represent different degrees of negative polarity in the phraseological dictionary are discussed. Different semantic effects related to the negative particle *ne* in idioms, as in *blizhnij svet — neblizhnij svet*, *k licu — ne k licu*, are examined. Finally, we account for the oppositeness as a regular model of the inner form that manifests itself in series of idioms like *k mestu i ne k mestu*, *star i mlad, ni sest' ni vstat'* etc.

TYPES OF SIMULATED EMOTIONAL EXPRESSIVE STATES IN THE RUSSIAN EMOTIONAL CORPUS

Kotov A. A. (kotov@harpia.ru) National Research Centre "Kurchatov Institute", Russia

People often simulate expression of emotions in communication without actual emotional arousal. We suggest that such simulation is forced by other hidden reactions and propose an initial classification. We also extend the architecture of a computer agent to make it able to produce simulated emotions.

GESTURE IDIOMS AND GESTURES: TYPES OF CORRESPONDENCE

Kozerenko A. (akozerenko@mail.ru) Russian Language Institute, Russian Academy of Sciences

The paper considers semantic analysis of Russian idioms, depicting gestures in their inner form. The relationship between the meaning of a gesture and that of the corresponding idiom is examined, as well as polysemy and synonymy relations between idioms, corresponding to the same or different gestures. Definitions of some idioms of the semantic field SADNESS, REGRET, DESPONDENCY are demonstrated. Statements made on the semantics of idioms are illustrated with examples of idiom usage in contemporary texts.

LINGUISTIC MOTIVATION FOR STATISTICAL TRANSLATION MODELS

Kozerenko E. B. (kozerenko@mail.ru) Institute for Informatics Problems, Russia

The paper deals with the problems of parallel texts alignment for enhancing the accuracy and adequacy of translation. Statistical and heuristic models of alignment and transfer are given. The solutions are proposed on the basis of a hybrid grammar, which includes linguistic rules and probabilities of language structures. The goal of the current development is the establishment of matches at the level of meaning, i. e. semantic matches. The meaning can be "packed" in different language structures, so the establishment of cross-language matches and inter-structural synonymy is of prime importance.

NONVERBAL DIALOG IN THE HISTORY OF KINESICS

Kreidlin G. E. (gekr@iitp.ru) Russian State University for the Humanities

The traditional distinction between synchronic and diachronic gesture studies, which has been the cornerstone of nonverbal semiotics and kinesics, is being partly erased if one regards the reflection of gestures in fiction. This paper analyses the descriptions of several somatic signs and nonverbal forms of dialog in the two books of J. Swift — "Gulliver's Travels" and "Gulliver's Erotic Adventures". It argues that these remarkable works of art implement both some of the most common 17th and 18th century gestures and Swift's philosophical and scientific ideas concerning public morals, social and personal activities, communication and notions of language and gestures as lingua franca. I mean to discuss nonverbal acts of that time, purposely performed or uncontrollably leaked, that enhance, improve or disguise verbal messages in the texts. Nonverbal behavior can replace, multi-

ply, or complement language, and Swift's books demonstrate these primary functions of nonverbal sign units vividly and convincingly. In Swift's time the gestural, or body language, as opposed to the natural languages has been considered common, plain, comprehensible, pure, forthright, and therefore the most effective in human communication. Face-to-face dialogs of the author's characters and their corporeal activity incorporate many nonverbal signs that most of his contemporaries regarded as universal. However, Swift mocks and even jeers sometimes at these prevalent viewpoints because he would not believe in the uniqueness and in the universality of the body language.

A CORPUS OF RUSSIAN DIALECTAL SPEECH: THE CONCEPT AND PARAMETERS OF EVALUATION

Kriuchkova O. I. (vpks@rambler.ru), **Gol'din V. E.** (goldinve@yandex.ru) Saratov State University

The concept and parameters of evaluation of a corpus of Russian dialectal speech are discussed based on the comparative assessment of two dialect corpora — dialect corpus within the National Corpus of the Russian Language (DC NCRL) and the Saratov Dialect Corpus (SarDC): the principles of selection of the dialect materials and the criteria of the dialect corpus representativeness; the principles of the speech continuum partition in the corpus; the parameters of textual fragments return; the forms of representation of the dialect texts in the corpus; the types and rules of annotation of the corpus textual basis; the parameters of the dialect texts meta-marking; the representation of non-linguistic information in the corpus; the possibilities of retrieval queries, optimal for dialect research. The paper proves that the dialect corpus cannot be based on the same model as the corpus of standard language because of the specific character of the dialect material. The dialect corpus must be modeled as a system of corpora of different dialects, representing the main dialect types of the Russian speech. According to the proportionality principle, the textual basis of the corpus of a separate dialect must be aimed at the modeling of communication in this specific dialect, reflecting the main types and forms of the dialect speech, as well as social differentiation of the dialect native speakers and genre and theme structure of the dialect communication.

A HIGH PRECISION METHOD FOR THE RECOGNITION OF SENTENCE BOUNDARIES

Kudinov A. S. (al.kudinov@corp.mail.ru), **Voropaev A. A.** (voropaev@corp.mail.ru), **Kalinin A. L.** (kalinin@corp.mail.ru) Project Search@Mail.Ru, Moscow

We present a machine-learning method of sentence boundary recognition. The approach successfully identifies punctuation marks, such as periods or question marks that are not sentence boundary markers. In spite of a relatively small initial learning set (which was prepared manually), the accuracy of this approach appears to be no less than 99 % when applied to an average web document. The method is based upon the decision tree technique combined with a tiny set of manually constructed rules that play the role of classification features. The rules are built using a dedicated declarative language, which is briefly described. A comparison of accuracy of the approach with two freely accessible software products is provided. According to our estimates, the algorithm provides good enough performance to be used in real-time environment such as indexer component of a web search engine. It can also be used to produce large learning sets to train faster machine learning models such as the maximum entropy model.

