Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной международной конференции «Диалог» (2019)

Выпуск 18

Computational Linguistics and Intellectual Technologies

Papers from the Annual International Conference “Dialogue” (2019)

Issue 18

УДК 80/81; 004
ББК 81.1
К63

Компьютерная лингвистика и интеллектуальные технологии:
По материалам ежегодной международной конференции «Диалог»
(Москва, 29 мая — 1 июня 2019 г.). Вып. 18 (25), 2019.

Сборник включает 64 доклада международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2019», 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.

Предисловие

18-й выпуск ежегодника «Компьютерная лингвистика и интеллектуальные технологии» содержит избранные материалы 25-й международной конференции «Диалог». На основании мнений нашего рецензентского корпуса для публикации в ежегоднике редколлегией были отобраны 64 доклада из ста работ, которые были приняты к представлению на конференции в 2019 году.

Работы в сборнике отражают те направления исследований в области компьютерного моделирования и анализа естественного языка, которые по традиции представляются на конференции:

- Компьютерные лингвистические ресурсы
- Компьютерный анализ документов (классификация, перевод, поиск, саммаризация, генерация, анализ тональности и т. д.)
- Глубокое обучение в NLP (методики применения, содержательная интерпретация)
- Компьютерный анализ Social Media
- Корпусная лингвистика и корпусометрия (методики создания, использования и оценки корпусов)
- Лингвистический анализ текста (морфология, синтаксис, семантика)
- Лингвистические онтологии и автоматическое извлечение знаний
- Мультимодальная коммуникация (включая лингвистический анализ речи)
- Модели общения и диалоговые агенты
- Компьютерная лексикография

В соответствии с традициями «Диалога», старейшей конференции по компьютерной лингвистике в России, отбор работ основывается на представлении о важности соединения новых методов и технологий анализа языковых данных с полноценным лингвистическим анализом. Диалог является де-факто крупнейшим форумом по проблемам создания современных компьютерных ресурсов, моделей и технологий для русского языка.

Одно из ключевых событий «Диалога» — подведение итогов технологических соревнований между разработчиками систем лингвистического анализа текстов, Dialogue Evaluation. В этом году состоялись четыре соревнования:

- автоматическая генерация заголовков новостей;
- автоматический анализ малоресурсных языков (для которых очень мало данных для машинного обучения);
- автоматическое разрешение анафоры и определение референциальных цепочек (различных упоминаний одного и того же объекта в тексте),
- автоматическое восстановление слов по контексту (гэппинг-эллипсис).

В сборник включены наиболее оригинальные работы участников этих соревнований.

Статьи в сборнике публикуются на русском и английском языках. При выборе языка публикации действует следующее правило:

- доклады по компьютерной лингвистике должны подаваться на английском языке. Это расширяет их аудиторию и позволяет привлекать к рецензированию международных экспертов.
- доклады, посвященные лингвистическому анализу русского языка, предполагающие знание этого языка у читателя, подаются на русском языке (с обязательной аннотацией на английском).

Несмотря на традиционную широту тематики представленных на конференции и отобранных в сборник докладов они не могут дать полной картины направлений «Диалога». Ее можно получить с помощью сайта конференции www.dialog-21.ru, на котором представлены обширные электронные архивы «Диалогов» последних лет и все результаты проведенных тестирований Dialogue Evaluation.

Мы обращаем внимание авторов и читателей сборника, что с 2018 года Редсовет отказался от печати сборника на бумаге, поскольку бумажный вариант пользуется все меньшей популярностью. Сборник, как и в прошлые годы, размещается на сайте конференции и индексируется Scopus.

Программный комитет конференции «Диалог»

Редколлегия сборника «Компьютерная лингвистика и интеллектуальные технологии»

Организаторы

Ежегодная конференция «Диалог» проводится при организационной поддержке компании ABBYY.

Учредителями конференции являются:

- Институт лингвистики РГГУ
- Институт проблем информатики РАН
- Институт проблем передачи информации РАН
- Компания ABBYY
- Филологический факультет МГУ
- Школа прикладной математики и информатики МФТИ

Международный программный комитет

Богуславский Игорь Михайлович	ИППИ РАН, Россия; Мадридский политехнический университет, Испания
Буате Кристиан	Университет Джозефа Фурье — Гренобль 1, Франция
Гельбух Александр Феликсович	Национальный политехнический институт, Мексика
Иомдин Леонид Лейбович	Институт проблем передачи информации РАН им. А. А. Харкевича, Россия
Кобозева Ирина Михайловна	МГУ им. М. В. Ломоносова, Россия
Козеренко Елена Борисовна	Институт проблем информатики РАН, Россия
Корбетт Гревил	Университет Суррея, Великобритания
Кронгауз Максим Анисимович	НИУ «Высшая школа экономики», Россия
Лукашевич Наталья Валентиновна	НИВЦ МГУ им. М. В. Ломоносова, Россия
Маккарти Диана	Кембриджский университет, Великобритания
Мельчук Игорь Александрович	Монреальский университет, Канада
Нивре Йоаким	Уппсальский университет, Швеция
Ниренбург Сергей	Университет Мэриленда, Балтимор, США
Осипов Геннадий Семёнович	Институт системного анализа РАН, Россия
Райгородский Андрей Михайлович	МФТИ, Школа прикладной математики и информатики, Россия
Раскин Виктор	Университет Пердью, США
Селегей Владимир Павлович	Компания ABBYY, МФТИ, Россия
Хови Эдуард	Университет Карнеги — Меллон, США
Шаров Сергей Александрович	Университет Лидса, Великобритания

Организационный комитет

Селегей Владимир Павлович, *председатель*	Компания ABBYY
Беликов Владимир Иванович	Институт русского языка им. В. В. Виноградова РАН
Браславский Павел Исаакович	Уральский федеральный университет
Добров Борис Викторович	НИВЦ МГУ им. М. В. Ломоносова
Захаров Леонид Михайлович	МГУ им. М. В. Ломоносова
Иомдин Леонид Лейбович	Институт проблем передачи информации РАН им. А. А. Харкевича
Кобозева Ирина Михайловна	МГУ им. М. В. Ломоносова
Козеренко Елена Борисовна	Институт проблем информатики РАН
Лауфер Наталия Исаевна	Компания Yandex
Ляшевская Ольга Николаевна	Институт русского языка им. В. В. Виноградова РАН
Пиперски Александр Чедович	РГГУ
Толдова Светлана Юрьевна	НИУ «Высшая школа экономики»
Федорова Ольга Викторовна	МГУ им. М. В. Ломоносова
Шаров Сергей Александрович	Университет Лидса

Секретариат

Родионова Ольга Игоревна, *координатор оргкомитета*	Компания ABBYY
Ульянова Анна Вячеславовна, *секретарь оргкомитета*	РГГУ

Рецензенты

Tania Avgustinova
Vladimir Benko
Anatoly Gersman
Diana Macartney
Preslav Nakov
Piek Vossen
Антонова Александра Александровна
Азарова Ирина Владимировна
Андрианов Андрей Иванович
Апресян Валентина Юрьевна
Артемова (Черняк) Екатерина Леонидовна
Архангельский Тимофей Александрович
Байтин Алексей Владимирович
Богданов Алексей Владимирович
Богданова-Бегларян Наталья Викторовна
Богуславский Игорь Михайлович
Бочаров Виктор Владиславович
Браславский Павел Исаакович
Васильев Виталий Геннадьевич
Галинская Ирина Евгеньевна
Галицкий Борис Александрович
Гельбух Александр Феликсович
Гращенков Павел Валерьевич
Губин Максим Вадимович
Даниэль Михаил Александрович
Добров Борис Викторович
Добровольский Дмитрий Олегович
Добрушина Нина Роландовна
Добрынин Владимир Юрьевич
Дроганова Кира Андреевна
Зализняк Анна Андреевна
Захаров Леонид Михайлович
Иванов Владимир Владимирович
Иомдин Борис Леонидович
Иомдин Леонид Лейбович
Катинская Анисья Юрьевна
Кибрик Андрей Александрович
Князев Сергей Владимирович
Кобозева Ирина Михайловна
Копотев Михаил Вячеславович
Коротаев Николай Алексеевич
Котельников Евгений Вячеславович
Котов Артемий Александрович
Кронгауз Максим Анисимович
Кутузов Андрей
Левонтина Ирина Борисовна
Леонтьев Алексей Петрович
Лобанов Борис Мефодьевич
Лукашевич Наталья Валентиновна
Лютикова Екатерина Анатольевна
Марков Александр Юрьевич
Мисюрев Алексей Владимирович
Недолужко Анна Юрьевна
Новицкий Валерий Игоревич
Пазельская Анна Германовна
Паперно Денис Аронович
Панченко Александр Иванович
Переверзева Светлана Игоревна
Пивоварова Лидия
Пиперски Александр Чедович
Подлесская Вера Исааковна
Смирнов Иван Валентинович
Смуров Иван Михайлович
Селегей Владимир Павлович
Слюсарь Наталия Анатольевна
Сорокин Алексей Андреевич
Тихомиров Илья Александрович
Толдова Светлана Юрьевна
Урысон Елена Владимировна
Усталов Дмитрий Алексеевич
Федорова Ольга Викторовна
Хохлова Мария Владимировна
Циммерлинг Антон Владимирович
Шаврина Татьяна Олеговна
Шаров Сергей Александрович
Шелманов Артём Олегович

Contents*

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

ПРАГМАТИКА В ИНТЕРПРЕТАЦИИ СФЕР ДЕЙСТВИЯ (НА МАТЕРИАЛЕ ПИСЬМЕННЫХ РУССКИХ ТЕКСТОВ)[1]

Апресян В. Ю. (valentina.apresjan@gmail.com, vapresyan@hse.ru)

Национальный исследовательский университет «Высшая школа экономики»,
Институт русского языка им. В. В. Виноградова РАН

PRAGMATICS IN THE INTERPRETATION OF SCOPE IN WRITTEN RUSSIAN TEXTS

Apresyan V. Ju. (valentina.apresjan@gmail.com, vapresyan@hse.ru)

National Research University Higher School of Economics,
Vinogradov Russian Language Institute of the Russian Academy of Sciences

The paper is a corpus study of pragmatic factors involved in disambiguating sentences with negation and universal quantifier in written Russian and English, such as *Ja ne pozval vseh svoih dal'nih rodstvennikov,* 'I haven't invited all of my distant relatives.' Ambiguity results from differences in scope. If negation scopes over the quantifier, we get partial negation: 'I have invited some, but not all of my distant relatives.' If negation scopes over the verb, we get total negation: 'I haven't invited any of my distant relatives.' Our study is based on Russian and English data extracted from a variety of corpora.

We demonstrate that despite syntactic differences, Russian and English rely on similar mechanisms of disambiguation via pragmatic reasoning. We show that quantifier 'all' has different interpretations with verb vs. quantifier negation: emphatic in the former case and quantificational in the latter. Contextual markers for each reading are consistent with this difference. V-negation occurs with demonstrative pronouns, negatively connoted nouns and temporal modifiers, which add emphasis (*I don't want to talk to all these idiots*; *I haven't eaten all day*), while Q-negation occurs in the context of quantitative verbs that consolidate the interpretation of quantity (*I haven't listed all the options*).

[1] Публикация подготовлена в ходе проведения исследования по проекту «Factors in resolving scope ambiguity» (№ 18-01-0007) в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)» в 2018–2019 гг. и в рамках государственной поддержки ведущих университетов Российской Федерации «5–100».

Certain pragmatically plausible readings are lexicalized in patterns, similar in the two languages and reflecting common background knowledge; e.g. *ne spat' vsju noch'* and *not to sleep all night* both mean 'not to sleep at all during the night'.

In both languages, Q-negation is more frequent than V-negation because of its semantic and pragmatic non-markedness. Q-negation is the default interpretation option which is changed to V-negation in the presence of V-negation markers. Due to syntax, in English its share is much higher than in Russian. Finally, we show that language speakers are able to infer intended scope readings in written language.

Key words: negation, universal quantifier, scope, ambiguity, disambiguation, pragmatics, implicature, background knowledge, construction, phraseme

1. Введение[2]

В работе рассматриваются прагматические факторы, которые влияют на интерпретацию неоднозначных русских предложений с предикатным отрицанием и квантором всеобщности[3] вида

(1) *Петя не позвал всех своих дальних родственников*

Как показано в [Падучева 1974], [Богуславский 1985], подобные предложения регулярно демонстрируют неоднозначность, вызванную возможностью разных сфер действия у предикатного отрицания. Если отрицание имеет сферу действия над глаголом, возникает интерпретация *тотального* отрицания: 'Петя не позвал никого из своих дальних родственников'. Если отрицание воздействует на квантор, возникает интерпретация *частичного* отрицания: 'Петя позвал не всех своих дальних родственников'.

Разным сферам действия соответствуют разные коммуникативные структуры предложения: согласно [Jackendoff 1972: 248–273], отрицание имеет сферу действия над *фокусом* предложения (в дальнейшей работе мы будем использовать в этом же смысле более принятое в отечественной лингвистике понятие *ремы*). Соответственно, при разных интерпретациях сфер действия возникают разные интонационные контуры: как показано в [Падучева 2005], [2011], смещенное отрицание, т. е. такое, у которого синтаксическая сфера действия не совпадает с семантической, привносит в высказывание контрастное фразовое ударение:

(2) *Петя не позвал ВСЕХ \ своих дальних родственников* (а позвал только часть)

[2] Пользуясь случаем, выражаю благодарность анонимным рецензентам за подробные и ценные замечания к первой версии работы, а также моему коллеге Г. А. Морозу за помощь с обработкой статистических данных.

[3] Вслед за [Падучева 1974/2009], [2011], мы используем этот термин в синтаксическом смысле — отрицание, которое стоит при предикате).

Феномену неоднозначных сфер действия семантических операторов посвящена большая лингвистическая литература [Jespersen, 1924]; [Klima, 1964]; [Montague, 1973]; [Hintikka, 1973]; [Cooper, 1979]; [Gil, 1982]; [Aoun and Li, 1989]; [Horn, 1989]; [Partee, 1993]; [Reinhart, 1997]; [Kiss, 2006]. Подробно изучена роль коммуникативной структуры предложения при разрешении неоднозначных СД; ср. работы [Jackendoff, 1972]; [Sgall, Hajičová, and Benesová, 1973]; [Partee, 1991]; [Hajičova, 1998]; [Koizumi 2009]. Известны работы, в которых рассматриваются синтаксические факторы [Kurtzman and MacDonald, 1993] и семантические принципы [Tunstall, 1998], влияющие на процессы интерпретации СД. В последнее время активно изучается роль просодии при определении сфер действия [Ionin, 2010]; [Syrett, Simon and Nisula, 2014]. Однако работы, посвященные изучению прагматических механизмов разрешения неоднозначности сфер действия, особенно на широком корпусном материале, нам неизвестны[4]. При этом представляется, что в письменных текстах, в которых отсутствует просодическая и отчасти коммуникативная информация, для разрешения неоднозначности носители языка опираются в первую очередь на прагматические компетенции.

2. Цели исследования

Целью исследования является изучение прагматических факторов, влияющих на интерпретацию неоднозначных русских предложений с предикатным отрицанием и квантором всеобщности в письменных текстах, в сравнении с аналогичным английским материалом.

3. Гипотезы

Работа отталкивается от следующих стартовых гипотез:

1) русский и английский язык различаются синтаксическими условиями, в которых возникает потенциальная неоднозначность СД в исследуемых предложениях, что связано с тяготением к синтаксически сентенциальному отрицанию в английском [Jespersen 1924], [Klima 1964], [Jackendoff 1969];

2) эти языки имеют одинаковые прагматические механизмы разрешения неоднозначности, которые опираются на фоновые знания и прагматические компетенции носителей языка;

[4] Уже после написания этой работы анонимный рецензент указал нам на интересное корпусное исследование [Tottie & Neukom-Hermann, 2010], в котором рассматривается английская конструкция *all …not*. В нашей работе исследуется конструкция, где отрицание предшествует квантору, таким образом, материал, так же как и многие выводы, несколько различается. Основные тенденции интерпретации сферы действия отрицания, отмеченные в указанной работе, связаны с синтаксическими особенностями субъектной именной группы, содержащей квантор, а не с прагматическими факторами.

3) большая часть потенциально неоднозначных фраз естественно интерпретируется либо с СД отрицания над квантором (Q-отрицание), либо с СД отрицания над глаголом (V-отрицание).

Кроме того, предварительный анализ русской и английской выборки позволил нам сформулировать еще два предположения, которые были проверены в ходе нашей работы:

4) Q-отрицание и V-отрицание различаются прагматически, и каждая из этих интерпретаций имеет свои лексические маркеры, соответствующие их прагматическим особенностям;

5) Q-отрицание существенно частотнее в английском.

Изначально мы также предполагали наличие в русском языке корреляции между интерпретацией и падежной формой именной группы с квантором (генитив для Q-отрицания vs. аккузатив для V-отрицания), однако это предположение не нашло подтверждения.

4. Методы и материалы исследования

Работа опирается на корпусное исследование неоднозначных контекстов. В качестве основного материала используется случайная выборка из примеров, полученных в ответ на запросы *не* + *Verb* + *весь* в Основном корпусе НКРЯ в текстах с 1990 года по настоящее время (http://www.ruscorpora.ru/search-main.html) и *not* + *Verb* + *all* в корпусе современного американского языка COCA (https://www.english-corpora.org/coca/).

Как видно из формулировки запроса, он ограничен предложениями, в которых квантор входит в группу дополнения или обстоятельства, но не подлежащего. Это связано с синтаксическим ограничением: в русском языке смещение семантической сферы действия предикатного отрицания на квантор маловероятно, если квантор входит в группу подлежащего. Большинство предложений, где *весь* относится к подлежащему, интерпретируется с СД отрицания над глаголом, ср.:

(3) *Сегодня весь цех не* [*вышел*] *на работу* (‘никто не вышел’)

Интерпретация с контрастной темой (*Сегодня* [*весь*] *цех не вышел на работу, а вышла только часть*) встречается очень редко. Таким образом, потенциальной неоднозначности практически не создается, и интереса для данной работы подобные предложения не представляют[5].

[5] В работе [Tottie & Neukom-Hermann, 2010] рассматриваются именно конструкции *all…not*, т.е. конструкции с квантором в группе подлежащего, в английском языке. Интересно, что основная тенденция, замеченная нами для конструкции *not…all*, подтверждается и на материале этой конструкции — а именно, превалирование Q-отрицания (хотя и существенно меньшее для квантора в группе подлежащего, нежели в группе дополнения или обстоятельства).

Основная выборка включает по 200 примеров из НКРЯ и COCA, размеченных по СД отрицания. Кроме того, используются иллюстративные примеры из НКРЯ, RuTenTen и EnTenTen (Sketch Engine). Данные НКРЯ, COCA, RuTenTen и EnTenTen также использовались для проверки гипотез, возникавших в ходе работы. Всего было рассмотрено более тысячи языковых примеров.

Помимо корпусного исследования, на материале русского языка был также проведен онлайн-эксперимент с использованием ресурса Toloka на Яндексе (https://toloka.yandex.com/), с целью проверки интуиции разметчика относительно интерпретации сфер действия.

5. Результаты

Ниже представлены возможные интерпретации СД изучаемых фраз, соответствующие им контекстуальные маркеры, частотное распределение интерпретаций и маркеров и интерпретация СД носителями языка.

5.1. Возможные интерпретации СД отрицания и их маркеры

В нашей выборке встретились следующие варианты интерпретаций исследуемых предложений: V-отрицание (СД отрицания над рематизированным глаголом), Q-отрицание (СД отрицания над рематизированным квантором), P-отрицание (СД отрицания над всей пропозицией), C-отрицание (СД отрицания над другой рематизированной составляющей) и ситуация неоднозначности (равная возможность разных СД); ср.:

(4) *Я не* [*СПАЛ*] *всю ночь*[6] */ I didn't* [*SLEEP*] *all night* 'Я вообще не спал' [V-отрицание]

(5) *Я не успел поговорить со* [*ВСЕМИ*] *кандидатами / I didn't have a chance to talk to* [*ALL*] *the candidates* 'Я успел поговорить не со всеми кандидатами' [Q-отрицание]

(6) *Если ты не* [*СКАЖЕШЬ ВСЮ ПРАВДУ*]*, между нами все кончено / If you don't* [*TELL ALL THE TRUTH*]*, we are through*, 'Если ты не скажешь правды вообще или если ты скажешь только часть правды, между нами все кончено' [P-отрицание]

(7) *Я не потратил все эти деньги* [*ЗРЯ*] */ I didn't spend all this money* [*FOR NOTHING*], 'Я потратил все эти деньги не зря' [C-отрицание][7]

---

[6] Скобками обозначается сфера действия отрицания.

[7] Такого рода СД не характерна для русского языка, но встречается в английском. В русском в таких контекстах предпочтительно отрицание непосредственно перед составляющей.

(8) *Не думаю, что они успеют все это сделать / I don't think they'll have time to do all that*, ситуация неоднозначности, прагматически равно возможны интерпретации 'Они не успеют сделать ничего' и 'Они успеют сделать не все'

Отличие примера (6) с P-отрицанием от примера (8) с неоднозначностью СД состоит в следующем. В (6) сформулированное условие выполняется как при полном отсутствии ситуации ('не сказать правду'), так и при ее частичном отсутствии ('сказать не всю правду'), и поэтому говорящий намеренно не выделяет ни одну из интерпретаций как приоритетную. Однако в (8) говорящий явно имеет в виду только одну интерпретацию, но в силу недостаточности контекста, читателю неизвестно, какую именно.

Интерпретации вида (6) и (7) редки в обоих языках, поэтому в дальнейшей части работы речь пойдет о V-отрицании, Q-отрицании и ситуациях неоднозначности между ними. В нашей выборке также встретилось некоторое количество примеров эксплетивного и риторического отрицания (*Если бы я не решил все задачи, я бы не пришел*; *Пока не съешь все, не приходи*), которые мы не рассматриваем, поскольку они не релевантны для наших целей[8].

5.2. V-отрицание vs. Q-отрицание

Помимо семантических и коммуникативных различий, V-отрицание и Q-отрицание различаются прагматически. Рассмотрим фразу (9):

(9) *И если не знать всего случившегося, то и не сообразишь, что этому лицу пришлось пережить* [Галина Щербакова. Моление о Еве (2000)]

Эту фразу можно интерпретировать и с V-отрицанием, и с Q-отрицанием. Если отрицание имеет СД над квантором, то последний имеет чисто *количественную* интерпретацию: 'Если знать не все детали того, что произошло, а только часть, трудно понять, что пришлось пережить этому лицу'. Однако если отрицание имеет СД на глаголом, интерпретация квантора носит *эмфатический* характер: 'Если не знать того очень значительного события, которое произошло, трудно понять, что пришлось пережить этому лицу'. В первом случае фразовое ударение падает на квантор: *Если не знать ВСЕГО случившегося*. Во втором случае просодически маркирован глагол или существительное: *Если не ЗНАТЬ всего случившегося* или *Если не знать всего СЛУЧИВШЕГОСЯ*. Функция квантора при V-отрицании скорее прагматическая; семантически он не является обязательным и может быть опущен: *Если не знать случившегося, трудно понять…*

Лексические маркеры Q-отрицания и V-отрицания согласуются с семантическими и прагматическими особенностями каждой из интерпретаций.

[8] Также называется pleonastic, abusive, formal negation [Jespersen 1940], [Vendryès 1950], [Jack 1977]. Об эксплетивном отрицании в русском см. [Brown 1999], [Barentsen 2014].

5.3. Маркеры Q-отрицания

Основной маркер Q-отрицания — предикаты с количественной семантикой. Это естественно, поскольку Q-отрицание задает количественную интерпретацию квантора. К предикатам с количественной семантикой мы относим глаголы, у которых количественный смысл возникает на каком-то шаге семантического разложения, например *перечислить, исчерпать, вмещать, закончить, включать*, а также их английские аналоги *list, cover, exhaust, complete, include*; ср.:

(10) *Не перечислить* [*всех*] *моих встреч с Маяковским*
[Борис Ефимов. Десять десятилетий (2000)]

(11) *Название книги не исчерпывает* [*всех*] *тем, затронутых автором*
[О. Вильченко. Феномены мышления, интуиции и памяти // «Наука и жизнь», 2008]

(12) *Hunter's letter did not include* [*all*] *of the details of the tentative agreement* (COCA)
'Письмо Хантера не включало всех деталей предварительного соглашения'

(13) *The codified Islamic law did not* [*cover*] *all cases* (COCA)
'Кодифицированное исламское право не покрывало всех случаев'

Практически все вхождения количественных предикатов, входящих в первую двадцатку по частотности встречаемости в исследуемой конструкции, интерпретируются с с Q-отрицанием. Для русского языка это предикаты *исчерпывать, вмещать* и *перечислить*, для английского *to cover* и *to include*. В НКРЯ встретился 91 пример *не исчерпывать* + *весь,* 26 примеров *не вмещать* + *весь* и 24 примера *не перечислить* + *весь*, все вхождения с Q-отрицанием. Как отмечается в [Падучева 2005], смещение отрицания возможно даже тогда, когда оно расположено справа от квантора: [*Всех*] *не перечислишь*. В COCA встретилось 44 примера *not cover* + *all* и 32 примера *not include* + *all*, также с Q-отрицанием.

Хотя семантически с этими предикатами возможно V-отрицание, прагматически оно маловероятно. *Исчерпывать, перечислять, вмещать* предполагают постепенное достижение некоторого результата или состояния, что располагает к интерпретации частичности (Q-отрицание). Трудно представить ситуацию, когда результат или состояние отсутствует полностью (V-отрицание): информационный объект или речевой акт не содержит *никакой* информации, а помещение не вмещает *ничего*. Поэтому невозможно не только V-отрицание этих предикатов, но и их употребление с отрицательными местоимениями: ??*Книга не исчерпывает никаких важных тем,* ??*Зал не вмещает никаких зрителей*, ?*Он не перечислил никаких свойств прямоугольного треугольника*. Для передачи таких отрицательных смыслов будут использованы другие лексические средства: *Книга не затрагивает никаких важных тем, Он не назвал никаких свойств прямоугольного треугольника*. Однако в метафорическом употреблении *вмещать* отчасти теряет семантику количественности и тотальное отрицание, в том числе V-отрицание, становится возможным: ср. *Моя голова*

не вмещает никаких абстрактных понятий, Моя голова не [*вмещает*] *всех этих абстрактных понятий.*

Помимо предикатов с собственно количественной семантикой, есть некоторые другие классы предикатов, которые частотны в сочетании с Q-отрицанием — это ментальные состояния (*знать, понимать*), импликативы (*успеть*), речевые акты (*сказать*). Все они в том или ином виде предполагают количественную оценку: ментальные состояния — объем информации, импликативы — количество и окончательность действий, речевые акты — объем информации. Соответственно, для них естественна интерпретация с Q-отрицанием:

(14) *Мы еще не знаем, что может сломаться, мы не знаем* [*всех*] *реакций*
[Елена Николаева. Наш коллега — робот // «Эксперт», 2015]

(15) *— В субботу не успели* [*все*] *закончить, — ответил Миша*
[Анатолий Рыбаков. Бронзовая птица (1955–1956)]

(16) *До сих пор не сказано* [*всей*] *правды* [Клара Скопина. Пауэрс: 43 года после провокации (2003) // «Наш современник», 2003.12.15]

В английском аналогичные предикаты также в основном встречаются с Q-интерпретацией: *He said he does not know* [*all*] *the details*; *I felt that he hadn't said* [*all*].

Однако в отличие от предикатов с собственно количественной семантикой типа *перечислить*, для которых, как показано выше, V-отрицание затруднено в силу прагматических причин, все прочие перечисленные классы предикатов не навязывают интерпретации Q-отрицания. Как будет показано ниже, при наличии соответствующих маркеров они способны интерпретироваться с V-отрицанием. При этом они являются наиболее частотными заполнителями конструкции с Q-отрицанием. Таким образом, их можно интерпретировать скорее как *колексемы* конструкции с Q-отрицанием [Stefanowitsch & Gries, 2003], нежели как маркеры Q-отрицания.

5.4. Маркеры V-отрицания

У V-отрицания есть два основных типа маркеров, в зависимости от прагматики фразы.

5.4.1. Эмфатические фразы с V-отрицанием

В эмфатических фразах количественная оценка, вносимая квантором *весь*, носит характер модальной рамки, и часто сопровождается отрицательной качественной оценкой, либо же полностью заменяется на нее:

(17) *Черт знает… Как же ты не* [*понимаешь*] *всего этого!*
[Константин Воробьев. Убиты под Москвой (1963)] —
количественная оценка 'не понимаешь чего-то значительного'

(18) *Она не* [*любила*] *все эти невыносимо сладкие пасхи и мазурки*
[З. Н. Гиппиус. Цыганка (1896)] — количественная оценка ‘много’ + качественная оценка ‘плохо’

(19) *Лично я не* [*доверяю*] *всем этим фитотерапевтам да гомеопатам* [Еремей Парнов. Александрийская гемма (1990)] — качественная оценка ‘плохо’

Здесь квантор *весь* имеет чисто прагматическую функцию и может быть опущен. Такое V-отрицание маркируется словами с отрицательной оценкой, часто указательным местоимением *этот* в значении отрицательной оценки. Введение эмфатического *этот* в контекст меняет интерпретацию с Q-отрицания на V-отрицание:

(20) *Я не хочу сегодня говорить со* [*всеми*] *студентами*
(‘хочу говорить не со всеми’, Q-отрицание)

(21) *Я не хочу сегодня* [*говорить*] *со всеми этими студентами*
(‘не хочу говорить ни с кем’, V-отрицание)

V-отрицание также маркируется устойчивыми сочетаниями типа *не хотеть, не любить, не верить* и пр.: *Не верю всем этим политикам*, *Не люблю всех этих актеров*[9]. Возможны и другие показатели отрицательной оценки:

(22) *Крот и мэр, по фамилии Мурцовкин, как бы не* [*замечая*] *всей этой презренной суеты, галантно беседовали в сторонке*
[Валерий Попов. Очаровательное захолустье (2001)]

(23) *При входе в городской транспорт не* [*расталкивай*] *всех локтями*
[Правила поведения в наземном транспорте (2000)]

В английском действуют аналогичные маркеры: *I don’t want to* [*talk*] *to all these idiots, Don’t* [*push aside*] *all these poor children, I don’t want to* [*listen*] *to all these disgusting details, I don’t* [*like*] *all those “keep calm” recommendations.*

Отрицательная оценка как маркер V-отрицания имеет прагматическое объяснение: естественно желание полного, а не частичного, отсутствия нежелательной ситуации.

5.4.2. Нейтральные фразы с V-отрицанием

Второй тип V-отрицательных предложений прагматически более нейтрален, и, хотя квантор *весь* в них также содержит оценку ‘много’, но сохраняет и свое прямое значение всеобщности. Он представлен во фразах, где *весь* входит в обстоятельство времени:

(24) *После вашего отъезда не* [*спала*] *всю ночь: думала, думала*
[Анатолий Алексин. Раздел имущества (1979)] = ‘всю ночь не спала’

В таких фразах квантор несет и семантическую, и прагматическую нагрузку и обычно не может быть опущен: **Я не ела день*, **Его не было дома ночь* vs. *Я не ела весь день, Я не спала всю ночь.*

---

[9] В цитированной работе [Tottie & Neukom-Hermann, 2010] отмечается существование в английском языке устойчивых выражений с Q-отрицанием типа *All is not lost.*

Фразы, где *весь* является частью темпорала, в русском языке по большей части интерпретируются с V-отрицанием, однако для них возможно и Q-отрицание. При этом V-отрицание и Q-отрицание во временных контекстах сочетаются с разными аспектуальными типами предикатов: V-отрицание с пунктивными, Q-отрицание — с длительными; ср. *Он не звонил все выходные* ‘Он ни разу не позвонил за выходные’ vs. *Я не говорила с ним всю ночь* ‘Мой разговор с ним длился не всю ночь’. В зависимости от временного периода и характера действия прагматически вероятна та или иная интерпретация: так, фраза *Я не говорила с ним всю ночь* скорее значит ‘Я говорила с ним не всю ночь’ (Q-отрицание), а фраза *Я не говорила с ним весь день* скорее значит ‘Я ни разу не говорила с ним за день’ (V-отрицание).

Если предикат допускает только пунктивную интерпретацию, он однозначно интерпретируется с V-отрицанием, как в примере (25), если только длительную — он однозначно интерпретируется с Q-отрицанием, как в примере (26):

(25) *Буря и волнение на море не* [*прекращались*] *всю ночь* [И. Резанов. Дым над Кракатау // «Вокруг света», 1984] = ‘всю ночь не прекращались’

(26) *Он топил печи, расчищал снег, таскал воду, колол дрова, но это не занимало* [*весь*] *день* [Марина Палей. Поминовение (1987)] = ‘занимало не весь день’

Маркерами V-отрицания в этом типе предложений являются лексикализованные сочетания предикатов и обозначений временных периодов, маркирующих прагматически вероятные ситуации полного отсутствия действия в течение периода времени: *Я не спал всю ночь, Ребенок не просыпался всю ночь, Она не ела весь день, Шторм не стихал всю ночь* и пр. Краткое исследование фраз с *не + V + всю ночь, не + V + весь день* и их английских аналогов (по 50 случайно выбранных фраз с каждым из темпоралов) дает следующее распределение.

В русском конструкция *не + V + всю ночь* интерпретируется с V-отрицанием в 49 случаях из 50, причем 44 фразы содержат частотное устойчивое сочетание *не спать всю ночь*. В выборке *не + V + весь день* 11 фраз оказались «шумом», а из оставшихся 39 фраз 32 также интерпретируются с V-отрицанием. Наиболее частотное сочетание — *не есть весь день* (8 вхождений). Однако фоновые ожидания относительно дневного времени более разнообразны, и встречаются также *не работать весь день, не быть дома весь день, не отвечать на телефон весь день*: по-видимому, полное отсутствие этих действий в течение дня воспринимается как достойное упоминания нарушение естественных ожиданий.

Если в прочих случаях русский и английский языки демонстрируют высокую степень параллелизма в лексической маркировке Q-отрицания и V-отрицания, то темпоральные обстоятельства обнаруживают несколько большее расхождение между языками. В целом, в английском языке наблюдается более разнообразная картина. В конструкции *not + V + all night* встретилось 29 предложений с V-отрицанием и 21 предложение с Q-отрицанием. Однако в 23 предложениях с V-отрицанием упоминается *not sleeping all night*, что доказывает лексикализацию этого контекста с V-отрицанием также и в английском языке. Другие лексикализованные примеры V-отрицания касаются природных явлений: *The rain didn’t [stop] all night, The wind didn’t [quiet down] all night*. В конструкции

not + V + all day 37 предложений из 50 интерпретируются с V-отрицанием. Как и в русском, самая частотная лексикализация V-отрицания — *not eat all day* (19 вхождений). Другие частые нарушения естественных ожиданий в дневное время касаются человеческой активности и погоды: *I have not [moved] all day* (10 вхождений) и *The rain did not [stop] all day* (8 вхождений).

5.5. Частотное распределение интерпретаций в русском и английском

Общие результаты частотного распределения интерпретаций представлены в таблицах ниже.

Таблица 1. Интерпретации русских фраз с предикатным отрицанием и квантором ***весь***

Интерпретация	Маркер	Кол-во	Общее кол-во	%
Q-отрицание	глаголы с семантикой количества	10	80	40
	глаголы с частой количественной оценкой	49		
	прочее	22		
V-отрицание	обстоятельство времени	14	64	32
	указательное местоимение	13		
	устойчивые сочетания с *не* вида *не любить*	23		
	слова с отрицательной оценкой	10		
	прочее	4		
Неоднозначно		23	23	11,5
P-отрицание	условие	5	7	3,5
	цель	2		
C-отрицание		1	1	0,5
Эксплетивное отрицание		25	25	12,5
Всего			200	100

Для оценки статистических данных был использован критерий хи-квадрат. Для английских данных значение хи-квадрат = 548,56, df = 5, p-value < 2,2e-16, для русских данных значение хи-квадрат = 151, df = 5, p-value < 2,2e-16. Таким образом, в обоих языках количественное распределение разных групп интерпретаций значимо отличается от случайного.

Можно видеть, что в обоих языках Q-отрицание встречается чаще; по-видимому, его следует признать семантически базовым для обоих языков. Это естественно: Q-отрицание прагматически нейтрально, квантор в нем используется в своем прямом значении, в то время как V-отрицание прагматически маркировано, и квантор в нем используется в основном в функции прагматикализованного показателя негативной эмфазы.

Таблица 2. Интерпретации английских фраз с предикатным отрицанием и квантором *all*

Интерпретация	Маркер	Кол-во	Общее кол-во	%
Q-отрицание	глаголы с семантикой количества	5	156	78
	глаголы с частой количественной оценкой	76		
	прочее	75		
V-отрицание	обстоятельства времени	3	20	10
	указательное местоимение	7		
	устойчивые сочетания с *не* вида *не любить*	1		
	слова с отрицательной оценкой	6		
	прочее	3		
Неоднозначно		8	8	4
P-отрицание	условие	1	2	1
	цель	1		
C-отрицание		2	2	1
Эксплетивное отрицание		12	12	6
Всего			200	100

Интересно, что в то время как Q-отрицание намного чаще встречается без специальных маркеров (т. е. предикатов с количественной семантикой), V-отрицание практически всегда лексически маркировано. Таким образом, без маркеров V-отрицания фразы в обоих языках скорее интерпретируются с Q-отрицанием. Однако привнесение в контекст негативной эмфазы (местоимения *этот*, слов с отрицательной оценкой типа *идиот* и устойчивых сочетаний с *не* типа *не любить*) практически гарантирует интерпретацию с V-отрицанием. В контекстах с темпоральным обстоятельством интерпретация определяется фоновыми знаниями говорящих и часто лексикализована (*не спать всю ночь* — V-отрицание)[10].

При этом между русским и английским языками есть большое количественное различие: в русском соотношение Q-отрицания и V-отрицания 40% vs. 32%, а в английском 78% к 10%. Кроме того, в русском языке Q-отрицание чаще лексически маркируется, чем в английском. Можно сделать вывод о том,

[10] Интересно, что в русском языке не обнаружилось влияния морфологии на интерпретацию, а именно корреляции между генитивом (как падежом, выражающим идею партитивности) и Q-отрицанием, с одной стороны, и аккузативом (как падежом, выражающим идею определенности) и V-отрицанием, с другой. Именные группы с квантором могут интерпретироваться обоими способами как в генитиве, так и в аккузативе. Как было сказано, дефолтной интерпретацией является Q-отрицание; введение в контекст маркеров V-отрицания приводит к смене интерпретации: *Она еще не знает* [*всех*] *деталей этого плана, Она еще не знает* [*все*] *детали этого плана* (Q-отрицание) vs. *Она еще не* [*знает*] *всех этих отвратительных деталей, Она еще не* [*знает*] *все эти отвратительные детали* (V-отрицание).

что в английском языке Q-отрицание представляет собой дефолтную интерпретацию для такого рода конструкций, в том время как в русском языке картина несколько сложнее.

Это объясняется синтаксическими различиями данных двух языков, а именно, существенно большей распространенностью сентенциального отрицания в английском, многократно отмечавшейся в литературе [Jespersen 1924], [Klima 1964], [Jackendoff 1969]. Доля Q-отрицательных интерпретаций в английском языке может, таким образом, повышаться за счет того, что в этом языке конструкция с синтаксически сентенциальным отрицанием является существенно более нейтральным, а иногда и единственно возможным способом выражения Q-отрицания; в русском языке, напротив, во многих случаях Q-отрицание выражается постановкой частицы *не* перед квантором.

Это различие можно заметить в переводах параллельного корпуса НКРЯ, где английские предложения с сентенциальным отрицанием часто переводятся на русский предложениями с отрицанием перед квантором:

(27) *You have not said* [*all*] *that you know* [J. R. R. Tolkien. The Lord of the Rings (1954)]

(28) *Вы рассказали не* [*все*]*, что знаете* [Дж. Р. Р. Толкин. Властелин колец: Две башни (М. Каменкович, В. Каррик, 1994)]

Преобладание сентенциального отрицания в английском приводит к ошибкам у англоязычных студентов, изучающих русский язык: они используют сентенциальное отрицание там, где в русском предпочтительна постановка отрицательной частицы перед квантором:

(29) ?*Я не понимала всё, что он сказал* (Russian Learner Corpus, Val (F, FL, IM) | eng | FL | 2011–2012), при более естественном *Я понимала не всё, что он сказал*

5.6. Эксперимент

Целью эксперимента была проверка предлагаемых нами интерпретаций СД в русской выборке[11]. В качестве стимулов было выбрано по 10 предложений из групп, размеченных нами как Q-отрицание, V-отрицание и неоднозначность, т. е. всего 30 предложений. Они были расположены в случайном порядке, который менялся для каждого участника. Для каждого предложения участник эксперимента должен был выбрать один из четырех возможных ответов: V-отрицание, Q-отрицание, «возможны обе интерпретации», «не могу ответить». Порядок ответов был произвольным и менялся для каждого информанта; ср. образец стимула:

(30) *Судя по всему, его не было всю ночь дома, или он просто не подходил к телефону*

[11] Для английской выборки, ввиду отсутствия аналогичной платформы, пришлось ограничиться проверкой с помощью интуиций информантов-носителей американского английского языка (Мадлен Креслин, Бела Шаевич).

Выберите один из ответов:

а. Он отсутствовал часть ночи
б. Его вообще не было дома ночью
в. возможны обе интерпретации
д. не могу ответить

В эксперименте приняло участие 6490 информантов. Результаты представлены в **Таблице 3**.

Таблица 3. Распределение ответов в разных группах СД[12]

Группа по СД	V-отрицание: *не [V] весь*	Q-отрицание: *не V [весь]*	Возможны обе интерпретации	Не могу ответить	Всего
Неоднозначность	2027 (31,2%)	3389 (52,3%)	915 (14%)	159 (2,5%)	6490
V-отрицание	**5323 (82%)**	502 (7,7%)	476 (7,3%)	189 (3%)	6490
Q-отрицание	568 (8,7%)	**5317 (82%)**	480 (7,3%)	125 (2%)	6490

Как видно из таблицы, предложенные нами интерпретации СД по большей части соответствуют интуициям информантов (82% «правильных ответов» для V-отрицания и Q-отрицания). В предложениях, которые мы определили как неоднозначные, представлен ожидаемо больший разброс интерпретаций: 31% приходится на V-отрицание, 52% на Q-отрицание и 14% участников считают, что возможны обе интерпретации. Преобладание Q-отрицания у неоднозначных фраз можно отнести за счет семантической базовости этой интерпретации. Опция «не могу ответить» во всех группах имеет приблизительно равное значение и, по-видимому, не является информативной.

6. Заключение

В ходе исследования было продемонстрировано, что в синтаксически разных языках действуют общие прагматические принципы разрешения неоднозначности СД. Прагматически естественные интерпретации отчасти лексикализованы, и у каждой из основных интерпретаций — Q-отрицания и V-отрицания — есть собственные лексические маркеры, соответствующие прагматике этих интерпретаций.

При этом можно установить некоторую иерархию факторов: в отсутствие маркеров V-отрицания фраза интерпретируется с Q-отрицанием; наличие маркеров V-отрицания меняет интерпретацию, за исключением случаев с темпоральными обстоятельствами. Последние либо лексикализованы с V-отрицанием (*не спать всю ночь*), либо требуют фоновых знаний для правильной интерпретации (*Не сиди с ним всю ночь* — Q-отрицание).

[12] Для оценки статистических данных был использован критерий хи-квадрат, со следующим результатом: значение хи-квадрат = 8540,2, df = 6, p-value < 2,2e-16. Таким образом, количественные различия в разных группах интерпретаций значимы.

Литература

1. *Aoun, Joseph, and Yen-hui Audrey Li* (1989). Scope and Constituency. Linguistic Inquiry. 20(2), p. 141–172.
2. *Barentsen, Adriaan, A.* (2014). Problemy opisanija sojuza poka. [Problems of the Description of the connective poka.] Die Welt der Slaven. 55, p. 377–410.
3. *Boguslavsky, Igor* (1985). Research in the semantics of syntax [Issledovanija po sintaksicheskoj semantike]. Moscow, Nauka.
4. *Boguslavsky, Igor* (1996). Sfera dejstvija leksicheskix edinic [Scope of lexical items]. Moscow.
5. *Brown, Sue* (1999). The syntax of negation in Russian. A Minimalist approach. CSLI. Publications. Stanford Monographs in Linguistics.
6. *Grice, Paul* (1975). Logic and conversation. In P. Cole & J. Morgan (ed.), Syntax and Semantics, 3: Speech Acts. pp. 41–58, New York: Academic Press.
7. *Hajičova, Eva* (1998). Topic-Focus Articulation, Tripartite Structures and Semantic Content. Kluwer, Dordrecht.
8. *Hintikka, Jaakko* (1973). Quantifiers vs. quantification theory. Dialectica, 27, p. 329–358. Reprinted in Linguistic Inquiry 5 (1974):153–177.
9. *Horn, Laurence* (1989). A Natural History of Negation. University of Chicago Press, Chicago.
10. *Ionin, Tania* (2010). The scope of indefinites: an experimental investigation. Natural Language Semantics, 18 (3). 295–350.
11. *Jack, G. B.* (1977). Negation in later Middle English prose. Archivum Linguisticum. 9, p. 58–72.
12. *Jespersen, Otto* (1924). The philosophy of Grammar. London, 1924.
13. *Jespersen, Otto* (1940). A modern English grammar on historical principles. Copenhagen: Munksgaard.
14. *Jackendoff, Ray* (1969). An interpretive theory of negation. In Foundations of Language, 5: 218–241.
15. *Jackendoff, Ray* (1972). Semantic Interpretation in Generative Grammar. Cambridge, Mass.: MIT Press.
16. *Kadmon, N., Roberts, C.* (1986). Prosody and scope: The role of discourse structure. CLS Proceedings.
17. *Koizumi, Y.* (2009). Processing the not-because ambiguity in English: the role of pragmatics and prosody. CUNY thesis.
18. *Kiss, K. É.* (2006). Quantifier Scope Ambiguities. In: The Blackwell Companion to Syntax. Everaert, M. and H. van Riemsdijk (eds). Blackwell.
19. *Klima, Ed.* (1964). Negation in English. In J. A. Fodor and J. J. Katz (eds.), The Structure of Language, p. 246–323.
20. *Kurtzman, H. S., and MacDonald, M. C.* (1993). Resolution of quantifier scope ambiguities. Cognition, 48, 243–279.
21. *Paducheva, Elena* (1974). On the semantics of syntax. Materials for the transformational grammar of Russian [O semantike sintaksisa. Materialy k transformacionnoj grammatike russkogo yazyka]. Moscow: Nauka.

22. *Paducheva, Elena* (2005). Effects of suspended assertion: global negation. [Effekty snyatoj utverditel'nosti: global'noe otricanie]. In The Russian Language in a Scientific light [Russkij yazyk v nauchnom osveshchenii]. 2 (10), p/ 17–42.
23. *Paducheva, Elena* (2014). Non-standard negation in Russian: external, displaced, global, radical. [Nestandartnye otricanija v russkom jazyke: vneshnee, smeshhennoe, global'noe, radikal'noe]. In Issues in Linguistics [Voprosy jazykoznanija]. 5, p. 3–23.
24. *Partee, Barbara H.* (1993). On the 'Scope of Negation' and polarity sensitivity. In Functional Description of Language: Proceedings of the Conference, Prague, November 24–27, 1992, ed. Eva Hajicova, 179–196. Prague: Faculty of Mathematics and Physics, Charles University.
25. *Reinhart, Tanya* (1997). Quantifier Scope: How Labour is Divided between QR and Choice Functions. In Linguistics and Philosophy 20, p. 335–397.
26. *Sgall, P., Hajičová, E. & Benesová, E.* (1973). Topic, Focus and Generative Semantics. Kronberg, Taunus: Scriptor.
27. *Stefanowitsch, Anatol & Stefan Th. Gries* (2003). Collostructions: Investigating the interaction between words and constructions. International Journal of Corpus Linguistics. 8.2:209–43.
28. *Syrett, K., Simon, G., & Nisula, K.* (2014). Prosodic disambiguation of scopally ambiguous quantificational sentences in a discourse context. Journal of Linguistics, 50(2), 453–493.
29. *Tottie, G.; Neukom-Hermann, A.* (2010). Quantifier-negation interaction in English: A corpus linguistic study of all...not constructions. In: Horn, L. R. The Expression of Negation. Berlin, New York, 149–185.
30. *Tunstall, Susanne Lynn* (1998). The interpretation of quantifiers: Semantics and processing. Doctoral Dissertations Available from Proquest. AAI9909228, https://scholarworks.umass.edu/dissertations/AAI9909228.
31. *Vendryès, Joseph* (1950). Sur la nègation abusive. Bulletin de la Société de Linguistique de Paris. 46, p. 1–18.
32. *Zeijlstra, Hedde* (2004). Sentential Negation and Negative Concord. Utrecht: LOT.