CONSTRUCTIONS WITH ABSTRACT NOUNS' IN AN ELECTRONIC DATABASE

Kustova G. I. (galinak03@gmail.com) Moscow State Pedagogical University

The paper discusses the types of abstract noun constructions and the types of information in an electronic dictionary (lexical database). The electronic dictionary includes «non-nominative» items which are used as predicates (e. g. *X в плену, в обмороке, в отчаянии, на тренировке, под арестом*), as sentence modifiers (***В заключении*** *он научился шить рукавицы*), as adverbial modifiers (*спрыгнул на ходу; ушел со службы* ***под предлогом болезни***), as parentheses (***во всяком случае,*** *он нам ничего не обещал*). The electronic dictionary includes such types of information on abstract noun constructions as the formal structure, the syntactic function, and the semantic type.

IDENTIFYING ROLE FUNCTIONS OF PEOPLE ON THE BASIS OF KNOWLEDGE STRUCTURES

Kuznetsov I. P. (igor-kuz@mtu-net.ru) Institute for Informatics Problems of the Russian Academy of Sciences, Moscow

The linguistic processor which extracts knowledge structures (information objects and their links) from natural language texts is considered.The development of the processor is connected with extracting implicit information, e. g. role functions of people. The proposed extraction methods are based on the analysis of knowledge structures. The methods are used for identification of role functions of people involved in criminal cases reported in law texts.

PRONOMINALIZATION OF SENTENTIAL ARGUMENT IN RUSSIAN

Letuchii A. B. (alexander.letuchiy@gmail.com) Russian Language Institute of Russian Academy of Sciences, Moscow

The article deals with the distribution of the three Russian pronouns referring to a sentential argument (e. g. — *Vasja ne priedet. — Ja* ***eto*** *znaju* '- Vasja will not come — I know **it**') — namely, ***eto***, ***tak*** and ***takoe***. Each of them has its particular distribution, including contexts where none of the other two pronouns can be used.

I show that the pronoun *takoe* is usually used in the context of negation and modal operators, but only rarely occurs in affirmative sentences with a verb in the indicative mood. The pronoun *eto*, contrary to *tak* and *takoe*, can be used with concrete descriptions of speech acts, including supplementary characteristics of speech, such as loudness, whereas *tak* is incompatible with these characteristics. The individual properties of the pronouns are reflected in their distribution in the corpus data. For instance, the proportion of infinitive clauses among the uses of the pronoun *takoe* is much greater than for the two other pronouns, which nicely agrees with the tendency of takoe to be used in modal contexts.

Finally, I show the difference between the uses of takoe where the pronoun refers to an NP and those where it refers to a sentential argument. In the former case, takoe always denotes a class of entities, whereas in the latter case, takoe can denote one particular content of a speech act. This difference has to do with different referential properties of NPs (objects) vs. propositions.

ON SOME NON-ASSERTIVE VERBS

Levontina I. (irina.levontina@mail.ru) Russian Language Institute, Russian Academy of Sciences

The meaning of the word is determined not only by the components it consists of but also by the status of each component in the logical structure of this word's meaning. The paper deals with a group of Russian verbs with a very peculiar logical structure and unusual syntactic properties. Their meaning is confined to non-assertive components, while the assertion is conveyed by the subordinate verb. In their semantic structure they are therefore similar to some discourse markers (particles, etc.). The verbs in question are *udat'sia*'manage', *ugorazdit'*, *udosuzhit'sja*, *spodobit'sja*, *zablagorassudit'sja*, *soizvolit'*, *soblagovolit'*, *posmet'* [≈dare], *imet' smelost'*, *vzjat'* (*vzjal i sdelal*) etc., most of them hard for translation. Some of such phrases can be approximately translated into English with the verb to do [*On spodobilsia prijti* ≈ He did come]. Thus the meaning of the sentence *On soblagovolil prijti* is confined to the message 'He came' and a combination of speaker's attitudes and expectations. Partly these verbs are negative polarity items (e. g. *udosuzhit'sja*), while others have positive polarity (e. g. *ugorazdit'*).

Special attention is given to the verbs *udosuzhit'sja* and *potrudit'sja*, which express the idea of being ready to make efforts. Interestingly, the meaning of these two verbs, including its logical structure, has been changing during the last 200 years. The paper demonstrates how their actual meaning has taken shape.

SPEECH REPORTING STRATEGIES IN RUSSIAN COMICS-BASED STORIES

Litvinenko A. (allal1978@gmail.com) Lomonosov Moscow State University

The paper considers the factors that influence the choice of speech-reporting strategies in Russian spoken discourse. 10 speakers were asked to produce stories based on a series of pictures that included empty speech 'bubbles'. The experiment resulted in 2 sets of stories, the first one being produced while looking at the pictures, and the second one — several hours later, without using the pictures. In order to be able to analyze the matching instances of reported speech from different speakers, we

marked 10 positions in the pictures, where speech was possible. We will show that not all such positions are actually used by the speakers to produce reported speech; that direct speech seems not to be a prevailing type, at least in this case; that there is no significant difference between telling and retelling a story as regards the choice of speech-reporting strategies. It is discussed that the importance of an episode for the story, the need to portray the characters and personal preferences in style should be considered as significant factors for a speaker choosing the most adequate form of speech reporting.