СЕМАНТИЧЕСКИЕ ТИПЫ ИМПЛИКАТУР И УСЛОВИЯ ИХ ВОЗНИКНОВЕНИЯ (НА МАТЕРИАЛЕ КОРПУСА ГАЗЕТНЫХ ЗАГОЛОВКОВ)[1]

Апресян В. Ю. (valentina.apresjan@gmail.com, vapresyan@hse.ru)

Национальный исследовательский университет «Высшая школа экономики», Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Орлов А. В. (alexander.orlov98@gmail.com)

Национальный исследовательский университет «Высшая школа экономики», Москва, Россия

SEMANTIC TYPES OF IMPLICATURES AND THEIR CONTEXTUAL TRIGGERS (BASED ON THE CORPUS OF NEWS HEADLINES)

Apresyan V. Ju. (valentina.apresjan@gmail.com, vapresyan@hse.ru)

National Research University “Higher School of Economics”, Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

Orlov A. V. (alexander.orlov98@gmail.com)

National Research University “Higher School of Economics”, Moscow, Russia

The paper aims at contributing to a typology of implicatures via their analysis in news headlines. By implicatures we mean cancellable implicit senses, irrespectively of whether they are inherent in lexical meanings or occur in certain contextual conditions. While generally implicatures are difficult to tie to a particular type of lexical environment, our analysis of headlines allows us to make a step in this direction. Headlines often use implicatures instead of assertions to convey information about the content of the article.

[1] Данное исследование частично финансировано из средств гранта 19-012-00291А, «Подготовка четвертого выпуска Активного словаря русского языка», рук. Б. Л. Иомдин.

Causal implicatures are the most frequent type in our sample. We study two types of causal implicatures. The first occurs in sentences with predicates that have a semantic argument of Cause, syntactically unexpressed in the sentence. If either the noun attribute or the noun itself contains an element of value judgment, it can be interpreted as filling the Cause argument of the predicate: to reward the hero (= 'to reward a person for heroism'), to punish the criminal (= 'to punish a person for the crime'). When Cause is thus expressed, it is an implicature and is cancellable: He rewarded the winner of the sports contest, yet not for the victory, but for volunteer work in a hospice. Another type of causal implicatures occurs in utterances with expressions of temporal sequence, such as after: After their quarrel she called it quits (= 'Because of their quarrel, she decided to break up with him'). While in some languages causal implicatures of temporal prepositions are grammaticalized as new lexical meanings, Russian temporal prepositions do not develop separate causal senses. This makes them an ideal context for causal implicatures, and headlines use posle 'after' to imply a causal relationship between the events described in the article, without committing the author to a definite statement to this effect. We also consider qualitative and factual implicatures which occur in certain specific contexts.

Key words: implicature, cause, semantic actant, syntactic actant, manipulation, Gricean maxim, relevance maxim, quantity maxim, scope

1. Введение

В работе исследуются контексты, способствующие возникновению импликатур в дискурсе. Под импликатурами мы понимаем, в соответствии с [Grice 1975], такие имплицитные смыслы, которые могут отменяться последующим контекстом. Именно отменяемость представляется нам главным отличительным свойством импликатур, противопоставляющим их другим видам смыслов[2].

В литературе принято противопоставлять conventional implicatures, т. е., импликатуры, задаваемые значением конкретных языковых единиц, и conversational implicatures, задаваемые более широким контекстом. Единого понимания того, какие именно элементы значения считать конвенциональными импликатурами, в лингвистике не существует. В [Levinson 1983:127–128] conventional implicature определяется как 'part of a lexical item's or expression's agreed meaning, rather than derived from principles of language use, and not part of the conditions for the truth of the item or expression'[3], т. е. как такой смысл, который не изменяет условий истинности высказывания (например, импликатура усматривается у союза *but* 'но'). Таким образом, под это определение попадают и компоненты

[2] Как известно, пресуппозиции сохраняются под отрицанием [Langendoen and Savin, 1971]: фразы Я знал, что он приехал и Я не знал, что он приехал содержат пресуппозицию — 'Он приехал'. Импликации [Karttunen 1971] под отрицанием меняются на противоположные смыслы, т. е. ведут себя как ассерции: Ему удалось решить задачу 'Он решил задачу', Ему не удалось решить задачу 'Он не решил задачу'.

[3] Похожее понимание предлагается в [Karttunen and Peters 1979].

значения, которые в некоторых подходах описываются как модальные рамки [Апресян 2009:514], т. е. смыслы, передающие *отношение* говорящего к действительности; ср. *Он съел всего два арбуза* vs. *Он съел целых два арбуза*.

В нашей работе мы понимаем под конвенциональными импликатурами только такие имплицитные компоненты лексического значения, которые подразумеваются в нейтральных контекстных условиях, но отменяются эксплицитно противоречащими контекстами. Такого рода смыслы Ю. Д. Апресян называет «слабыми смыслами»: «такой семантический компонент в ассертивной части толкования лексемы, который подавляется (вычеркивается) в толковании противоречащим ему семантическим компонентом, выраженным отдельной лексемой» [Апресян 2009:540]; ср. *Он проявил вкус* (= ‘хороший’) vs. *Он проявил плохой вкус*[4].

Относительно conversational implicatures в лингвистике также нет общепринятой точки зрения. Так, некоторые имплицитные смысловые компоненты высказывания, которые восстанавливаются адресатом из контекста (например, при эллипсисе, употреблении дейктических слов и пр.) некоторыми учеными определяются как *экспликатуры*: “An explicature is an ostensively communicated assumption that is inferentially developed from one of the incomplete conceptual representations (logical forms) encoded by the utterance” [Carston 2002], ср. также [Sperber & Wilson 1996].

Однако для нашей работы разграничение конвенциональных и разговорных импликатур, а также иные тонкие терминологические различия несущественны. Важными критериями для нас являются имплицитность и отменяемость смысла, т. к. именно эти свойства определяют функционирование импликатур в дискурсе. Поэтому в дальнейшем мы будем называть импликатурами любые *отменяемые имплицитные смыслы*, независимо от того, привносятся ли они значением лексических единиц, либо возникают на стыке лексического значения и контекстных условий.

Целью работы является исследование семантических типов импликатур, а также условий, в которых они возникают. Для других разновидностей имплицитных смыслов такие типологии существуют: так, известны триггеры пресуппозиций существования [Karttunen 1974], а также классы слов с указанием на наблюдателя [Апресян 1986], [Падучева 2013, 2018]. Однако импликатуры, в силу своей меньшей привязанности к конкретным лексическим единицам, труднее поддаются классификации. Одним из немногих хорошо известных контекстов возникновения скалярных импликатур является контекст с кванторами, описанный в [Levinson 1983], [Horn 1984].

В данной работе мы выделяем некоторые другие семантические типы импликатур и рассматриваем условия их возникновения. Изначальным материалом для нашего исследования послужил Корпус газетных заголовков (КГЗ), описанный в [Орлов 2019]. В КГЗ содержится множество примеров вводящих в заблуждение заголовков. Под вводящими в заблуждение понимаются такие

[4] Мы, однако, считаем, что слабые смыслы не входят в ассертивную часть значения: так, слово вкус в обсуждаемом значении мы бы толковали как ‘способность к эстетической оценке’, а не как ‘хорошая способность к эстетической оценке’.

заголовки, которые имеют несколько возможных прочтений, наиболее естественное из которых создает у читателя неверные ожидания относительно содержания статьи: а именно, что речь в статье пойдет о более интересной и значимой ситуации, чем та, что в действительности описана на странице издания.

Изначальная цель создания корпуса заключалась в сборе обучающего материала для нейросетей. Этот материал должен послужить основой для создания программного обеспечения, которое могло бы автоматически определять вводящие в заблуждение заголовки на русском языке. Первые попытки создания подобного ПО для английского при помощи алгоритмов, включающих работу нейросетей, описаны в [Anand et al. 2016], [Wei and Wan 2017].

Однако КГЗ представляет ценность и для лингвистов-теоретиков, в частности для исследователей, интересующихся механизмами возникновения импликатур. В корпусе была проведена классификация лингвистических приемов, которыми пользуются авторы для создания подобного рода заголовков. Одним из самых популярных приемов по частоте использования[5] является *ложная импликатура*, т. е. такая импликатура, которая задается контекстом заголовка, но опровергается содержанием статьи. Ценность работы с КГЗ заключается в том, что импликатуры в новостных заголовках не отличаются от общеязыковых по механизмам возникновения, но отмена импликатуры позволяет заметить ее наличие. В КГЗ встретились три основных семантических типа ложных импликатур — причинные, (контр)фактивные и качественные. Помимо КГЗ, мы используем для иллюстрации примеры из других источников, т. к. описываемые нами механизмы носят общеязыковой характер.

2. Причинные импликатуры

Причинные импликатуры являются наиболее частыми в нашей выборке (40 % от всех примеров с ложными импликатурами). Нам встретилось два типа причинных импликатур.

2.1. Причинные импликатуры, возникающие при невыраженности валентности причины и наличии оценочного компонента

Данный тип причинных импликатур возникает, когда в предложениях с предикатами, имеющими валентность причины, она не выражена стандартным способом, т. е. обстоятельственной группой с причинным предлогом. Ср. заголовок:

(1) *Путин наградил кричавшего «Слава Украине» хорвата* [15.07.2018, lenta.ru].

[5] На 15.04.2019 в КГЗ содержится 195 вводящих в заблуждение заголовков. Из них 38 (каждый пятый) содержат ложную импликатуру. См. более подробную статистику для каждого из приёмов и пример разметки заголовков в разделе Appendix.

Он апеллирует к ситуации, возникшей во время чемпионата мира по футболу, который проводился в 2018 году в России. Хорватский футболист Домагой Вида после победы в четвертьфинале выкрикнул «Слава Украине». В заголовке создается ложная импликатура ‘Хорвата наградили за то, что он кричал «Слава Украине»’. На самом деле футболист был награжден серебряной медалью в составе хорватской сборной за второе место в ЧМ. Если бы валентность причины была выражена ассертивно (группой *за + СУЩ*), заголовок был бы ложным. Однако поскольку прямое значение атрибутивной конструкции не указывает на причину, и причинный смысл имеет статус импликатуры, формально автор статьи не нарушает условий истинности.

Выражение валентности причины при помощи импликатуры, вводимой оценкой в объектной именной группе, представляет собой весьма распространенный общеязыковой феномен; ср. следующие фразы:

(2) *Он наказал дерзкого мальчишку*

(3) *Он наградил победителей*

В отсутствие стандартных средств выражения каузальной связи, оценочный компонент осмысляется как причинный и возникает импликатура: ‘наказал за дерзость’, ‘наградил за победу’. Таким образом, если в лексическое значение предиката входит валентность причины, но синтаксически данная валентность остается невыраженной, оценочные компоненты контекста воспринимаются как выражающие причину.

Если оценочных компонентов в высказывании нет, то претендентов на роль заполнителя валентности причины не образуется, и предложение воспринимается как эллиптическое:

(4) *Он наказал белобрысого мальчишку*

(5) *Он наградил семиклассников*

Отсутствие стандартного выражения валентности причины и осмысление оценочных компонентов значения как причинных создает именно импликатуру, а не ассертивный смысл. Ср. примеры, где импликатура, вводимая оценочным компонентом, отменяется последующим стандартным выражением причины:

(6) *Хотел наказать ведьму за то, что его жену погубила* [В. Бурлак],

отмененная импликатура ‘за то, что она ведьма’;

(7) *Поступок Курбского, но более всего его письма и невозможность наказать «беглого раба» за дерзость довели раздражительного и подозрительного царя до высшей степени злости и тиранства* [Н. И. Костомаров],

отмененная импликатура ‘за бегство’.

Помимо обозначений наград и наказаний, такого рода импликатуры возникают и у других семантических классов глаголов со значением эмоциональных или поведенческих реакций, вызванных чьими-то действиями и направленных на их субъекта, в условиях синтаксической невыраженности

валентности причины и наличии оценочного компонента в контексте. Сюда входят также гнев, жалость, месть, брань, критика, похвалы и благодарность: *жалеть*; *сердиться, злиться*; *заплатить, отплатить, мстить*; *(при)грозить*; *преследовать (за что-то), предать анафеме, подвергать гонениям*; *ругать, бранить, выговаривать, критиковать, пенять*; *(о)штрафовать*; *врезать, побить, выпороть*; *(от)благодарить*; *вознаградить, (по)жаловать* и др. Ср. *похвалить старательного студента за неожиданно творчески выполненное задание*, где возникает отменяемая группой с *за* импликатура ‘похвалить за старательность’.

При этом потенциал оценочных компонентов в создании причинных импликатур столь велик, что они могут возникать и тогда, когда в высказывании употреблен предикат, не имеющий валентности причины:

(8) *Этот негодяй ее предал* (предал, потому что негодяй)

(9) *Смышленый паренек хорошо справился с заданием* (хорошо справился потому, что смышленый)

Описанные нами причинные импликатуры носят контекстный характер и возникают на стыке лексического значения (в контексте предикатов со значением причины и оценочных слов) и синтаксических факторов (невыраженность валентности причины).

Сходное с этим явление описано в работе [Шмелев 1996:206], где обсуждаются эффекты индексальной референции при помощи качественных имен; ср. пример из данной работы:

(10) *Не подам руки грязнулям* (‘не подам потому, что грязнули’)
[Ю. Тувим, пер. С. Михалкова]

2.2. Конвенциональные причинные импликатуры у показателей времени

Другой тип причинных импликатур, часто встречающийся в новостных заголовках, — это конвенциональные импликатуры у показателей временного следования. Известно, что таксисные предлоги развивают причинные импликатуры:

(11) *После аварии она боится ездить на машине* (= ‘из-за аварии’)

В некоторых языках маркеры времени и пространства грамматикализуются в показатели логических отношений [Heine, Kuteva 2004:276]. Однако в русском языке каузальные связи, которые вводятся темпоралами типа *после*, отменимы.

(12) *После похода на концерт моей любимой группы, я перестал спокойно спать по ночам… Но я не думаю, что это как-то связано*

За счет того, что причинно-следственная связь у *после* является импликатурой, а не ассерцией, это слово, в отличие от большинства причинно-следственных союзов, может выражать опосредованную каузальную связь; ср. заголовок:

(13) *В Подмосковье девочка погибла после игры с братом*
[Российская газета, 14.01.2019]

Возникающая причинная импликатура ‘Игра с братом привела к гибели девочки’ верна, поскольку из статьи мы узнаём, что в результате активных движений детей во время игры на девочку упал шкаф и придавил её. Тем не менее, в данном контексте временной предлог *после* не может быть заменен на каузальные предлоги *из-за* или *в результате*, поскольку они выражают ассертивный смысл непосредственной причинно-следственной зависимости:

(14) ??*В Подмосковье девочка погибла из-за игры с братом*

(15) ?*В Подмосковье девочка погибла в результате игры с братом*

При этом обратная замена, т. е. обозначение прямой причинной связи темпоральным предлогом, возможна:

(16) *В мире уже зарегистрированы случаи заражения детей СПИДом в результате* (okпосле) *игр со шприцами, попадающими в бытовые контейнеры и на свалки* [Е. Любешкина. Обратная сторона упаковки // «Наука и жизнь», 2007]

СМИ часто используют предлог *после* для создания истинной импликатуры, чтобы избежать более обязывающего ассертивного выражения каузальной связи:

(17) *Юлия Латынина уехала из России после поджога своего автомобиля*
[09.09.2018, Republic]

3. (Контр)фактивные импликатуры

Под (контр)фактивными импликатурами мы понимаем разновидность импликатур, подразумевающих (не)совершение действия, о котором идет речь. Контрфактивную импликатуру можно проиллюстрировать следующим примером из КГЗ:

(18) *Матвиенко рассказала об отказе Путину и вспомнила о лучших годах жизни* [https://lenta.ru/news/2019/04/07/matvienko/]

Пример (18) содержит два типа манипуляции — игру слов, основанную на полисемии, и импликатуру. В отсутствие уточнения, *отказ* в данном контексте воспринимается как отвергнутое матромониальное предложение. Импликатура же подразумевает, что отказ был окончательным и действие не было совершено. Из статьи, однако, выясняется, что речь шла не брачном предложении, а о просьбе занять должность губернатора Санкт-Петербурга, и что после нескольких отказов Матвиенко согласилась (как известно, она некоторое время была губернатором).

Таким образом, отглагольное существительное *отказ*, как и глагол *отказать*, только имплицирует, что действие, о котором просили, не было произведено, поскольку этот смысл отменяется содержанием статьи.

В не-игровом дискурсе контрфактивная импликатура, вносимая глаголом *отказать* в форме СОВ ПРОШ, обычно истинна и подтверждается дальнейшим контекстом; ср. пример из НКРЯ:

(19) *Он пришел в строительную контору, просил аванс, и ему отказали. Последняя капля. Вышел на крыльцо и полоснул себя ножом по горлу* [С. Алексиевич],

импликатура: ‘он не получил аванс’.

Однако достаточно многочисленны и примеры, где речь идет об отмене отказа:

(20) *В педагогическом институте в Токио* [...] *ему сначала отказали из-за его ужасных синяков и распухшего носа, но он был настойчив, и в конце концов его допустили к экзаменам* [А. Геласимов],

отмененная импликатура ‘его не допустили к экзаменам’.

Таким образом, у глагола *отказать* смысл, что действие не состоялось, является импликатурой. Интересно, что у антонимичного ему глагола *согласиться* фактивная импликатура в форме СОВ ПРОШ, по-видимому, возникает не всегда или является более слабой: так, фраза (21) не предполагает, что действие, о котором просили, было сделано.

(21) *Правительство согласилось пойти на уступки* [«Еженедельный журнал», 2003.03.24] — *не предполагается, что к моменту речи правительство пошло на уступки*

С другой стороны, приблизительные конверсивы глагола *согласиться* глаголы *убедить* и *уговорить* в форме СОВ ПРОШ являются, как известно, перлокутивными, т. е. обозначают речевой акт с достигнутым результатом [Гловинская 1993, 2001]:

(22) *Его убедили остаться в Москве и поручили ему переводить на славянский язык греческие полемические книги* [В. О. Ключевский] = *‘он остался в Москве’*

(23) *Дмитрий Николаевич меня не пустил, уговорил остаться обедать, после чего засадил меня за карты* [Н. Варенцов] = *‘я остался обедать’*

При этом фактивный смысл в значении *убедить* и *уговорить* — не импликатура и не может быть отменен:

(24) **Его убедили/уговорили прийти, но он не пришел, при возможности*

(25) *Он согласился прийти, но не пришел*

Таким образом, близкие по семантике глаголы могут отличаться с точки зрения того, какой статус у них имеет смысл (не)совершения действия. Подобным образом, неречевые глаголы *увильнуть* и *уклониться*, во многом близкие по семантике к речевому глаголу *отказаться*, имеют другое семантическое устройство. Они являются импликативными [Karttunen 1971], и у них контрфактивный смысл также не может быть отменен:

(26) *Он увильнул от субботника / уклонился от участия в субботнике* = *‘Он не пришел на субботник’, при невозможности*

(27) **Он сначала увильнул от субботника / уклонился от участия в субботнике, но потом все же пришел*

4. Качественные импликатуры

Под качественными импликатурами мы подразумеваем такие импликатуры, которые содержат оценку (обычно положительную). Они часто возникают в контексте лексических единиц со значением умения. Приведем пример из КГЗ:

(28) *Режиссер «Черного лебедя» заговорил по-русски после Большого театра* [https://lenta.ru/news/2018/06/30/aronofsky/]

Как и в примере с *отказом*, данный заголовок содержит игру слов, основанную на полисемии (*заговорить* = ‘начать произносить’ или ‘приобрести умение говорить’[6]), а также, применительно ко второму пониманию, вводит импликатуру ‘Режиссер стал нормально говорить по-русски’[7]. При этом в статье речь идет лишь о том, что Даррен Аронофски написал в своем Instagram несколько хвалебных строк по-русски после просмотра спектакля, т. е. импликатура отменяется содержанием статьи. Подобного рода тривиальные качественные импликатуры характерны и для других слов со значением умений и способностей: *Он знает французский* (‘хорошо знает’), *Он разбирается в машинах* (‘хорошо разбирается’), *Он умеет кататься на горных лыжах* (‘хорошо умеет’), *Он говорит по-итальянски* (‘хорошо говорит’), *У него способности* (‘хорошие способности’). Все они отменяются показателями отрицательной оценки: *Он плохо знает французский*, *У него плохие способности* и т. д.

Как сказано выше, мы считаем, что данные импликатуры имеют ту же природу, что слабые положительные смыслы, описанные в работах Ю. Д. Апресяна применительно к словам *соображать, варить* (*У него котелок варит*), *вкус, класс* и некоторым другим, которые тоже могут быть отменены контекстом: *Он плохо соображает, У него котелок плохо варит, У нее плохой вкус* [Апресян 2009: 540]. Однако существует и большой класс оценочных слов со значением умений, у которых положительная оценка существенно более явно выражена и поэтому входит в ассерцию и не отменяется (*умелец, дар, талант, гений* и пр.).

Помимо обозначений умений, подобные импликатуры свойственны многим словам со значением профессий. Ср. пример *Я тоже артист, хотя плохой* [И. С. Тургенев], приводимый в Малом академическом словаре русского языка для иллюстрации противительного значения союза *хотя*. Интересно здесь то,

---

[6] Ср. лексикографическое описание глагола говорить в Активном словаре русского языка [Гловинская 2010].

[7] Интересно, что заголовок содержит и причинную импликатуру, вводимую таксисным предлогом после: ‘Режиссер овладел русским языком в результате визита в Большой театр’.

что странно сказать ??*Он артист, хотя хороший.* Видимо, это связано именно с положительной импликатурой, содержащейся в слове *артист*. То же можно сказать относительно слов *писатель, художник, поэт* и многих других; они возможны в контексте противительных союзов с показателями отрицательной оценки, но не положительной: *Он писатель / художник / поэт / скульптор / врач, хотя и плохой*, при странности ??*Он писатель / художник / поэт / скульптор / врач, хотя и хороший*.

5. Заключение

Изучение импликатур, особенно ложных, в новостных заголовках позволило нам заметить и сформулировать некоторые ранее не описанные классы импликатур, а также контекстов, где они возникают. В частности, были описаны условия, способствующие появлению некоторых видов причинных, (контр)фактивных и качественных импликатур. Хотя импликатуры — общеязыковой феномен, они особенно заметны в ситуациях, когда к ним привлекается внимание, а именно, в ситуациях отмены. Отмена же часто возникает именно в новостных заголовках. Это делает Корпус газетных заголовков, содержащий многочисленные примеры отменяемых содержанием статьи импликатур, полезным инструментом в изучении этого языкового феномена. В свою очередь, данная работа может быть полезна при разработке программного обеспечения, способного определять вводящие в заблуждение заголовки, поскольку в ней описываются типичные контексты возникновения отменяемых имплицитных смыслов.

Литература

1. *Anand A., Chakraborty T., Park N.,* (2016), We used neural networks to detect clickbaits: You won't believe what happened next!, available at https://arxiv.org/pdf/1612.01340.pdf.
2. *Apresjan Yu. D.* (1986), Deixis in lexis and grammar and the naive world model [Deiksis v leksike i grammatike i naivnaya model' mira], Semiotics and informatics [Semiotika i informatika], Num. 28, Moscow, pp. 272–298.
3. *Apresjan Yu. D.,* (2009), The studies in semantics and lexicography [Issledovaniya po semantike i leksikografii], Paradigmatics [Paradigmatika] Vol. 1, Yazyki slavyanskikh kul'tur, Moscow.
4. *Carston R.* (2002). Thoughts and Utterances: The Pragmatics of Explicit Communication, Blackwell Publishers Ltd., Wiley-Blackwell.
5. *Glovinskaya M. Ya.,* (1993), Semantics of speech verbs in terms of Speech Act Theory [Semantika glagolov rechi s tochki zreniya teorii rechevykh aktov]. The Russian language in its functioning: communicative and pragmatic aspects [Russkij yazyk v ego funktsionirovanii: Kommunikativno-pragmaticheskij aspekt], Nauka, Moscow, pp. 158–218.

6. *Glovinskaya M. Ya.,* (2001), Polysemy and synonymy in the tense-aspect system of Russian verbs [Mnogoznachnost' i sinonimiya v vido-vremennoj sisteme russkogo glagola], Azbukovnik, Moscow.
7. *Glovinskaya M. Ya.,* (2010), The lexicographic description of words with speech act meaning [leksikograficheskoe opisanie slov so znacheniem rechevykh aktov], The prospect of the Active Dictionary of the Russian Language [Prospekt aktivnogo slovarya russkogo yazyka], Yazyki slavyanskikh kul'tur, Moscow, pp. 391–436.
8. *Grice H. P.* (1975), Logic and conversation, Syntax and semantics, Num. 3: Speech acts, Academic Press, New York.
9. *Heine B., Kuteva T.,* (2004), World Lexicon of grammaticalization, Cambridge University Press, Cambridge.
10. *Horn L. R.,* (1984), Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature, Meaning, Form, and Use in Context: Linguistic Applications, Georgetown University Press, Washington, pp. 11–42.
11. *Langendoen D. T., Savin H.,* (1971), Studies in Linguistic Semantics, Irvington. pp. 54–60.
12. *Levinson S. C.,* (1983), Pragmatics. Cambridge: Cambridge University Press.
13. *Karttunen L.,* (1971), Implicative Verbs, Language, Vol. 47, Num. 2, pp. 340–358.
14. *Karttunen, L.* (1974). Presupposition and linguistic context. Theoretical Linguistics, 1, 181–194.
15. *Karttunen L., Peters S.,* (1979), Conventional implicature. Syntax and Semantics, Vol. 11: Presupposition, Academic Press, New York, pp. 1–56.
16. *Orlov A. V.,* (2019), The corpus of newspaper headlines [Korpus gazetnykh zagolovkov]. Coursework in Linguistics, National Research University "Higher School of Economics", Moscow.
17. *Paducheva E. V.,* (2013), Egocentric items in the language and interpretation regimes [Egotsentricheskie edinitsy yazyka i rezhymy interpritatsii], Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference "Dialog" [Kompyuternaya lingvistika i intellektual'nye tehnologii. Po materialam ezhegodnoj mezhdunarodnoj konferentsii "Dialog"], Moscow, pp. 486–503.
18. *Paducheva E. V.,* (2018), Egocentric items in a language [Egotsentricheskie edinitsy yazyka], Yazyki slavyanskikh kul'tur, Moscow.
19. *Sperber D., Wilson D.,* (1996), Relevance: Communication and Cognition, Blackwell Publishers Ltd., Wiley-Blackwell.
20. *Shmelyov A. D.,* (1996), Referential mechanisms of the Russian language [Referentsial'nye mekhanismy russkogo yazyka], Slavica Tampercia IV, Tampere.
21. *Wei W., Wan X.,* (2017), Learning to identify ambiguous and misleading news headlines, Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17), Melbourne, pp. 4172–4178.
22. *Апресян Ю. Д.,* (1986), Дейксис в лексике и грамматике и наивная модель мира, Семиотика и информатика, № 28, Москва, С. 272–298.
23. *Апресян Ю. Д.,* (2009), Исследования по семантике и лексикографии, Парадигматика, Т. 1, Языки славянских культур, Москва.

24. *Гловинская М. Я.,* (1993), Семантика глаголов речи с точки зрения теории речевых актов, Русский язык в его функционировании: Коммуникативно-прагматический аспект, Наука, Москва, С. 158–218.
25. *Гловинская М. Я.,* (2001), Многозначность и синонимия в видо-временной системе русского глагола, Азбуковник, Москва.
26. *Гловинская М. Я.,* (2010), Лексикографическое описание слов со значением речевых актов, Проспект активного словаря русского языка, Языки славянских культур, Москва, С. 391–436.
27. *Орлов А. В.,* (2019), Корпус Газетных Заголовков. Курсовая Работа бакалавра программы «Фундаментальная и компьютерная лингвистика», НИУ «Высшая школа экономики», Москва.
28. *Падучева Е. В.,* (2013), Эгоцентрические единицы языка и режимы интерпретации, Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции «Диалог». С. 486–503.
29. *Падучева Е. В.,* (2018), Эгоцентрические единицы языка. Языки славянских культур. Москва.
30. *Шмелев А. Д.,* (1996), Референциальные механизмы русского языка, Slavica Tampercia IV, Тампере.

Appendix

В корпусе представлено более 650 заголовков из 14 СМИ. В Таблице 1 статистика приводится только для СМИ, представленных в корпусе тридцатью и более заголовками. При интерпретации таблицы нужно учитывать, что один заголовок может относиться сразу к нескольким категориям. Все полученные проценты были округлены до ближайшего целого (0.5 округлялось вверх). Данные приводятся на 13.04.2019.

Под ложными в **Таблице 1** понимаются такие заголовки, для которых не существует прочтения, при котором они полностью соответствуют содержанию своей статьи, под честными — такие заголовки, которые не являются ни ложными, ни вводящими в заблуждение.

Список сокращений: Ле — lenta.ru, Р — Ридус, И — Известия, М — Мир Новостей, Ж — Жизнь, Ла — Life.ru, ПВ — Про Владимир

Таблица 1. Процент заголовков, которые находятся в корпусе по следующим поисковым запросам, %

СМИ		Ле	Р	М	Вз	Ве	Ла	ПВ	РБК
Способы создания вводящих в заблуждение заголовков	Нефактивная пресуппозиция	2	4	1	20	0	1	1	1
	Намеренная синтаксическая омонимия	2	5	0	0	0	0	1	0
	Намеренная семантическая омонимия	5	2	0	8	2	0	0	0
	Семантический сдвиг, без нетривиальной метонимии	9	5	11	12	4	7	2	1
	Нетривиальная метонимия	3	4	4	8	4	3	1	1
	De re вместо *De dicto*	1	1	0	4	0	0	0	0
	Гипербола	10	11	10	4	4	1	0	1
	Недоговаривание	7	11	3	16	1	6	4	10
	Ложная импликатура	8	5	6	8	2	3	1	3
	Не тема статьи	2	1	1	10	0	0	4	0
	ВСЕГО	**39**	**36**	**33**	**68**	**16**	**19**	**32**	**17**
Ложные заголовки		3	0	0	8	1	1	4	0
Честные заголовки		59	64	67	24	83	80	64	83

Примеры Разметки в КГЗ

Пример 1:

Заголовок: В России перестанут продавать поваренную соль
Издание: Известия
Дата: 10.09.2018
https://iz.ru/787157/2018-09-10/v-rossii-perestanut-prodavat-povarennuiu-sol
Приём: *de re вместо de dicto*
Краткое содержание статьи: В связи с новым ГОСТом с 1 сентября 2018 года производители обязаны писать на упаковках *пищевая соль,* и не *поваренная соль*

Пример 2:

Заголовок: В Госдуме заговорили об отмене пенсий
Издание: Lenta.ru
Дата: 13.08.2018
https://lenta.ru/news/2018/08/13/otmena/
Приём: *Нетривиальная метонимия*
Краткое содержание статьи: Один член Государственной Думы (Вячеслав Володин) в личном интервью заявил, что не исключает возможности отмены пенсии

LEARNING MULTI-PARTY DISCOURSE STRUCTURE USING WEAK SUPERVISION

Badene S. (sonia.badene@irit.fr),
Thompson K. (catherine.thompson@irit.fr),
Lorré J-P. (jplorre@linagora.com),
Asher N. (asher@irit.fr)

Discourse structures provide a way to extract deep semantic information from text, e.g., about relations conveying causal and temporal information and topical organization, which can be gainfully employed in NLP tasks such as summarization, document classification, sentiment analysis. But the task of automatically learning discourse structures is difficult: the relations that make up the structures are very sparse relative to the number of possible semantic connections that could be made between any two segments within a text; furthermore, the existence of a relation between two segments depends not only on "local" features of the segments, but also on "global" contextual information, including which relations have already been instantiated in the text and where. It is natural to try to leverage the power of deep learning methods to learn the complex representations discourse structures require. However, deep learning methods demand a large amount of labeled data, which becomes prohibitively expensive in the case of expertly-annotated discourse corpora. One recent advance in the resolution of this "training data bottleneck", data programming, allows for the implementation of expert knowledge in weak supervision system for data labeling. In this article, we present the results of our application of the data programming paradigm to the problem of discourse structure learning for multi-party dialogues.

Key words: Discourse Structure, Discursive relations, Weak Supervision, Attachment, Data Programming

ПРЕДСКАЗАНИЕ СТРУКТУРЫ МНОГОСТОРОННЕГО ДИСКУРСА С ПОМОЩЬЮ WEAK SUPERVISION

Баден С. (sonia.badene@irit.fr),
Томпсон К. (catherine.thompson@irit.fr),
Лорре Ж. П. (jplorre@linagora.com),
Ашер Н. (asher@irit.fr)

Ключевые слова: структура дискурса, дискурсивные отношения, weak supervision, связь, data programming

1. Introduction

Discourse structures are relational structures composed of *discourse units* (DUs), or instances of propositional content, and binary coherence relations over them,

conveying semantic (causal, temporal) and presentational (thematic, argumentative) information expressed by a text. We represent such structures as graphs containing a set of nodes representing the DUs and a set of labelled arcs representing the coherence relations. In the case of dialogues occurring between multiple interlocutors, extraction of their internal discourse structures can provide useful semantic information to "downstream" models used, for example, in the production of intelligent meeting managers or the analysis of user interactions in online fora.

Such representational schemes serve as annotation models of which discourse theorists have proposed several: (RST[1] [7], LDM[2] [11], Graphbank [15], DLTAG [5], PDTB[3] [12] and SDRT[4] [4]). Much of computational discourse-analysis is based on RST, which supposes that discourse structures are trees. However, this assumption becomes very difficult to maintain for dialogue [4] and even more unnatural for multi-party dialogue, which presents many examples of non-treelike structures [1] like the one shown in the **Figure 1**.

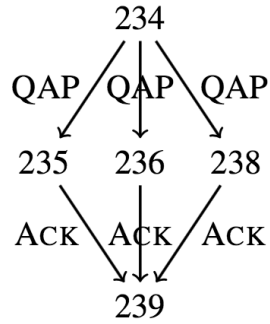


Figure 1: Interlocutors in segments 235, 236 and 238 all respond to a question asked at 234 (linked via edges labelled as Question-Answer Pair), and the interlocutor at segment 239 responds to each answer (linked with Acknowledgement)

Such examples motivate the SDRT annotation model, in which discourse structures are not assumed to be trees but rather directed acyclical graphs (DAGs) [2–4]. SDRT was used to annotate the STAC corpus[5], which is the corpus on which we trained a supervised deep learning model and a weakly supervised model in the discourse structure learning task in order to then compare them.

---

1 Rhetorical Structure Theory

2 Linguistic Discourse Model

3 The Penn Discourse Treebank

4 Segmented Discourse Representation Theory

5 link to STAC corpus: https://www.irit.fr/STAC/index.html

The data programming paradigm was introduced by Hazy Research in 2016 [14] along with a framework Snorkel [13] for using distant, disparate knowledge sources to apply noisy labels to large data sets, and then using those labels to train classic data-hungry machine learning (ML) algorithms. The data programming paradigm allows us to unify these noisy labels in order to generate a probability distribution for all labels for each data point. This set of probabilities replaces the ground-truth labels in a standard discriminative model outfitted with a noise-aware loss function and trained on a sufficiently large data set.

In this study the structure learning problem is restrained to predicting attachment between DU pairs. After training a supervised deep learning algorithm to predict attachments on the full STAC corpus, we then constructed for comparison a weakly supervised learning system in which we used 10% of the corpus as a development set. SDRT experts who annotated the corpus wrote a set of Labeling Functions (LFs) and tested them against this development set. We treated the remainder of the corpus as raw/unannotated data. After applying the LFs to the unannotated data and obtaining the results from the generative and discriminative models, we found that we gained in accuracy over ten points with respect to the supervised method. When the generative model was used in stand alone fashion, we gained almost 30 points in F1 score over the supervised method, which uses a state of the art deep learning model. When we think of the time it takes experts to hand-annotate dialogues, this means that the generative model and weak supervision may be far preferable to straight deep learning methods, in at least some cases.

2. Attachment Prediction: State of the Art

A discourse structure in SDRT is defined as a graph $\langle V, E_1, E_2, \ell, \text{Last}\rangle$, where: V is a set of nodes or discourse units (DUs) and $E_1 \subseteq V^2$ is a set of edges between DUs representing coherence relations. $E_2 \subseteq V^2$ represents a dependency relation between CDUs[6] and their constituent DUs. $\ell: E_1 \rightarrow R$ is a labeling function that gives the semantic type (an element of R) of an edge in E_1, and *Last* is a designated element of V giving the last DU relative to textual or temporal order.

The process of learning an SDRT structure for a dialogue or text has three natural steps:

1. Segment the text into DUs
2. Predict the attachments between DUs, i.e. identify the elements in E_1 and E_2
3. Predict the semantic type of the edge in E_1

In this paper, we focus on the second step, attachment prediction, the goal of which is to determine a substructure, $\langle V, E_1, E_2, \text{Last}\rangle$, of the complete discourse graph. This is a difficult problem for automatic processing: attachments are theoretically possible between any two DUs in a dialogue or text, and often graphs include

[6] SDRT also allows for Complex Discourse Units (CDUs), which are clusters of two or more DUs which can be connected as an ensemble to other DUs in the graph. The CDUs which are themselves graphs that can serve as arguments of coherence relations.

long-distance relations. The attachment problem as we have stated it is endemic to SDRT and theories that posit dependency structures for discourse. In RST the problem of attachment is less clear (see [8]).

[Muller et al. 2012] [9] is one of the first papers we know of on discourse parsing that targets the attachment problem. It targets a restricted version of an SDRT graph, a dependency tree in which CDUs are eliminated by a "flattening" strategy similar to the one we use below. This means that the target representations are of the form $\langle V^*, E^*_1, \text{Last}\rangle$, where V^* is V with the CDUs removed and E^*_1 results from running the flattening strategy on E_1. It trains a simple MaxEnt algorithm to produce structures probability distributions over pairs of elementary discourse units and exploits then global decoding constraints to produce the targeted structures. It reports F1 scores of 66.2% with the A* algorithm.

In terms of discourse structure prediction for multi-party dialogues, Perret et al. (2016) [10] targets a more elaborate approximation of the SDRT graphs: DAGs in which CDUs are eliminated from V and the relations on CDUs are distributed over their constituents. As with [9], this requires another reworking of E_1. Perret et al. then uses Integer Linear Programming (ILP) to encode both the objective function and global decoding constraints over local scores on multi-party dialogues and achieves an F-measure of 0.689 on unlabelled attachment.

3. The STAC Annotated Corpus

3.1. Overview

STAC is a corpus which captures strategic conversation between the players of an online version of the game Settlers of Catan. The full corpus contains 45 games, each of which is divided into an average of 57 dialogues. A dialogue begins at the beginning of a player's turn, and ends at the end of that player's turn. During the interim, players can bargain with each other or make spontaneous conversation. These player utterances are "linguistic" turns, whereas "announcements" made by the game Server regarding the game state or a certain player status are "non-linguistic" turns. Both types of turns are segmented into discourse units (DUs), and these units are then connected by a semantic relations of one of the 17 types admitted by SDRT. As a result, each dialogue contains a weakly connected DAG which is its discourse structure.

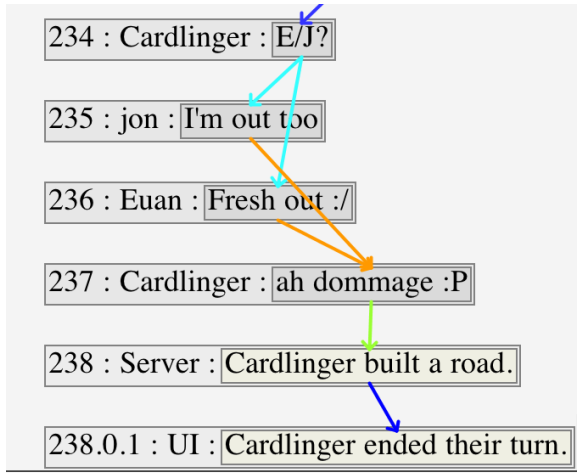


Figure 2: Excerpt of a STAC dialogue illustrating relations like Sequence (dark blue), Result (green), linguistic turns spoken by players and non-linguistic turns emitted by "Server" or "UI"

3.2. Data Preparation

The full STAC corpus includes 2,593 dialogues (or discourse structures), 12,588 linguistic DUs, 31,811 non-linguistic DUs and 31,251 semantic relations.

As discussed above, our task is to predict, for each dialogue, for each possible pair of DUs, whether the DUs are connected by a semantic relation, an operation which eventually yields a discourse structure for the dialogue. Before beginning our experiments, we implemented the following simplifying measures:

1. Roughly 56% of the total dialogues contain only non-linguistic DUs. These represent player turns in which no players bargain or chat with one another. The annotations in these dialogues are fairly regular given the purely mechanical succession of DUs, and are much less difficult and less interesting from a discourse analysis perspective. For this reason we ignore these non-linguistic-only dialogues for our prediction task.

2. In the corpus, shorter relations are more frequent than long-distance relations such that 67% of relations occur between adjacent DUs, and 98% of relations have a distance of 10 or less. (A relation of distance 10 stretches over 9 DUs between the source and the target DU.) In order to avoid a combinatory explosion of possible DU pairs, we restrict the relations we consider to a distance of 10 or less.

3. Out of the 17 possible relation types allowed by SDRT, we consider only the 4 most frequent: Question-answer-pair, Sequence (temporal), Result (causal), Continuation (thematic continuity). We retain about 70% of the

total relations. The reason for this choice will become apparent in following detailed discussion of labeling functions.

4. Because we are here only interested in predicting attachment between single DUs, following [9, 10] we "flatten" the CDUs by connecting all relations incoming or outgoing from a CDU to the "head" of the CDU, or its first DU. This results in shifts in the source and/or target DUs for about 40% of the relations.

5. In order to reduce run-time for each rule during development, we created "sandbox" sets for each relation type: smaller versions of the development set which ignored all candidate pairs except those which could possibly be attached by the relation type in question. We have a sandbox data set for the rules pertinent to a particular discourse relation and a larger sandbox data set for the rules for the four discourse relations that we examined.

After the above preparation, the STAC corpus as we use it in our learning experiments includes 1,130 dialogues, 12,509 linguistic DUs, 18,576 non-linguistic DUs and 22,098 semantic relations. (Here again we are only considering the 4 relations Question-answer-pair, Sequence, Result and Continuation.)

4. The Data Programming Pipeline: Experiments

4.1. Candidates and Labeling Functions

In constructing our weak supervision system, we took inspiration from the Snorkel implementation[7] of the data programming paradigm. The first step in the Snorkel pipeline is candidate extraction, followed by LF creation. Candidates are the units of data from which labels will be predicted; LFs are the simple expert-composed functions which will predict a label for each candidate. The prototypical Snorkel task is to predict whether there is a certain type of relation between two entities in a sentence within a text: candidates are pairs of entities extracted from sentences, and LFs are written using contextual information at the sentence level.

In the case of dialogue attachment prediction, we needed to find a way to give our LFs access to contextual information from the entire dialogue which they could apply to each candidate, or pair of DUs within a dialogue. We did this by fixing the order in which each LF would "see" the candidates such that it would consider adjacent DUs before distant DUs, and thus the LF would know its current position in a dialogue. We also allowed LFs to keep track of previously predicted relations to give them some information about dialogue history. Other information leveraged by the LFs included the DU raw text, speaker identities, the DU dialogue acts, DU types (linguistic or non-linguistic) and the distance between DUs.

[7] https://hazyresearch.github.io/snorkel/

```python
def LF_Result_L_L_case1(row):
    l=0
    if (any(x in row.target_text.lower() for x in resultWords)
                or any(x in row.source_text.lower() for x in resultWords)):
            l=1
    return l


def LF_Result_L_L_case2(row):
    l = 0
    if row.source_surface_act in ["Question", "Request", "Assertion"] \
    and (row.target_dialogue_act in ["Offer", "Counteroffer"] \
         or row.source_emitter == row.target_text.partition(' ')[0] or row.target_surface_act == "Request"):
        l=1
    return l

def LF_Result_L_L_case3(row):
    l = 0
    if row.source_emitter == row.target_emitter and row.source_addressee == row.target_addressee \
    and row.target_dialogue_act == "Counteroffer":
        l=1
    return l


def LF_Result_L_L_case4(row):
    l = 0
    if row.source_dialogue_act == "Counteroffer" and row.target_dialogue_act == "Refusal":
        l=1
    return l
```

Figure 3: *Result* connects a cause to its effect, i.e., the main eventuality of the first argument is understood to cause the eventuality given by the second. Here we show a sample of our rules written in python for the relation ***Result*** connecting two linguistic discourse units

As we are at present concerned only with predicting attachments, each LF returns a 1, a 0 or a −1 ("attached"/"do not know"/"not-attached") for each candidate. However, each of our LFs is written and evaluated with a specific relation type *Result*, *QAP*, *Continuation* and *Sequence* in mind. In this way, LFs also leverage a kind of type-related information. This makes sense from an empirical perspective as well as an epistemological one: an attachment decision concerning two DUs is tightly linked to the type of relation relating the DUs, and so when an annotator decides that two DUs are attached, he or she does so with some knowledge of what type of relation attaches them. Figure 3 shows a sample of our rules used for attachment prediction with the *Result* relation in mind.

4.2. The Generative Model

Once we have applied the LFs to all the candidates, we then move to the generative step. In Snorkel, the generative model unifies the results of the LFs, which is a matrix of labels given by each LF (columns) for each candidate (rows). Though the simplest approach to unification would be to take the majority vote among the LFs for each candidate, this is less effective in cases where all the LFs abstain or give "0". Further, this approach would not take into account the individual performances of the LFs. And so the generative model as specified in (1) provides a general distribution of marginal probabilities relative to n accuracy dependencies $\phi_j(\Lambda_i; y_i)$ between an LF λ_j and true labels y_i that depend on parameters theta θ_i.

$$p_\theta(\Lambda, Y) \propto exp(\Sigma_{i=1}^{m} \Sigma_{j=1}^{n} \theta_j \phi_j(\Lambda_i, y_i)) \quad (1)$$

The parameters are estimated without access to the ground truth labels by minimizing the negative log marginal likelihood of the output of an observed matrix Λ as in (2).

$$argmin_\theta - log\Sigma_Y(p_theta(\overline{\Lambda}, Y) \quad (2)$$

This objective is optimized by interleaving stochastic gradient descent steps with Gibbs sampling ones. The model thus uses the accuracy measures for the LFs in (1) to assign marginal probabilities that two DUs are attached to each candidate. In this model, the true class labels y_i are latent variables that generate the labeling function outputs, which are estimated via Gibbs sampling.

This calculation presupposes that the LFs are independent. However, the LFs are often dependent: one might be a variation of another or they might depend on a common source of distant supervision. If we don't take this into account, we risk assigning incorrect accuracies to the LFs. Getting users to indicate dependencies by hand, however, is difficult and error-prone. The generative model in Snorkel comes with the option of automatically selecting which dependencies to model without access to ground truth. It uses a pseudo-likelihood estimator, which does not require any sampling or other approximations to compute the objective gradient exactly. It is much faster than maximum likelihood estimation, because the estimator relies on a hyper-parameter ϵ that trades off between predictive performance and computational cost. With large values of ϵ no correlations are included and as it reduces the value progressively more correlations are added, starting with the strongest.

4.3. Discriminative Model

While the generative model outputs the marginal probabilities for each of the labels for each candidate, the discriminative model generalizes this output and augments the coverage of the LFs. While this may lead to a small reduction in precision, it is in exchange for a boost in recall.

We used BERT's [6] sequence classification model (source code on the link below[8]) with 10 training epochs and all default parameters otherwise. BERT, the Bidirectional Encoder Representations from Transformers, is a text encoder pre-trained using language models where the system has to guess a missing word or word piece that is removed at random from the text. Originally designed for automatic translation tasks, BERT uses bi-directional self-attention to produce the encodings and performs at the state of the art on many textual classification tasks. While in principle we could have used any discriminative model, as is suggested in the Snorkel literature, BERT gave us by far the best results on attachment prediction. For this reason we also used BERT as our model for supervised learning of attachment to compare its results with those of the weak supervision method.

[8] Link to BERT sequence classification model code: https://github.com/huggingface/pytorch-pretrained-BERT/blob/master/examples/run_classifier.py

5. Results and Analysis

In order to write a set of LFs/rules which adequately covered the data, we had to find a way to reasonably divide and conquer the myriad characteristics of the relations. We started by focusing on relation type: for each of the four most frequent relation types, we wrote a separate rule for each of the sets of endpoint types most prevalent for that relation. Result (RES) is the only relation type which was found between all four endpoint permutations: LL (linguistic source-linguistic target), LNL (linguistic source, non-linguistic target), etc. We used the relation behavior observed in our development/sandbox sets to write and revise the rules. The development set consisted of 3 games (10% of our data). The table 1 shows the performance of each rule on its own "sandbox" development set.

Table 1: Number of true positives, true negatives, false positives, false negatives and accuracy score for each LF when applied to the "sandbox" candidates from the STAC data

	TP	TN	FP	FN	Accuracy
QAP LL	294	1798	112	138	0.89
QAP NLNL	84	187	0	0	1.00
RES NLNL	739	2,929	13	55	0.98
RES LNL	13	2158	93	97	0.91
RES LL	25	316	19	37	0.85
RES NLL	2	139	0	2	0.98
Cont LL	16	9,818	110	106	0.97
Cont NLNL	613	3,254	0	1	0.99
SEQ NLL	90	658	2	14	0.97
SEQ NLNL	236	1,220	10	76	0.94

Table 2: Evaluations of the combination of the four LFs (QRSC)' attachment with the weakly supervised and supervised approaches on the sandbox data set

	Generative Model			Discriminative Model on Test	
	Dev	Train	Test	with Marginals	with Gold annotations
Precision	0.67	0.70	0.68	0.45	0.61
Recall	0.84	0.85	0.84	0.54	0.53
F1 score	0.75	0.77	0.75	0.49	0.57
Accuracy	0.92	0.93	0.92	0.84	0.88

We first evaluated the LFs for each discourse relation type individually on the development corpus, providing a measure of their coverage and accuracy on a subset of the data. Then we evaluated the generative model on the combination of the four LF types and the discriminative model on the test set of the corpus. The table 2 presents the results at the end of each step in our weak supervision system. To compare the two approaches, the discriminative model was first trained on the marginals provided

by our generative model, then on the "gold" annotations, which are manual annotations of the Stac corpus.

We find the results of the generative model striking as they are almost 20 points higher in F1 score concerning positive attachment over the discriminative model trained on the gold annotations. This shows the power of the rule based approach even when compared to a state of the art deep learning system. Another interesting point is that the discriminative model can still perform acceptably with the marginals data compared to its performance using the gold annotations; its accuracy is only 4 points lower and its F1 score is 8 points lower but still comparable with results in the literature, showing that the generative model has information to offer that the discriminative model can exploit well—thus opening up the possibility for transfer learning. Rather than naively treat these noisy training labels as ground truth, our noise-aware discriminative model gives a slight improvement in recall with a decrease in precision compared to the supervised approach. With respect to the individual LFs in isolation, we find that, apart from *QAP*, our rules for each relation type have an accuracy, precision and recall comparable to those for the supervised models. One reason for our lower precision for *QAP* may be attributable to a deficiency in the flattening procedure the effects *QAP* more frequently; in some cases the flattening algorithm re-attaches the *QAP* relation to a CDU head which was not in fact the component of the CDU which marked the question. What is interesting is the synergy between the rules such that when they all interact on the test data, they do very well on the generative model.

6. Conclusions and Future Work

Having chosen a single discriminative model for all our experiments, we were able to compare our weak supervision approach, in which we leveraged parts of the Snorkel system, with that of a standard supervised model on the difficult task of discursive attachment. Our approach allows us to model the discourse more precisely and to be generalized to other corpora. In contrast to a supervised algorithm, our results on the generative model are almost 30 points higher without covering all types of rules. In addition, we generate a lot of annotated data in a very short time.

In future work, we plan to enrich our weak supervision system by first covering all 17 types of SDRT relations on all data. We also plan to give LFs access to more sophisticated context which will take into account sequence-constraints in the attachments of the complete conversation and global structuring constraints. We will at that point be in a position to evaluate the structure of the global discourse by using our structured predictions as inputs to a maximum spanning tree (MST) for example. And in a broader scope, our experiments with this paradigm may suggest possible lines of inquiry into how weakly supervised methods might effectively capture the global structural constraints on discourse structures without decoding or elaborate learning architectures.

References

1. *Afantenos, S. et al.:* Discourse parsing for multi-party chat dialogues. Presented at the (2015).
2. *Asher, N.:* Reference to abstract objects in discourse. Springer Science & Business Media (1993).
3. *Asher, N. et al.:* Discourse structure and dialogue acts in multiparty dialogue: The stac corpus. In: LREC. (2016).
4. *Asher, N., Lascarides, A.:* Logics of conversation. Cambridge University Press (2003).
5. *Creswell, C. et al.:* Penn discourse treebank: Building a large scale annotated corpus encoding dltag-based discourse structure and discourse relations. Manuscript [fix this]. (2003).
6. *Devlin, J. et al.:* Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. (2018).
7. *Mann, W. C., Thompson, S. A.:* Rhetorical structure theory: Description and construction of text structures. In: Natural language generation. pp. 85–95 Springer (1987).
8. *Morey, M. et al.:* A dependency perspective on rst discourse parsing and evaluation. Computational Linguistics. 198–235 (2018).
9. *Muller, P. et al.:* Constrained decoding for text-level discourse parsing. Proceedings of COLING 2012. 1883–1900 (2012).
10. *Perret, J. et al.:* Integer linear programming for discourse parsing. In: Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: Human language technologies. pp. 99–109 (2016).
11. *Polanyi, L. et al.:* A rule based approach to discourse parsing. In: Proceedings of the 5th sigdial workshop on discourse and dialogue at hlt-naacl 2004. (2004).
12. *Prasad, R. et al.:* The penn discourse treebank 2.0. Annotation manual. The pdtb research group, (2007).
13. *Ratner, A. et al.:* Snorkel: Rapid training data creation with weak supervision. Proceedings of the VLDB Endowment. 11, 3, 269–282 (2017).
14. *Ratner, A. J. et al.:* Data programming: Creating large training sets, quickly. In: Advances in neural information processing systems. pp. 3567–3575 (2016).
15. *Wolf, F. et al.:* The discourse graphbank: A database of texts annotated with coherence relations. Linguistic Data Consortium. (2005).

ДИСКУРСИВНЫЕ СЛОВА В КОРПУСНОМ ИЗМЕРЕНИИ: *ОДНИМ СЛОВОМ* У ДОСТОЕВСКОГО И ЕГО СОВРЕМЕННИКОВ[1]

Баранов А. Н. (baranov_anatoly@hotmail.com),
Добровольский Д. О. (dobrovolskij@gmail.com)
Институт русского языка РАН, Москва, Россия

Рассматривается гипотеза о том, что дискурсивные слова характеризуют авторский стиль писателя. В качестве объекта исследования выбрано устойчивое словосочетание *одним словом* на материале представительных корпусов Достоевского, Толстого, Салтыкова-Щедрина, Тургенева и Гончарова. Проведенный анализ позволяет сделать вывод о том, что Достоевский и Салтыков-Щедрин отличаются от других писателей-современников как частотой использования *одним словом* в дискурсивной функции, так и разнообразием семантики этого выражения. Особенно интересен в этом отношении Достоевский, в прозе которого представлены все дискурсивные функции *одним словом*: интерпретация (собственно интерпретация, вывод, уточнение/пояснение), новая идея, регулятивные употребления (прерывание дискурса, маркирование трудностей в выборе номинации, маркирование смены номинации (изменение номинации может быть на базовую, альтернативную и обобщающую), введение чужой речи: как в виде прямой, так и не собственно прямой речи. Что касается недискурсивных употреблений выражения *одним словом*, то они распределены у рассматриваемых авторов более или менее равномерно.

Ключевые слова: дискурсивные слова, авторский стиль, корпусный подход, язык Достоевского

[1] Работа выполнена в рамках проекта № 18-012-90025, поддержанного РФФИ.

DISCURSIVE WORDS IN CORPUS DIMENSION: *ODNIM SLOVOM* IN THE WORKS OF DOSTOEVSKY AND HIS CONTEMPORARIES

Baranov A. N. (baranov_anatoly@hotmail.com),
Dobrovol'skij D. O. (dobrovolskij@gmail.com)
Russian Language Institute of the RAS, Moscow, Russia

The starting point of the present paper is the hypothesis that discursive words characterize the individual style of the author. The subject of the study is the fixed expression *odnim slovom* 'in one word' in the works of Dostoevsky, Tolstoy, Saltykov-Shchedrin, Turgenev and Goncharov. The analysis of representative corpora yields the conclusion that Dostoevsky and Saltykov-Shchedrin differ from other contemporaries both in the frequency of using this expression in the discursive function and in the variety of its semantics.

Particularly interesting in this respect is Dostoevsky, whose prose presents all discursive functions of the expression *odnim slovom*:

- interpretation (including interpretation proper, conclusion, and clarification/explanation),
- introducing a new idea,
- regulatory uses of *odnim slovom*, such as interruption of discourse, marking difficulties in choosing a nomination,
- marking the change of nomination, as well as
- the introduction of someone else's speech.

As for the non-discursive uses of the expression *odnim slovom*, they are distributed more or less evenly among the authors under consideration.

Key words: discursive words, individual style, corpus approach, Dostoevsky's language

1. Постановка проблемы

Изучению художественного стиля посвящены многочисленные исследования в литературоведении и в близких к нему лингвистических дисциплинах [Виноградов 1980]; [Лотман 1998]; [Томашевский 2002]; [Эйхенбаум 1969]. До последнего времени, однако, корпусные методы исследования художественного стиля в отечественной традиции практически отсутствовали. На этом фоне выделяются работы А. Я. Шайкевича и его соавторов, посвященные лингво-статистическому анализу представительных корпусов дискурсов различных типов, в том числе и художественных текстов [Шайкевич, Андрющенко, Ребецкая 2003]; [2013]; [2016]. В этих исследованиях в явном виде ставится задача определения сходства текстов, входящих в соответствующий корпус. Некоторые из этих характеристик сходства относятся и к художественному стилю.

Возможен, однако, и другой подход к корпусному исследованию стиля художественной прозы: в этом случае лингво-статистическая модель стиля основывается на анализе частотного распределения ограниченного круга лексем, важных с точки зрения общих тенденций стилевых изменений. Разумеется, в этом случае успех во многом зависит от выбора этих конкретных единиц. Известно, например, что к числу «слов Достоевского» относятся выражения *как бы* и *так* [Арутюнова 1999]; [2006], *по крайней мере* [Баранов 1996], *кстати* [Баранов, Добровольский 2018], *вдруг* и др.

Рассмотрим, насколько индивидуально-авторским оказывается употребление выражения *одним словом*, допускающего как дискурсивные, так и недискурсивные употребления.

Для исследования мы выбрали тексты Ф. М. Достоевского, а также Л. Н. Толстого, И. А. Гончарова, М. Е. Салтыкова-Щедрина и И. С. Тургенева. Эти писатели принадлежат к одному поколению, что позволяет предположить, что различия их языка объясняются особенностями индивидуальных стилей, а не спецификой языка эпохи.

Естественно начать изложение с описания контекстов употребления выражения *одним словом* и выделения его значений.

2. Структура значения выражения *одним словом*

Как уже отмечалось выше, словосочетание *одним словом* имеет дискурсивные и недискурсивные употребления. Последние представлены примерами типа (1):

(1) **а**. — Это всё вздор, — сказал Свидригайлов, намачивая полотенце и прикладывая его к голове, — а я вас *одним словом* могу осадить и все ваши подозрения в прах уничтожить. [Ф. М. Достоевский. Преступление и наказание]
б. Воспоминания толпою проходили перед ним, но были однообразны и исчерпывались *одним словом*: «ученье». [М. Е. Салтыков-Щедрин. Мелочи жизни]
в. — Да ты знаешь ли, — перебил ее Николай Артемьевич, — что я могу уничтожить тебя *одним словом*? Елена подняла на него глаза. — Да, сударыня, *одним словом*! [И. С. Тургенев. Накануне]
г. Между тем он одним взглядом, *одним словом* мог бы создать в ней глубокую страсть, к себе; но он молчит, он не хочет. [И. А. Гончаров. Обыкновенная история]
д. — Но человек может чувствовать себя неспособным иногда подняться на эту высоту, — сказал Степан Аркадьич, чувствуя, что он кривит душою, признавая религиозную высоту, но вместе с тем не решаясь признаться в своем свободомыслии перед особой, которая *одним словом* Поморскому может доставить ему желаемое место. [Л. Н. Толстой. Анна Каренина]

К недискурсивным относятся также употребления исследуемого выражения в сочетании с частицей *ни*: *Если вы теперь нас подслушивали, то должны же были заметить, что она* ***ни одним словом*** *<…> не поддержала меня перед*

князем (Ф. М. Достоевский. Дядюшкин сон). Это вполне понятно, поскольку семантика выражения *ни одним* связана с указанием на количество — типичный «пропозициональный» смысл.

Самая большая группа дискурсивных употреблений — это контексты **интерпретации**, разделяющиеся на собственно интерпретацию, вывод и уточнение/пояснение. Различия между ними относительно невелики. Собственно интерпретация отличается от вывода, грубо говоря, тем, что в контекстах вывода присутствует нечто вроде естественной, наивной аргументации. Ср. примеры (2а) и (2б) в сопоставлении.

(2) **а**. Прошу их не стесняться, веселиться, продолжать танцы, острю, смеюсь, *одним словом* — я любезен и мил. [Ф. М. Достоевский. Скверный анекдот]
б. У стола стоял господин в очень истрепанном сюртуке (он уже снял пальто, и оно лежало на кровати) и развертывал синюю бумагу, в которой было завернуто фунта два пшеничного хлеба и две маленькие колбасы. На столе, кроме того, был чайник с чаем и валялись куски черного хлеба. Из-под кровати высовывался незапертый чемодан и торчали два узла с каким-то тряпьем. *Одним словом*, был страшный беспорядок. [Ф. М. Достоевский. Идиот]

В (2а) аргументация в пользу идеи, что «я любезен и мил», отсутствует. Действительно, странно было бы считать аргументами описание просьбы «не стесняться, веселиться, продолжать танцы». С другой стороны, развернутое описание беспорядка на столе в (2б) в рамках наивной логики вполне может служить аргументом тезиса «был страшный беспорядок». Именно поэтому, выражение *одним словом* в (2б) можно заменить на *резюмируя, в силу сказанного* и т. п. Очевидно, что такая замена для контекстов типа (2а) невозможна[2].

Для выделения контекстов уточнения/пояснения также можно предложить критерий замены. В этом случае выражение *одним словом* с небольшой потерей смысла можно заменить на формы *точнее, уточняя, поясняя, говоря проще, яснее* и под. Ср. (3):

(3) Перед ними стоял человек, тормозящий русское дело, служащий польским интересам, *одним словом*, изменник отечеству, и ограничиться только тем, что прокричать ему только à bas, долой, вон!
[М. Е. Салтыков-Щедрин. Литература на обеде]

В примере (3) вполне были бы допустимы фразы типа *точнее, изменник отечеству* или *говоря яснее, изменник отечеству* и т. п. Еще одним дифференциальным признаком этого значения можно считать краткость пояснения, свойственную также и семантике вывода, но нехарактерную для собственно интерпретации.

Следующее значение *одним словом* вводит в дискурс **новую идею, или новый смысл**:

[2] Отметим, что приведенный критерий замены не является универсальным для выделения контекстов с семантикой вывода. Это требует особого обсуждения, выходящего за рамки данной работы.

(4) **а**. Да я сам был глуп и нетерпелив и всё дело испортил. Авдотье Романовне еще несколько раз и прежде (а один раз как-то особенно) ужасно не понравилось выражение глаз моих, верите вы этому? *Одним словом*, в них всё сильнее и неосторожнее вспыхивал некоторый огонь, который пугал ее и стал ей наконец ненавистен. [Ф. М. Достоевский. Преступление и наказание]
б. Всё хозяйство, главное — положение всего народа, совершенно должно измениться. Вместо бедности — общее богатство, довольство; вместо вражды — согласие и связь интересов. *Одним словом*, революция, бескровная, но величайшая революция, сначала в маленьком кругу нашего уезда, потом губернии, России, всего мира. [Л. Н. Толстой. Анна Каренина]
в. Посмотрите зимой в сумерки на улицу: свет борется со тьмою; иногда крупный снег вступает в посредничество, угождая свету своею белизною и увеличивая мрак своим облаком. Но человек остается праздным свидетелем этой борьбы: он приумолкает, приостанавливается; нет движения; улица пуста; домы, как великаны, притаились во тьме; нигде ни огонька; все предметы смешались в каком-то неопределенном цвете; ничто не нарушает безмолвия, ни одна карета не простучит по мостовой <…>. *Одним словом*, кажется, настала минута осторожности… а в самом деле эта минута есть, может быть, самая неосторожная в целом дне. [И. А. Гончаров. Счастливая ошибка]

По внутренней форме выражение *одним словом* не должно было бы вводить новое. Казалось бы, это значение плохо мотивировано внутренней формой. В то же время смысловая связь пропозиции, вводимой *одним словом*, с предшествующим текстом очевидна. Это еще одна мысль как добавление к сказанному на ту же тему. Как и в случае собственно интерпретации, распространенность, сложность этой новой мысли не лимитирована: это может быть краткая номинация *революция*, как в (4б), а может быть развернутое предложение, как в (4а) и (4в). Семантика *одним словом* в контекстах типа (4) — это добавление нового смысла.

Следующая большая группа контекстов — это **операторы, регулирующие дискурс**. Сюда относятся контексты прерывания дискурса (5) и маркирование с помощью *одним словом* трудностей в выборе номинации (6).

(5) — Но уверяю же вас, голубчик… помилуйте! — Mon ami! Mon enfant! — воскликнул он вдруг, складывая перед собою руки и уже вполне не скрывая своего испуга, — если у тебя в самом деле что-то есть… документы… *Одним словом* — если у тебя есть что мне сказать, то не говори; ради бога, ничего не говори; лучше не говори совсем… как можно дольше не говори… [Ф. М. Достоевский. Подросток]

(6) Извините меня, но я должен вам высказать, что слухи, до вас дошедшие или, лучше сказать, до вас доведенные, не имеют и тени здравого основания, и я… подозреваю, кто… *одним словом*… эта стрела… *одним словом*, ваша мамаша… [Ф. М. Достоевский. Преступление и наказание]

И (5), и (6) иллюстрируют различные техники регулирования дискурса. В примере (5) выражение *одним словом* указывает на то, что говорящий

не может или не хочет довести до конца предшествующую последовательность рассуждений. Одновременно говорящий формулирует что-то, что с его точки зрения более точно передает соответствующее коммуникативное намерение. В примере (6) оператор *одним словом* отражает трудности с выбором номинации, вводя ее различные варианты. Во всех случаях регулятивного употребления оператора *одним словом* это выражение одновременно указывает на хезитацию говорящего, отражая ту или иную степень затрудненности порождения, формулирования речевого акта. Внутренняя форма, как и в контекстах введения новой идеи, не мотивирует актуальное значение.

Следует отметить, что выделение контекстов использования выражения *одним словом* для прерывания дискурса сочетается с другими типами значений. Так, в примере (7) прерывание сочетается с выводом, а в примере (8) — с введением в дискурс новой идеи:

(7) Уверь кто-нибудь тогда честнейшего Степана Трофимовича неопровержимыми доказательствами, что ему вовсе нечего опасаться, и он бы непременно обиделся. А между тем это был ведь человек умнейший и даровитейший, человек, так сказать, даже науки, хотя, впрочем, в науке... Ну, *одним словом*, в науке он сделал не так много и, кажется, совсем ничего. Но ведь с людьми науки у нас на Руси это сплошь да рядом случается. [Ф. М. Достоевский. Бесы]

(8) Я там вам у нотариуса, что ли, или как там... *Одним словом*, я готов на все, выдам все документы, какие потребуете, все подпишу... и мы эту бумагу сейчас же и совершили бы, и если бы можно, если бы только можно, то сегодня же бы утром... [Ф. М. Достоевский. Братья Карамазовы]

Иными словами, семантика прерывания в некотором смысле ортогональна семантике интерпретации, вывода, введения новой идеи и т. п. Это другая характеристика дискурсивных употреблений.

Исследуемое выражение используется также для маркирования **смены номинации**: новая номинация может быть базовой (9), альтернативной (10) и обобщающей (11):

(9) **а**. Действительный статский советник Иван Ильич Пралинский всего только четыре месяца как назывался вашим превосходительством, *одним словом*, был генерал молодой. [Ф. М. Достоевский. Скверный анекдот]
б. — Григорий Литвинов, рубашка-парень, русская душа, рекомендую, воскликнул Бамбаев, подводя Литвинова к человеку небольшого роста и помещичьего склада, с расстегнутым воротом, в куцей куртке, серых утренних панталонах и в туфлях, стоявшему посреди светлой, отлично убранной комнаты, а это, прибавил он, обращаясь к Литвинову, это он, тот самый, понимаешь? Ну, Губарев, *одним словом*. [И. С. Тургенев. Дым]

(10) Перед ним, как бес перед заутреней, вертелся маленький человек не то армянин, не то грек, *одним словом*, существо, которое Прокоп, под веселую руку, называл «православным жидом». [М. Е. Салтыков-Щедрин. Дневник провинциала в Петербурге]

(11) Все это из самого полного внутреннего убеждения, что собственное лицо у каждого русского — непременно ничтожное и комическое до стыда лицо; а что если он возьмет французское лицо, английское, *одним словом*, не свое лицо, то выйдет нечто гораздо почтеннее, и что под этим видом его никак не узнают. [Ф. М. Достоевский. Дневник писателя]

В примере (9) номинация *генерал молодой* (с учетом инверсии) оказывается более типичной, чем человек, *всего только четыре месяца как* называвшийся *вашим превосходительством*. В примере (10) номинация *«православный жид»* альтернативна по отношению к номинациям *маленький человек, армянин, грек*. Между номинациями *французское лицо*, *английское лицо*, с одной стороны, и номинацией *не свое лицо* (как в примере (11)) прослеживается отношение «род—вид».

Следующий тип дискурсивных употреблений, встречающийся у рассматриваемых авторов довольно редко, — это **введение чужой речи**: как в виде прямой (12), так и не собственно прямой речи (13).

(12) — Хорошо, хорошо! Но скажите мне, почему вы узнали, что я такая женщина, с которой... ну, которую вы считали достойной... внимания и дружбы... *одним словом*, не хозяйка, как вы называете. [Ф. М. Достоевский. Белые ночи]

(13) — Тут есть, — тут именно существует особое обстоятельство, — подхватил Де-Грие просящим тоном, в котором все более и более слышалась досада. — Вы знаете mademoiselle de Cominges? — То есть mademoiselle Blanche? — Ну да, mademoiselle Blanche de Cominges... et madame sa mère... согласитесь сами, генерал... *одним словом*, генерал влюблен и даже... даже, может быть, здесь совершится брак. [Ф. М. Достоевский. Игрок]

Выражение *одним словом* в функции введения чужой речи (как и в функции маркера прерывания) может сочетаться с семантикой интерпретации, вывода, пояснения и т. п. Иными словами, контексты указанного типа выделяются на других основаниях — не по содержанию пропозиции, вводимой *одним словом*, а по авторству соответствующей реплики.

3. Количественное распределение выражения *одним словом*

В связи с поставленной задачей представляет интерес распределение контекстов употребления *одним словом* по авторам и по значениям. Сначала имеет смысл сравнить дискурсивные и недискурсивные употребления; ср. **таблицу 1**:

Таблица 1. Распределение дискурсивных и недискурсивных употреблений *одним словом*

	Достоевский	Салтыков-Щедрин	Толстой	Тургенев	Гончаров
Общее количество (абсолютная величина)	714	476	36	22	12
Дискурсивные употребления (абсолютная величина)	702	446	25	19	7
Дискурсивные употребления (относительная величина)[3]	2,42	1,64	0,13	0,26	0,07
Недискурсивные употребления (абсолютная величина)	12	30	11	3	5

Из **таблицы 1** хорошо видно, насколько Достоевский опережает всех других авторов в дискурсивных употреблениях исследуемого выражения. Салтыков-Щедрин оказывается в этом отношении близок Достоевскому, а они вместе противопоставлены остальным трем авторам. Интересно отметить, что хотя исследуемые писатели обнаруживают существенные различия в дискурсивных употреблениях выражения *одним словом*, в недискурсивной функции это словосочетание оказывается более или менее уравновешенным по всем авторам. В относительных величинах различия в частоте употребления (в порядке упоминания авторов в таблице) будут таковы: 0,03–0,1–0,06–0,04–0,05. Приведенные значения настолько малы, что их варьирование не является статистически значимым.

[3] Здесь и далее относительная величина считается на 10 тысяч словоупотреблений.

Таблица 2. Распределение дискурсивных употреблений *одним словом*

	Достоевский	Салтыков-Щедрин	Толстой	Тургенев	Гончаров
Собственно интерпретация	287 (40,9%)[4]	291 (65,2%)	11 (44%)	16 (84%)	5 (71,4%)
Вывод	203 (28,9%)	87 (19,5%)	7 (28%)	2 (10,5%)	0 (0%)
Уточнение / пояснение	26 (3,7%)	12 (2,69%)	3 (12%)	0 (0%)	0 (0%)
Новая идея	48 (6,8%)	32 (7,2%)	1 (4%)	0 (0%)	2 (28,6%)
Прерывание	107 (15,2%)	11 (2,46%)	3 (12%)	0 (0%)	0 (0%)
Трудность номинации	13 (1,8%)	7 (1,6%)	0 (0%)	0 (0%)	0 (0%)
Смена номинации на базовую	5 (0,7%)	1 (0,2%)	0 (0%)	1 (5,3%)	0 (0%)
Смена номинации на альтернативную	6 (0,8%)	2 (0,4%)	0 (0%)	0 (0%)	0 (0%)
Смена номинации на обобщающую	2 (0,28%)	2 (0,4%)	0 (0%)	0 (0%)	0 (0%)
Цитация прямой речи	1 (0,1%)	1 (0,2%)	0 (0%)	0 (0%)	0 (0%)
Несобственно прямая речь	4 (0,56%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)

Частотное распределение дискурсивных значений у различных авторов, приведенное в **таблице 2**, показывает, что активно используют различные дискурсивные функции выражения *одним словом* только Достоевский и Салтыков-Щедрин, причем наибольшее количество употреблений приходится на контексты собственно интерпретации (40,9% — Достоевский, 65,2% — Салтыков-Щедрин) и вывода, при этом у Достоевского вывод более частотен (28,9%), чем у Салтыкова-Щедрина (19,5%). Кроме того, Достоевский активно использует *одним словом* для прерывания дискурса (15,2% при чуть более 2% у Салтыкова-Щедрина). Процентное соотношение дискурсивных значений у других авторов не представляет интереса из-за низкой частоты употреблений данного выражения.

Полученные результаты подтверждаются и другим статистическим параметром — разнообразием семантики. Он вычисляется как процент реализаций каждого из значений от общего количества дискурсивных значений. Всего

[4] Здесь и далее процент контекстов от общего количества дискурсивных употреблений у данного автора.

в разделе 2 были выделены следующие значения (без учета недискурсивных употреблений): собственно интерпретация (1), вывод (2), уточнение/пояснение (3), новая идея (4), прерывание дискурса (5), трудность номинации (6), смена номинации на базовую (7), альтернативную (8), обобщающую (9), цитация прямой речи (10), несобственно прямая речь (11). Соответственно, если у автора представлены все значения исследуемого выражения, то есть 11, то его индекс разнообразия будет равен 100%.

Таблица 3. Разнообразие семантики ***одним словом*** по разным авторам

	количество значений	показатель разнообразия
Достоевский	11	100%
Салтыков-Щедрин	10	90%
Толстой	5	45%
Тургенев	3	27%
Гончаров	2	18%

Из **таблицы 3** следует, что в целом степень разнообразия коррелирует с количеством дискурсивных употреблений. Исключение составляет Толстой, который при достаточно низком количестве дискурсивных употреблений *одним словом* показывает довольно высокий индекс разнообразия. Это говорит в пользу того, что выявленные значения не были чисто авторскими и представлены в художественной прозе разных авторов второй половины XIX века.

***

Приведенный материал показывает, что выражение *одним словом* в дискурсивных употреблениях может рассматриваться как показатель авторского стиля. В этом отношении художественные стили Достоевского и Салтыкова-Щедрина весьма близки. В то же время по результатам анализа одного или нескольких дискурсивных слов нельзя делать окончательный вывод о характеристиках авторства. Так, в работе [Баранов, Добровольский 2018] показано, что частота употреблений слова *кстати* в дискурсивной функции объединяет по стилевым характеристикам Достоевского и Тургенева, противопоставляя их всем остальным рассмотренным авторам. Очевидно, что лишь анализ достаточно большого списка дискурсивных слов позволит сделать релевантные выводы о характере индивидуального стиля каждого из этих авторов.

Литература

1. *Арутюнова Н. Д.* (1999), Язык и мир человека. М., 1999. С. 846–869.
2. *Арутюнова Н. Д.* (2006), Неисследимые смыслы: просто так, да не так просто // Известия РАН. Серия литературы и языка. Т. 65, № 2, 2006. С. 14–22.
3. *Баранов А. Н.* (1996), Служебные слова как объект исследования авторской лексикографии (по крайней мере vs. по меньшей мере в художественных текстах Достоевского) // Слово Достоевского. М., 1996. С. 110–136.
4. *Баранов А. Н., Добровольский Д. О.* (2018) Кстати и некстати: к речевым практикам Достоевского // Русский язык в научном освещении, № 1 (35). С. 33–45.
5. *Виноградов В. В.* (1980) Избранные труды. О языке художественной прозы. М.: Наука.
6. *Лотман Ю. М.* (1998) Структура художественного текста // Ю. М. Лотман. Об искусстве. СПб.: Искусство.
7. *Томашевский Б. В.* (2002) Теория литературы. Поэтика. М.: Аспект Пресс.
8. *Шайкевич А. Я., Андрющенко В. М., Ребецкая Н. А.* (2003) Статистический словарь языка Достоевского. М.: Языки славянских культур.
9. *Шайкевич А. Я., Андрющенко В. М., Ребецкая Н. А.* (2013) Дистрибутивно-статистический анализ языка русской прозы 1850–1870-х гг. Том 1. М.: Языки славянских культур.
10. *Шайкевич А. Я., Андрющенко В. М., Ребецкая Н. А.* (2016) Дистрибутивно-статистический анализ языка русской прозы 1850–1870-х гг. Том 2. М.: Языки славянских культур.
11. *Эйхенбаум Б. М.* (1969) О прозе. Л.: Художественная литература.

References

1. *Arutyunova N. D.* (1999), Language and the world of human being [Yazyk i mir cheloveka]. Moscow, 1999, p. 846–869.
2. *Arutyunova N. D.* (2006), Inexplicable meanings: just like that, but not so simple [Neissledimyye smysly: prosto tak, da ne tak] // Proceedings of the Russian Academy of Sciences. A series of literature and language [Izvestiya RAN. Seriya literatury i yazyka]. Vol. 65, № 2, 2006, p. 14–22.
3. *Baranov A. N.* (1996), Sluzhebnye slova kak obyekt issledovaniya avtorskoj leksikografii (po krajnej mere vs. po men’shej mere v hudozhestvennyh tekstah Dostoevskogo) [Discursive words as an object of study of author’s lexicography (krajnej mere vs. po men’shej mere in Dostoevsky’s works)] // Word of Dostoevsky [Mir Dostoevskogo]. Moscow, 1996, p. 110–136.
4. *Baranov A. N., Dobrovol’skij D. O.* (2018), Kstati and nekstati: discourse practices in Dostoevsky’s works [Kstati i nekstati: k rechevym praktikam Dostoyevskogo], Russian language and linguistic theory [Russkiy yazyk v nauchnom osveshchenii], No. 1 (35), pp. 33–45.
5. *Eykhenbaum B. M.* (1969), On prose [O proze], Khudozhestvennaya literatura, Leningrad.

6. *Lotman Yu. M.* (1998), The structure of the artistic text [Struktura khudozhestvennogo teksta], On art [Ob iskusstve], Iskusstvo, St. Petersburg.
7. *Shaykevich A. Ya., Andryushchenko V. M., Rebetskaya N. A.* (2003), Statistical Dictionary of Dostoevsky [Statisticheskiy slovar' yazyka Dostoyevskogo], Yazyki slavyanskikh kul'tur, Moscow.
8. *Shaykevich A. Ya., Andryushchenko V. M., Rebetskaya N. A.* (2013), Distributive-statistical analysis of the language of Russian prose of 1850–1870-ies [Distributivno-statisticheskiy analiz yazyka russkoy prozy 1850–1870-kh gg], Vol. 1, Yazyki slavyanskikh kul'tur, Moscow.
9. *Shaykevich A. Ya., Andryushchenko V. M., Rebetskaya N. A.* (2016), Distributive-statistical analysis of the language of Russian prose of 1850–1870-ies [Distributivno-statisticheskiy analiz yazyka russkoy prozy 1850–1870-kh gg], Vol. 2, Yazyki slavyanskikh kul'tur, Moscow.
10. *Tomashevsky B. V.* (2002), Theory of Literature. Poetics [Teoriya literatury. Poetika], Aspekt Press, Moscow.
11. *Vinogradov V. V.* (1980), Selected Works. On the language of fiction [Izbrannyye trudy. O yazyke khudozhestvennoy prozy], Nauka, Moscow.

LANGUAGE MODEL EMBEDDINGS IMPROVE SENTIMENT ANALYSIS IN RUSSIAN

Baymurzina D. R. (dilyara.rimovna@gmail.com),
Kuznetsov D. P. (kuznetsov.den.p@gmail.com),
Burtsev M. S. (burtsev.m@gmail.com)

Neural Networks and Deep Learning Lab, Moscow Institute of Physics and Technology, Moscow, Russia

Sentiment analysis is one of the most popular natural language processing tasks. In this paper we introduce pre-trained Russian language models which are used to extract embeddings (ELMo) to improve accuracy for classification of short conversational texts. The first language model was trained on Russian Twitter dataset containing 102 million sentences, while two others were trained on 57.5 million sentences of Russian News and 23.9 million sentences of Russian Wikipedia articles. Although classifiers trained on top of language models perform better than in the case of utilizing of fastText embeddings of the same language style, we show that domain of language model also has a significant impact on accuracy. This paper establishes state-of-the-art results for RuSentiment dataset improving weighted F1-score from 72.8 to 78.5. All our models are available online as well as the source code which allows everyone to apply them or fine-tune on domain-specific data.

Key words: ELMo, embeddings from language model, text classification, sentiment analysis, Russian language

ПОВЫШЕНИЕ КАЧЕСТВА АНАЛИЗА ТОНАЛЬНОСТИ НА РУССКОМ ЯЗЫКЕ С ИСПОЛЬЗОВАНИЕМ ВЕКТОРНЫХ ПРЕДСТАВЛЕНИЙ ЯЗЫКОВЫХ МОДЕЛЕЙ

Баймурзина Д. Р. (dilyara.rimovna@gmail.com),
Кузнецов Д. П. (kuznetsov.den.p@gmail.com),
Бурцев М. С. (burtsev.m@gmail.com)

Лаборатория нейронных систем и глубокого обучения, Московский физико-технический институт (национальный исследовательский университет), Москва, Россия

Анализ тональности является одной из наиболее популярных задач обработки естественного языка. В данной работе мы представляем предобученные русские языковые модели, которые используются для получения векторных представлений слов при решении задачи классификации разговорных текстов. Одна языковая модель обучена на 102 миллионах предложений русского Twitter, а две другие — на 57,5 миллионах предложений русских новостей и 23,9 миллионах предложений из русских статьей Wikipedia. Несмотря на то, что классификаторы, обученные на векторных представлениях, извлеченных из языковых моделей, показывают результаты лучше, чем те, что обучены на векторных представлениях fastText соответствующего языкового стиля, мы показываем, что домен языковой модели также оказывает значительное влияние на качество классификации. В данной работе достигается новое наилучшее качество для набора данных RuSentiment, повышающее предыдущий результат с 72,8 значения взвешенной F1-метрики до 78,5. Все представленные модели и исходный код, в том числе для дообучения языковых моделей, доступны онлайн.

Ключевые слова: ELMo, embeddings from language model, классификация текстов, анализ тональности, русский язык

1. Introduction

Sentiment classification is an important part of chat-bots, from question answering helper on web-site to personal assistant that should track owner's mood and desires. The reason of the statement is that conversation with chat-bot should gratify a user but strongly in accordance to a situation.

There are three basic approaches to sentiment classification task: rule-based solution, machine learning (ML) models and neural networks (NN). Rule-based approach is the most popular because it does not require labelled datasets but only sentiment dictionaries. However, rule-based models often do not take into account context wider than two or three tokens. If it is possible to collect and annotate a domain-specific dataset, one can use supervised ML or NN models. While ML models are usually build upon embeddings of full text sample obtained from TF-IDF or count vectorizers, NN models assume character or token vector representations. Token embeddings could be obtained via many different methods including bag-of-words, GloVe [15], fastText [1]. However, token embeddings extracted from language models are becoming more and more popular. Language model embeddings allow to perform better even on small task-specific datasets which are often encountered in production.

Embeddings from Language Models (ELMo) [17] are vectors derived from bidirectional LSTM trained to solve the task of language modelling on a large text corpus. ELMo representations are deep and context-dependent. Internal states of the model can be combined and used similarly to other token embeddings like fastText but representation of each word is being formed by left and right context of this word. Language models require large text corpora and significant computational resources to be trained.

We have explored several discussions in Russian NLP community about actual performance of ELMo, and faced a lot of negative responses about accuracy of neural

models based on ELMo. Therefore, the paper has two main goals: first of all, we introduce three Russian language models pre-trained on Wikipedia articles, news and twits, and the second one is to compare performance of fastText and ELMo embeddings trained on corpora with different language styles. We demonstrate how the domain of language model influences on the accuracy of a classifier trained over obtained embeddings. Also we introduce the source code which allows to simply fine-tune ELMo on the domain specific data.

2. Related Work

A lack of studies on Russian sentiment analysis is caused by a lack of appropriate datasets. First of all, the largest sentiment lexicon is RuSentiLex [11] which latest version is dated by 2017 although neologisms appear regularly by borrowing from other languages or from positive and negative happenings in political, social and cultural life of Russia.

There are three common datasets for Russian sentiment analysis in academic research: aspect-oriented SentiRuEval 2015 [10], SentiRuEval 2016 [12] and RuSentiment [20]. In this paper we focus only on the second dataset, its description is set out in **section 3.2**.

All the word representations before ELMo were context-independent. Although some of them take into account sub-word information [1] or learn sense-depended word vectors to solve lexical ambiguity problem, none of the approaches consider context for word representation. Announced in [17] high performance of embeddings from language models applied to most of NLP tasks, specifically text classification, textual entailment, named entity recognition, question answering, coreference resolution and semantic role labelling opened a new room for research. In recently published paper [9] authors achieve state-of-the-art results on named entity recognition built upon Russian ELMo.

ELMo's achievements induced popularity of transfer learning approach when complex architecture pre-trained on language modelling task should be fine-tuned for solution of some other supervised problem [18].

3. Data

3.1. Language modelling data

The Russian language models corresponding to official language style were trained on Wikipedia[1] and Russian WMT News[2] while the Russian conversational language model was trained on Russian twits[3]. Clue characteristics of the datasets are presented in **Table 1**.

[1] https://ru.wikipedia.org/

[2] http://www.statmt.org/

[3] https://twitter.com/

Table 1: Data characteristics

Dataset	Number of words	Vocabulary size	Average number of words per sentence	File Size
Wiki	472 M	5.6 M	19.4	4.8 Gb
WMT News	1,133 M	4.1 M	19.6	12.0 Gb
Twitter	887 M	11.3 M	8.7	7.9 Gb

Preprocessed and cleaned WMT News sets are available for downloading, Wikipedia was spared from html-markup, and all hashtags and user logins were replaced by special tokens in Twitter. The vocabulary size for each dataset was set to 1 million frequency tokens. Finally, every dataset was splitted on training (98%) and validation (2%) samples.

3.2. Classification data

RuSentiment was published in 2018 [20] along with baseline results. The full dataset contains more than 30 thousands social media posts of average length 17 tokens, each post is related to one of five classes: positive, negative, neutral, speech and skip. Currently this is the largest publicly available dataset on Russian sentiment analysis. Around 21 thousands posts were randomly selected, and almost 7 thousands were pre-selected with an active learning-style strategy in order to diversify the data. We divide "random posts" subset on train and validation sets in a ratio of 9/1. The "pre-selected posts" set is not used in this paper. The test set is the same as in the original paper.

Linguists emit five Russian language styles: scientific, official, journalistic, artistic and colloquial. The first four styles and the last one differ a lot in terms of vocabulary and morphology. Therefore, we chose RuSentiment as the target dataset in this paper because the content relates to conversational style which often is not included to language modelling data while it is of current interest due to increasing popularity of chat-bots.

4. Experiments and Results

In this paper we explore the following token embeddings to cover different language styles:

- fastText embeddings trained on Russian Wiki and News corpora,
- fastText embeddings trained on Russian Twitter corpus,
- ELMo trained on Russian WMT News dataset,
- ELMo trained on Russian Wikipedia dataset,
- ELMo trained on Russian Twitter dataset,
- ELMo trained on Russian Twitter dataset and fine-tuned on RuSentiment.

300-dimensional fastText embeddings were trained with default parameters for skipgram model taking into account character n-grams from 3 to 6 characters.

4.1. Training and fine-tuning of language models

Language model consists of two main components: convolutional layers and 2 blocks of two recurrent layers. In the original implementation model receives as input indices of symbols in utf-8 encoding (from 0 to 255 plus three special symbols for padding, start and end of word). LSTM blocks pass forth and back over representations from convolutional layers, each block in its own direction similarly to bidirectional LSTM.

Training is being done in the similar to [6] and [8] way. An additional feed-forward layer followed by softmax is used to train language model. The model predicts words in direct and reverse orders for each LSTM blocks separately. The feed-forward layer is not used anymore after language model was fitted. To obtain context-dependent word representation weighted sum of word representations from all layers is used. Coefficients of this sum can be trained, and then can be different for all tasks. The upper layer also can be used similarly to TagLM [16] and CoVe [13]. Sentence representation is often formed as average or TF-IDF weighted sum [19] of word vectors.

This paper used model 4096/512 with 93.6 million of parameters[4]. The results of training language models on Wikipedia, WMT News, Twitter and fine-tuning of Twitter language model on RuSentiment data are presented in **Table 2**. Every language model was trained for 10 epochs in parallel on three 1080ti. Fine-tuning was conducted up to validation perplexity increase. The resulting perplexity of language model on "random posts" set of RuSentiment is 159.2 which was achieved after 4 epochs before overfitting began. The pre-trained language models were tested on full "random posts" set of RuSentiment. The resulting perplexity values are presented in **Table 2** in the last column. The language model trained on Twitter corpus performs best on RuSentiment dataset that was expected as language styles of corpora coincide.

Table 2: Results of training and fine-tuning ELMo

Data	Training time	Epochs	Perplexity on valid	Perplexity on RuSentiment
Wiki	6 days	10	43.692	17,364.89
WMT News	14 days	10	49.876	360.97
Twitter	10 days	10	94.145	172.25
Fine-tuning of Twitter on RuSentiment	15 min	4	159.2	—

Table 3 is presented for rough and fast estimation of the selected datasets similarity. As a metric of comparison, a perplexity of a bi-gram language model was chosen. The bi-gram model is to predict the conditional probability $P(w_n|w_{n-1})$ of a word w_n given the preceding word w_{n-1}. A KenLM [3] was used as an implementation

[4] https://allennlp.org/elmo, https://s3-us-west-2.amazonaws.com/allennlp/models/elmo/2x4096_512_2048cnn_2xhighway/elmo_2x4096_512_2048cnn_2xhighway_options.json

of the fast N-gram language model. The resulting perplexity values of bi-gram models trained on a corresponding dataset are diagonal elements of Table 3. Other elements show how accurately a bi-gram model from one specific domain (rows) predicts words of test set from another specific domain (columns). As shown in Table 3 the Twitter bi-gram language model predicts words of RuSentiment significantly better than those trained on WMT News and Wiki. Simultaneously, RuSentiment bi-gram model predicts words of Twitter dataset with quality comparable to model trained on Twitter.

Table 3: The perplexity of word bi-gram models on testing sets

Bi-gram model\Data	RuSentiment	WMT News	Twitter	Wiki
RuSentiment	116.67	4,847.68	**9,094.83**	**7,151.52**
WMT News	369,864.24	640.55	434,928.31	10,381.87
Twitter	**46,657.95**	1,740.06	6,762.07	8,330.85
Wiki	189,929.95	**1,583.86**	197,762.66	1,586.13

4.2. Training classifiers

There are two main approaches for text classification: convolutional and recurrent networks. Therefore, consider SWCNN [7] and BiGRU [2], [5] basic architectures of this paper.

The first model, shallow-and-wide convolutional neural network (SWCNN) illustrated in Fig. 1, sends non-trainable token embeddings to three convolutions with the same number of filters and different kernel sizes, each of which is followed by batch normalization layer [4], ReLU activation and global max pooling to reduce dimensionality. Pooled outputs are concatenated along the last dimension, and given to dense layer followed by batch normalization and ReLU activation. The output is given to classification dense layer also followed by batch normalization and softmax activation. Two dropout layers are placed directly before dense layers, and kernels are L2-regularized [14].

Bidirectional GRU (BiGRU) is demonstrated in Fig. 2. Non-trainable token embeddings are sent to bidirectional GRU layer which is followed by global max and average pooling. Pooled outputs are concatenated with two last states from BiGRU, and sent to dense layer followed by ReLU activation. Then output is given to the last classification dense layer followed by softmax activation. Two dropout layers are placed directly before dense layers, and kernels are also L2-regularized.

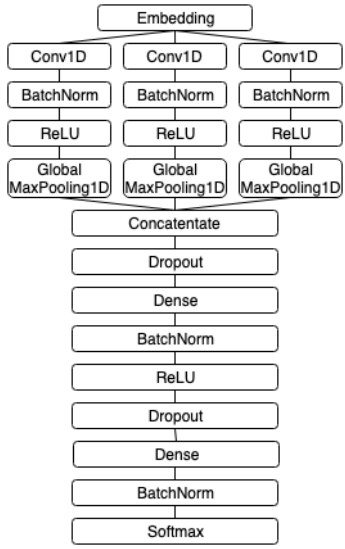


Figure 1: Shallow-and-wide CNN

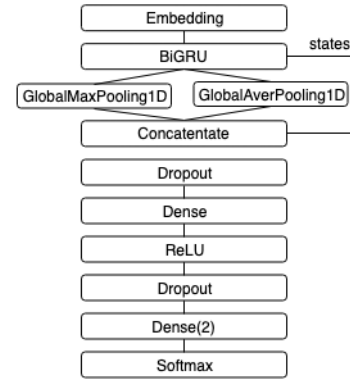


Figure 2: Bidirectional GRU

Baseline models are two networks of the above described architectures trained upon pre-trained fastText embeddings of dimensionality 300. The fastText skipgram model of official language style was trained on Russian Wikipedia and news corpora, fastText skpigram conversational style model was trained on Twitter dataset, both fastText models are available for downloading[5]. To explore domain-dependency of language models we also consider neural networks receiving token ELMo representations of dimensionality 1,024. The target metric is weighted F1-score, training is due to excess of patience limit.

All the experiments were conducted with the same parameters. Convolutional layers had 256 filters and kernels of sizes 3, 5, 7 while BiGRU layer had 256 units. The first dense layer had 100 units for both networks. Patience limit was set to 2, and maximum number of epochs was equal to 10. SWCNN models were strongly regularized with dropout rate of 0.5 and L2-coefficients 10^{-3} and 10^{-2} for convolutional and dense kernels. BiGRU model had dropout rate of 0.2, and L2-coefficient 10^{-6} for both recurrent and dense kernels.

[5] http://docs.deeppavlov.ai/en/latest/intro/pretrained_vectors.html

Table 4: Resulting scores on RuSentiment with different embeddings

Model	Embeddings	Validation F1-weighted	Test F1-weighted
Rogers et al. [20]	fastText VK	—	72.80
SWCNN	fastText Wiki+News	67.84	70.27
BiGRU	fastText Wiki+News	69.54	71.74
SWCNN	fastText Twitter	70.91	73.03
BiGRU	fastText Twitter	72.62	74.45
SWCNN	ELMo WMT News	70.27	72.42
BiGRU	ELMo WMT News	70.15	71.37
SWCNN	ELMo Wiki	68.11	71.28
BiGRU	ELMo Wiki	66.55	69.47
SWCNN	ELMo Twitter	75.40	78.50
BiGRU	ELMo Twitter	75.89	77.62
SWCNN	ELMo Fine-tuned	74.74	77.98
BiGRU	ELMo Fine-tuned	75.75	77.19

Each experiment was run for 4 times, the resulting averaged weighted F1-scores are presented in Table 4. For fastText embeddings BiGRU shows better than SWCNN results while for ELMo convolutional models outperform recurrent. Embedding models corresponding to official and journalistic language styles have almost the same scores with original paper [20] (71.7 weighted F1-scores when "pre-selected posts" were not used). Although fastText embeddings trained on Twitter dataset for both architectures beat not only baseline from [20] but all the models trained on domains of official (Wiki) and journalistic (News) styles, they are significantly transcended by conversational (Twitter) embeddings from language models. The best results (almost 6 points higher than previous state-of-the-art) are enriched by shallow-and-wide convolutional network trained on top of embeddings from Twitter language model.

5. Discussion

We have trained two popular architectures on 6 different embeddings of official, journalistic and conversational language styles. As the domain of target sentiment classification dataset is related to conversational language it was expected to obtain better results for conversational embeddings but the rate of the increase of scores is dramatic. Embeddings from language models not only appropriate but obligatory to be used in classification tasks if the domain of language model and target problem are close. Let us demonstrate several examples which support the statement in Table 5. One can pay attention to lexicon of the presented test samples, and which domain of language embeddings is closer than others.

Table 5: Examples of mistakes of models trained on top of different embeddings

Text sample	True label	ELMo News	ELMo Wiki	ELMo Twitter
василий зе бест!	*positive*	skip	skip	*positive*
вкусняшка, омном-ном	*positive*	neutral	skip	*positive*
полнейший зашквар назначать некогда хорошего футболиста сразу главным тренером «реала»	*negative*	neutral	neutral	*negative*
я променяла вас на диплом! а еще на министерское тестирование и гос экзамены!!я 0 числа уже с дипломом в зубах буду!!	*positive*	*positive*	skip	*negative*
все! завтра улетаю на евро- 0 в польшу болеть за сборную россии!	*positive*	*positive*	neutral	neutral
ну кто еще теперь задаст вопросы «зачем нами эта олимпиада?» «зачем нам спорт высоких достижений?». ведь можем же, когда захотим...	neutral	*negative*	neutral	*negative*

To summarize, we have introduced pre-trained Russian language models which allow to perform better, and to be evidential we have demonstrated how embeddings from language model outperform common fastText embeddings in Russian sentiment analysis task. Simultaneously, we have shown how significant the dependency of quality on the language model's domain is.