STATISTICAL CHARACTERISTICS OF SYNTAGMATIC SEGMENTATION OF UTTERANCES FROM THE VIEWPOINT OF EXPRESSIVE TEXT-TO-SPEECH SYNTHESIS

Lobanov B. (lobanov@newman.bas-net.by), **Getsevich Iu.** (mix1122@gmail.com) United Institute of Informatics Problems NAS Belarus

We describe the results of a statistical study of text segmentation into phrases that occurs during expressive reading of Russian fiction by a professional speaker(actor),The pupose is to find out whether part-of-speech tags could be used to predictbreaks between phrases in a sentence. The experimental material was Anton Chekhov's story. *A Hunting Drama*, presented in text (54 thousand words) and sound formats (an audio book with 7 hrs playing time). This material was divided into two parts: the initial segment of the tagged text of the story containing 420 sentences (ca. 6000 words) and the rest of the text (untagged). The untagged part was used for model evaluation. Prosodic phrases were manually tagged by a professional auditor — phonetician who listened to the text. The total number of tagged phrases in the initial 420 sentences was 1516 (of which 710 had pauses no longer than 100 msec and 380 had longer pauses). The average number of phrase breaks in a sentence was 3.6, while the average length of a phrase was 4 words. Pairs consisting of words belonging to 11 different parts of speech or POS-like morphological classes were investigated: adjective, adverb, conjunction, gerund, interjection, parenthetical word, noun, numeral, participle, pronoun, and finite verb. In addition to POS information, the statistical analysis takes account of punctuation marks appearing in the sentence (commas, hyphens, dashes, colons, semicolons and parentheses).. Quantitative distributions have als been obtained for phrase breaks occurring in the pairs: "punctuation mark — part of speech", "part of speech — punctuation mark", "space — part of speech", "part of speech — space". Potentials of using this data in expressive text-to-speech synthesis system are considered.

NON-STOCHASTIC LEARNING OF CROSS-LANGUAGE TRANSLITERATION RULES FROM A SMALL DATASET

Logacheva V. K. (logacheva_vk@mail.ru), **Klyshinskii E. S.** (klyshinsky@mail.ru) Keldysh IAM RAS

We present a language-independent method of generating rules for machine transliteration. The generation of rules is based on the analysis of a test dataset, which contains names written in the source language and their transliterations into the target language.

FACTORS OF REFERENTIAL CHOICE: COMPUTATIONAL MODELING

Loukachevitch N. V. (louk@mail.cir.ru),Lomonosov Moscow State University; **Dobrov G. B.** (wslc@rambler.ru), Lomonosov Moscow State University; **Kibrik A. A.** (kibrik@comtv.ru) Institute of Linguistics, Russian Academy of Sciences; **Linnik A. S.** (skylinnik@gmail.com), Lomonosov Moscow State University; **Khudiakova M. V.** (mariya.kh@gmail.com), Lomonosov Moscow State University

Referential choice between various referential expressions, such as descriptions, proper names, and pronouns, depends on a variety of factors. We present recent results of our modeling study into referential choice, based on the RefRhet corpus. The account of additional factors and the employment of mixed machine learning techniques enabled an improvement of referential choice prediction. This applies both to the two-way choice between full NP and pronoun and to the three-way choice "descriptive full NP vs. proper name vs. pronoun". We have demonstrated that the great majority of the factors taken into account are significant for modeling the referential choice.

CHARACTER NOMINATIONS IN ONTOLOGICAL PERSPECTIVE

Lukashevich N. Iu. (natalukashevich@mail.ru), **Kobozeva I. M.** (kobozeva@list.ru)
Lomonosov Moscow State University

The focus of this research is on ways to represent the meaning of character nominations –words naming either a person according to the person's traits of character, or the characteristic itself, and providing an insight into naïve psychology. An important feature of this lexical semantic group is that we attribute characteristics denoted by them to a person by generalizing from specific cases of the person's behaviour. Therefore the meaning of such words can be understood correctly only when both linguistic and extralinguistic information is taken into account. The paper analyses how knowledge in this sphere can be represented in an ontology.

AUTOMATIC RECOGNITION OF SPONTANEOUS UKRAINIAN SPEECH BASED ON THE UKRAINIAN BROADCAST SPEECH CORPUS

Liudovyk T. V. (tetyana.lyudovyk@gmail.com), **Robeiko V. V.** (valya.robeiko@gmail.com),
Pylypenko V. V. (valeriy.pylypenko@gmail.com)

The paper focuses on automatic recognition of spontaneous Ukrainian speech, introducing the Acoustic Corpus of Ukrainian Media Speech (ACUMS) Three configurations of a speech recognition system are considered. Special attention is paid to training basic and thematic acoustic and linguistic models as well as to the lexicon that contains word transcriptions reflecting spontaneous pronunciation.

The basic acoustic model was trained on recordings from approximately 2,000 speakers (52 hours). The basic language model was trained on ACUMS texts and on texts taken from Internet (400 Mb). Spontaneous variants of word transcriptions were obtained automatically based on standard Ukrainian pronunciation.

Experimental results show that clear normative speech is recognized 50% better than less intelligible speech with hesitations and reductions. Errors are due mainly to erroneous speech corpus annotation, non-vocabulary words (proper names in particular), spontaneous manner of pronunciation, short reduced words (conjunctions and prepositions), and a strong impact of language model on the algorithm searching for the best word sequence.

EXPLOITING DISTRIBUTIONAL SIMILARITY FOR LEXICAL ACQUISITION

McCarthy D. (diana@dianamccarthy.co.uk) Lexical Computing Ltd.