Acknowledgements

This work was supported by National Technology Initiative, and PAO Sberbank project ID 0000000007417F630002.

References

1. *Bojanowski, P. et al.:* Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics. 5, 135–146 (2017).
2. *Cho, K. et al.:* Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078. (2014).
3. *Heafield, K. et al.:* Scalable modified Kneser-Ney language model estimation. In: Proceedings of the 51st annual meeting of the association for computational linguistics. pp. 690–696, Sofia, Bulgaria (2013).
4. *Ioffe, S., Szegedy, C.:* Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. pp. 448–456 (2015).

5. *Johnson, R., Zhang, T.:* Supervised and semi-supervised text categorization using lstm for region embeddings. arXiv preprint arXiv:1602.02373. (2016).
6. *Jozefowicz, R. et al.:* Exploring the limits of language modeling. arXiv preprint arXiv:1602.02410. (2016).
7. *Kim, Y.:* Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882. (2014).
8. *Kim, Y. et al.:* Character-aware neural language models. In: Thirtieth aaai conference on artificial intelligence. (2016).
9. *Konoplich, G. et al.:* Named entity recognition in russian with word representation learned by a bidirectional language model. In: Conference on artificial intelligence and natural language. pp. 48–58 Springer (2018).
10. *Loukachevitch, N. et al.:* SentiRuEval: Testing object-oriented sentiment analysis systems in russian. In: Proceedings of international conference dialog. pp. 3–13 (2015).
11. *Loukachevitch, N. V., Levchik, A.:* Creating a general russian sentiment lexicon. In: LREC. (2016).
12. *Lukashevich, N., Rubtsova, Y. V.:* SentiRuEval-2016: Overcoming time gap and data sparsity in tweet sentiment analysis. In: Компьютерная лингвистика и интеллектуальные технологии. pp. 416–426 (2016).
13. *McCann, B. et al.:* Learned in translation: Contextualized word vectors. In: Advances in neural information processing systems. pp. 6294–6305 (2017).
14. *Ng, A. Y.:* Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In: Proceedings of the twenty-first international conference on machine learning. p. 78 ACM (2004).
15. *Pennington, J. et al.:* Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp). pp. 1532–1543 (2014).
16. *Peters, M. E. et al.:* Semi-supervised sequence tagging with bidirectional language models. arXiv preprint arXiv:1705.00108. (2017).
17. *Peters, M. E. et al.:* Deep contextualized word representations. arXiv preprint arXiv:1802.05365. (2018).
18. *Radford, A. et al.:* Improving language understanding by generative pre-training. URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf. (2018).
19. *Robertson, S.:* Understanding inverse document frequency: On theoretical arguments for idf. Journal of documentation. 60, 5, 503–520 (2004).
20. *Rogers, A. et al.:* RuSentiment: An enriched sentiment analysis dataset for social media in russian. In: Proceedings of the 27th international conference on computational linguistics. pp. 755–763 (2018).

BERT FINETUNING AND GRAPH MODELING FOR GAPPING RESOLUTION

Belkin I. (ilya.belkin-trade@yandex.ru)
Moscow Institute of Physics and Technology, Moscow, Russia

This paper reports our participation in the Automatic Gapping Resolution for Russian shared task (AGRR-2019) within Dialogue Evaluation 2019. Our team took the first place among other nine teams in all subtasks which includes gapping presence-absence classification, gap resolution and full annotation. The phenomenon of gapping is well theoretically studied. However, the problem of automatic gapping resolution is new and there is no baseline for it. We found it possible to bring this task into sentence classification and token tagging problems and solve them using recent advances in Natural Language Processing and deep learning. Training large language models with millions of parameters on small data became possible with the development of transfer learning methods. Using pretrained models for computer vision problems is straightforward and since BERT language model was realized it became possible to benefit from transfer learning in NLP. Our solution is heavily based on BERT, but we found that parsing gapping constructions, which are very structured, benefit from special postprocessing which includes modeling a gapping in the form of a directed graph. Our solution may be considered as the first public baseline for the task of automatic gapping resolution which is based on NLP modern practices.

Key words: gapping resolution, language model, transfer learning, ensembling, contextual embeddings

РАЗРЕШЕНИЕ ГЕППИНГА С ПОМОЩЬЮ ДООБУЧЕННОЙ ЯЗЫКОВОЙ МОДЕЛИ BERT И ГРАФОВОЙ МОДЕЛИ ЯЗЫКОВОГО ЯВЛЕНИЯ

Белкин И. (ilya.belkin-trade@yandex.ru)
Московский физико-технический институт, Москва, Россия

1. Introduction

Automatic analysis of natural language is complicated by many factors one of which is the presence of rare sentence-level constructions. Omission from a clause of some words may not affect the meaning of the sentence for humans but may puzzle automatic system. This construction is known in linguistics as ellipsis and may

take different forms. One of them is gapping, that occurs in coordinate structures and elides a repeated predicate, typically from the second clause, with its participants remaining expressed. Gapping has been largely studied from theoretic point of view in [3], [6], [7], [8]. To recover the full meaning of the sentence one has to find the position of the elided predicate and the head of the corresponding predicate. Here an example of this construction in Russian:

(1) *Дайте мне две пятерки, а я вам [дам] десятку.*

The word in brackets is omitted in original sentence. Its absence does not affect the meaning of the sentence for human reader. That means that inner representation of the sentence in some system of automatic analysis of natural language with and without recovered word should be similar. To deal with, such constructions should be detected and parsed appropriately. This led to the problem of automatic gapping resolution.

Automatic gapping resolution is a new problem not only for Russian but for other languages too. There are no open source tools or published experiments on this task. One of the reasons is the lack of data. Gapping is a very rare phenomenon and collecting a large enough corpus is a complicated task.

Automatic Gapping Resolution for Russian shared task (AGRR-2019) within Dialogue Evaluation 2019 is a pilot event for gapping resolution task for Russian held for the first time. The data provided to the participants of the Shared task was not published anywhere before. We were free in problem interpretation and method selection. Our team decided not to dive into theoretical studying of gapping as it may take a long time, but to bring the problem to a known task in Natural Language Processing and apply modern methods to solve it. Our approach has shown the best results among other participants and here we give an overview of our system.

The paper is organized as follows. First, we describe the shared task setup. Second, we present our method. Third, we report on the quality of our system and results achieved.

2. Shared task overview

AGRR-2019 is the first public benchmark for algorithms of automatic gapping resolution for Russian. Usually, when reporting method performance on some dataset, there is no need to describe what data and markup we have, what evaluation procedure we follow. These are standardized for particular benchmark and usually well-known. But it is not our case. The dataset and even exact problem standing are new. In this section we give an overview of key points of the shared task which are essential for further reading and understanding of our approach. First, we describe three tasks, which compose the problem. Than we give a description of the data and its markup. And lastly, we cover the evaluation procedure for each of tasks. For more information refer to [13].

2.1. Problem description

The meaningful statement of the problem was discussed in the introduction, and in this section, we give a more formal statement. The general problem of gapping resolution was represented to participants as three tasks, each of which was evaluated separately.

Binary presence-absence classification. For every sentence decide if there is a gapping construction in it.

Gap resolution. Predict the position of the elided predicate and the correspondent predicate in the antecedent clause.

Full annotation. In the clause with the gap predict the linear position of the elided predicate and annotate its remnants. In the antecedent clause find the constituents that correspond the remnants and the predicate that corresponds the gap.

Participants were free in method selection for each of tasks.

2.2. Data description

Participants were provided with three datasets: Train, Dev, Add. The first two were labeled manually and include 16,406 and 4,142 sentences respectively. Add dataset consists of 115,536 samples. It was obtained automatically by ABBYY Compreno [11] and, as were mentioned by organizers, may contain some mistakes. Test data was realized by the end of the AGRR-2019 and includes 20,951 samples. Each sentence with the gapping was marked and annotated with two remnants R1 and R2, their correlates in the antecedent clause cR1 and cR2, the position of the elided predicate V and the head of the correspondent predicate cV.

(2) *Тогда я **cV[** принял **cV] cR1[** ее **cR1] cR2[** за итальянку **cR2]**, а **R1[** его **R1] V[] cR2[** за шведа **cR2]**.*

Gap resolution task includes prediction of positions of tokens labeled as cV and V; participation in full annotation task requires to recover the whole markup for the sentence. Not all sentences contain both remnants, some contain only the first.

(3) *Но Христианин сказал Упрямому: — Нет, сосед, лучше ты последуй примеру Сговорчивого. Мы в самом деле там **cV[** получим **cV] cR1[** то, о чем я сейчас говорил **cR1]** , а сверх того — **V[] R1[** великую славу **R1]** .*

Others may contain more than one clause with the gap.

(4) *Индекс **cR1[** промышленного производства **cR2]** за январь-февраль 2008 г. **cV[** составил **cV] cR2[** 106,0% **cR2]** , **R1[** инвестиций в основной капитал **R1] — V[] R2[** 120,2% **R2]** и **R1[** оборота розничной торговли **R1] — V[] R2[** 116,3% **R2]**.*

Also, it is worth noting that the position of elided predicate is always the beginning of the second remnant R2, or, in sentences with only one remnant—the beginning of the first one R1. Sentences without gapping are taken into account only in binary presence-absence classification task.

2.3. Evaluation and metrics

Evaluation of participants' systems was performed on the Test dataset which was released two days before the deadline and did not contain markup. The main metric for binary classification task was standard f-measure. Gapping element annotations was measured by symbol-wise f-measure. E.g. if the gold standard offset for certain gapping element is 10:15 and the prediction is 8:14, we have 4 true positive chars, 1 false negative char and 2 false positive chars and the resulting f-measure equals 0.727.

3. Our approach

We consider the problem of gapping resolution as sequence labeling and sentence classification problem. Our solution is based on BERT and we follow model design and finetuning procedure as described in [5]. We use special post-processing procedure to obtain predictions which satisfy strict conditions observed from real data. And we use ensembling to clean automatically collected sentences.

3.1. Input Representation

Each sentence was converted to appropriate form for BERT model. This process includes WordPiece tokenization, adding special token [CLS] for sentence level classification as the first token and [SEP] as the last one, padding with [PAD] token to fixed length (128 in our experiments) and converting into indices of model vocabulary.

Also, in early experiments, we found that some additional preprocessing steps improve validation quality and accelerate convergence. First, we replace comma with point in decimals. And second, we replace dash with hyphen since the first one is not presented in the vocabulary.

3.2. Model

We follow model design proposals from the original paper [5]. By incorporating BERT with one additional linear layer model may be adapted for various tasks which include single sentence classification and single sentence tagging tasks. However, instead of training separate models for gapping absence-presence classification and full annotation tasks we combine these two models as shown in the **Fig. 1**. Distributed representation for the first token is used for binary sentence classification. Other tokens are classified into five classes. We do not predict the position of the elided predicate because it is always the beginning of R2 or R1. For more information on this see Data description section.

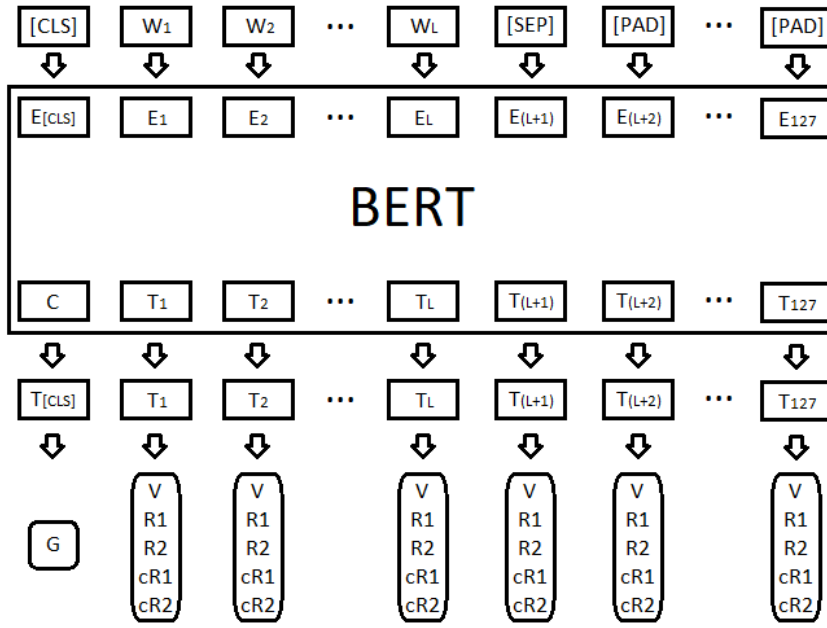


Fig. 1. Model architecture for gapping resolution problem

3.3. Training

Instead of training from scratch, we use pretrained weights [12] for BERT layers and finetune the model in three stages. First, we freeze BERT part and train only top linear layers for two cycles [9, 10]. In one cycle we linearly adjust learning rate up to 0.001 for the first 20% of cycle and then linearly decrease to the initial value of 1e-5. In the second stage, we train full model for two cycles. One cycle lasts two epochs and includes linear learning rate warmup for the first 10% of the time to 2e-5 and linear decreasing to the initial value. In the last stage, we train our model for two cycles with cycle size of three epochs and a linear warmup to 1e-6 for the first 30% of the cycle.

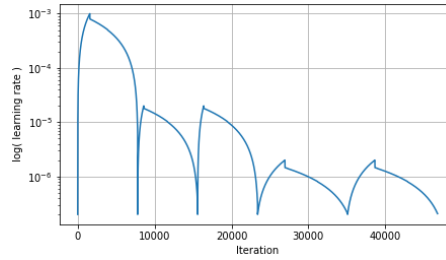


Fig. 2. Learning rate schedule of final model

Training procedure of models in ensemble is quite similar. First, we train only top layers for one cycle. Cycle size equals to two epochs, learning rate linearly increases to 0.001 during the first 20% of cycle. Finetuning stage includes three similar cycles with cycle size of two epochs and linear warmup of learning rate up to 2e-5 for the first 10% of cycle.

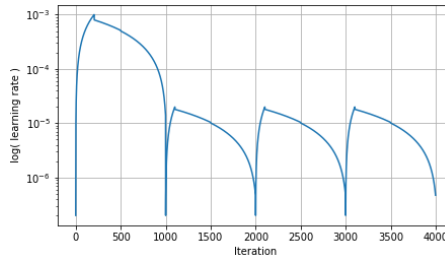


Fig. 3. Learning rate schedule of models in ensemble

3.4. Post-processing

To obtain resulting predictions we perform some additional steps on the raw outputs of the model. One possible way is to take argmax over output distributions for each token. However, this approach does not address some peculiar properties of the problem and lead to prediction errors which may be easily avoided.

By analyzing sentences with gapping, we found two such properties. First, possible tags at each position depend on all previous tags. Second, given a sentence the number of segments labeled as R1 and R2 are equal. To visualize possible sequences of tags we build a directed graph where each node corresponds to one or more sequential tokens of input. Nodes "_start" and "_end" correspond to [CLS] and [SEP] tokens respectively. Edges correspond to tokens with no label.

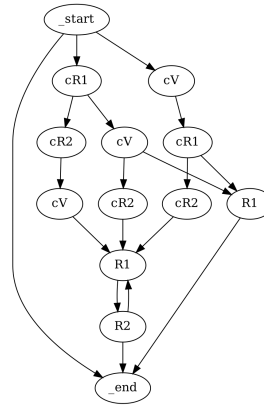


Fig. 4. Graph representation of gapping construction

Given a probability distribution over tags for each position of input we may compute a likelihood of any path in this graph. But we interested in the max-likelihood path. Fortunately, this may be easily found by Dijkstra algorithm [2].

3.5. Data cleaning

Since deep neural networks perform better with more data it is straightforward to concatenate Train and Add datasets and use all available data for training. But wrong labeled samples may harm convergence and lead to worse results and generalization. To deal with we use ensembling to filter the large dataset. In detail, we concatenate train and dev sets, shuffle them and split in five folds. Then we five times select one for validation and the rest for training. Since we use cycling learning rate scheduling it is straightforward to use checkpoint ensembling [4] to diversify ensemble by using weights of the network at the end of each cycle. For each train/validation split, we save three checkpoints and thus obtain 15 trained models. Then we filter the large dataset as follows. Given a sentence we calculate predictions of 15 networks and include this sentence to resulting training set if at least one condition holds: prediction of at least one model matches with its markup; predictions of at least 8 networks are the same, in this case, we omit markup provided and use ensemble answer as a ground truth. Following this procedure does not increase dataset diversity as networks are already able to give the right answer. However, even 'similar' samples may be good for training as they provide regularization and may increase generalization. Total amount of data used for training our final model was 126,202 sentences. We refer to this enlarged dataset as "Train + Add".

4. Results

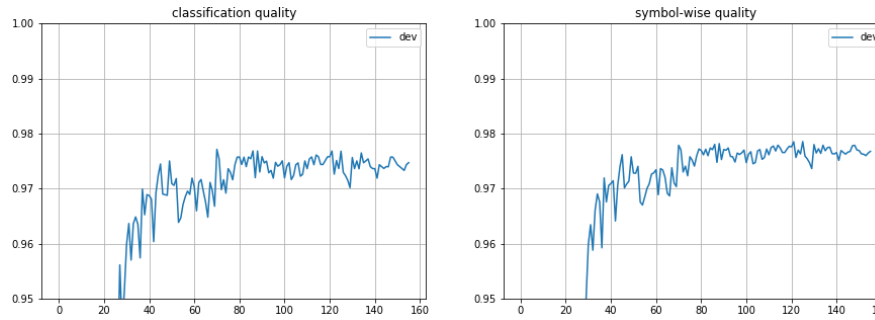


Fig. 5. Final validation metrics

In this section we report quality achieved on Train/Dev/Test datasets as well as give some analysis of model errors. Our result was the best among competitors and brought our team the first prize. The best quality on validation was achieved on 71 iteration and weights saved at the end of this iteration were used to obtain test predictions. Validation quality plot does not show signs of overfitting such as increasing of model error, but there is a large gap between train and validation/test quality.

Table 1. Results of final model on different subsets

	Binary classification quality	Gapping resolution quality
Train + Add	0.9988	0.9895
Train	0.9973	0.9784
Dev	0.9778	0.9313
Test	0.9590	0.8901

We have looked at some mistakes of our final model and observe that some of them are really ambiguous and difficult to avoid. However, it also makes some silly mistakes. We argue this is because of some overfitting. We give some examples of false negatives and false positives in Russian language and provide glosses(word-by-word translation) and translation into English to make sentences clear for non-Russian-speaking reader.

False negatives:

(5) *Russian: В жестах — нет песен, в музыке — ритма.*
Gloss: In gestures — no songs, in music rhythm.
English: There are no songs in gestures, in music — rhythm.

The word "нет" in Russian may be a particle or a verb depending on the context. In this sentence this is a verb and a part of gapping construction.

(6) *За 15 лет добросовестной службы в МЧС всё стало родным и все родными.*
Gloss: For 15 years of conscientious service in the Ministry of Emergency Situations, everything became relatives and all relatives.
English: We all became a family and everything became native for 15 years of conscientious service in the Ministry of Emergency Situations.

This example can illustrate model overfitting. In most sentences gapping construction is highlighted by punctuation, but not in this case

False positives:

(7) *Но к тому времени Джек был влюблен в Синди Пейдж, сейчас — миссис Джек Свайтек.*
Gloss: But by that time Jack was in love with Cindy page, now — Mrs. Jack Switek.
English: Jack was in love with Cindy Page, now Mrs. Jack Switek, by that time.

In this sentence "*Синди Пейдж*" and "*миссис Джек Свайтек*" are two names of the same person before and after wedding.

(8) *Раздался вой, а потом удаляющийся топот.*
Gloss: Rung howl, and then receding tramp.
English: There was a howl, and then the receding tramp.

This sentence is an example of markup mistake. It contains gapping but it is not labeled and thus considered as an algorithm error.

5. Conclusion

In this paper we present our winning solution of the shared task on automatic gapping resolution for Russian within Dialogue Evaluation 2019. This is the first time when the data for this problem was published and there are no public results for this task. To set up the first one we follow recent practices in natural language processing and training neural models in general. Our solution does not lean on gapping research, instead we follow data-driven approach. We think that combining these two approaches will lead to significant progress in gapping resolution.

References

1. *Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin* (2017) Attention Is All You Need, available at https://arxiv.org/abs/1706.03762.
2. *Dijkstra E. W.* (1959) A note on two problems in connexion with graphs, Numer. Math, Vol. 1, Iss. 1, P. 269–271.
3. *Elizabeth Coppock* (2001) In deffence of deletion, available at: https://www.linguistics.northwestern.edu/documents/award-winners/linguistics-undergraduate-award-past-winner-coppock.pdf.
4. *Hugh Chen, Scott Lundberg, Su-In Lee* (2017) Checkpoint Ensembles: Ensemble Methods from a Single Training Process, available at https://arxiv.org/abs/1710.03282.
5. *Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova* (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, available at https://arxiv.org/abs/1810.04805.
6. *John Robert Ross* (1970) Gapping and the order of constituents, Progress in linguistics: A collection of papers, 43:249–259.
7. *Jorge Hankamer* (1979) Deletion in coordinate structures, Garland Publishing, Inc., New York & London.
8. *Kyle Johnson* (2014) Gapping, available at: http://people.umass.edu/kbj/homepage/Content/gapping.pdf.
9. *Leslie N. Smith* (2017) Cyclical Learning Rates for Training Neural Networks, available at https://arxiv.org/abs/1506.01186.
10. *Leslie N. Smith* (2018) A disciplined approach to neural network hyper-parameters: Part 1 — learning rate, batch size, momentum, and weight decay, available at https://arxiv.org/abs/1803.09820.
11. https://www.abbyy.com/ru-ru/science/technologies/compreno/.
12. https://github.com/huggingface/pytorch-pretrained-BERT.
13. https://github.com/dialogue-evaluation/AGRR-2019.

АННОТИРОВАНИЕ ПРАГМАТИЧЕСКИХ МАРКЕРОВ В РУССКОМ РЕЧЕВОМ КОРПУСЕ: ПРОБЛЕМЫ, ПОИСКИ, РЕШЕНИЯ И РЕЗУЛЬТАТЫ[1]

Богданова-Бегларян Н. В. (n.bogdanova@spbu.ru)
Блинова О. В. (o.blinova@spbu.ru)
Мартыненко Г. Я. (g.martynenko@spbu.ru)
Шерстинова Т. Ю. (t.sherstinova@spbu.ru)
Зайдес К. Д. (kristina.zaides@student.spbu.ru)
Попова Т. И. (tipopova13@gmail.com)

Филологический факультет СПбГУ,
Санкт-Петербург, Россия

В статье описывается опыт аннотирования прагматических маркеров (ПМ) в двух русских речевых корпусах: «Один речевой день» (ОРД; диалоги) и «Сбалансированная аннотированная текстотека» (САТ; монологи). Для подготовки сплошной разметки ПМ было проведено 4 пилотных аннотирования на выборках из ОРД и САТ, что позволило сформировать итоговый список ПМ: 450 единиц, представляющих собой варианты 53 базовых структурных типов. В ходе обработки результатов пилотного аннотирования удалось получить предварительные данные о частоте встречаемости отдельных прагматических маркеров и их типов, а также о зависимости употребления ПМ от пола и уровня речевой компетенции говорящего. В результате обработки данных были получены частотные списки как самих ПМ, так и выполняемых ими функций.

Ключевые слова: русская повседневная речь, речевой корпус, прагматический маркер, корпусная разметка, монолог, диалог

[1] Исследование выполнено при поддержке гранта РНФ «Система прагматических маркеров русской повседневной речи» (проект № 18-18-00242).

PRAGMATIC MARKERS ANNOTATION IN RUSSIAN SPEECH CORPUS: RESEARCH PROBLEM, APPROACHES AND RESULTS

Bogdanova-Beglarian N. V. (n.bogdanova@spbu.ru)
Blinova O. V. (o.blinova@spbu.ru)
Martynenko G. Ya. (g.martynenko@spbu.ru)
Sherstinova T. Yu. (t.sherstinova@spbu.ru)
Zaides K. D. (kristina.zaides@student.spbu.ru)
Popova T. I. (tipopova13@gmail.com)

Philological Faculty of St. Petersburg State University,
St. Petersburg, Russia

The article describes the experience of pragmatic markers (PM) annotation in two Russian speech corpora: “One Speaker’s Day” (ORD; dialogues) and “Balanced Annotated Textotec” (SAT; monologues). To prepare an optimal PM annotation scheme, 4 pilot annotations were conducted on samples from ORD and SAT. It made it possible to form the final list of PM: 450 units, representing variants of 53 basic structural types. Processing the results of the pilot annotation allowed to obtain preliminary data on frequency of individual pragmatic markers and their types, as well as on the dependence of PM usage on sex and the level of speech competence of the speaker. As a result of statistical data processing, frequency lists of both PMs and their functions were obtained. The most commonly used in the dialogue are the PM *вот*, which is usually used as a «boundary marker» (G), and the PM *там*, which is usually used as a hesitative and/or rhythm-forming marker. In the monologue, the upper zone of the frequency list of the PMs is also full of boundary markers (G), marking the beginning/end of the monologue or serving as navigators in the text (*вот/ну вот, значит, так*). The most frequent types of PMs in dialogue are: X (hesitative markers), M (meta-communicative marker), GX (boundary/hesitative marker), K (xeno-indicator marker that introduces someone’s speech), RX (rhythm-forming/hesitative marker). In the list of the most frequent types of PMs in monologue speech, the markers of the type GX (boundary/hesitative marker) and X (hesitative marker) are in the lead. The analysis of the frequency lists of PMs showed that we can talk about statistically significant differences in the use of PMs in dialogue and monologue.

Keywords: Russian everyday speech, speech corpus, pragmatic marker, corpus annotation, monologue, dialogue

1. Введение

Обработка естественного языка, организованного в корпус, предполагает присвоение компонентам текстов знаков аннотации [Захаров 2005]; [Плунгян 2008]. Традиционные виды аннотации (морфологическая, синтаксическая и др.) реализуются в корпусах различного типа (см., например: [Gries, Berez 2017]), в том числе — с помощью различных программ (см., например: [Kuzmenko 2017]).

Однако существует разметка, провести которую автоматически крайне сложно. Речь идет о таких элементах структуры устного дискурса, которые уместно назвать *прагматемами*, или *прагматическими маркерами* (ПМ) [Богданова-Бегларян 2014]. ПМ активно функционируют в нашей речи: говорящий использует их, вербализуя трудности речепорождения, подыскивая нужное слово или производя метаязыковое комментирование сказанного. Автоматическое аннотирование ПМ затруднено прежде всего тем, что внешне они ничем не отличаются от значимых единиц и лишь в контексте реализуют свой новый статус, появляющийся в результате процесса *прагматикализации* (см. подробнее, например: [Bogdanova-Beglarian, Filyasova 2018]). Разметка ПМ в речевом корпусе не является тривиальной и требует значительной ручной работы экспертов-филологов (о путях решения возникающих проблем см.: [Zaides et al. 2018]).

2. Материал и методика

Материалом для выявления инвентаря ПМ русской устной речи и построения такой их типологии, которая была бы пригодна для аннотирования больших массивов данных, стали два корпуса (см. о начале этой работы: [Bogdanova-Beglarian et al. 2018]).

1. Корпус повседневной русской речи «Один речевой день» (ОРД), один из наиболее представительных на сегодняшний день ресурсов для анализа русского устного дискурса ([Русский язык… 2016]; [Bogdanova-Beglarian et al. 2016a, b, 2017]; [Богданова-Бегларян и др. 2017а]).
2. Корпус «Сбалансированная аннотированная текстотека» (САТ), включающий записи монологической речи, полученные от разных профессиональных групп носителей языка. Все тексты в САТ построены в рамках 4-х коммуникативных сценариев (чтение[2], пересказ, описание изображения, рассказ (см.: [Звуковой корпус… 2013], [2014], [2015]; [Богданова-Бегларян и др. 2017б]).

За основу словника ПМ была взята типология прагматем Н. В. Богдановой-Бегларян [Богданова-Бегларян 2014]. Слегка переработанный, данный словник был расширен за счет всех возможных структурных вариантов (расширений

[2] О неподготовленном *чтении* как разновидности спонтанного монолога см.: [Звуковой корпус… 2013].

базовой единицы), а также с учетом всех грамматических форм ПМ — для удобства их автоматического поиска по транскрипту и сведения в единую базу данных. В этот список попали как относительно частотные прагматические маркеры (*это самое, (ну) (я) не знаю, такой, короче, значит, (и) всё такое (прочее)* и др.), так и менее частотные, но регулярно употребляемые в устной речи (*типа того (что), боюсь (что), вроде, как бы, (и) всё такое (прочее)* и др.). Общий список изначально насчитывал 65 единиц. Тем самым был определен предварительный инвентарь единиц, подлежащих аннотированию в материалах монолога (САТ) и диалога (ОРД).

3. Подходы к аннотированию ПМ

Первичное аннотирование ПМ в корпусном материале проводилось непосредственно в среде ELAN, поддерживающей «привязку» разметки к определенному сегменту звукового сигнала. При этом было введено четырехуровневое аннотирование:

- уровень 1. PM — ПМ в той форме, как он представлен в транскрипте.
- уровень 2. Function PM — основные и дополнительные функции.
- уровень 3. Speaker PM — код говорящего.
- уровень 4. Comment PM — уровень комментариев.

Далее был пересмотрен и сокращен перечень функций ПМ; используемые для обозначения функций коды упрощены до однобуквенных, снято требование выделения основной функции:

- А — маркер-аппроксиматор,
- Г — разграничительный маркер (стартовый, финальный и навигационный),
- Д — дейктический маркер,
- З — все виды маркеров-заместителей (чужой речи, ряда перечисления или их частей),
- К — маркер-ксенопоказатель,
- М — маркер-метакоммуникатив,
- Ф — маркер-рефлексив,
- Р — ритмообразующий маркер,
- С — маркер самокоррекции,
- Х — хезитативный маркер.

Было введено также понятие базового варианта ПМ, который и указывался при аннотировании, что способствовало получению более однородной разметки.

Переработанная методика прошла успешную апробацию на материале второго этапа аннотирования [Bogdanova-Beglarian et al. 2018] и была адаптирована для разметки корпуса САТ.

4. Аннотирование ПМ в диалогической и монологической речи

Различия в подходах к аннотированию ПМ в речи разного типа вызваны тем, что в двух использованных корпусах данные представлены принципиально различно: в корпусе ОРД используется разметка, выполненная в среде ELAN (формат *.eaf), а корпус САТ не имеет многоуровневой разметки и представляет собой массив звуковых файлов и файлов расшифровки формата *.doc.

Для аннотирования ПМ в обоих корпусах были подготовлены две выборки материала. Для анализа *диалогической речи* было отобрано 149 эпизодов из корпуса ОРД, записанных от 98 информантов (вместе с их коммуникантами) (всего 308 905 словоупотреблений). В подкорпус вошли эпизоды «речевых дней» информантов разных профессиональных групп, преимущественно ситуации неформального общения. В группе информантов — 45 женщин (46 %) и 53 мужчины (54 %). Для исследования монологической речи из корпуса САТ были отобраны тексты разного типа, записанные от 34 информантов (всего 50 128 словоупотреблений). В подкорпусе представлена речь информантов, принадлежащих к двум профессиональным группам, — юристов и медиков.

4.1. Процедура пилотного аннотирования

Всего было осуществлено 4 этапа пилотного аннотирования материала.

Первое пилотное аннотирование было выполнено на выборке из корпуса ОРД объемом в 16 000 словоупотреблений параллельно 4-мя экспертами, по правилам, разработанным на подготовительном этапе. В ходе разметки использовался расширенный список функций ПМ, при этом аннотаторы выделяли главную из них и помещали соответствующий тег на первое место в боксе уровня «Function PM». Дополнительные функции перечислялись далее в алфавитном порядке. На уровне «Comment PM» отмечались некоторые дополнительные особенности употребления маркеров: например, редукция формы ксенопоказателя *говорит* до *грит* или *гыт* или особое интонационное оформление ПМ.

Сами теги представляли собой обозначения соответствующей функции. Возможные новые ПМ, а также различные варианты уже имеющихся в списке отмечались с помощью специальной пометы на уровне «Comment PM».

Анализ результатов первого пилотного аннотирования показал, что инструкция по разметке требует доработки. В ходе подготовки инструкции для *второго пилотного аннотирования* было решено использовать более короткий список функций ПМ и перечислять основные и дополнительные функции в одном ряду по алфавиту, поскольку практически каждый ПМ в устной речи оказывался полифункциональным, а иерархия выполняемых им функций при разметке выстраивалась не всегда однозначно и единогласно. Анализ результатов позволил оптимизировать методику и выработать более эффективную инструкцию для разметчиков [Bogdanova-Beglarian et al. 2018]. Переработанная методика прошла успешную апробацию на втором этапе аннотирования и сохранялась без существенных изменений на третьем и четвертом этапах.

Третье пилотное аннотирование было проведено на подкорпусе САТ (15 000 словоупотреблений). Оно выполнялось для предварительной оценки особенностей употребления ПМ в монологической речи и позволило сопоставить частоту употребления ПМ в зависимости от УРК[3] говорящего.

Четвертое пилотное аннотирование было проведено на подкорпусе ОРД (60 000 словоупотреблений). Выполнялось оно для предварительной оценки особенностей употребления ПМ в диалогической речи и позволило сделать некоторые выводы об особенностях использования ПМ в мужской и женской речи.

В конце каждого этапа пилотного аннотирования осуществлялась экспертная корректура прагматической разметки, пересматривался и дополнялся перечень выделяемых ПМ. На данный момент рабочий список вариантов ПМ насчитывает 450 единиц, представляющих собой варианты 53 базовых структурных типов.

5. Некоторые количественные характеристики ПМ в диалогической и монологической речи (сравнение ОРД и САТ)

Обработка результатов аннотирования ПМ в корпусном материале позволила получить данные о частоте встречаемости отдельных прагматических маркеров, а также о зависимости употребления ПМ от характеристик говорящего. Статистическая обработка результатов третьего и четвертого этапа пилотного аннотирования позволила получить выводы относительно наиболее употребительных ПМ и их функциях. Приведем некоторые из полученных данных.

5.1. Прагматические маркеры в ОРД и САТ

Частотные списки ПМ представлены в табл. 1, где приведены: ранги, частоты ПМ в абсолютных цифрах, доли конкретных ПМ от всех ПМ в выборке (в %), доли конкретных ПМ от всех слов выборки (в %) и ipm.

Размеры выборки ОРД (диалогическая речь) — 60 000 словоупотреблений. Размеры выборки САТ (монологическая речь) — 15 000 словоупотреблений.

Самыми употребительными в диалоге ПМ оказались: *вот*, чаще всего выступающий как дискурсивный маркер Г, и *там*, выступающий, как правило, в роли хезитативного и/или ритмообразующего маркера. Входят в эту зону также метакоммуникативы (М) *да* и *знаешь*, хезитативы (Х) *как бы*, *это*, *это*

---

[3] *Уровень речевой компетенции* определяется как степень свободы говорящего в выборе речевых средств, уровень его владения языковыми возможностями, его способность решать те или иные коммуникативные задачи. УРК коррелирует с двумя социальными характеристиками говорящего: высшее образование + профессиональное отношение к речи (преподаватели, актеры, лекторы, дикторы, политики…) → высокий УРК; высшее образование + непрофессиональное отношение к речи → средний УРК; отсутствие высшего образования + непрофессиональное отношение к речи → низкий УРК. Как убедительно показал анализ материала, реальные лингвистические корреляты имеют только полярные типы — высокий и низкий УРК [Звуковой корпус… 2013].

самое, *короче* и *так* и маркер-ксенопоказатель (К) *говорит* (чаще редуцированный). Частотность маркеров типа М в ОРД не случайна: в диалоге говорящие действительно постоянно вынуждены обращаться к собеседнику, так или иначе привлекая, а затем и удерживая его внимание, или передавать чужую речь.

Таблица 1 Наиболее частотные ПМ в ОРД и САТ

ранг	ПМ	f (ПМ)	доля от ПМ (%)	доля по выборке (%)	ipm
ОРД					
1	*вот*	149	14,06	0,25	2483
2	*там*	117	11,04	0,20	1950
3	*да*	82	7,74	0,14	1367
4	*говорит*	70	6,60	0,12	1167
5	*как бы*	60	5,66	0,10	1000
6	*это*	44	4,15	0,07	733
7	*это самое*	43	4,06	0,07	717
8	*знаешь*	41	3,87	0,07	683
9	*короче*	38	3,58	0,06	633
10	*так*	36	3,40	0,06	600
САТ					
1	*вот*	139	51,48	0,92	9232
2	*значит*	15	5,56	0,10	996
3	*так*	15	5,56	0,10	996
4	*там*	13	4,81	0,09	863
5	*как бы*	12	4,44	0,08	797
6	*ну вот*	12	4,44	0,08	797
7	*всё*	4	1,48	0,03	266
8	*и так далее*	4	1,48	0,03	266
9	*вот так вот*	3	1,11	0,02	199
10	*ну так*	3	1,11	0,02	199

В корпусе САТ верхняя зона частотного списка ПМ полна маркеров типа Г, маркирующих начало/конец монолога или служащих навигаторами по тексту: *вот/ну вот, значит, так*. Присутствуют в этой зоне и дейктические маркеры (Д) — *вот так вот*.

Для оценки различий между данными, полученными после составления частотных списков ПМ в ОРД и САТ, был использован тест Манна-Уитни; применялась программная среда R [R Core Team 2019]; сравнению подвергались значения ipm одних и тех же ПМ, см. **табл. 2**:

Таблица 2 Таблица для оценки различий между ОРД и САТ (фрагмент)

ПМ	ipm САТ	ipm ОРД
вот	9231,5870	2483,333
значит	996,2144	350
так	996,2144	600
там	863,3858	1950
как бы	796,9715	1000
ну вот	796,9715	350

В результате получены значения W = 2082, p-value = 1,077e-06, то есть $p < 0{,}001$. Таким образом, различия в употреблении ПМ между ОРД и САТ (т. е. в диалогической и монологической речи) можно признать статистически значимыми.

В ходе аннотирования помечались все встретившиеся в материале типы ПМ, как «чистые», так и «смешанные» (полифункциональные употребления) (АГ, АГХ, ГРХ, АФ и т. п.), отражающие общую полифункциональность ПМ, что весьма свойственно устной речи. Оказалось, что монофункциональных употреблений ПМ в ОРД существенно больше (68,7%), чем в САТ (37,4%). В ОРД чаще всего монофункциональными выступают такие ПМ, как Ф (рефлексив) (100,0%), М (метакоммуникатив) (93,0%) и З (маркер-заместитель) (91,7%). В САТ — К (ксенопоказатель) и Д (дейктический маркер) (по 100,0%). Среди полифункциональных преобладают ПМ, выполняющие, среди прочего, хезитационную функцию (АХ, АГХ, АКХ, ГХ, РХ и под.).

Наиболее употребительными в ОРД оказались прагматические маркеры Х (5283 ipm), М (3317 ipm) и ГХ (2417 ipm). В САТ — ГХ (6110 ipm), Х (4250 ipm) и АХ (2125 ipm). Видно, что среди маркеров, наиболее распространенных в САТ, преобладают ПМ хезитативного типа. В ОРД отчетливо преобладают метакоммуникативы (М): в диалоге говорящие часто вынуждены обращаться к собеседнику, так или иначе привлекая, а затем и удерживая его внимание, или передавать (пересказывать) чужую речь.

5.2. Функции прагматических маркеров в ОРД

Верхняя зона частотного списка типов ПМ в подкорпусе ОРД включает в себя следующие разновидности (см. **табл. 3**, в 23 случаях аннотаторы не смогли приписать ПМ функцию, такие случаи обозначены как NA).

Количество ПМ в речи отдельных информантов колеблется от 1 (0,09% от всего объема речевого материала по данному информанту) до 70 (И118; 6,6% от объема речевого материала информанта в выборке).

Таблица 3 Наиболее частотные типы прагматических маркеров диалогической речи

ранг	функция	кол-во	ipm	ранг	функция	кол-во	ipm
1	Х	317	5283	11	З	11	183
2	М	199	3317	12	Ф	9	150
3	ГХ	145	2417	13	ГМ	7	117
4	К	103	1717	14	МХ	7	117
5	РХ	70	1167	15	Р	6	100
6	АХ	52	867	16	АР	6	100
7	Г	33	550	17	ДХ	3	50
8	А	30	500	18	АРХ	3	50
9	NA	23	383	19	ГР	3	50
10	Д	20	333	20	ГРХ	3	50

В табл. 4 представлены данные, отражающие количество самых частотных функциональных типов ПМ в речи женщин и количество ПМ с теми же функциями в речи мужчин.

Таблица 4. Наиболее частотные типы ПМ в диалогической речи мужчин и женщин (фрагмент)

функция	ЖЕН	МУЖ
Х	216	101
М	143	56
ГХ	91	54
К	84	19
АХ	37	15
РХ	28	42
Г	24	9
А	17	13
Д	17	3
NA	11	12
З	8	3

Оценка статистической значимости различий в употребительности ПМ с разными функциями в речи мужчин и женщин выполнена с помощью критерия «хи-квадрат». Такая оценка показала, что различия являются статистически значимыми (X-squared = 273,18; $p < 0{,}001$), но аппроксимация может быть неправильной, то есть гипотеза о наличии значимых различий нуждается в дальнейшей проверке с привлечением большего объема размеченных данных. В речи женщин употребляются ПМ с 25 разными тегами функций, в речи мужчин — с 20 разными тегами, при этом 8 тегов в речи мужчин не встречается вовсе (АР, ГРХ, ДХ, КР, АГ, АГХ, ЗХ, МС).

5.3. Функции прагматических маркеров в САТ (зависимость от УРК говорящего)

Во всех частотных списках типов ПМ (общем и по трем УРК) первые две позиции уверенно занимают маркеры типа ГХ (пограничный маркер/хезитатив) и Х (хезитатив), ср.:

- общий список: 34,07 и 23,7 % соответственно;
- высокий УРК: 23,6 и 22,47 %;
- средний УРК: 38,89 и 19,44 %;
- низкий УРК: 39,45 и 27,52 %.

Видно, что поиск нужного слова или продолжения монолога, а также стремление просто выстроить связный текст — это главное в механизме спонтанного порождения устного текста, что вынуждает говорящего обращаться к специальным единицам — прагматическим маркерам соответствующих типов. Видно также, что доля таких ПМ возрастает по мере снижения УРК говорящего.

На третьем месте в трех списках из четырех (исключение — средний УРК) — маркер типа АХ, аппроксиматор/хезитатив, с помощью которого говорящий выражает и речевое колебание (чаще всего — поиск), и свою неуверенность в том, что подобрал нужное слово или верно выражает мысль. Количество таких ПМ в монологах достаточно велико:

- общий список: 11,85 %;
- высокий УРК: 15,73 %;
- низкий УРК: 11,93 %.

В речи информантов со средним УРК маркер АХ отошел на четвертую позицию, уступив место типу Г — маркерам начала/конца монолога или навигаторам по тексту.

Очевидно, что говорящие с любым УРК в равной степени испытывают трудности при спонтанном речепорождении и преодолевают эти трудности с помощью более или менее единого набора ПМ.

Анализ частотных списков ПМ в речи информантов САТ с разным УРК позволил сделать ряд наблюдений.

Так, в целом ПМ составили 1,8 % от общего массива слов в монологах-рассказах женщин-медиков (270 употреблений). Больше всего ПМ пришлось на группу информантов со средним УРК (группа Б, 1,39 %; 72 употребления), минимум — на группу с низким УРК (группа В, 0,92 %; 109 употреблений). Доля ПМ в группе информантов с высоким УРК (группа А) — 1,12 % (89 употреблений).

Во всех частотных списках (общем и по трем УРК) первое место уверенно занимает маркер *вот*, доля которого во всех случаях близка к 50 %: общий список — 51,48 %, высокий УРК — 48,31 %, средний УРК — 58,33 %, низкий УРК — 49,54 %.

Второе место в общем частотном списке ПМ занимает маркер *значит* (5,56 %), чаще свидетельствующий о низком УРК. Наши данные полностью подтвердили это предположение: употреблений *значит* в роли ПМ совсем не обнаружилось в речи информантов группы А, в группах же Б и В он занимает также второе место в соответствующих частотных списках (6,94 и 8,26 %), что

и обеспечило ему общее второе место по корпусу. Видно также, что по мере снижения УРК доля *значит* заметно возрастает.

Во всех четырех списках присутствуют и маркеры *как бы, ну вот, там, так* — по всей видимости, они более всего нужны любому говорящему для построения спонтанного монолога и менее всего при этом способны диагностировать УРК человека. *Так* и *там* в роли ПМ более всего представлены в речи информантов из группы В (низкий УРК — 7,34 и 6,42% соответственно; данные по всем монологам — 5,56 и 4,81%), Употреблений *как бы* и *ну вот* в роли ПМ больше всего в речи информантов из группы А (6,74 и 5,62%; общие данные — 4,44 и 4,44%). Средний УРК в рассматриваемом отношении ничем не примечателен.

Обращает на себя внимание также дейктический маркер *вот (…) вот*. В варианте *вот так вот* он присутствует в верхней зоне трех частотных списков ПМ: общем (1,11%), высокого УРК (2,25%) и низкого УРК (0,92%). В речи информантов группы Б (средний УРК) его в этой зоне не обнаружилось, в речи же информантов из группы В (низкий УРК) он представлен еще двумя структурными вариантами: *вот сейчас бы вот* и *вот эта вот* (по 0,92%).

6. Заключение

Проведенное исследование показало, что ПМ действительно представляют собой неотъемлемые элементы русского устного дискурса. В речи отдельных говорящих их доля может доходить до 6,6% от общего количества словоупотреблений, а в отдельных речевых фрагментах даже превышать долю значимых единиц.

Анализ частотных списков ПМ (общих по обоим корпусам и отдельных для разных групп говорящих) показал, что можно уверенно говорить о статистически значимых различиях в употреблении ПМ в диалоге и монологе.

Наиболее частотными функциями ПМ в речи всех групп информантов являются метакоммуникативная, разграничительная (дискурсивная), хезитативно-поисковая, и функция ксенопоказателя. Прагматические маркеры этих классов часто оказываются полифункциональными и реализуют ряд дополнительных функций.

Наиболее распространенным ПМ во всех частотных списках оказался *вот* (чаще — в разграничительной функции). В монологической речи высокую частоту встречаемости проявил маркер *значит* (как правило, в разграничительной или хезитативной функциях).

На основании полученных данных можно предварительно предположить, что частота использования ПМ в речи коррелирует с УРК говорящего: в среднем, чем он выше, тем меньше используется ПМ определенных типов, свидетельствующих о больших затруднениях говорящего в построении дискурса.

Все ПМ являются неизбежными элементами устной спонтанной речи, однако одни («хорошие» ПМ) не снижают качества речи, свидетельствуют об умении говорящего преодолевать естественные речевые сбои и не мешают восприятию и пониманию (высокий УРК), другие («плохие» ПМ) настолько ломают

структуру устного текста, что затрудняют понимание и свидетельствуют о низком качестве речи и неумении говорящего выстраивать связный устный текст (низкий УРК).

Наконец, пилотное аннотирование корпусного материала показало качественную неоднородность ПМ, проявляющуюся как в плане разнообразия выполняемых ими функций, так и в плане однозначности их выделения и отнесения этих единиц к прагматическим элементам устного дискурса, поэтому одной из перспективных задач предложенного направления исследования русской устной речи представляется выявление качественной дифференциации прагматических маркеров.

Литература

1. *Богданова-Бегларян Н. В.* (2014) Прагматемы в устной повседневной речи: определение понятия и общая типология // Вестник Пермского университета. Российская и зарубежная филология. — Вып. 3 (27), 2014. — С. 7–20.
2. *Богданова-Бегларян Н. В., Шерстинова Т. Ю., Блинова О. В., Мартыненко Г. Я.* (2017а) Корпус «Один речевой день» в исследованиях социолингвистической вариативности русской разговорной речи // Анализ разговорной русской речи (АРЗ–2017): Труды седьмого междисциплинарного семинара / Науч. ред. Д. А. Кочаров, П. А. Скрелин. — СПб.: Политехника-принт, 2017. — С. 14–20.
3. *Богданова-Бегларян Н. В., Шерстинова Т. Ю., Зайдес К. Д.* (2017б) Корпус «Сбалансированная Аннотированная Текстотека»: методика многоуровневого анализа русской монологической речи // Анализ разговорной русской речи (АРЗ–2017): Труды седьмого междисциплинарного семинара / Науч. ред. Д. А. Кочаров, П. А. Скрелин. — СПб.: Политехника-принт, 2017. — С. 8–13.
4. *Захаров В. П.* (2005) Корпусная лингвистика: Учебно-методическое пособие. — СПб.: СПбГУ, 2005. — 48 с.
5. *Звуковой корпус как материал для анализа русской речи* (2013) Коллективная монография. Часть 1. Чтение. Пересказ. Описание / Отв. ред. Н. В. Богданова-Бегларян. — СПб.: Филологический ф-т СПбГУ, 2013. — 532 с.
6. *Звуковой корпус как материал для анализа русской речи* (2014) Коллективная монография Часть 2. Теоретические и практические аспекты анализа. Том 1. О некоторых особенностях устной спонтанной речи разного типа. Звуковой корпус как материал для преподавания русского языка в иностранной аудитории / Отв. ред. Н. В. Богданова-Бегларян. — СПб.: Филологический ф-т СПбГУ, 2014. — 396 с.
7. *Звуковой корпус как материал для анализа русской речи* (2015) Коллективная монография. Часть 2. Теоретические и практические аспекты анализа. Том 2. Звуковой корпус как материал для новых лексикографических проектов / Отв. ред. Н. В. Богданова-Бегларян. — СПб.: Филологический ф-т СПбГУ, 2015. — 364 с.
8. *Плунгян В. А.* (2008) Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики // Русский язык в научном освещении. — № 16 (2), 2008. — С. 7–20.

9. *Русский язык повседневного общения: особенности функционирования в разных социальных группах* (2016) Коллективная монография / Отв. ред. Н. В. Богданова-Бегларян. — СПб.: ЛАЙКА, 2016. — 244 с.

References

1. *Bogdanova-Beglarian, N. V.* (2014), Pragmatems in Spoken Everyday Speech: Definition and General Typology [Pragmatemy v ustnoj povsednevnoj rechi: opredelenie pon'atia i obshchaja tipologia] // Perm University Herald. Russian and Foreign Philology [Vestnik Permskogo universiteta. Rossijskaja i zarubezhnaja filologia]. Iss. 3 (27), pp. 7–20.
2. *Bogdanova-Beglarian, N., Baeva, E., Blinova, O., Martynenko, G., Sherstinova T.* (2018), Towards a Description of Pragmatic Markers in Russian Everyday Speech // Speech and Computer. SPECOM 2018. Lecture Notes in Computer Science, vol. 11096. Springer, Cham / Karpov, A., Jokisch, O., Potapova, R. (eds.), pp. 42–48.
3. *Bogdanova-Beglarian, N., Blinova, O., Martynenko, G., Sherstinova T., Zaides, K.* (2018), Pragmatic Markers in Russian Spoken Speech: an Experience of Systematization and Annotation for the Improvement of NLP Tasks // Proceedings of the FRUCT'23. Bologna, Italy, 13–16 November 2018 / S. Balandin, T. Salmon Cinotti, F. Viola, T. Tyutina (eds.). FRUCT Oy, Finland, pp. 69–77.
4. *Bogdanova-Beglarian, N. V., Blinova, O. V., Sherstinova, T. Iu., Martynenko, G. Ja.* (2017a), Corpus "One Speaker's Day" in Studies of Sociolinguistic Variability of Russian Colloquial Speech [Korpus «Odin rechevoj den'» v issledovaniakh sociolingvisticheskoj variativnosti russkoj razgovornoj rechi] // Analysis of Spoken Russian (AR3–2017). Proceedings of the seventh interdisciplinary seminar [Trudy sed'mogo mezhdisciplinarnogo seminara]. St. Petersburg, pp. 14–20.
5. *Bogdanova-Beglarian, N., Filyasova, Yu.* (2018), Active Processes in Modern Spoken Language (Evidence from Russian) // Digital Transformation and Global Society. Third International Conference, Conference proceedings DTGS 2018, St. Petersburg, Russia, May 30 — June 2, 2018, Revised Selected Papers, Part II. Communications in Computer and Information Science (CCIS). Vol. 859 / D. A. Alexandrov, A. V. Boukhanovsky, A. V. Chugunov, Yu. Kabanov, O. Koltsova (eds.). Springer, Cham, pp. 391–403.
6. *Bogdanova-Beglarian N., Sherstinova T., Blinova O., Baeva E., Martynenko G., Ryko A.* (2016b), Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech // SPECOM 2016, Lecture Notes in Artificial Intelligence, LNAI, vol. 9811. Springer, Switzerland, pp. 659–666.
7. *Bogdanova-Beglarian, N. V., Sherstinova, T. Iu., Zajdes, K. D.* (2017b), Corpus "Balanced Annotated Text Library": Methodology Multi-Level Analysis of the Russian Monological Speech [Korpus «Sbalansirovannaja Annotirovannaja Tekstoteka»: metodika mnogourovnevogo analiza russkoj monologicheskoj rechi] // Analysis of Spoken Russian (AR3–2017). Proceedings of the seventh interdisciplinary seminar. Trudy sed'mogo mezhdisciplinarnogo seminara. St. Petersburg, pp. 8–13.
8. *Everyday Russian Language: Functioning Features in Different Social Groups* (2016), [Russkij jazyk povsednevnogo obshchenia: osobennosti funkcionirovania

v raznykh social'nykh gruppakh]. Bogdanova-Beglarian, N. V. (ed.). Collective Monograph, St. Petersburg, 244 p.

9. *Gries, S. T., Berez A. L.* (2017), Linguistic Annotation in/for Corpus Linguistics / N. Ide and J. Pustejovsky (eds.). Handbook of Linguistic Annotation. Berlin & New York: Springer, pp. 379–409.
10. *Kuzmenko, E.* (2017) Morphological Analysis for Russian: Integration and Comparison of Taggers / Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science, vol. 661. Springer, Cham, pp.162–171.
11. *Plungjan, V. A.* (2008), Corpus as a Tool and as an Ideology: on Some Lessons of Modern Corpus Linguistics [Korpus kak instrument i kak ideologia: o nekotorykh urokakh sovremennoj korpusnoj lingvistiki] // Russian Language in Scientific Description [Russkij jazyk v nauchnom osveshchenii]. 16 (2), pp. 7–20.
12. *R Core Team* (2019), R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.
URL: https://www.R-project.org/.
13. *Speech Corpus as a Base for Analysis of Russian Speech* (2013) Collective Monograph. Part 1. Reading. Retelling. Description [Zvukovoj korpus kak material dl'a analiza russkoj rechi: kollektivnaja monografia. Chast' 1. Chtenie. Pereskaz. Opisanie] / N. V. Bogdanova-Beglarian (ed.). St. Petersburg, 532 p.
14. *Speech Corpus as a Base for Analysis of Russian Speech* (2014) Collective Monograph. Part 2. Theory and Practice of Speech Analysis. Vol. 1. Some Features of Oral Spontaneous Speech of Various Types. Speech Corpus as a Base for Material for the Teaching of Russian as a Foreign Language [Zvukovoj korpus kak material dl'a analiza russkoj rechi. Kollektivnaja monografia. Chast' 2. Teoreticheskie i prakticheskie aspekty analiza. Tom 1. O nekotorykh osobennost'akh ustnoj spontannoj rechi raznogo tipa. Zvukovoj korpus kak material dl'a prepodavania russkogo jazyka v inostrannoj auditorii] / N. V. Bogdanova-Beglarian (ed.). St. Petersburg, 396 p.
15. *Speech Corpus as a Base for Analysis of Russian Speech* (2015) Collective Monograph. Part 2. Theory and Practice of Speech Analysis. Vol. 2. Speech Corpus as a Base for New Lexicographical Projects [Zvukovoj korpus kak material dl'a analiza russkoj rechi. Kollektivnaja monografia. Chast' 2. Teoretichskie i prakticheskie aspekty analiza. Tom 2. Zvukovoj korpus kak material dl'a novykh leksikograficheskikh proektov] / N. V. Bogdanova-Beglarian (ed.). St. Petersburg, 364 p.
16. *Zaides, K., Popova, T., Bogdanova-Beglarian, N.* (2018), Pragmatic Markers in the Corpus "One Day of Speech": Approaches to the Annotation // Computational Models in Language and Speech. Proceedings of Computational Models in Language and Speech Workshop (CMLS 2018) co-located with the 15th TEL International Conference on Computational and Cognitive Linguistics (TEL 2018). Vol-2303. Kazan, Russia, November 1, 2018 / Ed. by A. Elizarov, N. Loukachevitch. Kazan (Volga Region) Federal University, N. I. Lobachevsky, Institute of Mathematics and Mechanics, Kazan, Russia; Lomonosov Moscow State University, Research Computing Center, Moscow, Russia, pp. 128–143.
17. *Zakharov, V. P.* (2005), Corpus Linguistics: Teaching Aid [Korpusnaja lingvistika: Uchebno-metodicheskoe posobie]. St. Petersburg. 48 p.

KNOWLEDGE-BASED APPROACH TO WINOGRAD SCHEMA CHALLENGE

Boguslavsky I. M. (bogus@iitp.ru)[1,2],
Frolova T. I. (tfrolova@gmail.com)[1],
Iomdin L. L. (iomdin@gmail.com)[1],
Lazursky A. V. (lazursky@mail.ru)[1],
Rygaev I. P. (irygaev@gmail.com)[1],
Timoshenko S. P. (nyrestein@gmail.com)[1]

[1]A. A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia

[2]Universidad Politécnica de Madrid, Madrid, Spain

We propose a method to resolve anaphoric pronouns in the framework of Winograd Schema Challenge (WSC) by means of SemETAP—a knowledge-based semantic analyzer. WSC is a modern version of the famous Turing test. Its objective is to check a machine's ability to exhibit intelligent behavior indistinguishable from that of a human. In contrast to other approaches to WSC, which are based on machine learning, our method uses explicit knowledge. An important advantage of this approach is that it gives an opportunity to provide an explanation of the result understandable for humans. SemETAP interprets the text using both linguistic and extralinguistic (background) knowledge. The former is stored in the grammar and the dictionary of the ETAP-4 system, and the latter is provided by the SemETAP ontology, inference rules and the repository of individuals. We show how this knowledge is used for resolving WSC. At the moment, the performance of the algorithm is not high—54%. This is due to the incompleteness of the background knowledge supplied to the system. It is shown, however, that if the background knowledge is complete and accurate enough, the WSC test is resolved well and it is easily understandable why the system arrived at a particular conclusion.

Предлагается метод разрешения анафоры в рамках теста Winograd Schema Challenge (WSC) с помощью семантического анализатора SemETAP, основанного на знаниях. Тест WSC представляет собой современный вариант теста Тьюринга и предназначен для проверки того, в какой степени компьютер владеет фоновыми знаниями и некоторыми мыслительными операциями, свойственными человеку. В отличие от других подходов к WSC, использующих машинное обучение, наш метод основан на эксплицитных знаниях. Важное преимущество такого подхода состоит в том, что он позволяет дать обоснование полученного результата, понятное человеку. Для интерпретации текста SemETAP использует как лингвистические, так и внелингвистические (фоновые) знания. Лингвистические знания собраны в словарях и грамматике системы ETAP-4, а фоновые знания — в онтологии, массиве правил вывода и в базе индивидов. Мы показываем, какие знания и как используются для WSC-теста. Проведенная оценка алгоритма показала невысокий результат — 54 %. Это объясняется недостаточно

полными фоновыми знаниями, вложенными в систему. Тем не менее, показано, что, если фоновые знания системы достаточно детальны, WSC-тест дает хороший результат, обоснование которого легко понимается человеком.

Keywords: Winograd Schema Challenge, knowledge-based approach, knowledge representation, inference, Etalog language, anaphora resolution

1. Introduction

In this paper, we propose a knowledge-based method to tackle the problem known as Winograd Schema Challenge (WSC). This test is a modern version of the famous Turing test proposed in 1950 and since then playing an important role in the philosophy of artificial intelligence. Turing test is intended to check a machine's ability to exhibit intelligent behavior indistinguishable from that of a human. It consists in maintaining free conversation between a human and a computer through a text-only communication channel. The computer is considered as having passed the test if after a 5-minute conversation the human cannot reliably determine with whom he has conversed—with another human or with a computer. This test was strongly (and fairly) criticized for the central role of deception and trickery incorporated into the system. To pass the test, it was sufficient for the computer to fool the human into thinking he is dealing with another human by means of various tricks, puns, jokes, clever asides, emotional outbursts and the like. The weakness of the Turing test became especially obvious after it was successfully passed in 2014 by the chatbot Eugene Goostman who assumed a false identity of a 13-year old boy from Odessa. The ability to mislead the interlocutor, especially in the course of a short conversation, can hardly qualify as the most natural manifestation of computer intellect.

A better test was proposed in 2011 by Hector Levesque [Levesque 2011], cf. also [Levesque et al. 2011]. As opposed to the Turing test, the WSC test by Levesque requires an unambiguous answer to a series of questions that represent no difficulty for humans but need elementary background knowledge (or, in other words, that what is called "naïve picture of the world") and commonsense reasoning. These questions require to select the correct antecedent of an anaphoric pronoun in a sentence of a particular structure. Let us give two examples that rely on different types of knowledge.

(1) *The trophy does not fit into the brown suitcase because it's too [small/large].* What is too [small/large]?

(2) *Joan made sure to thank Susan for all the help she had [given/received].* Who had [given/received] help?

The questions are compiled in such a way that it is sufficient to replace one word with the other and the correct answer will be different. "With a very high probability, anything that answers correctly a series of these questions… is thinking in the full-bodied sense we usually reserve for people" [Levesque 2011]. However simple the test may seem, it is impossible to pass it if the computer does not possess basic knowledge

and reasoning capacities that any normal adult should possess. Hence, being able to make progress on this task enables us to move one step closer to building a machine that can truly understand natural language.

The rest of the paper is organized as follows. We review related work in **Section 2**. Most of the attempts to tackle WSC we are aware of are based on machine learning and statistics. In contrast, we are trying to advance within the knowledge-based framework. In **Section 3** we describe our approach, in **Section 4** we present semantic resources we are using. The key concept used for resolving WSC is the concept of semantic consistency, which is explained in **Section 5**. We report our experiments and analysis in **Section 6** and conclude in **Section 7**.

2. Related work

After the WSC was proposed in [Levesque 2011], it generated a lot of interest and in 2016 a first international WSC competition was organized within International Joint Conference on Artificial Intelligence in New York. The participants were offered a series of sentences similar to (1) and (2) above and related to different aspects of background knowledge. The prize could not be awarded to anybody. Most of the participants showed a result close to the random choice or even worse. The second competition scheduled for 2018 was canceled due to the lack of prospective participants. Thus, WSC sets a very high bar, which the current state-of-the-art can hardly overcome.

As of today, there have been several attempts to resolve WSC. All the authors recognize the necessity to take background knowledge into account but do it in very different ways. Most of the papers recur to some variant of machine learning including deep learning. Most attempts on solving WSC involve heavy utilization of annotated knowledge bases, rule-based reasoning, or hand-crafted features.

[Rahman and Ng 2012] employs the largest set of features derived from a variety of sources: narrative chains [Chambers and Jurafsky 2008], Google API, FrameNet information, heuristic polarity, machine-learned polarity, connective-based relations, semantic compatibility and lexical features.

[Haoruo Peng et al. 2015] develop the notion of Predicate Schemas, instantiate them with automatically acquired knowledge, and compile it into constraints that are used to resolve coreference. Specifically, two types of Predicate Schemas are introduced that cover a large fraction of the challenging cases. The first one specifies one predicate with its subject and object, thus providing information on the subject and object preferences of a given predicate. Example:

(3) *[The bee]*$_{e1}$ *landed on [the flower]*$_{e2}$ *because [it]*$_{pro}$ *had pollen.*
(The flower had pollen) IS MORE PROBABLE THAN (The bee had pollen).

The second type specifies two predicates with a semantically shared argument (either subject or object), thus specifying role preferences of one predicate, among roles of the other. Example:

(4) *[Bill]*$_{e1}$ *was robbed by [John]*$_{e2}$*, so the officer arrested [him]*$_{pro}$.
(Bill was robbed by John, The officer arrested John) IS MORE PROBABLE THAN (Bill was robbed by John, The officer arrested Bill).

These schemas are instantiated by acquiring statistics in an unsupervised way from multiple resources including the Gigaword corpus, Wikipedia, Web Queries and polarity information.

Trinh and Le 2018 propose a method for commonsense reasoning with neural networks, using unsupervised learning. No annotated knowledge bases or hand-engineered features are used. Large RNN language models are built that operate at word or character level. They are trained on unlabeled data taken in a number of massive and diverse text corpora, such as LM-1-Billion, CommonCrawl, SQuAD, Gutenberg Books. The authors claim that the diversity of training data plays an important role in test performance. The system successfully discovers important features of the context that decide the correct answer, indicating a good grasp of commonsense knowledge.

In contrast to this, the system described in [Quan Liu et al 2016], that obtained the highest score at the 2016 challenge, includes both supervised and unsupervised models. It makes use of the skip-gram model to learn word representations. The model incorporates several knowledge bases to regularize its training process, resulting in Knowledge Enhanced Embeddings (KEE). A semantic similarity scorer and a deep neural network classifier are then combined on top of KEE to predict the answers. The commonsense knowledge used is constituted by cause-effect relationship pairs automatically extracted from a large corpus. The pairs are composed of a verb and an adjective and belong to 4 types: "active V—positive A" (*win—happy*), "active V—negative A" (*rob—be arrested*), "passive V—positive A" (*be confident—not afraid*), and "passive V—negative A" (*be restricted—unable*). To combine context and commonsense knowledge for solving the WSC, the paper proposes to treat the commonsense knowledge as semantic constraints and learn KEE based on the generated constraints.

There are several papers that do not use machine learning but apply explicit knowledge representation and reasoning. The focus of the approach proposed in Schüller 2014 is on knowledge representation. To represent both the meaning of the text and the background knowledge, Roger Schank's graph framework is used. Inference is based on pragmatic effects described in Relevance Theory.

[Bailey et al. 2015] uses a series of axioms and inference rules. A mathematical framework for reasoning is introduced, based on the notion of correlation between the events. F and G are correlated if message F would cause the hearer to view G as more plausible, and message G would cause the hearer to view F as more plausible. For example, if we learn that "A fits into B", then it is more plausible that "B is big". It is supposed that such axioms can be acquired automatically from existing lexical and commonsense knowledge bases, such as WORDNET [Fellbaum 1998], FRAMENET [Baker, Fillmore, and Lowe 1998], VERBNET [Kipper-Schuler 2005], PROPBANK [Palmer, Gildea, and Kingsbury 2005], CONCEPTNET [Liu and Singh 2004), KNEXT [Schubert 2002], and the OPENCYC project (http://www.opencyc.org/doc).

The approach of [Sharma et al. 2015] is close to ours in the sense that it relies on a semantic parser to represent the meaning of the text and the background

knowledge. The parser produces graphs on which reasoning is performed. The concept of background knowledge adopted in the paper is somewhat unconventional. Knowledge is extracted from the web (or any other large text repository) on demand, individually for any processed sentence. The idea here is to extract sentences which contain commonsense knowledge required to answer the question about a given Winograd sentence. For example, for the initial sentence

(5) *The man couldn't lift his son because he was so weak*

and the question

(6) *Who was so weak?*

one needs to acquire the knowledge of the type "if X could not lift Y then X may be weak". This knowledge is looked for by creating string queries from the concepts in the sentence and the question and using them as queries to retrieve sentences from a text repository. For this example the query would be: ".*could not lift.*because.*weak.*". It is supplemented by another query obtained by substituting the verb for its synonym: ".*could not raise.*because.*weak.*" Practically, the process of knowledge hunting is nothing more than searching for sentences similar to the initial text in the hope to find a sentence in which the ambiguity in question would be resolved. In our example, such a sentence was found:

(7) *She could not lift it because she is a weak girl*,

in which a coreference resolver can determine that the two occurrences of *she* are coreferent. Each given sentence and the corresponding commonsense knowledge sentences are translated into semantic representation graphs, by using the K-Parser system. Semantic representation of the initial sentence (5) is compared with the semantic representation of the sentence (7) found in the corpus, which results in the inference that *he = man*. Rules and constructs for reasoning are formulated within the Answer Set Programming framework.

One could doubt that what the system extracts from the text repository is really deep background knowledge and not just a text similar to the initial one. Regardless, this method is closer to commonsense reasoning than many statistical or machine learning methods commonly used in NLP and in particular Natural language understanding (NLU).

3. Our approach

Nowadays, when computational linguistics is dominated by a powerful machine learning mainstream, the frameworks proposed beyond this mainstream have to justify their choice. We assume that computational linguistics is a fundamental branch of science at the intersection of linguistics and artificial intelligence. Its aim is to describe natural language by means of computer modeling. This is a contrast to NLP, which primarily aims at the development of useful applications. We believe that in many cases the choice of the paradigm is determined by the task the researchers are facing. In many cases, a linguistic model based on knowledge has a higher explanatory power

than a model obtained by machine learning, at least given the current state of technology. If we wish to learn which words are closer in meaning and which are farther, distributive semantics will be a good choice. However, if our task needs defining what the adjective *intelligent* exactly means and how it differs from its synonyms *clever* or *smart* (and any complete model of semantics should necessarily include this information), then machine learning will hardly solve our problem. If our target is to build a concrete useful application (e.g. a syntactic parser or a system of machine translation), it is quite probable that machine learning should be a paradigm of choice. If, however, what we want is not just to describe a certain linguistic phenomenon or a fragment of language but also to understand it and for instance to be able to compare it to a similar phenomenon of another language or of the same language but at a different period of time, then a machine learning-based model will not be the optimal one, since it will not be as transparent as a knowledge-based model can be.

Similar considerations can be heard from the computer science camp. Erik Mueller, a well-known specialist in artificial intelligence and commonsense reasoning and one of the key authors of IMB Watson, in his recent book "Transparent Computers: Designing Understandable Intelligent Systems" insists that computers should be more transparent, open and understandable. We should be able to understand why they arrived at a particular conclusion or why they behaved in a certain way. Accordingly, "intelligent systems should be able to reason like people, they should use concepts familiar to people and combine them in ways that make sense to people" [Mueller 2016]. Computers should explain their reasoning so as to help us decide whether to accept or reject the system's advice. It is important for computers to be transparent, because transparency promotes understanding, is educational, makes it easier to fix problems, improves customer satisfaction, and builds trust. Neural networks are like black boxes. You don't know what they are doing and how they come to the conclusion. At the same time, symbolic techniques, such as Cyc, Event Calculus or OntoSem, are close to be transparent. They represent knowledge symbolically, and they reason symbolically, in a way people can understand.

Coming back to modeling commonsense reasoning and, in particular, WSC, we are entering the field where transparency of the solution is especially desirable. Let us look at one of the examples mentioned above. To resolve the pronoun in sentence (2) above, [Haoruo Peng et al. 2015] uses the background information that the agent of robbing is often an object of arrest. Such information can be extracted from large corpora by means of statistical processing. However, even if this information is available, it only tells us what happens more often but does not provide any explanation or hint of why this should be the case. If a system is modeling commonsense reasoning, it is reasonable to expect some motivation of its decisions, which presumably requires availability of explicit knowledge. For example: ARREST is a curtailment of freedom of a person who committed a criminal action or is suspected thereof. On the other hand, ROBBING is a criminal action. This knowledge makes the assumption of coreference between the agent of ROBBING and the object of ARREST very plausible.

This is the way that we are trying to follow. We adopted what is often called *knowledge-based approach*: explicitly representing knowledge in a formal language, and providing procedures to reason with that knowledge. A major obstacle is that the

bulk of background knowledge of humans is not formalized. As of now, it is impossible to build a unified model to cover all this knowledge due to its boundlessness. The attempts to automatically extract background knowledge needed for reasoning are worthy of respect but, as far as we know, they did not deliver tangible result so far, if we are speaking about minimally non-trivial knowledge.

We assume that the formalization of this knowledge can be achieved along the line of incrementally building different fragments of the "naïve picture of the world" so that in the long run the whole picture is covered. This is a long way to go, and we are at its beginning.

This scenario does not seem unfeasible to many researchers. Cf. [Levesque et al. 2012]: "WSC allows for incremental progress: we can begin with simple lexical analyses of the words in the sentences, and then progress all the way to applying arbitrary amounts of world knowledge to the task … In addition, the schema can be grouped according to domain. Some examples involve reasoning about knowledge and communication; others involve temporal reasoning or physical reasoning. Researchers can choose to work on examples in a particular domain, and to take a test restricted to that domain". "Hand crafting microtheories" is one of the methods for creating commonsense knowledge bases, along with statistical and corpus-based machine learning techniques and crowd sourcing, which were invited for presentation at the Twelfth International Symposium on Logical Formalization of Commonsense Reasoning (http://commonsensereasoning.org/2015/cfp.html).

As of today, there have been several attempts of formalization of different fragments of human background knowledge (microtheories), beginning with very narrow scenarios (such as breaking an egg and pouring it into a bowl [Morgenstern 2001]) and ending with larger pieces of the "naïve picture of the world", such as emotions, interpersonal relations, naïve psychology, causality, change of state, etc.—[Gordon and Hobbs 2004]; [Gordon and Hobbs 2011]; [Gordon, Hobbs et al. 2011]; [Hobbs and Gordon 2008]; [Hobbs and Gordon 2010]; [Hobbs and Gordon, 2014]; [Hobbs, Sagae et al. 2012]; [Montazeri and Hobbs, 2011], [Montazeri and Hobbs 2012]). However, so far their number is absolutely insufficient for covering texts of general semantics.

We made one more step along this road. Out of a collection of WSC texts[1], we extracted a subcollection which requires knowledge on mental predicates, translated it into Russian and built a semantic description of the corresponding fragment of the human naïve picture of the world. This description is implemented as a set of inference rules (see **Section 4** below). The experiments that we carried out showed that in many cases this description ensures correct resolution of the WSC test, provided that the key elements of the text are covered by the description. It is essential that this formalization is not geared specifically to the WSC test. It is applicable in a wide range of tasks, such as question answering, extraction of implicit knowledge from the text, recognition of textual entailment, machine comprehension, etc.

[1] https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WSCollection.html

4. Inference rules

In our experiments, we used the SemETAP semantic analyzer. In previous publications, we described its different aspects [Boguslavsky et al. 2015], [Boguslavsky 2017], [Rygaev 2017], [Boguslavsky et al. 2018] and will not repeat them here. Let it only be reminded that:

- SemETAP is an option of the ETAP-4 linguistic processor and reuses its non-semantic modules (morphological analysis, syntactic dependency parsing, and normalization).
- Semantic analysis makes use of linguistic data and extralinguistic information (background knowledge). The linguistic data are provided by the Combinatorial Dictionary and the Grammar, and the background knowledge is stored in the Ontology, Repository of Individuals and the set of inference rules SemRule.
- Inference rules is a crucial component of SemETAP. We proceed from the assumption that the depth of understanding is growing with the number of inferences we can draw from the text. In many cases, a decomposition of the concept meaning helps produce additional inferences and thus achieve a deeper understanding.
- Two levels of semantic structure are distinguished. Basic semantic structure (BSemS) interprets the text in terms of ontological concepts. Enhanced semantic structure (EnSemS) extends Bsems by means of a series of inferences.
- From the formal point of view, semantic structures of both types are represented in the RDF format, i.e. as sets of triples of the type (*Ontoelement-1 relation Ontoelement-2*), where *Ontoelement-i* is a variable or a constant denoting a concept or an instance and *relation* is an object or data property of the ontology that holds between *Ontoelement-1* and *Ontoelement-2*.

For the purposes of this paper, EnSemS is the most important representation, since it makes explicit all the inferences that the knowledge available permits to make from the text and the context. The inferences, in their turn, may help select the more appropriate antecedent in the WSC sentence.

The rules that generate EnSemS, which we call inference rules, are mostly written in the Etalog language [Rygaev 2018]. At the time of writing this paper (February 2019), the number of Etalog rules has reached 408. There are two major types of inference rules—rules for concepts and rules for relations. Concept rules mostly decompose the meaning of the concept explicating components that are relevant from the inference perspective. For example, in describing events, special attention is paid to the following aspects:

a) preconditions that need to have taken place; for example, preconditions of the event “Ivan bought a book from Masha” are Masha’s having the book and Ivan’s having money. If Ivan refused to close the door, the precondition is that someone had asked him to do so.
b) objectives of its participants; for example, Ivan’s goal in the event “Ivan asks Masha what time it is it” is to make Masha tell him the time.
c) results of the event, both obligatory and possible; for example, the result of the event “Ivan bought a book from Masha” is Ivan’s having the book. If Ivan lost the book, it results in Ivan’s not having the book any more, the precondition being that he had it before.

d) if the event is complex, what are its subevents; for example, the event "Peter exchanged his book for Ivan's apple" consists of two acts of givings: Peter gave his book to Ivan and Ivan gave his apple to Peter.
e) presuppositions that the event may have; for example, if Ivan does not know that Peter failed the exam, it is still the case that Peter's failure at the exam did take place.
f) assessment of different components of the event from the point of view of its participants or the speaker; for example, the event "Ivan defeated Peter" is beneficial for Ivan and unbeneficial for Peter.

Fig. 1 presents the screenshot of the rule for the concept `Refusing` written in Etalog.

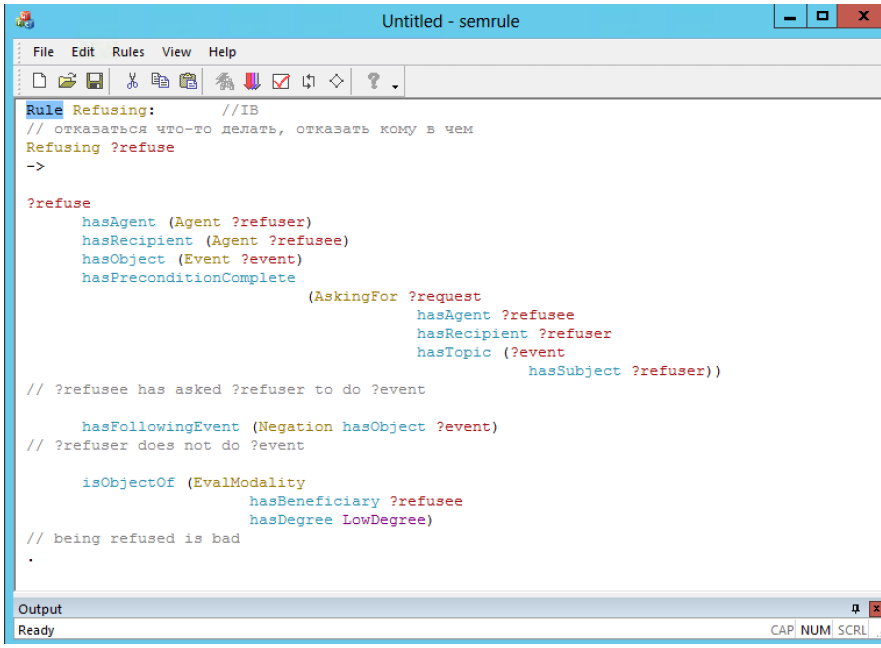


Fig. 1. Inference rule `Refusing`

It says that `Refusing` has an agent (`?refuser`), a recipient (`?refusee`) and an object (`?event`) that the agent refused to do. The precondition of `Refusing` is that `?refusee` has previously asked `?refuser` to do `?event`. The result of `Refusing` is that `?refuser` does not do `?event`. Besides, being refused is bad for `?refusee`.

Among the inference rules that describe relations, noteworthy is the group of rules that define modal and temporal relationships between events. These are: `hasResult`, `hasPossibleResult`, `hasPrecondition`, `hasPossiblePrecondition`, `hasSpeakerCommitment` (introduces factive complements: *Ivan does not know that Peter failed the exam*), `hasPreventedEvent` (introduces the consequence that did not take place: *Ivan forgot to call a taxi*), `hasSubEvent`,

`hasPossibleSubEvent` (introduces a subevent that may, but not necessarily, take place: *chewing* is a possible subevent of *eating*), `hasSubEventFinal` (introduces the final stage of a complex event), `hasSyncEvent` (the event is synchronized in time and modality with another event), `hasPossibleSyncEvent` (the event may be synchronized in time and modality with another event), `hasSyncAntiEvent` (the event excludes another event at the same time: *sleep—be awake*), `simultaneously` (the event is synchronized with another event in time but not in modality).

5. In search of semantic consistency

One of the first approaches to semantic consistency was proposed in [Apresjan 1974/1995:13–15]. It was applied to the task of word sense disambiguation. A sentence interpretation is considered semantically consistent when the repetition of semantic elements is maximal. This idea was illustrated by example (8)

(8) *Xorošij konditer ne žarit xvorost na gazovoj plite* 'a good pastry-cook does not fry pastry straws in a gas-stove'

Some of the words here are ambiguous. The word *xvorost* is ambiguous between 'pastry straws' and 'dry tree branches fallen on the ground'. *Plita* means either 'heating device for cooking food' or 'flat piece of solid material'. *Konditer* may mean 'a person that cooks pastry' or 'an owner of the confectionary'. Obviously, in all the three cases one should select the first of the senses indicated ('pastry straws', 'heating device for cooking food', and 'a person that cooks pastry'), because all these senses contain the semantic element 'cooking/cooked food'.

For WSC, one cannot use this approach directly. Often, competing interpretations differ not so much in the composition of semantic elements as in their organization. The same elements are organized in different predications, that is structures composed of a predicate and its arguments. Therefore, we modified slightly the notion of semantic consistency. We will consider sentence interpretation *Int1* more consistent than *Int2*, if it contains more identical predications. It is essential that we check consistency not on the initial text, nor on its syntactic structure and not even on the Basic Semantic Structure. We are searching the Enhanced Semantic Structure, because it contains the full body of inferred predications, and all of them should be taken into account. In more detail, the algorithm for determining consistency will be presented below. Here, we will illustrate its idea with a concrete example.

Here is one of the classical WSC pairs:

(9) *James asked Robert for a favor but he refused.*

(10) *James asked Robert for a favor but he was refused.*

We will proceed as follows. For each sentence, we will build two variants which differ in the antecedent selection. Then we will produce EnSemS for both and check which of them manifests a higher degree of consistency in the sense defined above. Let us begin with sentence (9).

(9a) *James asked Robert for a favor but Robert refused.*

(9b) *James asked Robert for a favor but James refused.*

Sentences (9a) and (9b) are composed of the same semantic elements, but organized differently. The key component for comparing (9a) and (9b) is *refusing*. As we saw above (Fig. 1), its meaning contains a reference to a precondition. A1's refusal to do A2 presupposes that A3 asked A1 to do A2 before. Hence, (9a) contains the element 'ask for' twice: 1) it makes part of the first part of the sentence ('James asked Robert for a favor') and 2) it is part of the precondition of refusal in the second part of the sentence ('somebody asks Robert for something'). These predications are identical up to unification. This means that they coincide, if the same variables are instantiated by the same expressions (James --> somebody, favor --> something). It's easy to see that in (9b) the corresponding predications do not unify: the addressee of the first occurrence of 'asking for' is Robert, while in the second occurrence it is James. So, (9a) has identical (up to unification) propositions, and (9b) does not have them. Consequently, (9a) is more consistent than (9b) in the sense indicated above. In the same way, one can show that (10a) is less consistent than (10b):

(10a) *James asked Robert for a favor but Robert was refused.*

(10b) *James asked Robert for a favor but James was refused.*

However, this time the difference between the predications in the less consistent sentence is due to different agents of 'asking for' and not addressees: in (10a) it is James in the first request, and Robert in the second one.

5.1. Algorithm for consistency check

The algorithm for determining consistency is based on the idea of similarity between two nodes in the graph. Similarity is a numerical value which can range from −1 (two nodes are absolutely different) to 1 (two nodes are identical).

The algorithm takes a pronoun node and calculates its similarity to each potential antecedent node. The one with a higher similarity is selected. Similarity is calculated in the following way:

1. Similarity of a node to itself is 1.
2. Similarity of two different constant nodes is −1.
3. Similarity of two nodes which have incompatible classes is −1.
4. Similarity of two nodes which have compatible classes is calculated based on their environment in the graph, namely:
    a. For each incoming and outgoing functional relation of the pronoun node find a corresponding relation of the antecedent node and match their values using the same algorithm,
    b. If there is no corresponding relation on the antecedent node then assume the similarity is 0.
    c. Return a total similarity of all the relations divided by the number of the relations.

Here is the (simplified) calculation for the case in (9):

Pronoun node	Antecedent node	Calculation
(Agent isRecipientOf(AskingFor hasAgent(Agent)))	(Human hasGivenName "Robert" isRecipientOf(AskingFor hasAgent(Human hasGivenName "James")))	Classes match. The value of the only relation (isRecipientOf) match with similarity = 1. So the total similarity is 1/1 = 1.
(Agent isRecipientOf(AskingFor hasAgent (Agent)))	(Human hasGivenName "James" isAgentOf(AskingFor hasRecipient(Human hasGivenName "Robert")))	Classes match. There is not match for the only relation (isRecipientOf) hence its value is 0. So the total similarity is 0/1 = 0.

6. Experimental results

We carried out two series of experiments. In the first series, we processed the sentences of the WSC mental subcollection. These sentences were open to us when we were writing inference rules. They served as the testing bed of the model. In most of these sentences all the antecedents were identified correctly. Here are some of these sentences:

Okun' proglotil červja, on byl golodnyj.
The perch swallowed the worm, it was hungry. It = the perch

Okun' proglotil červja, on byl vkusnyj.
The perch swallowed the worm, it was tasty. It = the worm

Petr dal konfetu Ivanu, potomu čto on byl goloden.
Peter gave Ivan a candy, because he was hungry. He = Ivan

Petr dal konfetu Ivanu, potomu čto on byl ne goloden.
Peter gave Ivan a candy, because he was not hungry. He = Peter

Ivan vo vsem podražaet Petru, on ego obožaet.
Ivan imitates Peter in everything, he_1 adores him_2. He_1 = Ivan, he_2 = Peter

Ivan vo vsem podražaet Petru, on sil'no na nego vlijaet.
Ivan imitates Peter in everything, he_1 has a strong influence upon him_2.
He_1 = Peter, he_2 = Ivan

Petr postučal v dver' Ivana, no on ne otvetil.
Peter knocked at Ivan's door, but he did not reply. He = Ivan

Petr postučal v dver' Ivana, no on ne polučil otveta.
Peter knocked at Ivan's door, but he did not receive a reply. He = Peter

Ivan poprosil Petra ob odolženii, no on otkazal.
Ivan asked Peter for a favor, but he refused. He = Peter.

Ivan poprosil Petra ob odolženii, no emu otkazali.
Ivan asked Peter for a favor, but he was refused. He = Ivan.

Ivan oral na Petra, potomu čto on byl rasstroen.
Ivan was shouting at Peter, because he was upset. He = Ivan

Ivan utešal Petra, potomu čto on byl rasstroen.
Ivan was comforting Peter, because he was upset. He = Peter

Ivan obižal Petra, tak čto my ego zaščitili.
Ivan offended Peter so we defended him. He = Peter.

Ivan obižal Petra, tak čto my ego nakazali.
Ivan offended Peter so we punished him. He = Ivan.

We can draw two conclusions from this experiment. First: if the background knowledge provided is detailed and accurate, the WSC test can be passed in most cases, and the explanation of the result is easily understandable by humans. Second: it is time- and effort-consuming to compile all the information needed, including a) entering new ontology concepts, b) linking Russian words with the ontology, and c) formulating inference rules. As is often the case with rule-based systems, the result is rather fragile. The rules should be complete and accurate, in order to achieve the expected result.

The second series of experiments was an evaluation proper. We compiled a test corpus of 20 new sentences belonging to the same mental domain. Many of the mental predicates of the test corpus were not represented in the concept dictionary or in the inference rules. These were added to the concept dictionary and supplied with semantic descriptions before running the test, but—naturally—without having access to the test corpus. The result of the testing was rather low—54%, slightly above the random benchmark (50%).

The evaluation results are summarized in Table 1. The table shows for each test sentence whether an antecedent for a pronoun was identified correctly (1) or not (0). For sentences with two pronouns each antecedent identification result is displayed separately. We compare two approaches: the one based on the syntactic constraints and the semantic one described in this paper.

The syntactic approach for anaphora resolution developed in our labs [Inshakova 2019] gives quite good results (precision and recall around 70% depending on the test corpus). But, as was explained above, purely syntactic approach is not suitable for WSC sentences, which are specifically built in such a way that syntactic constraints are helpless and background knowledge is required. Hence the syntactic approach

on our test WSC sentences gives us 50% of correctly identified antecedents (exactly the chance level). This was expected and can be an indication that the WSC sentences are composed correctly.

On the other hand we expected that the semantic approach would give us noticeably better results. Regretfully, it did not happen. The result of 54% is only marginally better than the syntactic approach and a random selection. In some cases the system could not decide between two antecedents and a random choice had to be applied (those are denoted by 0.5 in the table). In other cases the antecedent was selected incorrectly.

The error analysis shows that in all the cases the failure to identify the antecedent correctly was due to knowledge incompletely provided to the system. This can be partly explained by limited time for the preparation of the evaluation. But, more importantly, it is not clear yet what is the amount of effort needed to produce a description complete enough.

Table 1. Evaluation results

Test sentence	Anaphora resolution approach	
	Syntactic	Semantic
Vrač propisal Petru očki, potomu čto on ploxo vidit. The doctor prescribed Peter glasses, because he has poor eyesight	0	0
Vrač propisal Petru očki, potomu čto on proveril ego zrenie. The doctor prescribed Peter glasses, because he checked his eyesight	1 + 1	1 + 1
Kolja posovetoval Petru otdoxnut', potomu čto on očen' ustal. Kolya advised Peter to have a rest, because he was very tired	1	0.5
Kolja posovetoval Petru otdoxnut', no potom on peredumal. Kolya advised Peter to have a rest, but later on he changed his mind	0	0.5
Petr dal den'gi Ivanu, potomu čto on bogat. Peter gave money to Ivan, because he was rich	1	1
Petr dal den'gi Ivanu, potomu čto on beden. Peter gave money to Ivan, because he was poor	0	0
Petr odolžil den'gi Ivanu, no on ix ne vernul. Peter lent money to Ivan, but he did not give it back	0	0.5
Petr odolžil den'gi Ivanu, potomu čto on xotel pomoč emu. Peter lent money to Ivan, because he wanted to help him	1 + 1	0.5 + 0.5
Petr pobedil Kolju, potomu čto on xorošo igral. Peter defeated Kolya because he played well	1	0.5
Petr pobedil Kolju, potomu čto on ploxo igral. Peter defeated Kolya because he played poorly	0	0.5
Petr pomog Kole s zadaniem, potomu čto on dobryj. Peter helped Kolya with the task, because he is kind	1	0

Test sentence	Anaphora resolution approach	
	Syntactic	Semantic
Petr pomog Kole s zadaniem, potomu čto on poprosil ego pomoč. Peter helped Kolya with the task, because he asked him to help	1 + 0	1 + 1
Petr obvinil Ivana, no ego opravdali. Peter accused Ivan, but he was acquitted	0	0.5
Petr obvinil Ivana, no potom on požalel ob etom. Peter accused Ivan, but later he regretted it	1	1
Petr zaviduet Ivanu, potomu čto on xorošo tantsuet. Peter is envious of Ivan, because he is a good dancer	0	0.5
Petr ne zaviduet Ivanu, xotja on xorošo tantsuet. Peter is not envious of Ivan, although he is a good dancer	0	0
Džon rasserdilsja na Billa, xotja on dobryj. John got angry at Bill, although he is kind	1	0.5
Džon rasserdilsja ne Billa, xotja on ne vinovat. John got angry at Bill, although he was not guilty	0	0.5
Vasja umoljal Ivana ostat'sja doma, no on ne soglasilsja. Vasya begged Ivan to stay at home, but he refused	1	1
Vasja umoljal Ivana ostat'sja doma, no on ne dobilsja uspexa. Vasya begged Ivan to stay home, but he was unsuccessful	0	0
Average	**0.50**	**0.54**

7. Conclusion

This paper proposes a knowledge-based framework for solving Winograd Schema Challenge (WSC). We use the general semantic parser SemETAP that represents the sentence at two semantic levels: Basic Semantic Structure shows the semantics of the isolated sentence, and the Enhanced Semantic Structure supplies it with inferences made on the basis of available knowledge. Background knowledge is implemented by means of inference rules written in the Etalog inference language. SemETAP can be used for a wide range of tasks that require explicit representation of implicit knowledge. Experiments show that if the background knowledge provided is detailed and accurate, the WSC test can be passed in most cases, and the explanation of the result is easily understandable by humans.

8. Acknowledgements

This work was supported by the RSF grant No. 16-18-10422-P, which is gratefully acknowledged.

References

1. *Apresjan Yu. D.* (1974), Leksicheskaya semantika. Sinonimicheskie sredstva yazyka [Lexical semantics. Synonimic means of language]. Moscow: Nauka; Second edition: Selected works: In 2 vol. Vol. I. Moscow: Shkola «Yazyki Russkoi Kul'tury», 1995.
2. *Bailey D., A. Harrison, Yu. Lierler, V. Lifschitz, and J. Michael* (2015), The Winograd schema challenge and reasoning about correlation. In: Working Notes of the Symposium on Logical Formalizations of Commonsense Reasoning.
3. *Baker, C. F.; Fillmore, C. J.; and Lowe, J. B.* (1998), The Berkeley FrameNet Project. In: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1 (ACL/COLING), 86–90.
4. *Boguslavsky I., V. Dikonov, L. Iomdin, A. Lazursky, V. Sizov, S. Timoshenko.* (2015), Semantic Analysis and Question Answering: a System Under Development. In: Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2015), p. 62.
5. *Boguslavsky I.* (2017), Semantic Descriptions for a Text Understanding System. In: Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2017), p. 14–28.
6. *Boguslavsky I., Frolova T., Iomdin L., Lazursky A., Rygaev I., Timoshenko S.* (2018), Semantic analysis with inference: high spots of the football match. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018", Moscow, May 30—June 2.
7. *Chambers N., D. Jurafsky* (2008), Unsupervised learning of narrative event chains. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies , pages 787–797.
8. *Fellbaum, C., ed.* (1998), WordNet: An Electronic Lexical Database . MIT Press.
9. *Gordon, A. and Hobbs, J.* (2004), Formalizations of Commonsense Psychology. AI Magazine 25(4):49–62.
10. *Gordon, Andrew S., and Jerry R. Hobbs* (2011), "A Commonsense Theory of Mind-Body Interaction", in Proceedings of the 10th Symposium on Logical Formalizations of Commonsense Reasoning, AAAI Spring Symposium Series.
11. *Haoruo Peng, Daniel Khashabi, and Dan Roth.* (2015), Solving hard coreference problems. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 809–819.
12. *Hobbs, Jerry R., and Andrew Gordon* (2008), "The Deep Lexical Semantics of Emotions", Proceedings, LREC 2008 Workshop on Sentiment Analysis: Emotion, Metaphor, Ontology, and Terminology, Marrakech, Morocco, May 2008.
13. *Hobbs, Jerry R., and Andrew Gordon.* (2010), "Goals in a Formal Theory of Commonsense Psychology", in A. Galton and R. Mizoguchi (eds.), Formal Ontology in Information Systems: Proceedings of the Sixth International Conference (FOIS 2010), IOS Press, Amsterdam, pp. 59–72.
14. *Hobbs, J., and Gordon, A.* (2014), Axiomatizing Complex Concepts from Fundamentals (invited paper). Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2014), April 6–12, 2014, Kathmandu, Nepal.

15. *Hobbs, Jerry R., Alicia Sagae, and Suzanne Wertheim.* (2012), "Toward a Commonsense Theory of Microsociology: Interpersonal Relationships", in M. Donnelly and G. Guizzardi (eds.), Formal Ontology in Information Systems: Proceedings of the Seventh International Conference (FOIS 2012), IOS Press, Amsterdam, Netherlands, pp. 249–262.
16. *Inshakova, E.* (2019), An anaphora resolution system for Russian based on ETAP-4 linguistic processor (this volume).
17. *Kipper-Schuler, K.* (2005), VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon . Ph.D. Dissertation, University of Pennsylvania.
18. *Levesque H.* (2011), The Winograd Schema Challenge. In: AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning.
19. *Levesque, H., Davis, E., Morgenstern, L.* (2011), "The Winograd Schema Challenge", In: Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning.
20. *Liu, H., and Singh, P.* (2004), Conceptnet: A practical commonsense reasoning toolkit. BT Technology Journal 22:211–226.
21. *Montazeri, Niloofar, and Jerry R. Hobbs.* (2011), "Elaborating a Knowledge Base for Deep Lexical Semantics", in J. Bos and S. Pulman (eds.), Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011), January 2011, pp. 195–204.
22. *Montazeri, Niloofar, and Jerry R. Hobbs.* (2012), "Axiomatizing Change-of-State Words", in M. Donnelly and G. Guizzardi (eds.), Formal Ontology in Information Systems: Proceedings of the Seventh International Conference (FOIS 2012), IOS Press, Amsterdam, Netherlands, pp. 221–234.
23. *Morgenstern, Leora.* (2001), MidSized Axiomatizations of Commonsense Problems: A Case Study in Egg Cracking. Studia Logica, April 2001, Volume 67, Issue 3, pp. 333–384
24. *Schubert, L.* (2002), Can we derive general world knowledge from texts? In: Proceedings of the Second International Conference on Human Language Technology Research (HLT), 94–97. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
25. *Mueller E.* (2016), Transparent Computers: Designing Understandable Intelligent Systems. Createspace Independent Publishers.
26. *Palmer, M.; Gildea, D.; and Kingsbury, P.* (2005), The proposition bank: An annotated corpus of semantic roles. Computational Linguistics 31(1):71–106.
27. *Quan Liu, Hui Jiang, Zhen-Hua Ling, Xiaodan Zhu, Si Wei, Yu Hu.* (2016), Combing Context and Commonsense Knowledge Through Neural Networks for SolvingWinograd Schema Problems. arXiv:1611.04146v1 [cs.AI] 13 Nov 2016.
28. *Rahman A., V. Ng.* (2012), Resolving complex cases of definite pronouns: the Winograd schema challenge. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 777–789. Association for Computational Linguistics.
29. *Rygaev I.* (2017), Rule-based Reasoning in Semantic Text Analysis. Proceedings of the Doctoral Consortium, Challenge, Industry Track, Tutorials and Posters @ RuleML+RR 2017 hosted by International Joint Conference on Rules and Reasoning 2017 (RuleML+RR 2017).