Lexical acquisition has been dubbed the bottleneck of large scale robust natural language processing applications for at least two decades. There is now a substantial body of research dedicated to this important subfield of computational linguistics. Since the 1990s, researchers have turned to corpora for automatic lexical acquisition, rather than rely on extraction from existing online lexical resources. This allows for coverage of new domains, genres and languages without existing resources and where available resources do not provide sufficient coverage or require tailoring to the specific text type. A large body of lexical acquisition from corpora uses distributional similarity whereby the similarity between two words is calculated from the extent that the words have similar contexts of occurrence. Distributional similarity approaches are used for smoothing unseen events using data from seen events. They are also used as an approximation of semantic similarity since there is a strong tendency for words that exhibit similar distributional behaviour to share in their underlying semantics. This paper provides a summary of research that I, along with various collaborators, have conducted using distributional similarity to automatically acquire sense frequency information, selectional preferences and estimates of semantic non-compositionality of putative multiwords.

MULTIPLE NARRATORS IN VARLAM SHALAMOV'S TEXTS

Mikheev M. Iu. (m-miheev@rambler.ru) Lomonosov Moscow State University

I examine the author's point of view in 135 prosaic texts taken from the Kolyma Tales (KT) by Varlam Shalamov. I consider certain characteristics of non-trivial cases that might be called I-narration and He-narration (first- and third-person narration) considering not just the narrator's perspective alone, but also whether that person is called by another name by others or if the person

remains nameless. The result: the stories in the primary cycle contain a few types of narration at different levels, but by the end of the KT, the multiple incarnations of the author start to decrease and the text gradually approaches traditional autobiography.

ILLUSTRATIVE GESTURES AS MARKERS FOR DISCOURSE MACROSTRUCTURE

Nikolaeva Iu. (lis_julia@list.ru) Lomonosov Moscow State University

The paper explores interrelations between discourse structure and gestures accompanying oral narration. It shows how illustrative gestures the reveal discourse macrostructure. Certain issues of speech production and comprehension are discussed with regard to nd the role played by the gesture.

MEANINGS, DIATHESES AND ONTOLOGICAL CATEGORIES OF THE RUSSIAN WORD *VPECHATLENIE* 'IMPRESSION'

Paducheva E. V. (elena.paducheva@yandex.ru) VINITI, Russian Academy of Sciences

The Russian word *vpechatlenie* 'impression' is usually included in the class of emotions, as well as the verb *vpechatljat* 'to make impression'. But derivational relationship between the noun and the verb remains unclear: dictionaries explicate the meaning of the verb *vpechatljat* with the help of the verb phrase *proizvodit' vpechatlenie* 'produce impression', which does not help. The noun *vpechatlenie* is characterized by an idiosyncratic combinability (non-attested by other nouns of emotion) and an irregular polysemy. In this paper *vpechatlenie* is treated as motivated not by the verb *vpechatljat*, but by the verb *vpechatlet'* 'to produce an imprint', which existed in the Russian language up to the beginning of the 19th century but later disappeared. This verb belongs to the class of image creation verbs, such as depict (something as something), represent (something as something), etc. It used to have an uncommon diathesis: *Avpechatlel na /v* Y-*e obraz* Z X-*a* = 'A created on /in Y the image <imprint> Z of X'. Or, take a non-agentive variant: X *vpechatlel na /v* Y-*e svoj obraz* Z = 'X created on /in Y its image <imprint> Z'. The participant X is, as a rule, the consciousness of a human being.

The verb *vpechatlet'* makes all the relationships transparent. It becomes possible (i) to reveal the derivational patterns corresponding to the different meanings of *vpechatlenie* and to assign ontological categories to these meanings; (ii) to describe combinability of the word as an effect of its ontological categories; (iii) to uncover semantic relationships between different meanings.

In this way we get an account of the unique position of the word *vpechatlenie* among the nouns of emotion. Still the language of the Internet demonstrates that the word *vpechatlenie* experiences a pressure from its neighbors and gradually acquires the combinability characteristic of prototypical nouns of emotion, namely, of the names of states. In particular, the verb phrase *ispytat' vpechatlenie*, lit. 'experience impression', becomes frequent, by analogy with *ispytat' udovol'stvie* 'pleasure', *ispytat' radost'* 'joy', etc.

A METHOD OF SENTIMENT ANALYSIS IN RUSSIAN TEXTS

Pazel'skaia A. (pazelskaya@i-teco.ru), **Solov'ev A.** (a.solovyev@i-teco.ru) «I-Teco», Moscow, Russia

This paper presents an overview of methods of sentiment analysis. It also describes our experience of building a system for detecting sentiment in natural Russian texts (mass media). The system uses rule-based approach, calculating sentiment within a simple clause on the basis of word sentiment, output of a Natural Language Processing (NLP) module, and rules of sentiment combination.

Word sentiment is determined in sentiment dictionaries created and regularly updated by experts (more than 15000 words and collocations by now). The system uses separate dictionaries for different parts of speech: nouns, verbs, adjectives, adverbs, verbal and non-verbal collocations. Every word and collocation in the dictionary is marked for its sentiment polarity and sentiment strength.

The NLP module provides morphological and syntactic information (NPs, complex verbs, syntactic roles, clause types and boundaries, etc.). This information is further used to combine word sentiment and to identify sentiment of subject and object within a clause, as well as of the clause as a whole and of the monitored object within the clause.

The system is regularly tested by experts on new mass media texts, it shows about 80% recall and 90% precision.

GENERIC TERMS IN EVERYDAY VOCABULARY AS A SPHERE OF SUBTLE DIFFERENCES BETWEEN SERBIAN AND CROATIAN

Piperski A. Ch. (apiperski@gmail.com) Lomonosov Moscow State University

There are significant differences in the everyday vocabulary of Serbian and Croatian. The speakers are aware of diverging specific terms (e. g., words for 'spoon', 'glasses', 'passport'), but they fail to notice some diverging generic terms (words for 'kitchenware', 'cutlery', 'writing supplies'). This is explained by the fact that generic terms show considerable amount of variation even within one language and cannot serve as markers of identity.