30. *Rygaev I.* (2018), Etalog—a natural-looking knowledge representation formalism // Proceedings of ITaS 2018 School and Conference (http://itas2018.iitp.ru/media/papers/1570472169.pdf).
31. *Schüller P.* (2014), Tackling Winograd schemas by formalizing relevance theory in knowledge graphs. In: Fourteenth International Conference on the Principles of Knowledge Representation and Reasoning.
32. *Sharma A., Nguyen Ha Vo, Somak Aditya, and Chitta Baral.* (2015), Towards addressing the Winograd schema challenge-building and using a semantic parser and a knowledge hunting module. In IJCAI, pages 1319–1325.
33. *Trieu H. Trinh, Quoc V. Le.* (2018), A Simple Method for Commonsense Reasoning. arXiv:1806.02847v1 [cs.AI] 7 Jun 2018.

COMPARING MODELS OF MORPHEME ANALYSIS FOR RUSSIAN WORDS BASED ON MACHINE LEARNING

Bolshakova E. I. (eibolshakova@gmail.com)

Moscow State Lomonosov University; National Research University Higher School of Economics, Moscow, Russia

Sapin A. S. (alesapin@gmail.com)

Moscow State Lomonosov University, Moscow, Russia

The paper reports on the experimental comparison of several machine learning models proposed in recent years for automatic morpheme segmentation of Russian words, including conditional random fields (CRF), sequence-to-sequence neural network (Seq2seq), convolutional neural network (CNN) model, as well as a new model we have developed with the aid of gradient boosted decision trees (GBDT). For more complete research, in our experiments we have also evaluated the semi-supervised method of Morfessor. All the morpheme analysis models being compared are briefly described in the paper, some of them perform only segmentation of words into morphs, the other produce segmentation with classification of resulted morphs. Since for Russian language linguistics rules for splitting words into morphs (and also the classification of some morphs) may differ, the experiments were performed for two data sets differing in labeling, which are obtained respectively from CrossLexica's dictionary and Tikhonov's dictionary. The experimental evaluation has shown that two best models of morpheme segmentation with classification, namely GBDT and CNN models have comparable quality, giving about 86–94% of word-level accuracy.

Keywords: morphological segmentation, morpheme analysis of Russian words, machine learning models for morphology, morpheme segmentation with classification

СРАВНЕНИЕ МОДЕЛЕЙ МОРФЕМНОГО РАЗБОРА ДЛЯ РУССКОГО ЯЗЫКА, ОСНОВАННЫХ НА МАШИННОМ ОБУЧЕНИИ

Большакова Е. И. (eibolshakova@gmail.com)
МГУ имени М. В. Ломоносова; НИУ ВШЭ, Москва, Россия

Сапин А. С. (alesapin@gmail.com)
МГУ имени М. В. Ломоносова, Москва, Россия

1. Introduction

Morpheme segmentation is the task of breaking words into constituent morphs (root and affixes), which are the smallest meaningful units of texts. This task was studied since early years of natural language processing (NLP), but not intensively, and significant progress in its solution is related with the use of machine learning techniques [4], [5], [10], [11]. Nevertheless, the task is far from complete solution, it is especially difficult for languages with rich morphologies (such as Russian) with many affixes of various types and meanings.

Nowadays, the problem of automatic morpheme segmentation became more topical, since information about morphemic structure of words is already in use in various NLP applications and tasks, including machine translation, recognition of semantically related words (cognates, paronyms, etc.), constructing word embeddings (by using morphemes as meaningful units instead of symbol N-grams), handling rare and out-of-vocabulary words (by deriving their meaning based on distributional word vectors representations) and so on.

The problem of morpheme analysis is investigated in two variants, which we consider in this paper:

- only segmentation, that is splitting a word into morphs or morpheme-like units, for example: *пре-крас-н-ый, beauti-ful*;
- segmentation with classification (categorization) of segmented morphs, by recognizing and labeling their types (the main types are Prefix, Root, Suffix, Ending), for example: *пре*:PREF/*крас*:ROOT/*н*:SUFF/*ый*:END, *beauti*:ROOT/*ful*:SUFF.

The earliest known method of morpheme segmentation was proposed by Z. Harris in [8] and based on letter variety statistics, which is counted on dictionary words. Morpheme boundaries are detected at locations in a word where the predictability of the next letter in the word is low. The method was tested with a small English dictionary (about 1000 words) and showed 61% of precision.

The last years, the most known solution for morpheme segmentation is implemented in Morfessor system [4, 5, 10], which exploits unsupervised machine learning methods to be trained on a large text collection. Evaluation of the method for English,

Finnish, and Turkish showed about 70% of F-measure for detected morpheme boundaries. The method is especially useful when labeled text data needed for supervised learning are absent or insufficient. Recently, a semi-supervised method was developed, which uses some labeled data in addition to the text collection.

The task of morpheme segmentation with classification of segmented units is more relevant for morphologically rich languages (such as Russian), but it is almost unexplored, mainly because for a long time the amount of data required for supervision was not sufficient.

For Russian, the first known work on morpheme segmentation with classification through machine learning was undertaken in the morphological processor [3], whose functionality among other features provides morpheme analysis of words. The task was solved by classifying letters of words according to the main types of morphs, with conditional random fields (CRF) method. After training on the labeled data of Wiki and CrossLexica's dictionary [2], the method showed classification accuracy up to 79.5%.

In two more recent works, neural machine learning models for morpheme segmentation of Russian words were proposed. In [1], the sequence-to-sequence neural network model (Seq2seq) trained on data of Tikhonov's dictionary [12] is described. The model implements only segmentation and outperforms Morfessor's model trained on Russian text collection Librusec1.

The work [11] successfully applied convolutional neural network (CNN) for morpheme segmentation of Russian words and classification of segmented morphs. The CNN model was trained and evaluated on Tikhonov's dictionary, achieving accuracy of classification up to 88% and F-measure about 98% for morpheme boundaries, thus outperforming the previous models for Russian.

For further research of various supervised machine learning methods with respect to morpheme analysis task for Russian, we have developed one more model of segmentation with classification, applying gradient boosted decision trees (GBDT) [7] for the task2. Our choice of this method was based on the following considerations. First, compared with neural network methods, decision trees are simpler and more interpretable. They may also be applied to tasks of sequence tagging (morpheme segmentation may be treated as sequence labeling), and GBDT method is powerful enough due to boosting.

The main purpose of this paper is to describe an experimental comparison of our GBDT model and three above-mention supervised models3 for Russian on the same training data, both for morpheme segmentation task and for segmentation with classification. For this reason, we exploited two data sets with somewhat different labeling, since in some cases linguists have no consensus about how to correctly split Russian words into morphs (and also about classification of some morphemes). The first dataset contains about 23,000 segmented words taken from CrossLexica [2], the second is obtained from electronic version of Tikhonov's dictionary [12] with 90,000 segmented words.

---

1 http://lib.rus.ec

2 https://github.com/alesapin/GBDTMorphParsing

3 In experiments we have used either open source code of them or available trained models.

For both data sets, the quality of segmentation has been evaluated (with F-measure of morph boundaries), as well as classification accuracy for the models performing morpheme classification. For completeness, we have also included in the comparison semi-supervised segmentation method of Morfessor [10]. The experimental evaluation has shown that GBDT and CNN models are the best models of morpheme segmentation with classification, having comparable quality of 86–94% (depending on the datasets) measured in word-level accuracy of classification.

The paper starts with brief description of the morpheme segmentation models for Russian being compared, our new GBDT model is described in more details. Then the differences between the training data sets are explained, and results of experimental evaluation for the models are presented and discussed.

2. Morpheme Segmentation Models under Comparison

Morfessor [4], [5], [10] presents a family of statistical morpheme segmentation methods based on the maximum a posteriori estimation principle (MAP). Initially, the purely unsupervised method was developed, which works out the best variant of breaking words into morpheme-like segments for a given large unlabeled text. A semi-supervised improvement of the method was proposed later, it refines word segmentation by additional usage of yet segmented data [10]. For English and Finnish, the best results were obtained when training on text collection with 200 thou. words and additional dictionary of 10 thou. segmented words. For Morfessor 2.0, F-measure for detected morpheme boundaries increased to 77–80% for English, Finnish, and Turkish (for Russian, the method was trained but not evaluated).

The model [3] created for Russian and based on CRF method [9] performs morpheme segmentation with classification. The task was considered as sequence labeling for letters of a given word, by classifying them to four main types of morphs: Prefix, Root, Suffix, Ending. The built CRF classifier accounts for several features of a letter being classified, features of the word being segmented (its morphological tags), as well as Harris' values [8] (that are local maximums of letter frequencies counted for various positions in the words). The classification implies detecting boundaries between morphs of different classes, but cannot perform complete segmentation, since resulted segments may contain several successive morphs of the same type. For example, letters of the noun *душевность* (*soulfulness*) are classified as follows:

д	*у*	*ш*	*е*	*в*	*н*	*о*	*с*	*т*	*ь*
R	R	R	S	S	S	S	S	S	E

Successive suffixes *евн-* and *ост-* are not separated, complete segmentation should be the following: *душ*:ROOT/*евн*:SUFF/*ост*:SUFF/*ь*:END. Another weak point of such classification is inability to distinguish postfix *ся/сь* from endings. The CRF classifier trained on segmented words from Wiki and CrosLexica's dictionary has achieved 74.2% of word-level classification accuracy (percent of completely correctly analyzed words).

The sequence-to-sequence neural network model (Seq2seq) implemented in [1] for morpheme segmentation task considers this task as sequence transduction and uses ideas of encoder-decoder, originally applied for machine translation. For Russian,

the model was trained on the data of Tikhonov's dictionary [12], demonstrating 93.95% of F-measure for detecting morpheme boundaries, thus outperforming by this measure both Morfessor (trained on Librusec text collection) and the considered CFR model (but Seq2seq model does not perform morpheme classification).

Significantly better quality of both segmentation and morpheme classification was achieved by convolutional neural networks (CNN) model [11], which was also trained and tested on Tikhonov's dictionary. The best reported results are 98.10% for F-measure on morpheme boundaries and 88.71% of word-level accuracy. The implemented model is quite complicated, it ensembles three CNN models with different random initializations and additionally makes use of morpheme memorizing techniques. The authors tried to add long-short memories (LSTM) layers to the network, but this did not improve the quality of the model. Compared with the above-described CRF model, it detects boundaries between successive prefixes and suffixes and also produces a more detail classification of segmented morphs, including postfix *ся/сь*, linking letter in multi-root words, and also hyphen. For this purposes, the model classifies letters of a word into 22 classes based on BMES (BIOES) labeling scheme (often used in named entities recognition task). The classes account for cases of beginning (B), middle (M), and ending (E) positions of letter within roots, suffixes, prefixes, as well as their Single letter variants. The following example shows classification for letters of the word *учитель* (*teacher*) (its segmentation is *уч*:ROOT/*и*:SUFF/*тель*:SUFF):

у	*ч*	*и*	*т*	*е*	*л*	*ь*
B-ROOT	E-ROOT	S-SUFF	B-SUFF	M-SUFF	M-SUFF	E-SUFF

It should be noted, that before evaluation the resulted morpheme classification (performed by the CNN ensemble) some segmented words are corrected by an auxiliary procedure, which fixes incorrect sequences of morph types (in particular, if a suffix is located before a root).

Our model developed for morpheme segmentation with classification is based on decision trees with gradient boosting (GBDT) [7]. Similar to CNN model, our classification of morphemes includes the main types: prefix, root, suffix, ending, and also postfix *ся/сь* (*мы-ть-ся*), linking letter for multi-root words (such as *пар-о-ход*), hyphen (*по-нашему*). In contrast to CNN model, for purposes of segmenting successive suffixes, prefixes, roots we label only beginnings of them (since the set of BMES labels is redundant for the task of morpheme segmentation). Thereby, the model classifies letters of word into 10 classes, an example for the word *учитель* (*teacher*) is given below (B-ROOT and B-SUFF encode the beginning of root and suffix, respectively):

у	*ч*	*и*	*т*	*е*	*л*	*ь*
B-ROOT	ROOT	B-SUFF	B-SUFF	SUFF	SUFF	SUFF

For classification, our GBDT model takes into account both features of the letter being classified and features of its word. The former includes the letter itself (represented in one-hot encoding format), is it a vowel, the position of the letter in the word, its occurrence frequency in training data set, and also Harris' values [8].

Since the gradient boosting method is not oriented to sequence labeling tasks, in order to account for information about sequencing of letters in our task (to be more

precise, to account for influence of the previous and subsequent letters on the class of the current letter), a window of small size is used in the model: 5 letters on the left and 5 letters on the right are accounted as features (we assume that there are no more long dependencies between the letters).

Among features of the word our model includes some its morphological tags: part of speech, case, number, gender, time (if any), and stem length (we obtain all of them from the morphological parser CrossMorphy [3] with non-contextual method of morphological disambiguation).

Similar to work [11] we also have elaborated a heuristic procedure correcting some errors of GBDT classification. In particular, for word *рождаться* the segmentation obtained by GBDT: *р*:ROOT/*о*:PREF/*жд*:ROOT/*ать*:END/*ся*:POSTFIX will be corrected as *рожд*:ROOT/*ать*:END/*ся*:POSTFIX. The procedure relies on obvious rules of morphotactics for Russian: any word should begin with a prefix or root, a root may go after prefix, a suffix may go after root or another suffix, and so on.

Our GBDT model was implemented with Catboost library [6] that does not require to manually encode categorical features (such as parts of speech, etc.) into one-hot encoding. Besides implemented GBDT model, for experiments we used available implementations of the other described models4 for morpheme segmentation and classification.

3. Data Sets for Training and Evaluation

For Russian language there are two data sets with words spitted into classified morphs and thus suitable for training supervised machine learning models—an electronic version of Tikhonov's dictionary [12] with 96046 words and a subset from CrossLexica's dictionary [2] with 23426 segmented words. In both data sets, the same morpheme types (prefix, root, suffix, ending, postfix) is used, and successive prefixes and suffixes are labeled. Beside the size of the data sets, they differ significantly with respect to several important features.

First, many words represented in both data sets have different segmentation into morphs (and even the classification of some morphs), because there is no full agreement between linguists about rules of morpheme analysis. The authors of these dictionaries applied slightly different rules for splitting words into morphs and their classifying, and they also pursued different principles of forming the dictionaries. In particular, in Tikhonov's data set word *бывать* is segmented as *бы*:ROOT/*ва*:SUFF/*ть*:SUFF and in CrossLexica's set, as *бы*:ROOT/*ва*:SUFF/*ть*:END. In this case morph *ть* of verb infinitive is interpreted either as suffix or as ending, but this is unresolved question in Russian linguistics. Unlike Tikhonov's dictionary, where many prefixes are not separated from roots because of their "lexical cohesion", in CrossLexica's data all possible prefixes are usually separated even for words of foreign origin. For example, the word *продуктивный* (*productive*) in Tikhonov's data

[4] https://github.com/aalto-speech/morfessor
http:/gihub.com/alesapin/XMorphy
http:/gihub.com/kpopov/morpheme_seq2seq
https://github.com/AlexeySorokin/NeuralMorphemeSegmentation

set is segmented as *продукт*:ROOT/*ивн*:SUFF/*ый*:END, while in CrossLexica's set: *про*:PREF/*дукт*:ROOT/*ив*:SUFF/*н*:SUFF/*ый*:END (since this word has the common root with word *индуктивный—inductive*). The considered difference in morpheme segmentation are mainly explained by different functions of the dictionaries: Tikhonov's dictionary was originally built as a derivational dictionary, while CrossLexica's segmented and labeled data were created for constructing pairs of morpheme paronyms (that are words with the same root but differing in affixes and having close meanings, such as *массивный* and *массовый—massive* and *mass*).

Moreover, the considered dictionaries differs in lexicon. Tikhonov's dictionary contains many obsolete words that now hardly appear in texts, for example: *снитковый*, *лядунка*. There are no such words in CrossLexica, but it includes many words of modern lexicon, such as *эксклюзивный* (*exclusive*), *целлюлит* (*cellulite*), and so on. At the same time, in CrossLexica's data set there are no multi-root words and words with hyphen.

It is unclear a priory, which of these dictionaries is more relevant and suitable for machine learning (on our opinion, for various NLP application may be useful models trained on various data sets). For this reason, we have evaluated all above-described supervised machine learning models separately on both these data sets, except the CRF model (since only the model trained on CrossLexica's data is available). Thus, we have seven supervised models for comparison.

For evaluation, we took the available Seq2seq and CNN models pre-trained on Tikhonov's dictionary, the other models were trained and evaluated in the following way. The data for training and testing were randomly divided in proportion 80:20, the training samples were 76,836 and 18,740 words, for Tikhonov's dictionary and CrossLexica respectively. Each trained model was then evaluated with remaining samples from the same dictionary. Morfessor's semi-supervised model was trained on certain part of Librusec text collection (about 100M tokens) and was evaluated on the corresponding testing data.

4. Results of Experiments and Discussion

Our experiments with training GBDT model showed that the best evaluation scores were achieved for 10,000 iterations for the Tikhonov's dictionary and 5,000 for the CrossLexica's data, and the optimal depth of decision trees turned out to be 10.

All the machine models under comparison perform segmentation into morphs, and we have evaluated them by BPR metric (boundary precision and recall). Precision is the ratio of the number of true morpheme boundaries to the number of all revealed boundaries, while recall is the ratio of the number of true boundaries to the total number of boundaries. F-measure is computed as mean harmonic of the recall and precision. The results are given in **Table 1**.

Table 1. Comparison of Morpheme Segmentation for Russian

Model	CrossLexica's Dictionary			Tikhonov's dictionary		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Morfessor	93.30	75.40	83.40	94.7	73.70	82.90
CRF	96.05	70.93	81.60	—	—	—
Seq2seq	94.62	93.92	94.27	94.07	93.83	93.95
CNN	98.68	98.75	98.72	**97.86**	**98.35**	**98.10**
GBDT	**98.84**	**99.26**	**99.05**	97.76	98.26	98.01

It turned out that CNN and GBDT models are the best for morpheme segmentation, they have very close scores and thus comparable quality. For Tikhonov's dictionary, CNN model has slightly better scores than GBDT, but the latter slightly outperforms CNN on CrossLexica's data. Seq2seq model shows quite good results, but this is the average quality. Morfessor and CRF model mainly lose in recall (because of undersegmentation) and thus in F-measure. CRF model demonstrates the worst result in F-measure (because it does not detect boundaries between successive suffixes), despite the fact that CRF method is suited for sequence labeling tasks.

In evaluation of morpheme segmentation with classification only CRF, CNN and GBDT models can participate, but we have compared only two best models. For comparison, we have evaluated classification accuracy for letters (the ratio of correctly recognized classes of letters to the number of all letters), as well as accuracy for words (the ratio of completely correctly segmented words with true classes of their segmented morphs). Table 2 presents results of the evaluation. Two evaluated models again have close scores. GBDT model insignificantly wins on CrossLexica's data, and for Tikhonov's dictionary we have the opposite situation.

However, it should be noted that while applying CNN model, 180 words from 4,686 testing words of CrossLexica's data set and 1,871 from 19,210 testing words of Tikhonov's data (that is respectively 4% and 10% of all the testing words) have been corrected by the auxiliary procedure, which fixes wrong morpheme types in the output of neural network classifier. In contrast with CNN, the analogous correction procedure of GBDT have fixed only 42 words of CrossLexica's set and 804 for Tikhonov's data (0.2% and 4% respectively). Therefore, in Table 2, we present accuracy scores obtained without the correction of the classifiers ("Uncorrected" columns in the Table). One can notice that despite GBDT is a more simple method than CNN, this method itself (without correction) works better CNN.

Table 2. Accuracy of Morpheme Segmentation with Classification

Model	CrossLexica's Dictionary			Tikhonov's dictionary		
		Words			Words	
	Letters	Corrected	Uncorrected	Letters	Corrected	Uncorrected
CNN	97.88	93.23	90.48	**96.64**	**88.71**	82.62
GBDT	**98.39**	**94.20**	**93.85**	96.40	86.54	**86.24**

Expert analysis of errors in morpheme segmentation shows that the most frequent ones are related with wrong boundaries between root and suffix, such as *печеч*:ROOT/*к*:SUFF/*а*:END instead of correct *печ*:ROOT/*ечк*:SUFF/*а*:END for the word *печечка* (*little bake*), the error occurs in both the models. Another frequent error is different segmentation of suffixes, for example: *возбуждение* (*excitation*): воз:PREF/бужд:ROOT/ени:SUFF/е:END and воз:PREF/бужд:ROOT/ен:SUFF/и:SUFF/е:END, the former is erroneous for CrossLexica's data, and the latter, for Tikhonov's data. Indeed, the problem of identificating boundaries between succsessive morphemes is especially difficult for Russian suffixes.

Table 3. Examples of wrong segmentation

Word	Model	Data	Wrong segmentation	Correct segmentation
препираться	GBDT	TN	*препир*:ROOT/*а*:SUFF/*ть*:SUFF/*ся*:PF	*препира*:ROOT/*ть*:SUFF/*ся*:PF
фанатский	GBDT	CL	фан:ROOT/ат:SUFF/ск:SUFF/ий:END	*фанат*:ROOT/*ск*:SUFF/*ий*:END
помои	CNN	TN	*помо*:ROOT/*и*:END	*по*:PREF/*мо*:ROOT/*и*:END
пришить	CNN	CL	*при*:PREF/*ши*:ROOT/*ть*:END	*при*:PREF/*ш*:ROOT/*ить*:END

Some examples of errors in complicated cases of segmentation are given in Table 3: in "Data" column, symbol TN denotes Tikhonov's data set, while CL, CrossLexica's data set. One more interesting error is the following segmentation: *по*:PREF/*душ*:ROOT/*еч*:SUFF/*н*:SUFF/*ый*:END (given by CNN on CrossLexica) instead of *под*:PREF/*уш*:ROOT/*еч*:SUFF/*н*:SUFF/*ый*:END, for word *подушечный* (*pillow's*): it seems as the model misunderstands this word as *подушевой* (*per person*).

Since trained decision trees is interpretable, we can know significance of the features have been exploited in our GBDT model (their weights, as contribution to the results). The most important features are the following: the letter being classified (10.89%), the next 2 letters (11.79% and 7.32%), 3 preceding letters (11.41%, 9.19%, and 5.74%). The Harris' values are also important, giving 4.19% for the initial part of a word and 3.12% for the end part. Morphological features of the word are less important, but give 9.22% in total.

5. Conclusions and Future Work

We have experimentally evaluated and compared five various machine learning models of morpheme segmentation for Russian, exploiting for their training two data sets with different labeling of constituent morphs. Two models of morpheme segmentation with classification, namely GBDT (the gradient boosted decision trees) and CNN (convolutional neural network) have showed the best and comparable results, and thus they may be used in various NLP experiments with Russian text. It seems that in the achieved quality of segmentation is close to possible limits for machine

learning. Nevertheless, the problem of gold standard for Russian morpheme segmentation requires additional research.

As for our GBDT model, it seems reasonable to study some new features accounting for in machine learning, in order to further improve the model, supposedly by using statistics of affixes. We also plan to extend and refine the data sets for training and to combine machine learning with additional procedures based on linguistic rules.

References

1. *Arefyev N. V., Gratsianova T. Y., Popov K. P.* (2018), Morphological Segmentation with Sequence to Sequence neural network, Computational linguistics and Intellectual Technologies: Proceedings of the Int. Conference "Dialogue 2018", Moscow, pp. 82–91.
2. *Bolshakov I. A.* (2013), CrossLexica—Universum of links between Russian words [CrossLexica—universum svyazey mezshdu russkimi slovami], Business Informatics [Biznes Informatica], No 3 (25), pp.12–19.
3. *Bolshakova E. I., Sapin A. S.* (2018), A Morphological Processor for Russian with Extended Functionality, Analysis of Images, Social Networks and Texts: 6th Int. Conference AIST 2017, Moscow, Revised Selected Papers, LNCS, 10716, Springer, Cham, pp. 22–33.
4. *Creutz M., Lagus K.* (2006), Morfessor in the Morpho Challenge, PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes, Venice, Italy.
5. *Creutz M., Lagus K.* (2007), Unsupervised models for morpheme segmentation and morphology learning, ACM Transactions on Speech and Language Processing, 4(1), Article 3.
6. *Dorogush A. V., Ershov V., Gulin A.* (2018) CatBoost: gradient boosting with categorical features support, available at: //arXiv preprint arXiv:1810.11363.
7. *Friedman J. H.* (2002), Stochastic gradient boosting, Computational Statistics & Data Analysis. Vol.38, pp. 367–378.
8. *Harris S. Zellig.* (1967), Morpheme boundaries within words: Report on a computer test, Transformations and Discourse Analysis Papers 73, pp. 68–77.
9. *Lafferty J., McCallum A., Pereira F. C. N.* (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data, Proceedings of the 18th International Conference on Machine Learning, Williamstown, pp. 282–289.
10. *Smit P., Virpioja S., Gronroos S., Kurimo M.* (2014), Morfessor 2.0: Toolkit for statistical morphological segmentation, Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the ACL, Gothenburg, pp. 21–24.
11. *Sorokin A., Kravtsova A.* (2018) Deep Convolution Networks for Supervised Morpheme Segmentation of Russian Language, Proceedings of the Conference on Artificial Intelligence and Natural Language, St-Petersburg, Springer, Cham, pp. 3–10.
12. *Tikhonov A. N.* (1990) Word Formation Dictionary of Russian language [Slovoobrazovatel'nyi slovar' russkogo yazyka], Moscow, Russkiy yazyk.

MULTILINGUAL PARALLEL CORPORA AS A SOURCE FOR QUANTITATIVE CROSS-LINGUISTIC GRAMMAR RESEARCH (THE CASE OF VOICE CONSTRUCTIONS)

Bonch-Osmolovskaya A. A. (abonch@gmail.com),
Nesterenko L. V. (lnesterenko@hse.ru)

National Research University "Higher School of Economics", Moscow, Russia

Multilingual parallel corpora make possible the application of quantitative methods in cross-linguistic research. Due to the lack of appropriate resources, this has not become a widespread technique among linguists, but the studies based on this idea tend to emerge. In our work, we focus on the application of logistic regression for the research of passive voice constructions with an overtly expressed agent. The study is conducted on the data extracted from a multilingual parallel corpus that was created for this purpose. The issue we find noteworthy about voice alternation is the motivation for choosing active instead of passive, i.e. when a person would say 'This essay was written by Mary' instead of 'Mary wrote this essay'. Relying on theoretical studies, we selected a bunch of features claimed to be important for this kind of choice and used them for training logistic regression models. As a result, based on the model coefficients we can detect which features appear to be passive triggers.

Key words: multilingual parallel corpora, quantitative methods, logistic regression, feature selection, passive voice, semantic roles

МУЛЬТИЯЗЫЧНЫЕ ПАРАЛЛЕЛЬНЫЕ КОРПУСА КАК РЕСУРС ДЛЯ КВАНТИТАТИВНЫХ МЕЖЪЯЗЫКОВЫХ ИССЛЕДОВАНИЙ В ГРАММАТИКЕ (НА ПРИМЕРЕ ЗАЛОГОВЫХ КОНСТРУКЦИЙ)

Бонч-Осмоловская А. А. (abonch@gmail.com),
Нестеренко Л. В. (lnesterenko@hse.ru)

Национальный исследовательский университет «Высшая школа экономики», Москва, Россия

Ключевые слова: мультиязычные параллельные корпуса, логистическая регрессия, анализ признаков, страдательный залог, семантические роли

1. Introduction

Parallel corpora are known to be a powerful resource for research and NLP engineering. Mostly, they have been used in machine translation and development of multilingual applications. However, since linguists pointed out the benefits of parallel corpora usage in typological studies [Dahl 2007], [Wälchli 2007], researchers tend to set aside traditional techniques of collecting language data and perform their experiments on parallel texts. There are few typological studies, e.g. [Mayer, Cysouw 2012], [Cysouw 2014], [Östling 2015], [Östling 2016], [Asgari, Schütze 2017], [Bonch-Osmolovskaya, Nesterenko 2018], presenting application of quantitative methods to multilingual parallel data. Even though the idea of using parallel data and quantitative analysis gains popularity, the lack of suitable data, such as deeply annotated multilingual parallel corpora containing texts of different genres, slows down the progress in this field [Nesterenko 2019].

In this paper, we show how a multilingual parallel corpus and machine learning techniques can be used for the cross-linguistic analysis of a grammatical phenomenon, namely, passive voice constructions with overtly expressed agent. Despite the fact that there is vast variety of literature on voice constructions [Melchuk, Kholodovich 1970], [Keenan, Dryer 1981], [Frajzyngier 1982], [Shibatani 1985], [Comrie 1988], [Shibatani 1988], [Fox, Hopper 1994], [Givón 1994], [Kallulli 2006], [Tsunoda, Kageyama 2006], only a few papers include quantitative [Berez, Gries 2010] or corpus-based analysis [Jisa et. al. 2002], [Plecháčková 2007], [Xiao 2007].

The sentences like *Mary wrote this essay* and *This essay was written by Mary* represent the same situation that has different language encoding. The question is what linguistic factors motivate the passive occurence instead of active. We assume that different combinations of factors that might influence the choice between active and passive are relevant for different languages. Modelling of this problem demands the use of multilingual parallel corpora and the samples from the same context environment. That is important because it allows to evaluate how a fixed set of factors works for different languages. Using a corpus that contains parallel texts in 9 European languages and feature annotation tranfer prodedure, for each language we train a logistic regression model that predicts the probability of passive occurrence in particular context relying on a set of linguistic features. The resulting parameters of the models determine which features have stronger impact on the passive occurrence, and help to reveal differences between language strategies for choosing passive.

Logistic regression is widely applied for modelling of grammatical phenomena on monolingual data and annotation transfer is a common procedure in studies on parallel data, in our paper we combine these two methods, which, as far as we know, were previously used only in separation.

The paper is organized as follows. First, in **Section 2**, we discuss the factors that might influence the passive occurrence and determine a set of features used for training logistic regression models. Then, in **Section 3** we describe our multilingual corpus and emphasize some essential points in the process of its development. In **Sections 4** and **5** we describe the dataset and the models for the passive occurrence prediction. Finally, in **Section 6** we discuss the obtained results.

2. Factors determining the choice of voice construction

According to the framework of St. Petersburg typology group the crucial concept for voice description is diathesis. It is defined as "a pattern of mapping of semantic arguments onto syntactic functions (grammatical relations)" [Kulikov 2011]. The basic transitive diathesis has the following structure: the first (macro)role Actor is mapped onto the grammatical relation of Subject, while the semantic (macro)role Undergoer is mapped onto the grammatical relation of Direct Object (for the macroroles framework see [Foley & Van Valin 1984]).

Table 1. Active and passive diathesis

	Active			Passive	
Semantic argument level	X (Actor)	Y (Undergoer)	⇒	X (Actor)	Y (Undergoer)
Syntactic functional level	Subject	Direct Object		Direct Object/—	Subject

In this paper, we will consider only the passives with overtly expressed agent and leave agentless passives for future research.

The semantic features of the participants play an important role in voice alternation. There are examples which demonstrate that if there is no prototypical agent in a clause, then the initial active construction cannot be reformulated as passive with overtly expressed agent.

(1) a. *$250000 won't buy this kind of house any more.*
b. **This kind of house won't be bought by $250000 any more.*
[Shibatani 1985]

(2) a. *The refugees have {seen/witnessed} some traumatic events.*
b. *Some traumatic events have been {seen/witnessed} by the refugees.*
c. *{This country/The last decade} has {seen/witnessed} some traumatic events.*
d. **Some traumatic events have been {seen/witnessed} by {this country/the last decade}.*
[Langacker 2006]

In [Keenan, Dryer 1981] authors formulate a similar constraint based on the notion of a patient. They emphasize that the verbs without a patient are not easily passivizable. These assumptions indicate that besides other features we should focus on characteristics of the participants.

A detailed description of semantic roles and parameters that characterize participant relations and might be relevant for the explanation of voice alternation can be found in [Lehmann 2006]. Lehmann discribes the roles of participants regarding their empathy, involvement and control. The features involvement and control form a continuous scale where the roles can be placed. Involvement indicates if the situation is inconceivable without a particular participant, and control parameter presupposes that the participant has control over the situation or the participant is being controlled. The empathy parameter is based on the empathy hierarchy and adds another dimension to Lehmann's system.

One of the generalizations in [Keenan, Dryer 1981] tells us that some languages, mostly Asian, like Chinese or Vietnamese, distinguish two types of passive constructions depending on negative vs. positive nature of subject affectedness. There are no indications if this might influence the choice of passive occurrence, but we include this factor in our set of features in order to check whether it has any importance for passives in European languages.

In Jespersen's list of factors motivating passive occurrence [Jespersen 1924], among others, we find "The passive turn may facilitate the connection of one sentence with another". Shibatani [Shibatani 1985], referring to [Dixon 1979], indicates that passives can create a syntactic pivot so that syntactic transformations can take place. For our study, we engage the idea of connection between sentences in two ways. First, we take into consideration the presence of contextual refereces to the verb arguments. Second, we pay attention to the fact whether the sentences are in contrast/opposition relation.

All in all, our set of features include semantic roles and their characteristics (empathy, involvement), characteristic of the verb meaning (whether it denotes a negative action), contextual fatures such as contrast/opposition relation between sentences and mention of the arguments in the previous context. The elaboration of features we describe in more detail in **Section 4**.

3. The Corpus

In our study, we use a corpus that consists of the collection of J. K. Rowling Harry Potter books series from 1 to 7 in 9 languages: English, German, Swedish, French, Italian, Spanish, Russian, Czech and Bulgarian. The size of the text data per language is about 1 million tokens. In order to be used productively in research, a multilingual parallel corpus should satisfy some requirements. First, an essential attribute of a parallel corpus is alignment, because it allows to compare the same parts of the texts in different languages and, if necessary, transfer their characteristics, e.g., from one pivot language to the others. In our corpus, texts are aligned pairwise at sentence and word level. For sentence alignment, we processed the texts with Gale & Church algorithm [Gale, Church 1991] and for the word alignment, we used the Efmaral toolkit [Östling, Tiedeman 2016]. Another thing, crucial for effective data extraction from a multilingual parallel corpus is morphological and syntax annotation in a unified format. The unified annotation format simplifies data extraction and the search for words or phrases with the same grammatical properties across all languages in the corpus. For these purposes, we used a UDPipe parser [Straka, Straková 2017], that runs on models trained for different languages within the Universal Dependencies tags set [Nivre et al. 2016].

For our experiment, we extracted corresponding active and passive sentences in 9 languages relying on the alignment and UDPipe annotation. After that, the English samples were manually annotated with the features we previously defined as relevant for active and passive distinction and this annotation was transferred to the data from all other languages. In the next section we describe the dataset and the feature annotation more precisely.

4. The Dataset

For our experiments (the methods itself we describe in Section 5), we designed the dataset according to the basic principles of machine learning. Sentences are considered as objects of classification, linguistic factors build up feature vectors for learning, and there are two class labels — active and passive.

Most of the sentences from the corpus have alignments in all languages of the sample, i.e., each translational unit consists of translational equivalents in 9 languages, and all of them are marked for the voice construction type. There are 236 fully aligned samples and there are partially aligned additional samples that we had to use in order to avoid high class imbalance in our the data. The distribution of active and passive constructions in the datasets and the total number of samples is presented in Table 2.

Table 2. The proportions of active and passive sentences by language

	EN	DE	SE	IT	ES	FR	RU	CZ	BG
Active (full alignment)	169	195	196	183	212	186	216	217	187
Passive (full alignment)	67	41	40	53	24	50	20	19	49
Additional passive (partial alignment)	103	23	52	53	50	51	40	37	52
Total	339	259	288	289	286	287	276	273	288

EN — English, DE — German, SE — Swedish, IT — Italian, ES — Spanish, FR — French, RU — Russian, CZ — Czech, BG — Bulgarian

Linguistic features of the samples were annotated manually. The existence of fully aligned sets of the sentences allowed us to focus only on the English data and transfer the values of features to the data of the other languages. Alignment is not the only reason why we can easily project the features from one language to the other ones. The characteristics we use are applicable to the situation and its participants, i.e., those that are not affected by the translation and hold the same for all languages. Since our goal is not to distinguish between active and passive constructions, like in the task of morphological or syntax parsing, but to predict the probability of passive occurrence, we do not use features that directly indicate the presence of passive or active (e.g., oblique case, verb form).

The features we selected for our experiment refer to the semantic level of language representation. Formulating the features, we adopted the scheme of the basic transitive diathesis from [Kulikov 2011] and most of the features are assigned to X and Y.

Table 3. The basic transitive diathesis

Semantic argument level	X (Actor)	Y (Undergoer)
Syntactic functional level	Subject	Direct Object

There are two binary features that represent the ***semantic role status*** of X and Y. One feature encodes if X is a prototypical **actor** and the other one encodes if Y

is a prototypical **undergoer**[1]**. For example, in a sentence** *John hit the dog*, 'John' is a prototypical agent and 'the dog' is a prototypical undergoer, in this case both features get 1 as their value. In the sentence *John saw the dog*, 'John' is an experiencer and 'the dog' is a theme, both features are of value 0.

For encoding of ***involvement*** feature we used idea of continuum proposed in [Lehmann 2006], we decided to assign the value of 1 to the most central participants and 0 value was assigned to the most peripheral participants. The participants located in the middle of continuum got values between 1 and 0, depending to which end of the continuum they were closer. Involvement is assigned for both X and Y.

For the ***empathy*** encoding we used the hierarchy proposed in [Lehmann 2006], see **Figure 1**.

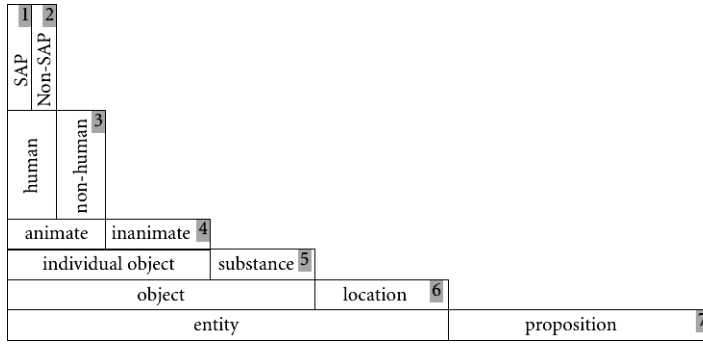


Fig. 1. The empathy hierarchy [Lehmann 2006]

According to this hierarchy there are seven degrees of empathy, where 1 is the highest value that is attested to speech act participants and 7 is the lowest value attested to propositions.

Besides the empathy itself, we include two features that express relation of the X empathy and the Y empathy**.** The first one is the ***absolute difference of X and Y empathy values***, the other one is the ***ratio of the absolute difference and the sum of X and Y empathy values***.

We also include a feature that represents if a verb has a ***negative*** component in its meaning, i.e. verbs like *attack* or *interrupt* get 1 as a feature value.

The next two features are a contextual, they encode, whether the arguments of the verb ***were mentioned*** in 3 preceding sentences, again, separetely for X and Y.

(3) When my sister was six years old, she was attacked, set upon, by three Muggle boys.

The ***contrast*** feature unlike the other ones is related to the situation as a whole. Let us consider some examples, which represent the cases of a positive value for this feature.

(4) a. If you continue behaving like that, *you will be punished by the headmaster.*
b. She tried to reach him, but *he was swallowed by the darkness*.

[1] This coincides with the scheme of control feature described in [Lehmann 2006].

If the target situation is somehow opposed to the other situation or two situations are in relation of contrast then this feature gets a value of 1 and 0 otherwise.

As a result, we end up with a list of twelve features:

1) X is actor
2) X's empathy
3) X was mentioned
4) X's involvement
5) Y's empathy
6) Y was mentioned
7) Y' involvement
8) Y's undergoer
9) Contrast
10) Empathy distance
11) Empathy distance 2
12) Negative action

Further we describe the models trained on these features and show, how they influence the choice of passive occurrence in the languages from our sample.

5. The Models

The goal of this experiment is to determine which of the factors appear to be the most important triggers for the passive occurrence, and to what extent the languages from our sample differ in their mechanism for choosing between active and passive constructions. To meet that goal, for each language in the sample we train a logistic regression model that predicts the probability of passive construction occurrence.

To avoid overfitting, for evaluation we used stratified cross-validation with 10 folds and checked these results by the permutation test [Ojala, Garriga 2010], which shows if the classification result is significant compared to a model trained on data with randomly assigned class labels. Since the datasets for most of the languages are unbalanced we decided to use the "class_weight" parameter of logistic regression function from the sklearn library in Python and assign weights to classes. In table 4 we present the evaluation results.

Table 4. Cross-validation results with permutation test scores

Metric	EN	DE	SE	IT	ES	FR	RU	CZ	BG
F1(macro average)/ Accuracy	0.88/ 0.88	0.82/ 0.85	0.83/ 0.85	0.85/ 0.86	0.79/ 0.82	0.83/ 0.86	0.7/ 0.77	0.72/ 0.77	0.84/ 0.85
PTS p-value	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009

PTS — Permutation test score, EN — English, DE — German, SE — Swedish, IT — Italian, ES — Spanish, FR — French, RU — Russian, CZ — Czech, BG — Bulgarian

All the results have low p-values of the permutation test, it indicates that our results are significant and not random. The F1 and accuracy values for Russian and Czech seem to be rather low, what questions the reliability of the results we get for these models. That might be explained by the fact of the dataset unbalance. However, the German data have similar class ratio but the results are much better, which makes us think that Russian and Czech data probably need different set of features for modelling this problem.

The **Figure 2** represents the feature coefficients distribution in the languages, we included only those features that appeared to be statistically significant in the models, i.e., they had p-values < 0.05[2]. The insignificant features are assigned to a zero coefficient value.

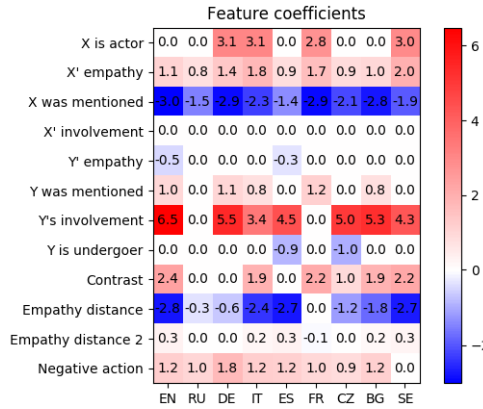


Figure 2. The model coefficients

The model coefficients reflect the impact of the features: the positive coefficient tells us that the feature contributes to the passive occurrence positively, negative coefficients indicate a positive impact on active occurrence. The features with values close to zero have very little impact on the classification.

In the final section we discuss the results and make conclusions.

6. Discussion and conclusion

As we can see from the Figure 2 the languages reveal similar coefficient distributions, but still they are not the same. Definitely there are features that have no impact in all or almost all languages, those are *X involvement*, *Y empathy*, *Y was mentioned*, *Y undergoer*, *Empathy distance 2*. The features relevant for majority of the languages are *Y involvement* (very strong), *X empathy* and *Negative action,* which are the triggers for passive occurrence, and also X was *mentioned*, which is a pro-active feature. The feature *X actor* is a strong passive trigger in German, Swedish, Italian, and French,

[2] The features Y involvement in Italian and Y empathy in Spanish were nearly significant but increased the evaluation results, so we also included they in the final models.

Contrast is a pro-passive feature for English, Italian, French, Czech, Bulgarian and Swedish, and *Empathy distance* is an pro-active feature for English, Italian, Spanish, Czech, Bulgarian and Swedish.

At this point, we do not make any claims about the nature of passives, and the results should be treated as preliminary. We demonstrated, how logistic regression can be applied to the multilingual parallel data for exploration of a language phenomenon. The determined set of language-independent features appeared to be relevant for prediction of passive occurrence in the majotiry of the considered languages. Except for Russian and Czech, the models have relatively good predictive power (the mean accuracy rate is about 82–85%). The robustness of current models should be tested under other conditions. The future research should include similar experiments based on 1) larger and more balanced datasets 2) multilingual parallel texts of other genres (e.g., movie subtitles, reports of Europarlament) 3) non-parallel corpus data. The current set of features may reveal different results for that data.

References

1. *Asgari E., Schütze H.* (2017), Past, Present, Future: A computational investigation of the typology of tense in 1000 languages, https://arxiv.org/abs/1704.08914
2. *Berez A. L. and Gries S. T.* (2010), Correlates to middle marking in Dena'ina iterative verbs. International Journal of American Linguistics, 76(1), pp.145–165.
3. *Bonch-Osmolovskaya A. A., Nesterenko L. V.* (2018), Networks as an instrument for "searches" and "findings" in multilingual parallel corpora [Seti kak instrument poiska i nakhodok v multijazychnykh parallelnykh korpusakh], EVRika! Papers on "searches" and "findings" to the anniversary of E. V. Rakhilina [EVRika! Sbornik statej o poiskakh i nakhodkakh k jubileju E. V. Rakhilinoj], Labirint, Moskva, pp. 305–320.
4. *Comrie B.* (1988). Passive and voice. Passive and voice, 16, pp. 9–24.
5. *Cysouw M.* (2014), Inducing semantic roles. Perspectives on semantic roles. Perspectives on semantic roles. Luraghi S., Narrog H. (eds.). Amsterdam: Benjamins, pp. 23–68.
6. *Dahl Ö.* (2007), From questionnaires to parallel corpora in typology. STUF-Sprachtypologie und Universalienforschung, 60(2), pp. 172–181.
7. *Foley W. A., Van Valin, R. D. Jr.* (1984), Functional Syntax and Universal Grammar. Cambridge: Cambridge University Press.
8. *Fox B. A., Hopper P. J.* (1994), Voice: Form and function (Vol. 27). John Benjamins Publishing.
9. *Frajzyngier Z.* (1982), Indefinite agent, passive and impersonal passive: a functional study. Lingua, 58(3–4), pp. 267–290.
10. *Gale W. A., Church K. W.* (1991), Identifying Word Correspondences in Parallel Texts. HLT, 91, pp. 152–157.
11. *Givón T.* (1994). Voice and inversion (Vol. 28). John Benjamins Publishing Company.
12. *Guyon I., Weston J., Barnhill S., Vapnik, V.* (2002), "Gene selection for cancer classification using support vector machines", Mach. Learn., 46(1–3), pp. 389–422.

13. *Jespersen, O.* (1924), The Philosophy of Grammar. London: Allen & Unwin.
14. *Jisa H., Reilly J., Verhoeven L., Baruch E., Rosado E.* (2002), Passive voice constructions in written texts: A cross-linguistic developmental study. Written Language & Literacy, 5(2), pp. 163–81.
15. *Kallulli D.* (2006), A unified analysis of passives, anticausatives and reflexives. Empirical issues in formal syntax and semantics, 6, pp. 201–225.
16. *Keenan, E. L., Dryer, M. S.* (1981). Passive in the world's languages. Linguistic Agency, University of Trier.
17. *Kulikov L.* (2011), Voice typology. In The oxford handbook of linguistic typology. Oxford University Press, pp. 368–398.
18. *Lehmann, C.* (2006), Participant roles, thematic roles and syntactic relations. Voice and Grammatical Relations: In Honor of Masayoshi Shibatani, 65, pp. 153–174.
19. *Langacker, R. W.* (2006), Dimensions of defocusing. Voice and grammatical relations: In Honor of Masayoshi Shibatani, pp. 115–137.
20. *Manninen S., Nelson D.* (2004), What is a passive? The case of Finnish. Studia Linguistica, 58(3), pp. 212–251.
21. *Mayer T., Cysouw M.* (2012), Language comparison through sparse multilingual word alignment. Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH. Association for Computational Linguistics, 2012, pp. 54–62.
22. *Melchuk I. A. & Kholodovich A. A.* (1970), On the theory of voice [K teorii grammaticheskogo zaloga] Asian and African studies [Narody Azii i Afriki], 4, pp. 111–124.
23. *Nesterenko L. V.* (2019), Multilingual parallel corpora: Alternative resource of language data for typological studies, usage perspectives and problems [Multijazychnye parallelnye korpus: novyj istochnik dannykh dlya tipologicheskikh issledovanij, perspektivy ispolzovanija i problemy], Problems of Linguistics [Voprosy jazykoznanija], accepted
24. *Nivre J., de Marneffe M.-C., Ginter F., Goldberg Y., Haji J., Manning C. D., McDonald R., Petrov S., Pyysalo S., Silveira N., Tsarfaty R., Zeman D.* (2016), Universal dependencies v1: A multilingual treebank collection. Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), pp. 1659–1666.
25. *Ojala M., Garriga, G. C.* (2010), Permutation tests for studying classifier performance. Journal of Machine Learning Research, 11(Jun), pp. 1833–1863.
26. *Östling R.* (2015), Word order typology through multilingual word alignment. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Vol. 2: Short papers, pp. 205–211.
27. *Östling R.* (2016), Studying colexification through massively parallel corpora. The lexical typology of semantic shifts, 58, pp. 157–176.
28. *Östling, R., Tiedemann, J.* (2016). Efficient word alignment with markov chain monte carlo. The Prague Bulletin of Mathematical Linguistics, 106(1), pp. 125–146.
29. *Plecháčková J.* (2007), Passive Voice in Translation: A Corpus-Based Study (MA dissertation), https://is.muni.cz/th/ew9h5/diplomka_final_version.pdf
30. *Shibatani M.* (1985), Passives and related constructions: A prototype analysis. Language, pp. 821–848.