RELATIVE CLAUSES IN SPOKEN RUSSIAN AND ELSEWHERE: A CORPUS APPROACH

Podlesskaia V. I. (podlesskaya@ocrus.ru) Russian State University for the Humanities

The paper addresses the problem of discrepancy between syntactic and prosodic grouping in Russian relative clauses. Basing on oral corpora systematically annotated for prosodic details, the paper demonstrates structural and prosodic "autonomy" of relative clauses from their heads, which previously remained unnoticed in the literature on relativization, which is mainly based on written data.

EXPLORING SEMANTIC ORIENTATION OF ADVERBS

Potemkin S. B. (potemkin@philol.msu.ru), **Kedrova G. E.**(kedr@philol.msu.ru) Faculty of Philology, Lomonosov Moscow State University

Sentiment analysis often relies on a semantic orientation lexicon of positive and negative words. Determining the semantic orientation of words is necessary for correct estimation of the content of statements in the media, Internet, in the writings and speech. Qualitative adverbs expressing evaluation, intensity, direction of action are important as the modifiers of the main sentence predicate. In this paper we propose a method for extracting a seed set of adverbs from a collection of pairs of antonym. A model based on the representation of a set of synonyms from the Russian lexicons as a graph, and determination the semantic orientation of the adverbs concerning three main dimensions of the semantic differential are also demonstrated. The assessment of performance of the method in comparison with the dictionary data shows the effectiveness of the method obtained.

SOME PECULIARITIES OF THE SYNTACTIC STRUCTURE OF RUSSIAN PROVERBS: A STUDY OF ONE-PREDICATE SENTENCES

Renkovskaia E. (jennyrenk@rambler.ru) ABBYY

The paper discusses some peculiarities of the syntactic structure and word order in Russian proverbial sentences with one verb as a predicate. We argue that the syntactic structure of proverbs is dependent on their general pragmatic purposes. The paper focuses on the syntactic features that make proverbs a specific type of Russian sentences.

GENDER IDENTIFCATION OF THE AUTHOR OF A SHORT MESSAGE

Romanov A. S. (alexx.romanov@gmail.com), **Meshcheriakov R. V.** (mrv@keva.tusur.ru) Tomsk state university of control system and radioelectronics

Gender identification of the author of a short message (20–200 characters) is studied. The paper describes a set of experiments with short message texts performed using a support vector machine approach. The task is viewed as a classification problem with two possible alternatives: male and female. Important features of short messages to be considered when determining the author's gender are singled out. The database of electronic communications collected for research included 41780 posts by 15 men and 15women. Experiments used a software system Avtoroved developed by the paper's authors. Altogether, about 50

text attributes at the level of symbols, words, sentences and their combinations were studied. As a result, relevant characteristics of short messages were identified: unigrams and trigrams of symbols, function words, punctuation and emoticons. The total accuracy of gender identifications was 0.74.

A CORPUS-BASED STUDY OF MORPHOLOGICAL VARIABILITY: VARIATION IN GENDER FORMS OF RUSSIAN NOUNS

Savchuk S. O. (savsvetlana@mail.ru) Russian Language Institute, Russian Academy of Sciences

The paper presents the results of a corpus-based study on gender variation in Russian nouns. The list of variants was composed by analyzing textbooks and dictionaries compiled at the beginning and the 2nd half of the 20th century. The total number of gender variants. including outdated and substandard ones, exceeds 600. The variants are classified according to their morphological and semantic features. The next stage of the research is focused on gender variants within the group of indeclinable nouns. The usage of every lexeme from the list was analysed in the texts of the Russian National Corpus, all gender variants was registered in the database and the correlation between variants was determined. The comparison of corpus data with the data derived from dictionaries made it possible to find out the changes in correlation between variants within the studied period and to formulate some trends in variants functioning.

OBJECT IDENTIFICATION IN PROBLEM OF AUTOMATIC DOCUMENT PROCESSING

Seryi A. S. (32112.alien@gmail.com), **Sidorova E.** (lena@iis.nsk.su) Institute of Informatics Systems SB Russian Academy of Sciences

The paper presents an approach to automation of filling of an information system with the data obtained as a result of automatic document processing. The extracted data must be standardized as a network of information objects of a certain format. The backbone of such technique is to build so called focus set for every information object found in a text. Focus set for a single information object consists of all of the relations between this object and other input entities. There are several separate data processing stages: the search for duplicates, direct search, the search for similars and the search via the focus sets technique. A degree of data reliability is also provided. Thus an obsolescence of information, occurrence of the inexact and duplicated data, and conflict of new data with legacy information is taking into consideration.

THE PROPER PLACE OF MEN AND MACHINES IN LANGUAGE TECHNOLOGY: PROCESSING RUSSIAN WITHOUT ANY LINGUISTIC KNOWLEDGE

Sharov S. (s.sharoff@leeds.ac.uk), University of Leeds, UK, **Nivre J.** (joakim.nivre@lingfil.uu.sv), Uppsala University, Sweden

The paper describes several experiments aimed at designing tools for processing Russian texts, namely for Part-Of-Speech tagging, lemmatisation and syntactic parsing, exploiting exclusively statistical approaches without coding any linguistic rules specifically for Russian. While not claiming any new ground for machine learning research, the results demonstrate the possibility to create state-of-the-art tools for Russian in very short time using only machine learning and no hard-coded linguistic knowledge. One of the results of this study is a set of publicly available resources which can be used in standard pipelines for processing Russian. However, they also demonstrate hidden costs associated with the use of purely statistical methods and the need to integrate linguistic parameters into statistical procedures.