31. *Shibatani M.* (1988), Passive and voice (Vol. 16). John Benjamins Publishing.
32. *Straka M., Straková J.* (2017), Tokenizing, POS-tagging, lemmatizing and parsing UD 2.0 with UDPipe. Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pp. 88–99.
33. *Tsunoda T., Kageyama T.* (2006). Voice and grammatical relations: in honor of Masayoshi Shibatani (Vol. 65). John Benjamins Publishing.
34. *Wälchli B.* (2007), Advantages and disadvantages of using parallel texts in typological investigations. STUF — Sprachtypologie und Universalienforschung, 60(2): 118–134.
35. *Xiao R. Z.* (2007), What can SLA learn from contrastive corpus linguistics? The case of passive constructions in Chinese learner English. Indonesian JELT, 3(1), pp. 1–19.

REFERENTIAL CHOICE IN MULTIMODAL COMMUNICATION[1]

Budennaya E. V. (jane.sdrv@gmail.com)
Institute of Linguistics, RAS, Moscow, Russia

This article deals with an application of referential markup to a large multimodal resource "Russian Pear Chats and Stories", annotated for vocal, oculomotor, manual and cephalic channels. Despite a large number of works on referential choice, it has never been investigated within the framework of multimodal communication. For this purpose, a special annotation scheme in the ELAN environment is proposed, allowing one to annotate different types of referential units and to conduct a simultaneous tracking of referential expressions (full NPs, pronouns, demonstratives, zeroes, etc) with accompanying verbal and non-verbal units. The analysis of three recordings (overall duration equals to 141 minute), where the new referential annotation was introduced in addition to the existing multimodal markup, reveals a range of understudied peculiarities of the referential choice. It was found that the role of the Commentator in the conversation entails a significantly larger amount of constructions with a zero subject pronoun, compared to the monologue discourse of the Narrator and the Reteller. The analysis of referential expressions and accompanying pointing gestures complied with more general data previously obtained on the English material and showed that nouns are significantly more often accompanied by a pointing stroke than personal pronouns, while demonstratives occupy an intermediate position between nouns and personal pronouns as units potentially accompanied by a gesture.

Keywords: referential choice, referential expression, full NP, personal pronoun, demonstrative pronoun, pointing gesture, multimodality

1. Introduction. Multimodal communication as a subject of study and the resource "Russian Pear Chats and Stories"

Human communication is a simultaneous interaction of verbal and non-verbal components, or channels. Their joint analysis has only recently become possible within the multimodal (multichannel) approach [Kress 2002]; [Kibrik 2010]; [Knight 2011]. Its development went along with the creation of multimodal resources compiled invast repositories of media files annotated for various channels of communication [Kibrik, Podlesskaya 2009]; [Brône, Oben 2015]; [Kibrik 2018a].

This article lies within the framework of multimodal studies and is based on one of the largest resources in terms of annotated channels, "Russian Pear Chats and Stories" (for details, see the website www.multidiscourse.ru). It consists of 24 sessions

[1] The study was supported by RFBR, research project #18-00-01485

of natural communication in groups of four participants discussing the Pear Film [Chafe 1980]. Each of the four participants has a fixed role: the Narrator (N), the Commentator (C), the Reteller (R) and the Listener (L). At the preliminary stage, N and C watch the film, then N tells the plot of the film for R. No interruption is allowed until N completes the story. This is followed by a conversation stage, in which C adds details to the first story of N and R asks questions that helps him/her understand the plot better. After that, R retells the plot of the film to L, who joins the rest of the participants only at the last final stage. For details, see [Kibrik 2018b].

The main mark-up of the "Russian Pear Chats and Stories" corpus is carried out with ELAN software (https://tla.mpi.nl/tools/tla-tools/elan) and includes the annotation of vocal, oculomotor, manual and cephalic channels. In addition, on a subcorpus of three sessions (04), (22), (23), a referential annotation is performed. This paper is particularly devoted to this aspect. At present, there are many works which discuss potential factors of referential choice [Arnold 2001]; [Kaiser, Truswell 2008]; [Kibrik et al. 2016], but this is the first time it is explored in the context of multimodal communication. In this regard, a brief review of the phenomenon of referential choice and its aspects studied in this framework will be given in Section 2; Section 3 will focus on basic principles of the applied annotation. Section 4 will provide summary results of different referential units' behavior throughout the corpus. Finally, Section 5 will examine one application of referential annotation in the multichannel approach—the analysis of referential choice and accompanying pointing gestures.

2. Referential choice: types of annotated referential expressions

Referential choice is the choice of a language expression which refers to any definite object or phenomenon. The same referent can be marked with a full NP, a personal pronoun, a reflexive, a zero form, etc. Among all language expressions with a specific definite reference, two basic types of units can be distinguished: anaphoric and deictic.

Anaphoric units are language expressions that are impossible to interpret unambiguously without referring to previously known contextual information. The majority of works on referential choice are dedicated to this type reference (see [Kibrik 1996]; [Kaiser 2013]; [Kibrik et al. 2016]; [Sauermann, Gagarina 2017] inter alia). The list of anaphoric language expressions is large and includes:

- third person pronouns:

(1) (h) (q) (a) **[on]** XX sobiraet /–gruši,
'**[He]** is picking pears'

- demonstrative pronouns:

(2) Nu potomu čto/**[tot]** nemnožko byl v\šoke;
'Well as **[that** *one*[2]**]** was a bit shocked'

[2] Here and throughout italics is used in English translations for Russian zero forms.

- definite pronouns:

(3) /volosy **[u vsex]** dostatočno dlinnyje;
'The hair **[of everybody]** is quite long'

- indefinite pronouns:

(4) nu-u n'= || **[kto-to]** /vyše,
[kto-to] \niže^{w}
'Well **[some]** are taller, **[some]** are shorter'

- zeroes:

(5) a potom **[Ø$_{pro}$]** spuskajets'a po /lestnice
'and then [*he*] goes down the stairs'

All these units belong to a wider category of reduced reference [Kibrik 2011]. This category is opposed to full reference, i.e. constituents with a noun or numeral head, which act as antecedents for subsequent anaphoric units:

(6) (a) (m) velosiped **[mal'čiku]$_i$** /velik
(h) [on]$_i$ jedet ne v /sedle,
'The bike is too large **[for the boy]**, [he] does not ride the saddle'

(7) Bylo **[dve /napolnennyx]**, **[odna]** byla \pustaja.
'**[Two]** were full, **[one]** was empty.'

In this study, both reduced anaphoric expressions and their full antecedents are annotated within a wider category of anaphoric reference.

Another type of reference is deixis. Unlike anaphora, where the referent's interpretation relies on previous context, deictic referents are identified through visual attention. The linguistic expressions most commonly used for of deictic reference include first and second person pronouns (both overt and zero forms), demonstrative pronouns and demonstrative adverbs. In this study the list of annotated deictic expressions is limited to first- and second-person subject pronouns. In a number of languages, including Russian, one may choose not to express these pronouns explicitly (see Example 9) and the choice between overt and zero subject forms is a complex issue which depends on a range of heterogeneous factors.

(8) (/Možno **[ja]** rasskažu^{h}?
'Can **[I]** tell?'

(9) **[Ø$_{pro}$]** \Voobšče ne pomn'u etogo!
'[*I*] absolutely do not remember it'

Although the dominant referential pattern in Russian is the one in which a subject pronoun is explicitly expressed, its omission is also quite common and ranges from one-fourth to one-third of all occurrences [Kibrik 1996]; [Grenoble 2001]. Factors that influence the presence or absence of Russian subject pronouns have been investigated (see, for example, [Seo 2001]; [Zdorenko 2010]), but at present they remain neither definitively classified nor fully understood.

Most studies of the referential choice focus on anaphoric devices and do not include deixis. This is partly due to the fact that in the case of deictic subject reference, the referent of the linguistic expression is depends upon a predetermined binary opposition permitted only in languages which allow subject omission, a process significantly different from the selection of anaphoric expressions. Nevertheless, since the corpus "Russian Pear Chats and Stories" implies further use for general linguistic research, both types of reference are included in the current markup. It is assumed that the corresponding annotations will help to establish additional prosodic and kinetic factors associated with referential choice.

In recent years, researchers have developed numerous resources that investigate coreference in anaphoric relationships, e.g. ARRAU [Poesio, Artstein 2008], GREC [Belz et al. 2010], WSJ MoRA Corpus [Kibrik 2016 et al.]; RuCor [Toldova et al. 2014]. The markup of these corpora is based on special procedures considering the distance between the anaphor and the antecedent, as well as their discourse, grammatical and syntactic properties [Kibrik et al. 2016], provided by automatic extraction tools (see, for example, MMAX [Müller, Strube 2007], RuCor Annotation tool [Toldova et al. 2016]). Unlike previously cited works, this article focuses on the relations between different types of referential expressions and various kinetic channels of communication. This issue has been partially investigated with regard to Germanic and Turkic languages [Gullberg 2006]; [Debreslioska et al. 2013]; [Azar, Özürek 2015], but this is the first time Russian corpus data is brought into the picture. Since the annotation of all communication channels in "Russian Pear Chats and Stories" was performed in ELAN and the current task involves further analysis of non-verbal channels on a par with with accompanying referential units, referential annotation at the current stage was also carried out in ELAN. Accordingly, certain solutions were modified to incorporate embedded and zero units, since by default ELAN does not allow such kinds of annotations. The following solutions will be presented below.

3. Principles of referential annotation

The entire referential annotation of the "Russian Pear Chats and Stories" corpus is divided into two markups dealing with anaphoric and deictic expressions, respectively. Within the annotation, anaphoric and deictic units form two main independent tiers (refAnaphora/refDeixis) which both depend on the Words[3] tier (part of the vocal annotation). Each of these two tiers is a parent for several tiers where a range of referential parameters is annotated (see below). For the third person zero pronoun, the following explicitly expressed word in the Words tier is annotated as a parent item. Sometimes referential expressions form embedded constituents: several syntactically related (either with coordination or subordination) anaphoric expressions form a compound expression with a definite reference:

[3] On the intermediate tier refAnSpread for anaphoric expressions, which is located between Words and refAnaphoraN tiers in the hierarchy, see below.

(10) {sw} /po-mojemu eto bylo **[[/dva mal'čika]$_i$ i [devočka]$_j$]$_k$**
'I think there were [[two boys]$_i$ and [a girl]$_j$]$_k$'

(11) /Potom v kadre pojavl'ajets' **[[(h) (a) \mal'čik [na velosipede]$_i$]$_j$**
'Then in the shot [a boy [on a bike]$_i$]$_j$ appears

Since ELAN does not allow creating a tier with different levels of annotation, all referential expressions that do not contain embedded constituents (=discrete annotations of the refAnaphora tier) depend on the refAnSpread tier. Each of the annotations of the refAnSpread tier represents a group of one or several syntactically related expressions refAnaphora by default. At present, based on the recordings #4, #22, and #23 of the "Russian Pear Chats and Stories" corpus, the maximum number of embedded constituents was equal to three items. For this reason, in the annotation scheme the refAnSpread tier is by default a parent to the three child tiers, refAnaphora1, refAnaphora2 and refAnaphora3, respectively. When no embedded referential are found, only the first refAnaphora1 tier is annotated.

At present, the referential annotation does not allow automatic extraction of coreference units. The elaboration of an appropriate tool for this purpose is left for future research. However, other parameters potentially affecting the referential choice are taken into account. In the current version of the referential markup, the following parameters are annotated, each of them forming a separate layer that depends either on the refDeixis or on the refAnaphoraN[4] tier, according to the type of the referential expression:

1) For deictic expressions:
   a. Explicitness (*refDeiExpr* tier): explicitly expressed/not expressed (Overt/Zero). For a first or second zero pronoun, the following explicitly expressed word in the *Words* tier serves as a parent element.
   b. Person (*refDeiPerson* tier): first/second person (1/2)
   c. Number (*refDeiNumber* tier): singular/plural (SG/PL)
   d. Grammatical role (*refDeiSynt* tier): subject/direct object/indirect object/other (Subj/DirObj/IndirObj/Other)

2) For anaphoric expressions (*refAnaphora1–3*):
   a. Type of reference (*refAnType* tier): Full/ Reduced
   b. Type of referential expression (*refAnExpression* tier; depends on the *refAnType* tier): NP with a noun/numeral head (NomP/NumP) for full reference VS third person pronoun/ demonstrative pronoun/ definitive pronoun/indefinite pronoun/zero for reduced reference (Pron3/Dem/Def/Indef/Zero). In the case of a preceding preposition, a corresponding expression with a Prep-tag is to be chosen.
   c. Gender (*refAnGender* tier): male/female/neutral/mixed/other (M/F/N/Mixed/Other). 'Mixed' is used for a compound expression which refers to several different entities (as shown in Example 10).

[4] N equals to 1, 2 or 3, due to the number of embedded units.

'Other' is used when it is impossible to define the gender of paired elements, such as *nožnicy* 'scissors', *br'uki* 'pants', etc.

d. Number of the referent (*refAnNumber* tier): singular/plural (SG/PL)

e. Syntactic expression (*refAnSynt* tier): subject/direct object/indirect object/other (Subj/DirObj/IndirObj/Other).

Apart from this, each *refAnaphoraN* and *refDeixis* tier has an dependent tier with comments where the type of relation for compound elements (Coordination / Subordination), a preceding definite (With Def) or indefinite pronoun (With Undef), or relative clause (With REL) can be marked, along with emphasis (Emphasis) or contrastiveness (Contrast).

An example of referential annotation is presented below in **Figure 1**:

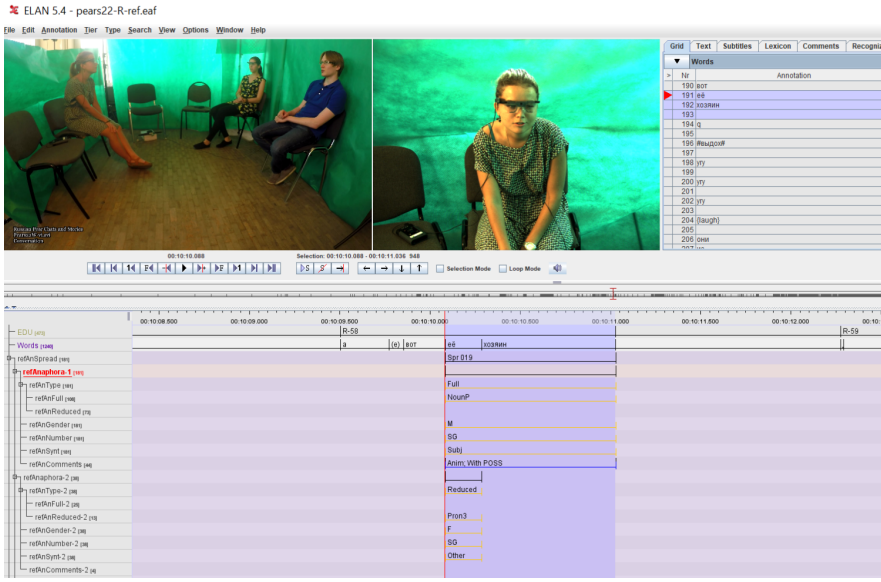


Fig. 1. Referential annotation of the compound unit [[*jeje*] *xoz'ain*] '[[her] owner]'

4. The distribution of referential units in the corpus

According to the demo subcorpus of 3 sessions (9 videos—22N, 22C, 22R; 4N, 4C, 4R; 23N, 23C, 23R, total duration equals to 141 minute), among the most frequent anaphoric referential expressions full NPs with a noun or numeral head (FullNP) account the highest proportion (52%). They are followed by personal pronouns (Pron, 34%) and zero forms (Zero, 13%). Demonstrative pronouns are extremely rare (Dem, 1%):

Table 1. The distribution of different anaphoric expressions in the "Russian Pear Chats and Stories" corpus

	4N	4C	4R	22N	22C	22R	23N	23C	23R	Total
FullNP	161	103	187	144	92	154	121	123	133	1,218 (52%)
Pron	119	61	152	86	63	88	112	72	58	811 (34%)
Dem	2	1	2	0	2	2	2	5	2	18 (1%)
Zero	54	27	68	27	19	32	18	38	25	308 (13%)

This ratio is consistent with other results on Russian oral discourse (see, for example, [Grenoble 2001]). From a statistical perspective, no clear relationship between the percentage of referential expressions and the speaker's role (N/C/R) was revealed at this stage. However, a number of features (in particular, a rather high percentage of pronouns in 23N, compared to other speakers, see **diagram 1**) draws attention and may be the subject of further research.

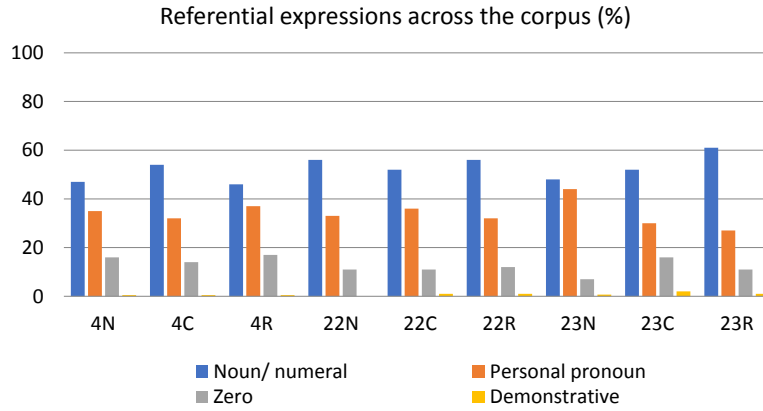


Diagram 1. The distribution of different anaphoric expressions according to the speaker's role (N/C/R)

The analysis of deictic subject reference revealed the expected advantage of the pattern with an explicitly expressed pronoun (67%) over the zero one, see **Table 2**.

Table 2. Deictic subject reference in the "Russian Pear Chats and Stories" corpus

	4N	4C	4R	22N	22C	22R	23N	23C	23R	Total
Overt	18	4	11	20	7	7	9	4	5	85 (67%)
Zero	8	6	3	5	7	2	6	3	2	42 (33%)

It was also found that commentators used a significantly larger percentage of constructions with a zero subject pronoun, compared to other participants:

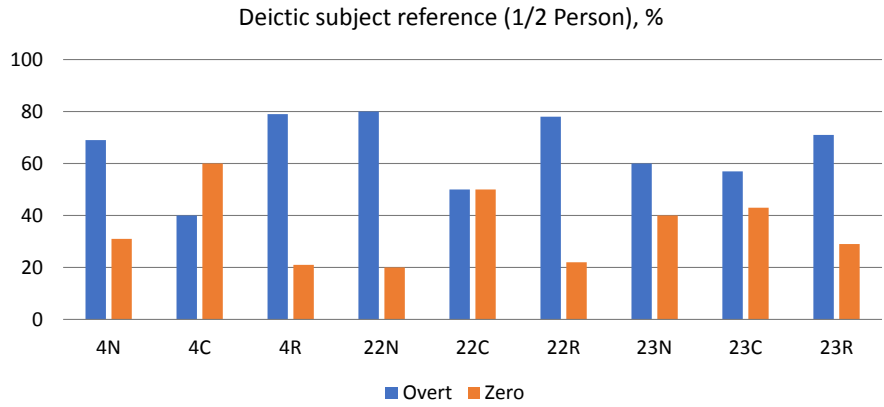


Diagram 2. The distribution of personal deictic pronouns according to the speaker's role (N/C/R)

Statistical analysis (Fisher's exact test, p-value <0.05) confirmed that C dropped pronouns significantly more often than N and R. Apparently, the role of the Commentator, who participated only in a spontaneous conversation process and never produced a more structured monologue, was associated with more conditions for the discourse subject ellipsis (for more details on this phenomenon, see [Zdorenko 2010]). In this regard, the empirical remark that in Russian "the absence of the first person subject pronoun is related to the neutralization of the speaker's role" [Fougeron, Breillard 2004: 159] can be proven: compared to N and R, C focused to the least extent on himself/herself. This additionally contributed to a greater percentage of patterns with a zero subject pronoun.

5. Application of referential annotation: referential expressions in interaction with the manual channel

Studies conducted on several European languages indicate that new and cognitively less accessible referents are more likely to be expressed by full NPs with a noun head [Chafe 1994] and accompanied by a gesture [Levy, Fowler 2000]. In contrast, pronouns and zeros are associated with the most cognitively accessible information that does not imply additional expression on a non-verbal level. Our research was aimed at testing this hypothesis on Russian-language data, focusing particularly on pointing gestures.

On a semantic level, this type of gesture connects most closely to specific definite linguistic entities, and therefore to referential expressions [Kibrik 2011: 41]. At present, studies on relations between reference and gestures exist ([Gullberg 2006; Debreslioska et al. 2013]), yet there are rather few works on particular types of gestures and accompanying referential expressions, and all of them are based on languages others than Russian ([Sluis, Krahmer 2007] on Dutch, [Azar, Özyürek 2015] on Turkish).

In the course of our analysis, all pointing gestures were extracted from the previously conducted manual annotation of the corpus [Litvinenko et al. 2016]. Afterward, strokes—the most semantically significant phases of the gestures—were aligned with referential expressions that overlapped with them in time, according to the principle of "minimal overlapping" [Fedorova et al. 2015]. Many gestures accompanied the verbs of movement and did not correspond to explicitly expressed referential expressions. Although these gestures underwent no further analysis, the summary **table 3** incorporates them as well.

Statistical analysis of the data obtained showed that in most cases pointing strokes were aligned with full NPs with a noun and numeral head (χ-square, p-value < 0.01). Next are demonstrative pronouns, which overlap with a pointing stroke more often than personal pronouns (χ-square, p-value $= 0.05$). This conclusion is consistent with similar observations on other languages but specifies the behavior of pronouns and demonstratives.

It was also found that the "prototypical" object, accompanied by a pointing gesture, is a singular noun (see **Table 4**). No correlation between the presence of a pointing gesture, on the one hand, and the referent's gender and grammatical role, on the other hand, has been detected.

Table 3. Types of different referential expressions (nouns/numerals; personal pronouns and demonstratives, including demonstrative pronouns within full noun phrases with a noun head), aligning with strokes of pointing gestures

	4N	4C	4R	22N	22C	22R	23N	23C	23R	Total
Pointing gestures	78	14	30	52	30	40	54	19	39	356
NounP/ NumP	37 (45%)	3 (14%)	14 (47%)	20 (38%)	12 (37%)	15 (38%)	21 (39%)	8 (42%)	22 (56%)	153 (44%)
Pron	4 (5%)	2 (14%)	7 (23%)	3 (6%)	7 (23%)	4 (10%)	7 (13%)	2 (10%)	8 (21%)	44 (14%)
Dem_all	2 (2%)	0 (0%)	2 (7%)	0 (0%)	2 (7%)	1 (3%)	0 (0%)	1 (5%)	2 (5%)	11 (3%)

Table 4. Singular and plural referential expressions aligned with pointing gestures' strokes (%)

	TOTAL	Aligned with pointing strokes	%	p-value (χ-square)
Full NP SG	871	121	14%	0.05
Full NP PL	380	37	9%	

This data can serve as an aid to the analysis of human communication. In particular, obtained correlations can help determine the most probable referential expression in case of a sound loss.

6. Conclusion

The article has demonstrated the application of referential markup to a large multimodal corpus. A specially designed annotation scheme was presented, allowing for a simultaneous analysis of referential expressions and accompanying non-verbal means of communication. It was shown that the addition of a new referential component could contribute to studies of both referential choice and the interaction of different communication channels. Namely, it was found that the role of the Commentator in the process of spontaneous dialogue entails a significantly larger percentage of constructions with zero subject pronouns, compared to the monologue discourse of the Narrator and the Reteller. The analysis of referential expressions and accompanying pointing gestures confirmed data which were previously obtained on the English material [Levy, Fowler 2000] and showed that in most cases pointing strokes were accompanied by full NPs with a noun or numeral head and almost never accompanied by personal pronouns. Demonstratives took an intermediate place in this hierarchy.

Further research on the relationship between referential expressions and speakers' nonverbal behavior (e.g., eye and head movements) will likely contribute to establishing other unnoticed peculiarities of human communication.

References

1. *Arnold J.* (2001), The Effect of Thematic Roles on Pronoun Use and Frequency of Reference Continuation, Discourse Processes 31, pp. 137–162.
2. *Azar Z., Özürek A.* (2015). Discourse Management: Reference tracking in speech and gesture in Turkish narratives, Dutch Journal of Applied Linguistics 4 (2), pp. 222–240.
3. *Belz A., Know E., Viethen J., Gatt A.* (2010). Generating referring expressions in context: the GREC task evaluation challenges, E. Krahmerand, M. Theune (eds.), Empirical Methods in Natural Language Generation, Springer, Berlin, pp. 294–328.
4. *Brône G., Oben B.* (2015), What you see is what you do. On the relation between gaze and gesture in multimodal alignment, Language and Cognition 7 (4), pp. 485–498.
5. *Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D. & Wellner P.* (2005), The AMI Meeting Corpus: A PreAnnouncement, Proceedings of the Second international conference on Machine Learning for Multimodal Interaction, 28–39.
6. *Chafe W.* (ed.) (1980), The pear stories: Cognitive, cultural, and linguistic aspects of narrative production, Ablex, Norwood.
7. *Chafe W.* (1994), Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing, University of Chicago Press.
8. *Chinchor N., Robinson P.* (1997), MUC-7 named entity task definition, Proceedings of the7th Conference on Message Understanding, Fairfax, VA, 29.
9. *Debreslioska S., Özürek A., Gullberg M., Perniss P.* (2013). Gestural viewpoint signals referent accessibility. Discourse Processes, 50(7), pp. 431–56.

10. *Fedorova O. V., Kibrik A. A., Korotaev N. A., Litvinenko A. O., Nikolaeva Ju. V.* (2016), Temporal coordination between gestural and speech units in multimodalcommunication [Vremennaya koordinatsiya mezhdu hestovymi i rechevymi edinitsami v mul'timodal'noy kommunikatsii], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii], RGGU, Moscow, pp. 159–170.
11. *Fougeron [Fužeron] I., Breillard J. [Žan Brejar]* (2004), 'Mestoimenie "ja" I postroenie diskursivnyx svjazej v sovremennom russkom jazyke' [The pronoun "ja" and the construction of discourse links in modern Russian], T. M. Nikolaeva (ed.), Verbal'naja i neverbal'naja opory prostranstva mežfrazovyx svjazej, Jazyki slavjanskoj kul'tury, Moscow, pp. 147–166.
12. *Granstrom B., House D., Karlsson I.* (eds.) (2002), Multimodality in language and speech systems, Kluwer, Dordrecht.
13. *Grenoble L.* (2001), Conceptual reference points, pronouns and conversational structure in Russian, Glossos, 1(1).
14. *Grishina E. A.* (2017), Russian gestures from a linguistic perspective [Russkaya zhestikulyatsiya s lingvisticheskoy tochki zreniya], Jazyki slavyanskoy kul'tury, Moscow.
15. *Gullberg M.* (2006), Handling discourse: Gestures, reference tracking, and communication strategies in early L2, Language Learning, 56 (1), pp. 155–196.
16. *Kaiser E.* (2013), Looking beyond personal pronouns and beyond English: Typological and computational complexity in reference resolution, Theoretical Linguistics 39 (1–2), pp. 109–122.
17. *Kaiser E., Trueswell J.* (2008), Interpreting pronouns and demonstratives in Finnish: Evidence for a form-specific approach to reference resolution, Language and Cognitive Processes 23(5), pp. 709–748.
18. *Kendon A.* (1967) Some functions of gaze-direction in social interaction, Acta Psychologica, Vol. 26 (1), pp. 22–63.
19. *Kibrik A. A., Podlesskaja V. I.* (eds.) (2009), Night Dream Stories: A corpus study of spoken Russian discourse [Rasskazy o snovidenijax: Korpusnoe issledovanie ustnogo russkogo diskursa], LRC, Moscow.
20. *Kibrik A. A.* (2010), Multimodal linguistics [Mul'timodal'naya lingvistika], Yu. I. Aleksandrov, V. D. Solov'yev (eds.), Cognitive studies [Kognitivnyye issledovaniya], Vol. IV, Institute of psychology, Moscow, pp. 134–152.
21. *Kibrik A. A.* (2011), Reference in discourse, Oxford University Press, Oxford.
22. *Kibrik A. A., Khudyakova M. V., Dobrov G. B., Linnik A., Zalmanov D. A.* (2016), Referential choice: predictability and its Limits, Frontiers in Psychology 7:1429.
23. *Kibrik A. A.* (2018a), Russian multichannel discourse. Part I. Setting up the problem [Russkiy mul'tikanal'nyy diskurs. Chast' I. Postanovka problemy], Psikhologicheskiy zhurnal, Vol. 39 (1), pp. 70–80.
24. *Kibrik A. A.* (2018b), Russian multichannel discourse. Part II. Corpus development and avenues of research [Russkiy mul'tikanal'nyy diskurs. Chast' II. Razrabotka korpusa i napravleniya issledovaniy], Psikhologicheskiy zhurnal, Vol. 39 (2), pp. 78–89.

25. *Knight D.* (2011), Multimodality and active listenership: a corpus approach. Corpus and discourse, Bloomsbury, London.
26. *Krasavina O., Chiarcos C.* (2007), PoCoS: Potsdam coreference scheme, Proceedings of the Linguistic annotation workshop, Prague, (Association for Computational Linguistics, Stroudsburg, PA), pp. 156–163.
27. *Kress G.* (2002), The multimodal landscape of communication, Medien Journal, Vol. 4, pp. 4–19.
28. *Levy E., Fowler, C.* (2000), Grounding references in perception, D. McNeill (ed.), Language and gesture, Cambridge Univeristy Press, New York, pp. 215–234.
29. *Müller C., Cienki A., Fricke E., Ladewig S. H., McNeill D. & Tesendorf S.* (eds.) (2013), Body—Language—Communication: An International Handbook on Multimodality in Human Interaction, Mouton, Berlin.
30. *Müller C., Strube M.* (2006). Multi-level annotation of linguistic data with MMAX2, S. Braun, K. Kohn, J. Mukherjee (eds.), Corpus Technology and Language Pedagogy. New Resources, New Tools, NewMethods, Peter Lang, Frankfurt, pp. 197–214.
31. *Poesio M., Artstein R.* (2008), Anaphoric annotation in the ARRAU corpus, Proceedings of the 6th International Conference on Language Resources and Evaluation, Marrakech.
32. *Sauermann A., Gagarina N.* (2017), Grammatical Role Parallelism Influences Ambiguous Pronoun Resolution in German, Frontiers in Psychology 8.
33. *Seo S.* (2001). The frequency of null subjects in Russian, Polish, Bulgarian and Serbo-Croatian: An analysis according to morphosyntactic environments, Ph.D. thesis, Dept. of Slavic languages and literatures, Indiana University.
34. *Sluis, Ielka van der, Krahmer E.* (2007), Generating Multimodal References, Discourse Processes 44 (3), pp. 145–217.
35. *Toldova S. Ju., Roytberg A., Nedoluzhko A., Kurzukov M., Ladygina A., Vasilyeva M., Azerkovich I., Grishina Y., Sim G., Ivanova A., Gorshkov D.* (2014), Evaluating Anaphora and Coreference Resolution for Russian. Papers from the Annual International Conference "Dialogue": Computational Linguistics and Intellectual Technologies [Komp'juternaja lingvistika i intellektual'nye tehnologii. Po materialam ezhegodnoj Mezhdunarodnoj konferencii "Dialog"] 13 (20), RGGU, Moscow, 2014, pp. 681–695.
36. *Zdorenko T.* (2010), Subject omission in Russian: A study of the Russian national corpus, S. Gries, S. Wulff, M. Davies (eds.), Corpus-linguistic applications: Current studies, new directions, Rodopi, Amsterdam, pp. 119–131.

APPLYING AN AUTOMATIC FTD CLASSIFIER TO THE ANNOTATION OF THE GICR CORPUS

Bulygin M. V. (bulyginmv1996@gmail.com)
Radboud University, Nijmegen, Netherlands

Sharoff S. A. (s.sharoff@leeds.ac.uk)
Russian State University of Humanities, Moscow, Russia;
Leeds University, Leeds, UK

This paper addresses the task of automatic genre classification for Russian within the Functional Text Dimensions (FTD) framework. Our aim in this study was to build the optimate FTD classification model to annotate web texts from the GICR corpus. For training data, we used an extended GICR dataset. We used the Support Vector Machine method with linear kernel for classification and converted training data to lower case to increase accuracy. During our research we experimented with several classification parameters, such as types of features, C-value and feature filtering to determine the best option for the classification model of the GICR dataset. The resulting model was able to achieve satisfactory classification accuracy and was used for GICR annotation. We also looked at the most significant features for each FTD in our best performing model and compared them to the most frequent words in which these features occur. Finally, we applied our model to segments of the GICR and looked at the FTD components in these segments.

Key words: Functional Text Dimensions, genre classification, feature selection, Web corpora annotation

ПРИМЕНЕНИЕ АВТОМАТИЧЕСКОГО FTD КЛАССИФИКАТОРА ДЛЯ АННОТАЦИИ КОРПУСА ГИКРЯ

Булыгин М. В. (bulyginmv1996@gmail.com)
Университет Радбауд, Неймеген, Нидерланды

Шаров С. А. (s.sharoff@leeds.ac.uk)
Российский государственный гуманитарный университет, Москва, Россия; Университет Лидса, Лидс, Великобритания

1. Introduction

Language corpora evolved dramatically since the introduction of the first corpora. We started with a 1-million-word corpus that was collected manually and had restricted annotation, and nowadays we have massive language corpora that usually contain at least over 100 million words and have different sophisticated technics for corpus annotation, like POS-tagger, morphology and syntax analyzer and parser, etc. In some cases, we might even see megacorpora, with billions of words in them. For example, General Internet-Corpus of Russian (GICR) [Piperski et al., 2013] represents a variety of texts from the Russian web and is comprised of 20 billion words. It is obvious that such corpora cannot be constructed and annotated manually. The solution of this problem lies in the automatization of the process. Hence the various machine learning techniques are introduced to this field.

In this study we are researching the field of automatic genre classification for Russian. The importance of such classification is immense, especially for web corpora, where texts are collected by a web crawler, and the main purpose of the texts is not always clear.

Accurate genre classification can ease a user's navigation through the corpus, and allow scientists to research the difference in language use in various language subclasses.

Another problem that occurs when working with big web corpora is choosing the suitable genre classification system. Here, we are looking for the system that would incorporate in itself the balance between distinguishing ability and an adequate amount of genre labels. From the perspective of theoretical text typology, which tries to cover all the possible text variations, the number of genres in a language is extremely high. For example, [Gorlach, 2004] lists around 2,100 genres for English and [Adamzik, 1995] differs over 4,000 genres for German. Such classification systems satisfy the theoretical necessity of describing all types of texts, but are absolutely impractical. The classification system for corpus annotation needs to have a reasonable number of genre labels, in order to collect an adequate sub-corpus for each genre and to be convenient for the users [Sharoff, 2018]. The classification system for a web corpus should also reflect the diversity of Internet texts. Web texts have a strong tendency for hybridism between genres and new types of texts appear on the web all the time [Santini et al., 2010]. Ideally, our classification system should be able to consider and represent all of that information.

In our study we adopt the Functional Text Dimension (FTD) [Sharoff, 2018] approach for genre classification. We chose this classification framework because it has great coverage ability comparable with that of a long list of genre systems, while maintaining a relatively short list of genre labels. This result is achieved through the introduction of functional dimensions of text instead of discrete genre labels. In total there are 18 functional dimensions, which represent different language functions. In the majority of genre systems a text is believed to belong to only one specific genre, but in the FTD framework texts are described in each functional dimension independently. Thus, one text can score positive results in several dimensions simultaneously. The FTD classification system was designed to describe any text found on the web. Since each FTD represents a specific language function, we can use their combination to describe text hybridism and possible new genres, which can often be found in web texts.

To build a classifier we need a reliable manually annotated training corpus. In the FTD framework annotators are presented with a key question for each functional dimension. Depending on their answer, a text is classified as strongly (scored as 2), partially (scored as 1) or not at all (is a default score 0) belonging to the functional dimension.

2. Data

One part of our training data is a piece of the GICR corpus [Piperski et al., 2013], which was collected and annotated in terms of FTD in [Sharoff, 2018]. This corpus was later extended with around 500 new annotated texts from GICR by Serge Sharoff. In this study, we are using the resulting corpus as our training data.

The GICR corpus consists of texts from a variety of genres from the Russian web, such as blogs, news sites, social media etc. The annotated corpus was split on training and testing data. The split was approximately 90% of texts on training to 10% of texts on testing (see Table 1).

Table 1: Size and composition of the GICR dataset

Data set	Documents	Words
Training data	1,800	2,249,818
Testing data	140	163,923
Total	1,940	2,413,741

Manual corpus annotation is an essential, but extremely time-consuming task. Because it demands a great amount of human and time resources our dataset is limited. That is why some of the functional dimensions that naturally occur less frequently are not present in our dataset.

We are training our models to classify texts in 10 functional dimensions[1]: A1 argum, A4 fiction, A7 instruct, A8 news, A9 legal, A11 person, A12 commpuff, A14 research, A16 info, A17 eval (see Table 2)[2].

Table 2: Description of training dataset in terms of FTD

FTD	A1	A4	A7	A8	A9	A11	A12	A14	A16	A17
Size	0.14	0.05	0.04	0.25	0.05	0.12	0.17	0.1	0.12	0.09

Also due to the lack of necessary annotation in our dataset, we are not training our models to find texts with partial belonging to the functional dimension. Thus, all of the texts that score positive results are treated equally and marked with the FTD score 1.

[1] The numeration and labels are taken from [Sharoff, 2018].

[2] For complete description of FTDs and how they are defined see [Sharoff, 2018].

3. Experiments

In our research we conducted several experiments involving various features for classification and different hyperparameters of the classification model. The aim of these experiments was to find the most accurate model, which could be used to classify texts of the GICR corpus.

3.1. Classification method

For all models in this study we use the Support Vector Machine (SVM) method for classification. It is a very popular method that is used for various tasks in NLP, such as sentiment analysis [Mullen, Collier, 2004], language recognition [Campbell et al., 2004] and text classification [Sassano, 2003]. SVM also proved to be the best performing method for classification with a dataset similar to ours [Bulygin, Sharoff, 2018].

We conducted our research with SVM in the Python scikit-learn library [Pedregosa, 2011]. During our work we experimented with different kernels of SVM and also with case of letters in texts. The scikit-learn library has 4 build-in SVM kernels: rbf (radial basis function), poly (polynomial), sigmoid, and linear. The linear kernels outperform all the other kernels by a high margin. In this paper we only show the results for the models with linear kernel.

We also looked into the influence of the letter case in training data on the classification performance. The models with the letter case kept as in the original texts generally perform worse than models with letter case converted to lower case, with the only exception being in the functional dimension A12, which contains promotional texts. We attribute this exception to the fact that texts in A12 often contain phrases in upper ('screaming') case, which is a specific feature of this FTD. However, the model without case conversion is unreliable on the complete corpus and in this study we only present results for models with case converted to lower.

3.2. C-value

In the linear SVM model the parameter that is in charge of the strength of regularization is the C parameter. Using low values of C will cause the model to adjust to the majority of data, while using a higher value of C would make model put more attention into the correct classification of each data point [Guido, Muller, 2017]. In our research we build models with different C values. While the default C value is 1, we also looked at models with the C value equal to 10 and 100 (see **Table 3**). We also experimented with a C value less than 1, but the results of these models were much less accurate and they are not present in this paper.

3.3. Feature selection

Feature selection is a process in which a subset of features is selected from all features of training data. The best subset of features contains the least number of features that contribute most to the prediction model [Guyon, Elisseeff, 2003]. Feature selection allows to avoid the overfitting of the model, to reduce training time and

to simplify the model. In our study we implement the basic approach to feature selection. We apply document frequency and only use features that appear at least in 10% of texts of the training data. Models after feature selection have significantly less features than before. For example, a model with character 5-gram features has 235,492 features before selection, and 3,793 features after feature selection.

3.4. Features

In this study we chose character n-grams for features in our models. We made this decision because character n-grams are very useful for text classification [Zhang et al. 2015] and also character n-grams show the best performance in research with training datasets similar to ours [Bulygin, Sharoff, 2018], [Sharoff, 2010].

One of the properties of character n-grams as features is that they can contain not only lexical information about a text, but also morphological, which helps the model to perform better. For our experiment we built models with bigrams, trigrams, 4-grams and 5-grams as features (see **Table 3**). We used scikit-learn preprocessing tools for tokenization and vectorized features using tf-idf technique.

4. Evaluation

For each model we provide 2 metrics: precision and recall. The precision metric tells us how many of the classified documents were classified correctly, while the recall metric shows how many of the texts from the testing data were classified accurately. Only through combination of these metrics one can assess the overall performance of classification.

We named models according to features and to parameters that were set for that model. Thus, model named 'svm-5gr-C10-nfs' should be understood as the model that uses the SVM classification method and character 5-grams as classification features, with C parameter set to 10 and does not use feature selection methods.

Table 3: Evaluation of classification accuracy of models with various parameters

FTD / model	metric	A1	A4	A7	A8	A9	A11	A12	A14	A16	A17
svm-2gr-C1-nfs	precision	0.0	0.0	1.0	0.95	1.0	0.71	0.90	1.00	0.0	1.0
	recall	0.0	0.0	0.50	0.88	0.71	0.31	0.90	0.73	0.0	0.06
svm-3gr-C1-nfs	precision	0.79	0.75	1.0	0.93	1.0	0.71	1.0	0.92	0.75	1.0
	recall	0.50	1.0	0.50	0.86	0.86	0.31	0.88	0.92	**0.35**	0.56
svm-4gr-C1-nfs	precision	0.88	0.75	1.0	0.94	1.0	0.83	0.95	1.0	0.75	1.0
	recall	0.54	1.0	0.50	0.92	0.86	0.31	0.90	0.87	0.16	0.71
svm-5gr-C1-nfs	precision	0.83	1.0	1.0	**0.97**	1.0	0.83	1.0	0.92	0.78	1.0
	recall	0.45	0.67	0.50	0.86	0.86	0.31	0.88	0.92	**0.41**	0.56
svm-3gr-C10-nfs	precision	0.79	0.75	1.0	0.94	0.86	0.38	0.89	0.93	0.53	0.92
	recall	0.58	1.0	0.50	0.92	0.86	0.31	0.85	0.93	0.47	0.65

model \ FTD	metric	A1	A4	A7	A8	A9	A11	A12	A14	A16	A17
svm-5gr-C10-nfs	precision	0.91	0.60	1.0	0.98	0.86	0.46	1.0	1.0	0.80	0.92
	recall	**0.77**	1.00	0.50	0.90	0.86	0.38	0.90	0.93	0.42	0.71
svm-3gr-C100-nfs	precision	0.75	0.75	1.0	0.94	0.86	0.33	0.89	0.93	0.56	0.92
	recall	0.58	1.0	0.50	0.92	0.86	0.31	0.85	0.93	0.47	0.65
svm-5gr-C100-nfs	precision	0.86	0.60	1.0	0.98	0.86	0.46	1.0	1.0	0.80	0.92
	recall	0.73	1.0	0.50	0.90	0.86	0.38	0.90	0.93	0.42	0.71
svm-3gr-C10-fs	precision	0.67	0.75	1.0	0.93	1.0	0.50	0.87	**1.0**	0.53	0.92
	recall	0.55	1.0	**0.67**	0.86	0.86	0.31	0.81	**1.0**	0.47	0.69
svm-5gr-C10-fs	precision	0.68	0.50	1.0	0.95	1.0	0.36	0.88	0.86	0.53	**0.92**
	recall	0.59	1.0	0.67	0.88	0.86	0.25	0.88	**1.0**	**0.53**	0.69

The first four models in Table 3 are basic SVM models with no additional parameters that have different classification features. Out of those four models, the best performing feature appears to be character trigram features and character 5-gram features, where 5-grams slightly outperform trigrams. For the following experiments with SVM parameters we used both these features.

Next, we tested models with different C values. The results show that the increase of C value leads to a better recall score of the model, but lowers the precision. Therefore, the most optimal C value would be 10. Such models are the most balanced, and take into account both precision and recall metrics.

In the end, we implemented feature selection to our best performing models. The reduction of the features helped the model to increase recall score for A7, A14 and A16 functional dimensions. However, these models lost some precision points for some FTDs. In the following experiments with feature extraction and classification of GICR's segments we are going to use 'svm-5gr-C10-fs' model. It is one of our best models and feature selection makes 5-gram interpretation more efficient.

5. Analysis of features

Most of the classifiers used in Machine Learning are a so-called 'black box', because we do not know for sure how the parameters and weights were assigned for the model. However, some of the classifiers, including SVM, are able to show the most valuable features of the model. This can shed some light on how the fitting of the model is performed.

We collected the most valuable features of the 'svm-5gr-C10-fs' model for each of the FTD present in our training corpus. We also provide the most frequent words, where these features appear (see Table 4).

Table 4: The most significant classification features for each FTD

FTD	Features	Words
A1	' соци', 'оказа', ' прич', 'ителя', ' наро', ' нам ', ' бога', ' нет ', ' чем ', ' они '	социальной, оказания, причем, представителя, международного, богатства, показатели, доказательства, оказались, причин, заместителя, народа
A4	' ка , ', ' прос', ' , - ', ' , и ', ' а , и", ' его ', 'казал', ' глаз', ' , не", '. он '	человека, просто, его, сказал, глаза, века, казалось, показал, некоторые
A7	'нстру', ' доба', 'добав', 'форма', ' вас ', ' жела', ' поро', 'если ', ' если', ' запр'	конструкции, добавить, год, информации, желание, пород, если, запрос, инструментов, порой, запрещено
A8	'ября ', ' новы', 'сказа', 'сообщ', ' моск', ' сооб', 'явил ', 'заяви', 'аявил', ' заяв'	сентября, новых, сказал, московских, сообщения, заявил, октября, основы, появились
A9	'ветст', 'должн', 'етств', ' стат', ' зака', 'мать ', 'стать', ' рабо', ' или ', 'федер'	должны, соответствии, статьи, заказ, принимать, работы, ответственности, должностных, федерального, статус
A11	'много', ' лет ', ' свое', ' нас ', 'перед', 'лись ', ' . к ', 'стно ', ' меня', ' мне '	оказались, появились, остались, находились, известно, совместно
A12	'знако', ' азмер ', ' крас', 'овый ', ' для ', ' прод', 'наком', ' высо', ' сайт', 'ство '	признаков, красоты, красный, красивый, новый, знаком, продукции, высокой, сайте, количество, большинство
A14	' она ', ' зада', 'ений ', 'больш', ') . ', ' язык', ' расс', 'ости ', ' боль', ' . и '	задачи, языка, рассмотрения, больше, отношений, решений, изменений, расследования, деятельности
A16	' зако', 'начал', 'изма ', 'закон', ' . а ', 'нный ', ' . в ', ' прин', 'века ', ' жела'	закона, начала, механизма, данный, принять, человека, желание, организма, современный единственный
A17	' игра', ' хотя', 'хотя ', 'смотр', 'разу ', 'овски', ' мало', ' без ', 'стати', ' книг'	играть, рассмотрения, сразу, московский, кстати, книги, разумеется, банковские

As shown in Table 4, features that are the most significant for the FTD are appearing in words that are often associated with texts of this functional dimension. For example, 'заявил' in A8 and 'должностных' in A9.

6. Applying FTD classifier to the GICR corpus

The GICR corpus contains over 2 million documents with over 20 billion words. The corpus is split into several segments, based on the source of the document. We chose the 'svm-5gr-C10-fs' for the GICR classification, since 5-grams and the C value of 10 turned out to be the best overall performing classification options. We also chose to filter features, because it speeds up the classification and models with a reduced number of features are usually more reliable.

Before we applied our 'svm-5gr-C10-fs' model to the segments of the GICR corpus, we decided to test it on the raw dataset of the GICR texts from the 'livejournal' segment, since our model was only tested on our training corpus. For this experiment, we randomly picked 100 texts from 'livejournal' subcorpus and annotated them manually. Then we evaluated the 'svm-5gr-C10-fs' model on these texts. The overall

precision averaged around 75% and overall recall was 51%. For the two most represented functional dimensions in 'livejournal' segment A1 and A11 these metrics scored 75% precision, 47% recall for A1 and 83% precision, 62% recall for A11. We considered the performance of this model adequate, so we applied it for the classification of the GICR corpus (see Table 5).

Table 5: Classification of GICR segments in terms of FTD

FTD / segment	A1	A4	A7	A8	A9	A11	A12	A14	A16	A17
Livejournal.com	0.39	0.02	0.003	0.09	0.002	0.42	0.01	0.001	0.02	0.05
Blogs.mail.ru	0.52	0.02	0.01	0.004	0.0004	0.39	0.01	0.0003	0.006	0.04
magazines.russ.ru	0.16	0.33	0.0	0.003	0.01	0.24	0.0	0.06	0.14	0.05
News	0.04	0.0001	0.0002	0.92	0.003	0.003	0.002	0.002	0.02	0.004
Vk.com	0.71	0.04	0.003	0.01	0.001	0.19	0.03	0.0003	0.01	0.007
Total GICR	0.45	0.02	0.007	0.11	0.002	0.30	0.03	0.005	0.02	0.05

The classifier was able to mark most of the texts in each segment, though some of the texts were left unlabeled. These texts are not taken into account in Table 5.

In the 'livejournal' segment we see the dominance of A1 (argumentative blogs) and A11(personal stories) functional dimensions. Both of these FTDs are common for social networks, and it is not a surprise that they compose most of the 'livejournal' segment of GICR.

The 'blogs.mail.ru' segment is quite similar to the 'livejournal' segment, as it also has A1 and A11 FTDs comprising it. However, it is expectable, since both sites are platforms for blogs and, hence they have similar type of texts.

The 'magazines.russ' segment consists of various journals with different style of articles. This can be seen in our results. This segment is the most well-rounded, with no dimension being severely dominant.