INTERLINGUISTIC PUNS IN RUSSIAN JOKES

Shmeleva E. Ia. (eshkind@mail.ru), **Shmelev A. D.** (shmelev.alexei@gmail.com) Vinogradov Institute of Russian Language, Russian Academy of Sciences

The paper deals with a certain mechanism of verbal humor used in Russian jokes, namely, interlingual puns, that is, the contrast between two linguistic expressions of different languages that sound alike. The interlingual puns, entailing the interplay between languages, are based on interlingual homonymy or paronymy and are the products of a transaction between languages. The paper describes various types of interlingual puns.

REFLECTING ACCENTUATION IN THE RUSSIAN MORPHOLOGICAL DICTIONARY OF THE MULTIFUNCTIONAL LINGUISTIC PROCESSOR ETAP-3

Sizov V. G.(sizov@iitp.ru), **Podlesskaia O. I.** (olga@iitp.ru) Laboratory of Computational Linguistics, Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences

Our work is aimed at the introduction of accentual information into the morphological dictionary of the multifunctional linguistic processor ETAP-3. A special formal description language has been created, and special rules for most of the basic accentual schemes have been designed. Special algorithms have been written for morphological analysis and synthesis.

ANAPHORA RESOLUTION OF THE THIRD-PERSON PRONOUN IN TEXTS FROM NARROW SUBJECT DOMAINS WITH GRAMMATICAL ERRORS AND MISTYPINGS

Skatov D. S. (ds@dictum.ru), **Liverko S. V.** (liverko@dictum.ru) Dictum Ltd, Nizhny Novgorod, Russia

Third-person pronoun anaphora resolution in texts from Internet sources (forum comments, opinions) belonging to specific subject domains (cars, household appliances etc.) is discussed. A concrete solution is offered. High precision with acceptable recall (and vice versa) is illustrated by an example of opinions on cell phones.

SOFTWARE FOR AUTOMATED STATISTICAL ANALYSIS OF PHONETIC UNITS FREQUENCY IN RUSSIAN TEXTS AND ITS APPLICATION FOR SPEECH TECHNOLOGY TASKS

Smirnova N. (nsmirnova@speechpro.com), **Chistikov P.** (chistikov@speechpro.com) Speech Technology Center (http://www.speechpro.com)

Currently the development of most speech technology applications is based on the use of pre-recorded speech data produced by one or several speakers. The principal requirement to the speech corpus is sufficient coverage of speech units involved in a specific task. The type of units may differ depending on the approach adopted.

The most popular way of obtaining speech material is through making speakers read some text, since read speech allows strict control over unit coverage (phonetic, prosodic and the like).

For the purpose of automating and facilitating the acquisition of text corpora of desired phonetic composition and coverage, a special tool "TextAnalyser" has been developed. The software is primarily intended for the development of automatic speech recognition and synthesis systems. It makes use of an electronic dictionary containing 180 000 Russian word forms and is based on an automatic transcription tool developed for the Russian TTS system. It allows the generation of texts with required phonetic coverage, the assessment of several types of phonetic unit frequencies in Russian texts (monophones, diphones, triphones, syllables) and the reduction of data redundancy. TextAnalyser was applied for statistical analysis of a large text corpus in Russian comprising 460 965 words (2 500 288 phonemes). As a result of text processing, frequencies of occurrence were obtained for all relevant kinds of Russian-language phonetic units. In the paper we present ordered monophone and diphone frequency lists. The obtained monophone statistics is compared to previously published data.

SELECTION AND PREPARATION OF TERMS FOR THE RUSSIAN-ENGLISH THESAURUS OF COMPUTATIONAL LINGUISTICS

Sokolova E. G. (minegot@rambler.ru), **Semenova S. Iu.** (sonya_sem@mail.ru), Russian State University for Humanities, **Zagorul'ko Iu. A.** (zagor@iis.nsk.su), **Kononenko I. S.** (irina_k@cn.ru), A. P. Ershov Institute of Informatics Systems SB RAS, **Zakharov V. P.** (vz1311@yandex.ru), Saint-Petersburg State University, **Krivnova O. F.** (okrivnova@mail.ru), Lomonosov Moscow State University

The initial phase of the development of Russian-English thesaurus on terminology in the field of computational linguistics is described. One of the first tasks is the choice of candidate sources of terms allowing for the bilingual nature of the electronic resource.

Other problems to be solved are those of terminology extraction and selection of basic term list as well as the study of peculiarities of representation of terms and relations between them. The diversity of the field of computational linguistics, its interdisciplinary nature and the lack of Russian terminological sources and term definitions due to certain lagging of the field in Russia as compared to the English-speaking countries — all these factors explain the kind of decisions made at this stage. One of them concerns the use of the Russian-language corpus of papers presented at the International Conference "Dialogue" (2000–2010). This corpus proved to be a helpful source of terms in real use. Besides, dictionaries as well as indices and glossaries of textbooks and manuals have been examined in order to derive definitions. As an additional source of terms for the Russian part of the thesaurus the English-language terminological sources have been utilized and their terms and definitions translated into Russian. This is especially important for the terms in some empirical and technologically advanced subfields, such as speech technologies.

CASE AS A CHARACTERISTIC OF IDENTITY UNDER ELLIPSIS IN RUSSIAN

Testelets Ia. G. (yakov_ts@mail.ru) Russian State University for Humanities

In Russian, all elliptical operations except N'-ellipsis require the identity of case values in NPs of the antecedent and the elliptical gap, and identity of role or grammatical relation does not suffice when cases are different. Ellipsis follows the six-case model, the peripheral cases, like partitive or 'the case of expected object', pattern with their more standard counterparts. With direct and indirect object NPs, however, case values may be different in the antecedent and the gap, e. g. with recipients, addressees, and the genitive of negation.

ON THE DYNAMIC SEMANTICS OF THE WORD «МИФ»

Trub V. M. (trub44@ukr.net) The Institute of Ukrainian language of the National Academy of Ukraine

One of the main problems of modern semantic research is polysemy. The way to explain polysemy consists in the representation of meaning of a polysemantic word as a result of semantic transitions. The derivative meanings can be explained as a result of transferring attention on one of the components of the initial meaning and suppressing another meaning. We illustrate this principle by the Russian polysemantic word миф 'myth'. It is shown that different meanings of this noun can be interpreted as a result of focusing attention on the different components of its initial definition such as the content of the myth, the disproof of this content, the positive axiological evaluation of different aspects of this content.

CONSESSIVE CONJUNCTION *KHOTIA* 'THOUGH' AND "CANCELLED EXPECTATION"

Uryson E. V. (Uryson@gmail.com) Russian Language Institute, Russian Academy of Sciences

The paper is focused on the semantics of the concessive conjunction khotja 'though'; cf. (1) *Khotja pogoda byla ochen' plokhaja (P), oni kazhdyj den' kupalis*' (Q) 'Though the weather was very bad (P), they bathed every day (Q)'. Different definitions of the meaning 'khotja' are found in the literature. In typology it is generally agreed that its main components are implication and negation: *Though P, Q* = 'usually if P, then not-Q; in this case P and Q'. In traditional Russian grammar it is commonly supposed that the basic semantic component of ***khotja*** 'though' is "cancelled expectation": situation P in the subordinate clause induces the expectation not-Q, and this expectation fails in the main clause. Both definitions are adequate for examples like (1). Some new material enables to choose between them. I analyze sentences like (2) *Khotja bol'shinstvo potukhshikh vulkanov — eto gory konusoobraznoj formy (P), ne vsjakaja takaja gora — byvshij vulkan (Q)* 'Though most of the extinct volcanoes are conical mountains (P), not every conical mountain is an extinct volcano (Q)'; (3) *Khotja Fjodor zarabatyvaet bol'she Ivana (P), on tozhe ne mozhet soderzhat' semju (Q)* 'Though Fjodor earns more than Ivan (P), he cannot provide for his family either (Q)'. I show that the main semantic component of 'khotja' is "cancelled expectation". In cases like (1) "cancelled expectation" is due to our knowledge of usual casual relations between situations (fixed in frames or scenarios). In cases like (2)–(3) "cancelled expectation" is not due to any frames or scenarios but, rather, to some properties of human consciousness. For instance, (2) presupposes a stable association between volcanoes and conical mountains. This association is the basis of our "expectation" and the consessive conjunction

khotja marks that this expectation fails. (3) presupposes a comparison of two people, and in such a context situation P induces our "expectation", which fails in the main clause (cf. Grice's informativity postulate). Thus, in these cases "cancelled expectation" is due to our standard line of reasoning. The concessive conjunction *khotja* is a means for marking that reasoning of this type is wrong. Some linguistic problems of description of the concessive conjunction *khotja* are discussed.

ENANTIOSEMY IN RUSSIAN PHRASEOLOGY

Voznesenskaia M. M. (voznes-masha@yandex.ru) Russian Language Institute, Russian Academy of Sciences

The phenomenon of enantiosemy in Russian phraseology, its sources and types are considered. Enantiosemy is shown to be connected with different interpretation of an inner form (antiphrasis, implications of different kinds etc.). In some cases enantiosemy arises as a result of pragmatic interpretation of a corresponding situation (emotive connotations).

RVAT' ZUBY AND *MYT' DEN'GI*: ONE USE OF SIMPLE IMPERFECTIVES IN RUSSIAN

Zalizniak A. A. (anna.zalizniak@gmail.com), Institute of linguistics, Russian Academy of Sciences, **Mikaelian I. L.** (irina-mikaelian@yandex.ru), Penn State University

The article discusses a specific stylistic effect produced by the use of simple (non-prefixed) imperfective verbs when, under special conditions, they substitute for prefixed imperfective verbs, as in myt' den'gi used instead of otmyvat' den'gi ("to launder money"), varit' truby, instead of svarivat' ("to weld pipes"), etc.

There are three main features that characterize this effect: (1) The verb and its complement constitute a more or less idiomatic expression. (2) The simple imperfective belongs to a lower or substandard register. Thus žeč' gorški is a low-register use with respect of obžigat' gorški ("to fire pots"); myt' den'gi is cruder than otmyvat' den'gi, and pisat' pulju is even lower than raspisyvat' pulju ("to play a game of preference"). (3) Simple imperfectives signal hermetic or jargon-like expressions: they are used in the speech of professional communities or communities of interest, cf. varit' truby instead of svarivat', gruzit' film instead of zagružat' ("to download a film)."

Such pairs of imperfective verbs are not very numerous, but they constitute an open list, which proves that the Russian verb system possesses a mechanism that generates this kind of "quasi" non prefixed verbs.

EXCLAMATIVES IN RUSSIAN: A CORPUS STUDY

Zevakhina N. (natalia.zevakhina@gmail.com) Lomonosov Moscow State University

The paper investigates the exclamative use of wh-words in the National Corpus of Russian Language. Exclamation is a verbally uttered speaker-oriented emotion occurring when the state of affairs in the real world violates the speaker's expectations. The paper shows that, basically, Russian wh-words as exclamatives can be split into four groups according to four criteria: (i) independent exclamative use; (ii) 'anaphoric' use (i. e., with reference to phrases within a particular discourse); occurrence in special syntactic constructions (iii) with the particle *tol'ko*, (iv) with the particle *vot*. The first group of wh-exclamatives satisfies all of the criteria, the second group complies with (ii)-(iv), etc. We also discuss the exclamative use of Russian wh-words in sentential arguments and lexico-grammatical properties of matrix predicates allowing for such contexts that rather exhibit a continuum than clear-cut classes.

SCRAMBLING TYPES IN THE SLAVIC LANGUAGES

Zimmerling A. V. (meinmat@yahoo.com) Moscow State University for the Humanities

The paper discusses the types of scrambling in the Slavic languages and in Universal Grammar. It is argued that all kinds of scrambling may be explained as instances of optional movement. Scrambling types are classified on the basis of final and initial movement domains in the clausal complex where sentence categories move. Slavic languages have all four theoretically possible scrambling types of non-clitic elements and both types available for clitic elements. The diagnostic features of clitic scrambling are described for the first time.

Author Index

Alekseev A. 32
Avgustinova T. 42
Baranov A. N. 53
Belikov V. I. 63
Benigni V. 72
Berdichevskii A. 89
Bergel'son M. B. 99
Bichineva S. 107
Bocharov V. 107
Bogdanova N. V. 116
Bol'shakov I. A. 131
Boriskina O. O. 142
Borisova E. G. 153
Braslavskii P. I. 160
Bylinina E. G. 169
Chetverkin I. I. 177
Chistikov P. G. 699
Corbett G. G. 1
Cotta Ramusino P. 72
Davydov A. G. 187
Dobrov G. B. 520
Dobrovol'skii D. O. 53
Erekhinskaia T. N. 196
Fedorova O. V. 207
Frolova T. I. 219
Gel'bukh A. F. 131
Getsevich Iu. S. 316, 495
Giliarova K. A. 232
Gol'din V. E. 412
Granovskii D. 107
Grashchenkov P. 250
Grishina E. A. 258
Hovy E. 3
Iagunova E. V. 274
Ianko T. E. 289
Iomdin B. L. 304
Iomdin L. L. 316
Ionov M. 250
Kalinin A. L. 422
Karpenko M. P. 327
Karpova O. S. 341
Kedrova G. E. 603
Khudiakova M. V. 520
Kibrik A. A. 520
Kibrik A. E. 4
Kiseleva K. L. 354
Kiselev Iu. A. 160
Kiselev V. V. 187
Klyshinskii E. S. 509
Kobozeva I. M. 530
Kochetkov D. S. 187
Kononenko I. S. 713
Kotov A. A. 364
Kozerenko A. D. 375
Kozerenko E. B. 384
Kreidlin G. E. 401
Kriuchkova O. Iu. 412
Krivnova O. F. 713
Kudinov A. S. 422
Kustova G. I. 434
Kuznetsov I. P. 447
Letuchii A. B. 460
Levontina I. B. 473
Linnik A. S. 520
Litvinenko A. O. 484
Liudovyk T. V. 541
Liverko S. V. 686
Lobanov B. M. 316, 495
Logacheva V. K. 509
Loukachevitch N. V. ... 32, 177, 520
Lukashevich N. Iu. 530
Maliutina S. 250
McCarthy D. 19
Meshcheriakov R. V. 621
Mikaelian I. L. 769
Mikheev M. Iu. 814
Nikolaeva Iu. 553
Nivre J. 658
Okat'ev V. V. 196
Os´mak N. A. 116
Ostapuk N. 107
Ovchinnikova T. E. 153
Paducheva E. V. 560
Pazel´skaia A. G. 575

Piperski A. Ch. 304, 588
Pivovarova L. M. 274
Podlesskaia O. Iu. 219, 672
Podlesskaia V. I. 594
Potemkin S. B. 603
Protasov S. V. 327
Pylypenko V. V. 541
Rakhilina E. V. 341
Renkovskaia E. A. 610
Reznikova T. I. 341
Robeiko V. V. 541
Romanov A. S. 621
Russo M. M. 304
Ryzhova D. A. 341
Savchuk S. O. 628
Semenova S. Iu. 713
Seryi A. S. 646
Sharov S. 658
Shmelev A. D. 829
Shmeleva E. Ia. 829
Sidorova E. A. 646
Sizov V. G. 672
Skatov D. S. 686
Smirnova N. S. 699
Sokolova E. G. 713
Solov'ev A. N. 575
Somin A. A. 304
Stepanova M. 107
Testelets Ia. G. 726
Titova A. S. 196
Trub V. M. 738
Uryson E. V. 747
Uspenskaia A. M. 207
Voropaev A. A. 422
Voznesenskaia M. M. 760
Zagorul'ko Iu. A. 713
Zakharov V. P. 713
Zalizniak A. A. 769
Zevakhina N. A. 782
Zimmerling A. V. 796

Научное издание

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной
Международной конференции «Диалог»

Выпуск 10 (17). 2011

Ответственный за выпуск **В. Л. Талис**
Вёрстка **К. А. Климентовский**

Подписано в печать 12.05.2011
Формат 152 × 235
Бумага офсетная
Тираж 200 экз. Заказ № 264

Издательский центр «Российский
государственный гуманитарный университет»
125993, Москва, Миусская пл., д. 6
Тел.: +7 499 973 42 06

Отпечатано с готового оригинал-макета в типографии
ООО «Издательско-полиграфический центр Маска»
117246, Москва, Научный пр-д, д. 20, стр. 9