The 'news' segment is composed of articles from ria.ru, lenta.ru and rosbalt.ru. The most represented dimension here is A8. A portion of the texts in this segment is from the A1 (argumentative blogs) FTD, which is common for news texts.

The 'vk' segment is a social network segment. It is also dominated by the A1 and A11 functional dimension.

An interesting question that comes up during the FTD research is how functional dimensions correlate with linguistic features. We used a script that extracts linguistic features on the A1 and A8 subsets. The script was adapted for Russian by Serge Sharoff from MultiDimensional Analysis [Biber, 1988].[3] The subsets were classified by our model from the 'livejournal' segment of GICR. In our experiment we looked at two features: the verbs in the present tense and in the past tense. The results show that verbs in the present tense are much more common in the A1 dimension than in the A8 dimension, with median values 0.03490829 and 0.02912898 respectively.

[3] https://github.com/ssharoff/biberpy

However, the verbs in the past tense are more frequent in the A8 dimension with median value 0.03572108 and much less frequent for the A1 dimension with median value 0.01967835. This opens the possibility to compare the use of language in argumentative opinion pieces vs reporting news. More research is still required.

7. Conclusion

In this paper we presented an experiment during which we tested several classification features and parameters to find the optimal classification options for the GICR dataset. The resulting model uses character 5-gram features, has C-value of 10 and uses the feature selection technique, where features are filtered by document frequency. This model was used for the annotation of the segments of the GICR corpus. Furthermore, we looked at the most significant features of our model for each FTD and compared these 5-grams t0 the most frequent words of the training corpus, in which these features can be found.

In further studies we would like to continue our experiments with GICR annotation. One of the possible lines of research is the correlation between text-internal linguistic features and text-external genre classification. The original idea comes from Douglas Biber's Multi-Dimensional analysis [Biber, 1986]. The MD analysis was also implemented for the English web texts in [Biber, Egbert, 2016]. Similar research was conducted for Russian in [Katinskaya, Sharoff, 2015], where the researchers used FTD classification and compared it to the MD analysis. This study showed very promising results and we would like to apply this knowledge to the GICR corpus.

Another interesting research area concerns related text classification tasks. We have not experienced considerable issues with detecting spam, most of it was classified as A12, Promotion. However, it'd be very interesting to investigate deviations from the prototypes (such as a newspaper report) as caused by spam in the social networks.

References

1. *Adamzik K.* (1995), Textsorten—Texttypologie, Eine kommentierte Bibliographie, Nodus, Münster.
2. *Biber D.* (1986) Spoken and written textual dimensions in English: resolving the contradictory findings. Language, Vol. 62, pp. 384–414.
3. *Biber D., Egbert J.* (2016) Register Variation on the Searchable Web: A Multi-Dimensional Analysis. Journal of English Linguistics, 44(2), pp. 95–137.
4. *Bulygin M., Sharoff S.* (2018) Using Machine Translation for Automatic Genre Classification in Arabic. In Proc Dialogue, Russian International Conference on Computational Linguistics.
5. *Campbell W., Singer E., Torres-Carrasquillo P., Reynolds D.* (2004) Language recognition with support vector machines. In Proc. ODYS, pp. 41–44.
6. *Görlach M.* (2004), Text types and the history of English, Walter de Gruyter.
7. *Guido S., Muller C.* (2017), Introduction to Machine Learning with Python, O'Reilly Media, Inc, pp. 57–58.

8. *Guyon I., Elisseeff A.* (2003), An introduction to variable and feature selection, J. Mach. Learn, pp. 1157–1182.
9. *Katinskaya A., Sharoff S.* (2015), Applying Multi-Dimensional Analysis to a Russian Webcorpus: Searching for Evidence of Genres. In Proc BSNLP, Sofia.
10. *Mullen T., Collier N.* (2004) Sentiment analysis using support vector machines with diverse information sources, in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 412–418.
11. *Piperski A., Belikov V., Kopylov N., Selegey V., and Sharoff S.* (2013), Big and diverse is beautiful: A large corpus of Russian to study linguistic variation, In Proc 8th Web as Corpus Workshop (WAC-8).
12. *Santini M., Mehler A., Sharoff S.* (2010), Riding the rough waves of genre on the web, Genres on the Web: Computational Models and Empirical Studies, Springer, Berlin/New York.
13. *Sassano M.* (2003), Virtual examples for text classification with support vector machines, In Proceedings of Empirical Methods in Natural Language Processing, pp. 208–215.
14. *Sharoff S., Wu Z., Markert K.* (2010) The Web Library of Babel: evaluating genre collections. In Proc Seventh Language Resources and Evaluation Conference, LREC, Malta.
15. *Sharoff S.* (2018), Functional text dimensions for annotation of web corpora, Corpora.
16. *Zhang X., Zhao J., LeCun Y.* (2015), Character-level convolutional networks for text classification, In Advances in Neural Information Processing Systems, pp. 649–657.

A SIMPLE FINGERPRINT APPROACH TO EXTRACTING THE GLOBAL PROSODIC PROPERTIES FROM FIELD DATA[1]

Chechuro I. Yu. (ilyachechuro@gmail.com)[1],
Lyashevskaya O. N. (olesar@yandex.ru)[1, 2]

[1]National Research University Higher School of Economics, Moscow, Russia;
[2]Vinogradov Institute of the Russian Language RAS, Moscow, Russia

The paper reports a method to create a speaker's prosodic fingerprint based on the global characteristics of the pitch movement. Prosodic fingerprint is the distribution of f0 in the low, middle, and high ranges and the distribution of pitch movements from one range into other [Šimko et al. 2017]. This fully automated method can be used to classify the records and to provide the reference level for more sophisticated analysis of the pitch movement and intonation strategies. We evaluate the method by applying it to the spontaneous Russian spoken data recorded in different regions. We model the correlation between the fingerprint and sociolinguistic features such as age, gender, and region. The results of this analysis allow to formulate several sociolinguistic hypotheses that can further be tested with a more detailed analytic technique.

Key words: prosodic fingerprint, speaker's prosodic portrait, pitch movement, unigram fingerprint, delta fingerprint, Russian prosody

[1] The article was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) and supported within the framework of a subsidy by the Russian Academic Excellence Project '5-100'.

ПРОСТОЙ МЕТОД ПРОСОДИЧЕСКОГО ОТПЕЧАТКА ДЛЯ АНАЛИЗА ОБЩИХ СВОЙСТВ ИНТОНАЦИИ НА МАТЕРИАЛЕ ПОЛЕВЫХ ДАННЫХ

Чечуро И. Ю. (ilyachechuro@gmail.com)[1],
Ляшевская О. Н. (olesar@yandex.ru)[1, 2]

[1]Национальный исследовательский униветситет «Высшая школа экономики», Москва, Россия; [2]Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

В статье обсуждается применение метода создания просодического отпечатка говорящего на основе общих характеристик движения основного тона. Просодический отпечаток—это распределение f0 в нижнем, среднем и верхнем диапазонах и распределение движений высоты тона из одного диапазона в другой [Šimko et al. 2017]. Этот полностью автоматизированный метод может использоваться для классификации записей в корпусе и получения представления о фоне, с которым будут сравниваться данные при дальнейшем, более сложном анализе стратегий интонирования. Мы применили метод к спонтанным русскоязычным данным, записанным в разных регионах. Разработаны модели анализа зависимости между данными просодического отпечатка и социолингвистическими характеристиками, такими как возраст, пол и регион. Результаты проведенного нами анализа данных позволяют сформулировать ряд социолингвистических гипотез, которые впоследствии могут быть проверены с использованием более глубоких методов анализа.

Ключевые слова: просодический отпечаток, просодический портрет говорящего, движение частоты основного тона, русская просодия

1. Introduction

The notion of the speaker's prosodic portrait [Kibrik 2009], [Kibrik, Fedorova 2018] is important as it assumes the existence of certain intrinsic, or neutral, properties of a particular speaker's voice. [Kibrik and Fedorova 2018] includes in the representation of the portrait features such as the f0 range (minimal and maximal f0 values), the standard level of the elementary discourse unit onsets, the level of fallings in the final and non-final positions, and the level of rises (and fallings on post-accent syllables) characteristic of the comma and "three dots" intonation. More detailed representations may contain hundreds of features. [Hönig et al. 2015] create fingerprints of recordings rather than speakers and include as many as 167 features. However, all such representations rely heavily on the manual (or partially automated) annotation of the underlying linguistic and perceptive information.

In this study, we propose a simple automated approach to creating primary speaker's prosodic fingerprint based on the global characteristics of the pitch movement. Prosodic fingerprint is the distribution of f0 in the low, middle, and high ranges and the distribution of pitch movements from one range into other [Šimko et al. 2017]. Using the pitch data extracted with the Praat software, we create fingerprints of 26 Russian speakers from different regions of Russia. We then conduct a case study and illustrate how this technique can be used to help formulate primary hypotheses regarding the relations between pitch movement and the sociolinguistic features of speakers such as age, gender, and region. This study is a part of a larger research project dealing with the regional aspects of the Russian intonation previously, in which data from different regions of Russia and ex-USSR are collected and documented.

Our study is complementary to the qualitative studies of Russian intonation that suggest the classifications of intonation patterns and their functional interpretations such as [Bryzgunova 1977], [Odé 1989], [Kodzasov & Krivnova 2001], [Korotaev & Podlesskaya 2008], [Kodzasov 2009, Grammatchikova et al. 2014], [Vol'skaya 2014], [Podlesskaya 2017], [Yanko 2017], [Korotaev 2018]. We use a quantitative approach, which makes our study more in line with the studies of [Skrelin and Volskaya 2006, 2008], and those based on the corpus of One Speaker's Day [Stepanova et al. 2008] and CORUSS [Kachkovskaia et al. 2016]) that aim at modeling the speakers' behaviour rather than propose overall generalizations. The study reported here is an attempt to automatically investigate the structure of the field data in order to formulate primary research hypotheses based on the observed phenomena.

Using the recordings made in four different regions of Russia (Krasnoyarsk, Moscow, Nakhodka and Novosibirsk), we analyze the pitch movement in spontaneous speech of the native speakers of regional Standard Russian. For each speaker, we recorded three samples of spontaneous speech: an interview, a dialogue and a retell of the Pear movie [Chafe 1980]. With the help of linear mixed effect modelling, we explore the correlation between the shape of the fingerprints and the biological sex, age and place of residence of the speakers.

The remainder of this paper is organized as follows. Section 2 discusses the experimental settings, and Section 3 addresses the data sampling. Section 4 presents the method to collecting the fingerprints based on the distribution of the pitch values. Section 5 reports the statistical analysis of the data, and in Section 6, a discussion of the obtained results is provided.

2. Participants and Experimental Conditions

All participants are monolingual native speakers of Russian born in Krasnoyarsk, Novosibirsk, Nakhodka, and Moscow. Krasnoyarsk and Novosibirsk represent Siberia, Moscow—Standard Russian and Nakhodka—Far East (the city population of which usually originates from different regions of the ex-USSR and is highly mixed). At the moment of the recordings, all the participants lived in their home regions or have recently moved to Moscow to study at the university (1st year students in the beginning of the 1st semester). All regional participants were divided into two age groups: from 25 to 40 years old vs. 45 years old and older. This division was made in order

to balance the sample; in the analysis presented in this paper age was used as a numeric and not as a categorical variable. In each age group, there were two male and two female participants. The speakers from Moscow were represented by two females from the lower age group.

Each recording has been taken from two participants. In all pairs, the interlocutors knew each other relatively well (they were classmates, friends or relatives) and belonged to the same age and social group.

The spontaneous dialogues were recorded in the "fieldtrip" conditions in a quiet room using a recorder that supports .WAV format with no compression. The recordings made in Moscow, including those with the regional respondents, were made with a professional recorder and individual headset microphones for each speaker.

The experiment began with setting up the recording devices and instructing the participants. This stage took from 5 to 10 minutes. During this time, the participants could talk to each other freely and simultaneously get used to the recording equipment and the experimental environment.

2.1. Tasks for the Participants

There were three types of tasks. In the first task, the participants had to tell a small story about their life (e.g. parents and family, school, favorite teachers, hometown, etc.). The second task was an experiment with a map based on [Usacheva 2017], in which two participants, the instructor and the follower, were given a map of the Moscow Zoo printed on an A2 sheet and a set of objects (coins, pencils, dices, etc.) to place on the map. The experimenter placed objects on the Zoo map in front of the instructor and the instructor had to explain the positions of these objects to the follower so that the follower could repeat it on his map. During the experiment, the speakers communicated using mobile phones. The third part of the experiment implied retelling the Pear Movie [Chafe 1980] that was presented to the participants on the screen of the experimenter's laptop.

3. Data Sampling and Annotation

Each speaker in the dataset was represented by 40 randomly selected utterances. The utterances were extracted from each recording type using the following proportion: 15 recordings from the interview, 15 recordings from the experiment, and ten recordings from the pear story. The length of the recordings was not normalized. The pitch values were extracted from each recording with a 10 ms step. The pitch values were extracted with the standard functions of Praat. The pitch range (maximal and minimal pitch values) was defined for each speaker separately.

The data have been annotated in Praat. The first tier contained the boundaries of the speech units defined by pauses on the oscillogram. The parts of the recordings that contained sounds other than the participant's speech (experimenters' instructions, random noises) or were technically problematic to analyze (e.g. distortion or low volume units) were marked on a separate tier. These parts of the recordings were not used in the analysis.

4. Constructing the fingerprints

The analysis of the data presented in this section partly adopts the approach introduced in [Šimko et al. 2017]. Smoothing and wavelet transformations were omitted in the analysis. For each of the speakers, the pitch range was defined as the difference between the minimal and the maximal pitch values in all 40 recordings, with the exclusion of 5% of the observations: 2.5% with the minimal values and 2.5% with the maximal values [Fig. 1]. This type of range narrowing lowers the probability of including octave jumps and other artefacts into the analysis. Then, the pitch values were normalized by the z-score. The remaining range (95% observations) was divided into three equal parts that were coded, respectively, with −1 (Low), 0 (Medium) и 1 (High), which correspond to the commonly used division of the pitch range into Low, Medium and High [Keijsper 2003], [Odé 1989].

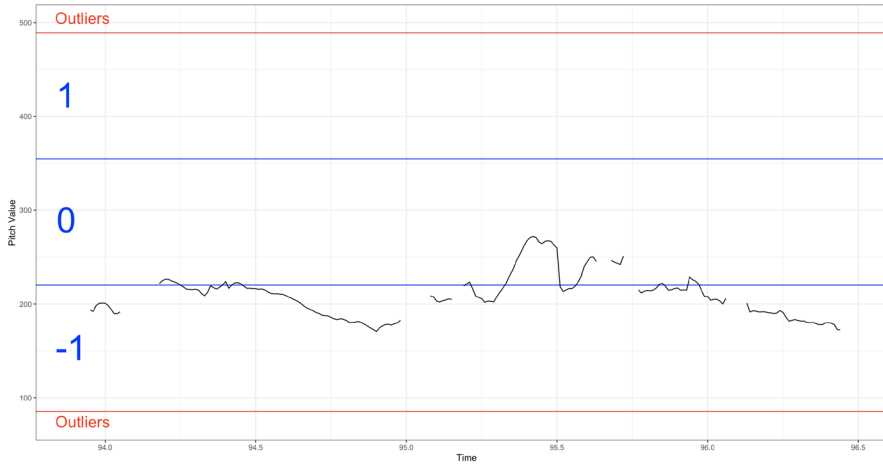


Fig. 1. Pitch sub-ranges in a recording sample

At the second step of annotation the transitions between the −1, 0 and 1 levels were encoded. The data were annotated as follows: if the points N and N+1 (taken with a 10 ms interval) are in the same pitch level, we interpret this as no-change in pitch shape and code it as 0. If the points are in different sub-ranges, the transition is coded as the difference between the levels: −2 (High to Low), −1 (Medium to Low, High to Medium), +1 (Low to Medium, Medium to High) или +2 (Low to High). This annotation was designed in order to distinguish substantial pitch movements from its minor fluctuations within a single sub-range.

The two types of the annotation were used to compose four datasets: the distribution of observations by the sub-ranges, the transitions between the sub-ranges, and the number of observations in each sub-range and the transitions of each type.

Each line of the dataset corresponded to one pitch value and contained the information about the speaker's name, their place of living, biological sex, age, text type, sentence ID from 1 to 40, time on the recording the observation corresponds to, the sub-range value −1, 0 or 1 (further called unigram) or the transition value −2, −1, 0, 1, 2 (further called delta). Upon this dataset, we created a new one, where each line

corresponded to a sentence and the rows contained the meta information about the text and the number of unigrams or deltas of each type in this sentence.

Due to the size of the datasets of the first type (the size of each dataset in the .csv format was over 300 megabytes) and the limited resources of the personal computer used for statistical modelling, the data of the first dataset were not used in the current study. Nevertheless, we plan to use these data for statistical modelling using specifically designed systems with a better performance. The analysis presented in this paper was conducted using only the second type of datasets.

5. Data Analysis

The preliminary analysis of the data was conducted using histograms of the unigrams and deltas values per speaker. Figure 2 illustrates the distribution of unigrams per speaker, the histograms are colored by the region:

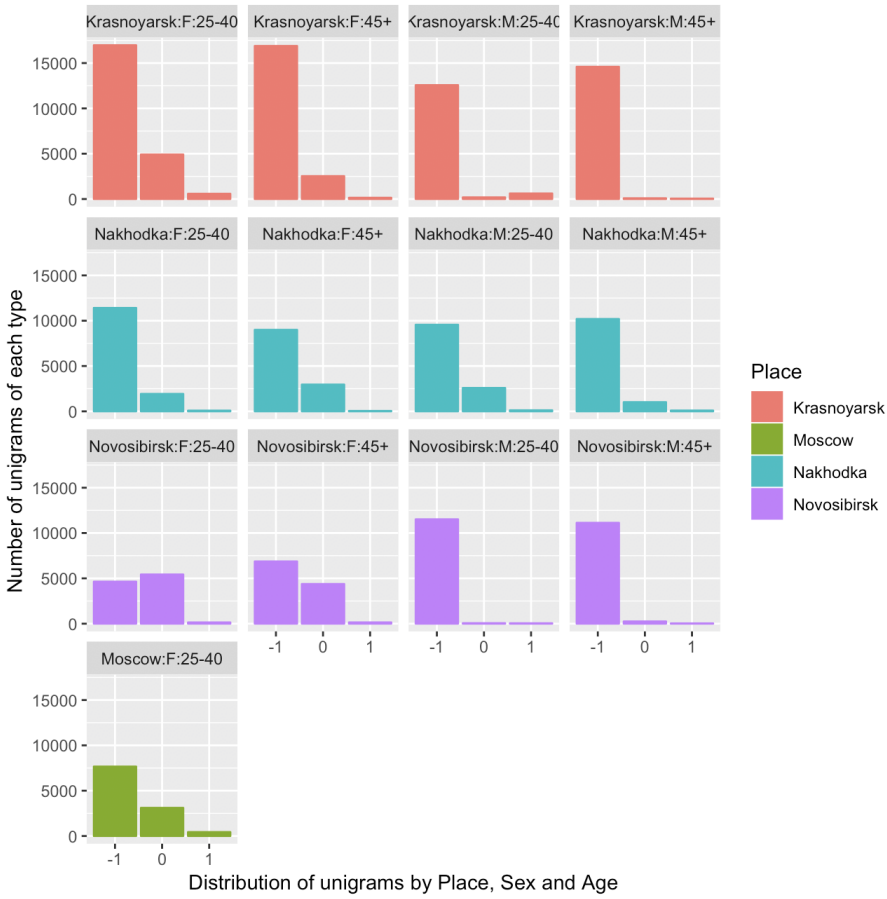


Fig. 2. The Distribution of the z-scored Unigram Values by Place, Sex and Age

The shape of the histograms in Figure 2 suggests that the difference between male and female respondents may play a significant role in Novosibirsk and Krasnoyarsk with women having a greater proportion of "0" unigrams, while in Nakhodka this difference is less pronounced and both sexes are similar. The histograms also indicate that the number of "1" unigrams may not be of much significance and the opposition can be viewed as "–1" vs. "not –1" values.

Figure 3 illustrates the distribution of deltas by speaker. The bar for "0" deltas is intentionally omitted since its relative size did not allow to observe the "1" and "–1" in men and women (the range for this bar is roughly between 10,000 and 20,000). The shape of the histograms suggests that the main difference between male and female respondents is in how often the "–1" and "1" deltas occur in their recordings. Again, in Nakhodka this difference is less pronounced than in other regions. Apart from three speakers, the number of –2 and 2 deltas is imperceptible and the main opposition appears to exist between 0 vs. –1 and 1 deltas. The data can thus be coded as "0" vs. "not 0" deltas indicating the presence and the absence of pitch movement.

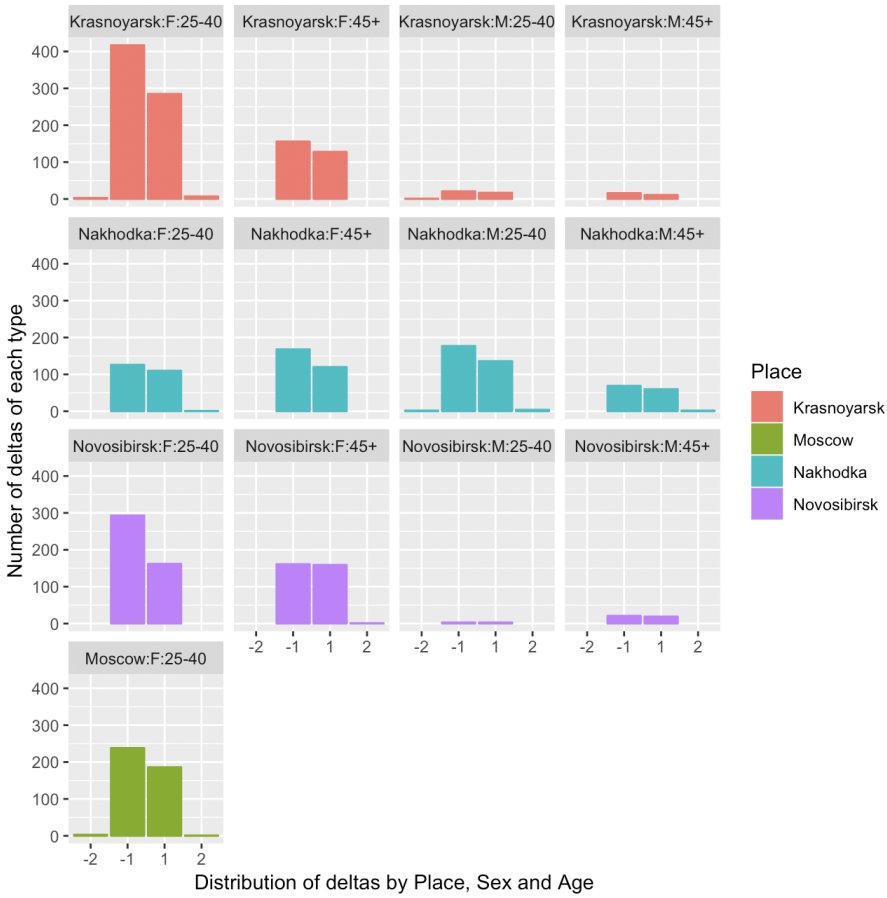


Fig. 3. The Distribution of the z-scored delta Values by Place, Sex and Age

In the following analysis of the data we use linear mixed effect modeling to explore the effects of biological sex, age, place of origin and type of text (dialogue vs. monologue) on the distribution of unigrams and deltas. We fitted the following models. The first model predicted the proportion of "–1" to "not –1" unigrams on the basis of biological sex, age, place of origin of the speakers, the type of the text and the speaker identity as a random effect lmer(X1Prop ~ Sex*Age + Place + TextType + (1 | Speaker_ID)). The second model predicted the same value on the basis of sex, age and the type of the text with the place of origin as a random effect lmer(X1Prop ~ Sex*Age + TextType + (1 | Place)). Then, we fitted two models with the same predictors for the proportion of "0" deltas to the "not 0" deltas: lmer(X0Prop ~ Sex*Age + Place + TextType + (1 | Speaker_ID)) and lmer(X0Prop ~ Sex*Age + TextType + (1 | Place)).

The predicted value in the first model was the proportion of "–1" unigrams to the "not –1" values. The controlled variables were biological sex, age, place of origin and type of text (dialogue vs. monologue) and the speaker ID as a random effect. The stepwise regression model selection with backward elimination has shown that the only significant variable are sex (p-value = 0.000244) and text type (p-value = 0.009803) with random intercepts by speakers. The effect of the two variables is provided in Fig. 4.

Figure 4 shows that the proportion of "–1" unigrams is significantly lower in females than in males. Similarly, the same proportion to a smaller degree is observed with respect to the text type. The factors of place and age turned out to be insignificant.

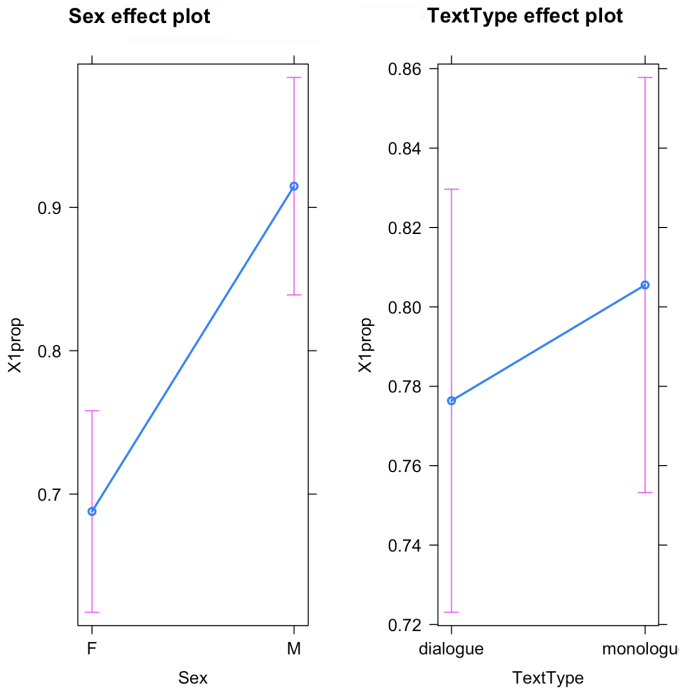


Fig. 4. The effect of biological sex and text type on the proportion of "–1" unigrams to "not –1" unigrams

The second model was delta-based and predicted the proportion of "0" deltas to the "not 0" deltas. The set of predictors was the same as in the first model. The backward model selection has shown that the only significant predictors is biological sex. Figure 5 illustrates that the amount of pitch movement in male participants is significantly lower than that in female participants.

The unigram-based model we fitted next was similar to the one described above with the only change being made to the random effect structure: instead of fitting random intercepts for speakers, we used random intercepts for different places. The backward stepwise regression model validation has shown that the significant predictors are sex, age and text type with the age-related change in pitch use being significant in both sexes.

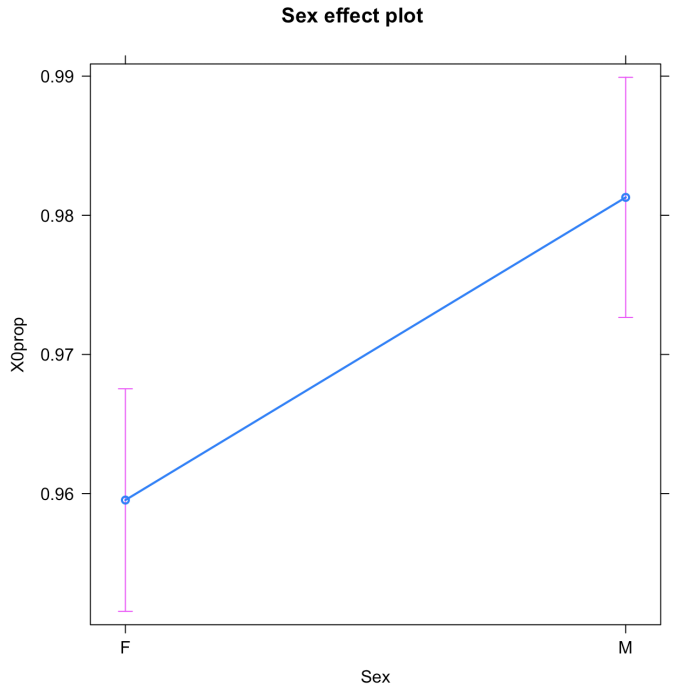


Fig. 5. The effects of biological sex and age on the proportion of "0" and "not 0" deltas

Similar changes were made to the delta-based model. The significant effects turned out to be the same as in the unigram-based model. Figure 6 and Figure 7 illustrate the effect of age with respect to biological sex. The effect of age is significant in both genders and is more pronounced in women (the older female respondents use less pitch movement than the younger ones). The models thus suggest that though there is no global effect of age as suggested by the first two models, it does exist within each city separately.

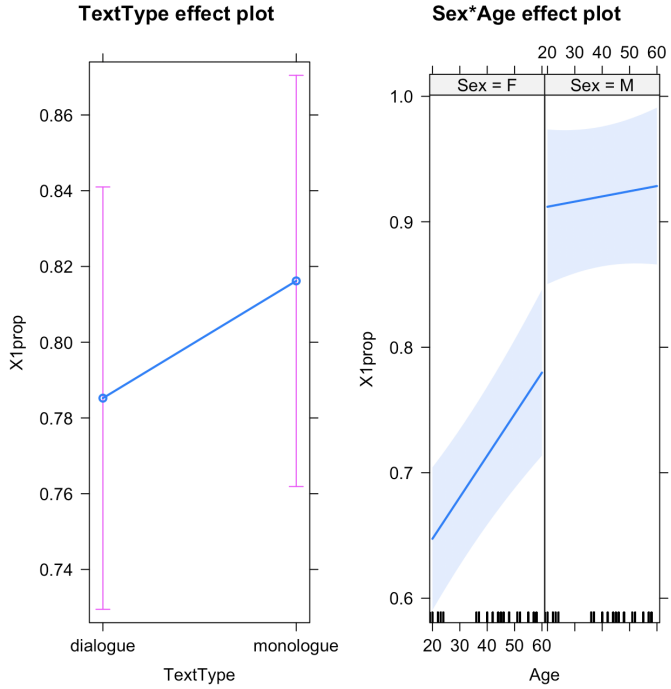


Fig. 6. Effect of age by sex and text type on the proportion of "–1" unigrams

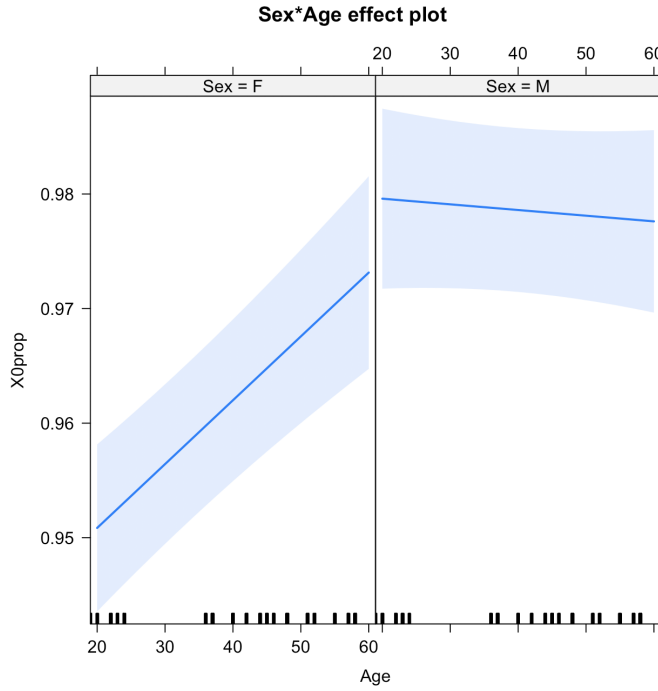


Fig. 7. Effect of age by sex on the proportion of "0" deltas

The last pair of models that we have fitted predicted the proportion of −1 unigrams and 0 deltas on the basis of the interaction between sex, age and place: *lm(formula = X0prop ~ Sex:Age:Place)* and *lm(formula = X1prop ~ Sex:Age:Place)*. The data for Moscow were removed from the dataframe since they only correspond to one age group and one biological sex.

Both models suggest that there is an age-related change in men in all regions (p-values < 0.01 in both models for all regions), while for women it is attested only in Krasnoyarsk (for "0" deltas, p-value = 0.007) and in Novosibirsk (for "−1" unigrams, p-value = 0.002). Figure 8 and Figure 9 illustrate the effect of age and sex on the proportion of −1 unigrams and 0 deltas.

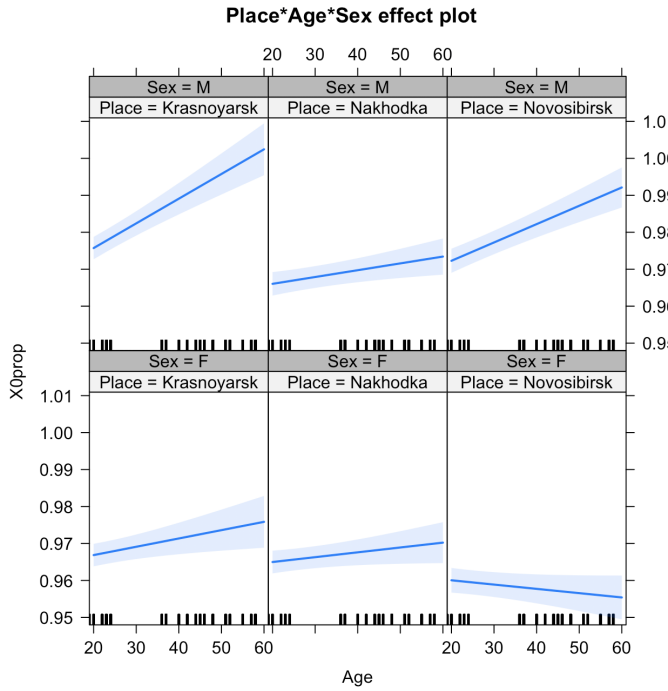


Fig. 8. Effect of age and sex by region on the proportion of "−1" unigrams

The last two models also allow to hypothesize that there is a major difference between the Siberian cities and Nakhodka: while in Novosibirsk and Krasnoyarsk the differences between speakers of different sexes are relatively clear, in Nakhodka men and women appear to use pitch more similarly. Another possible interpretation of this result may be that the biological sex in Nakhodka does play a role but our current annotation system does not track these differences.

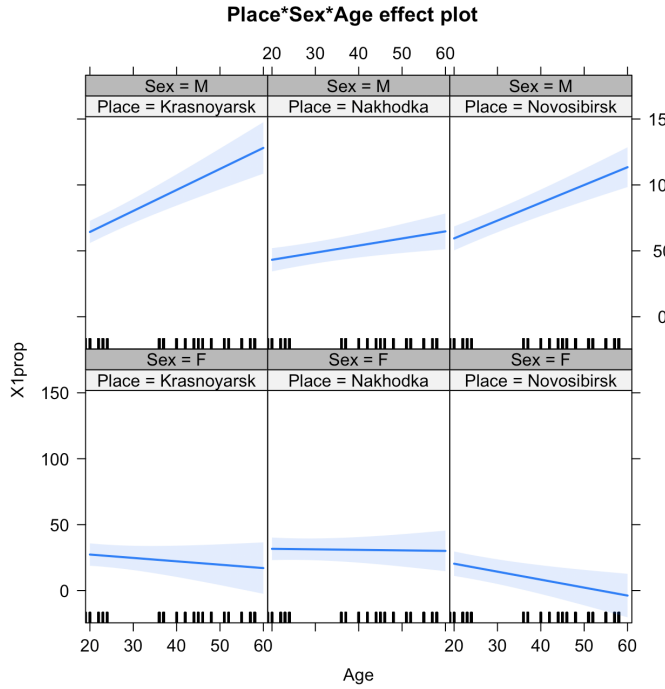


Fig. 9. Effect of age and sex by region on the proportion of "0" deltas

6. Discussion

The results of the regression analysis can be interpreted as follows. First, the first unigram-based regression model has shown that male and female speakers use the available pitch range differently. While males mostly use the "Low" part of the range, women use other parts of the range more often. Linguistically, this means that men may use less pitch movement in their speech, which may be related to a more expressive function of pitch in male speech (significant pitch changes are rare and therefore more noticeable). The first delta-based model has shown that male speakers cross the sub-range boundaries significantly less often than females, which supports the hypothesis of the comparably lower pitch use in their speech.

Another possible explanation of these results is that men divide their pitch range differently than women and our version of the tripartite division of the pitch range is not sensitive enough to track the pitch changes. The pitch movement of males may occur within a single sub-range (e.g. within "–1") and the remaining part of the range will be reserved for the rare utterances with an extreme degree of expression. There are thus two possible scenarios: (a) males use pitch movement more rarely than females and (b) the pitch movement in males has a lower amplitude in males but it is not necessarily less frequent. Both interpretations, however, suggest that there is a major difference between male and female speech and only differ in the nature of these

differences, which means that the use of the pitch range that can be tracked automatically using a relatively simple technique. These results are interesting in the sense that they contradict the findings reported in the previous studies [cf. Skrelin, Volskaya 2006, 2008], where no gender-related differences have been reported. The difference between genders that is observed in our study may most likely be explained by the different choice of measured parameters. While in the previous studies the models measured and predicted the proportions of pitch curves of different types in different speakers, we model the overall amount of pitch movement regardless of the particular contours. The second reason may be that the previous studies did not consider regional variation, which appears to have an impact on the prosodic portraits of male and female speakers. In our study, gender-related differences were observed in Novosibirsk and Krasnoyarsk, while in Nakhodka they were not attested. This hypothesis, however, requires additional testing with the use of a larger data sample from each region.

Interestingly, the models with place as a random effect suggest that there is an age-related difference in male speakers with older speakers having a different amount pitch movement. Thus, though the age-related differences are not seen globally, they exist within each city. From the linguistic point of view, this means that older men use pitch differently from the younger ones but there is no such tendency in women. It may also mean that the parts of the pitch range get re-organised with the increase age and the available range starts to be used differently.

The regional difference between Nakhodka and other cities tracked by the last two models allows to hypothesize the existence of an areal comparative concept, namely of the difference between men and women in the intensity of the pitch use. This means that different regions of Russia may differ with respect to whether men intonate somehow differently than women or not. Another possible interpretation of this result may be that the biological sex in Nakhodka plays a role but our current annotation system does not track these differences. Both results, however, suggest that the regions of Russia differ with respect to how men and women use pitch.

7. Conclusion

In this paper, we proposed a simple fully automated method of portraying a speaker's pitch usage. We created fingerprints of 26 speakers from different regions of Russia and conducted statistical analysis of these data. The regression analysis has shown that even though our representation of data is very simplistic, it may reveal some significant correlations with the sociolinguistic features. The results of this analysis allow one to formulate several sociolinguistic hypotheses that can further be tested with a more detailed analytic technique and a larger data sample. The first hypothesis regards the differences in the use of the available pitch range in male and female speakers and the frequency of pitch change in the speakers of different biological sex. The second hypothesis is related to the age-related differences in the pitch use in male speakers. Finally, the third hypothesis concerns differences in pitch use between men and women across different regions.

8. Supplementary Materials

Data used in the analysis and source code for the R scripts are available at: https://github.com/author/screenedrepository.

References

1. *Bryzgunova E. A.* (1977), Sounds and intonation of Russian speech [Zvuki i intonatsiia russkoi rechi]. Moscow.
2. *Chafe W. L.* (1980), The pear stories: Cognitive, cultural, and linguistic aspects of narrative production, Norwood, NJ: Ablex.
3. *Féry, C.* (2017), Intonation and Prosodic Structure. CUP.
4. *Grammatchikova E. V., Knyazev S. V., Luk'yanova L. V., Pozharitskaya S. K.* (2014), The rhythmic structure of the word and the place of realization of the tonal accent in regional variants of the modern standard Russian [Ritmicheskaya struktura slova i mesto realizatsii tonal'nogo aktsenta v regional'nykh variantakh sovremennogo russkogo literaturnogo yazyka], in Issues of theoretical and applied phonetics. Collection of articles in honor of O. F. Kryvnova [Aktual'nyje voprosy teoreticheskoj i prikladnoj fonetiki. Sbornik statej k jubileju O. F. Krivnovoy], Moscow, Buki Vedi.
5. *Hönig F., Batliner A., Nöth A.* (2015), How many speakers, how many texts—the automatic assessment of non-native prosody. In Proceedings of SLaTE 2015, Leipzig, September 4–5, 2015.
6. *Keijsper, C. E.* (2003), Notes on intonation and voice in modern Russian. Studies in Slavic and General Linguistics, 30, 141–214.
7. *Kachkovskaia T., Kocharov D., Skrelin P. A., Volskaya N. B.* (2016), CoRuSS-a New Prosodically Annotated Corpus of Russian Spontaneous Speech, Proceedings of LREC-2016.
8. *Kibrik A. A.* (2009), A speaker's prosodic portrait, Kibrik A. A., Podlesskaja V. I. (eds.), Night Dream Stories: A corpus study of spoken Russian discourse. Moscow: Jazyki slavjanskikh kul'tur.
9. *Kibrik A., Fedorova O.* (2018), A «Portrait» Approach to Multichannel Discourse, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018), Miyazaki, Japan.
10. *Kodzasov S. V.* (2009). Studies in Russian prosody [Issledovanija v oblasti russkoj prosodii]. Moscow: Jazyki slavjanskikh kul'tur.
11. *Kodzasov S. V., Krivnova O. F.* (2001) General Phonetics [Obschaya fonetika]. Moscow.
12. *Korotaev N. A.* (2018), How intonation structures spoken narratives: non-final phase contexts [Intonatsionnaja struktura ustnogo rasskaza v kontekste nezavershennosti], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog-2018", Issue 17 (24), Moscow.

13. *Korotaev N. A., Podlesskaya V. I.* (2008), Prosody of clause-combining in Russian: A corpus-based study [Frazovaja akcentuacija v složnyx predloženijax s postpozitivnym pridatočnym v russkom jazyke: opyt ispol'zovanija ustnogo korpusa s prosodičeskoj razmetkoj], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2008", Issue 7 (14), Moscow, pp. 234–240.
14. *Odé, C.* (1989), Russian intonation: a perceptual description, Series: Studies in Slavic and General Linguistics, Vol. 13, Amsterdam–Atlanta, Rodopi.
15. *Podlesskaya V. I.* (2017), "Ja skazhu tebe s poslednej prjamotoj": Direct and indirect speech viewed through the prism of prosodically annotated corpus data [«Ja skažu tebe s poslednej prjamotoj»: prjamaja i kosvennaja reč' po dannym korpusa s prosodičeskoj razmetkoj], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2017", Issue 16 (23), Vol. 2, Moscow, pp. 355–371.
16. *Šimko J., Suni A., Hiovain K., Vainio M.* (2017), Comparing languages using hierarchical prosodic analysis, Proceedings of Interspeech-2017, Stockholm, Sweden, Interspeech, ISCA, Baixas, pp. 1213–1217.
17. *Skrelin P., Volskaya, N.* (2006), Russian read and spontaneous speech: prosodic data analysis, Proceedings of Nordic Prosody 2006.
18. *Skrelin P., Volskaya N.* (2008), Prosodic model for Russian, Proceedings of Nordic Prosody 2008.
19. *Stepanova S. B., Asinovsky A. S., Bogdanova N. V., Rusakova M. V., Sherstinova T. Ju.* (2008), Speech corpus of the Russian everyday communication "One Speaker's Day": basic conception and current state [Zvukovoj korpus russkogo jazyka povsednevnogo obschenija «Odin rechevoj den'»: kontseptsija i sostojanie formirovaniya], Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue", Issue 7 (14).
20. *Usacheva M.* (2017), Experiments in the field: the experience of describing Permian languages [Eksperimenty v pole: opyt opisaniya permskikh yazykov]. A talk given at the workshop on Minor languages in the big linguistics [Malyje jazyki v bol'shoj lingvistike], Moscow, Russia, 2–3 November 2017.
21. *Vol'skaya N. B.* (2014), Intonation and language contact: pragmatic aspects of within- and cross-language interference [Intonatsija i jazykovoj kontakt: pragmaticheskij aspekt vnutri- i mezhjazykovoj interferentsii], Proceedings of XLII Int. philological conference, Selected papers, St.-Petersburg.
22. *Yanko T. E.* (2008), Intonation strategies of Russian speech in comparative aspect [Intonatsionnyje strategii russkoj rechi v sopostavitel'nom aspekte], Moscow, Jazyki slavjanskikh kul'tur.

CLASSIFICATION MODELS FOR RST DISCOURSE PARSING OF TEXTS IN RUSSIAN

Chistova E. V. (chistova@isa.ru)
FRC CSC RAS, Moscow, Russia;
RUDN University, Moscow, Russia

Shelmanov A. O. (shelmanov@isa.ru)
Skoltech, Moscow, Russia,
FRC CSC RAS, Moscow, Russia

Kobozeva M. V. (kobozeva@isa.ru),
Pisarevskaya D. B. (dinabpr@gmail.com),
Smirnov I. V. (ivs@isa.ru)
FRC CSC RAS, Moscow, Russia

Toldova S. Yu. (toldova@yandex.ru)
NRU Higher School of Economics, Moscow, Russia

The paper considers the task of automatic discourse parsing of texts in Russian. Discourse parsing is a well-known approach to capturing text semantics across boundaries of single sentences. Discourse annotation was found to be useful for various tasks including summarization, sentiment analysis, question-answering. Recently, the release of manually annotated RuRSTreebank corpus unlocked the possibility of leveraging supervised machine learning techniques for creating such parsers for Russian language. The corpus provides the discourse annotation in a widely adopted formalisation—Rhetorical Structure Theory. In this work, we develop feature sets for rhetorical relation classification in Russian-language texts, investigate importance of various types of features, and report results of the first experimental evaluation of machine learning models trained on Ru-RSTreebank corpus. We consider various machine learning methods including gradient boosting, neural network, and ensembling of several models by soft voting.

Key words: RST, word embedding, discourse parsing, machine learning on annotated corpus, feature selection

КЛАССИФИКАЦИЯ РИТОРИЧЕСКИХ ОТНОШЕНИЙ ДЛЯ ДИСКУРСИВНОГО АНАЛИЗА ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

Чистова Е. В. (chistova@isa.ru)
ФИЦ ИУ РАН, Москва, Россия,
Российский университет дружбы народов, Москва, Россия

Шелманов А. О. (shelmanov@isa.ru)
Сколтех, Москва, Россия, ФИЦ ИУ РАН, Москва, Россия

Кобозева М. В. (kobozeva@isa.ru),
Писаревская Д. Б. (dinabpr@gmail.com),
Смирнов И. В. (ivs@isa.ru)
ФИЦ ИУ РАН, Москва, Россия

Толдова С. Ю. (toldova@yandex.ru)
НИУ ВШЭ, Москва, Россия

Ключевые слова: дискурсивный анализ, теория риторических структуру, векторные представления слов, отбор признаков, обучение на размеченном корпусе

1. Introduction

There are many natural language processing tasks that require the analysis of text beyond the boundaries of single sentences. Recently, researches have started to approach this problem by leveraging discourse parsing, which made it a very prominent research topic. One of the most widely adopted discourse models of text is Rhetorical Structure Theory (RST), developed by W. Mann and S. Thompson [18]. RST represents a text as a tree of discourse (rhetorical) relations (“Cause”, “Condition”, “Elaboration”, “Concession”, “Sequence”, “Contrast”, etc.) between text segments – discourse units (DUs). These units can play various roles inside a relation: nuclei contain more important information, while satellites give supplementary information. The leaves of the tree are so called elementary discourse units (EDUs), usually clauses. Discourse trees in RST integrate both shallow and deep discourse structure. Discourse units on different levels are combined by the same set of relations. The well-known applications of automatic discourse parsing include the systems for summarization [9], sentiment analysis [25], question-answering [11], natural language generation [21], and dialog parsing [1].

This work is devoted to the problem of developing a system for rhetorical parsing of Russian texts. Recently, the release of manually annotated Ru-RSTreebank corpus [20] unlocked the possibility to use machine learning techniques for this task. In particular, we consider the tasks of classification of discourse relations between DUs into rhetorical types, as well as determining the nuclearity of DUs in a relation.

The contributions of this paper are the following:

- We investigate importance of various types of features for discourse relation classification in Russian-language texts and develop a feature set for this task.
- We report the results of the first experimental evaluation of machine learning models trained on Ru-RSTreebank corpus.
- We publish the models and the code for evaluation.

The rest of the paper is structured as follows: **Section 2** presents the background and related work on discourse parsing. **Section 3** briefly describes the manually annotated corpus of rhetorical structures Ru-RSTreebank. **Section 4** examines features, classification models, and feature selection procedure. **Section 5** describes the experimental evaluation of the developed methods, results of feature importance investigation, and results of error analysis. Section 6 concludes the paper and outlines the future work.

2. Background and Related Work

One of the early attempts at data-driven discourse parsing [26] rely to a large extent on syntactic features. The authors leverage lexicalized syntactic trees, probabilistic models, and a bottom-up parser for segmenting and building sentence-level discourse trees. In [22], syntactic features and POS tags are used as features in a shift-reduce discourse parser driven by an averaged perceptron. In HILDA parser [8] the feature set is extended with information about discourse markers, punctuation, and word-level n-grams. In some other works, it is suggested using also syntax and discourse production rules [6], [16], POS tags of the head node and the attachment node, as well as the dominance relationship between DUs, and the distance of each unit to their nearest common ancestor [5]. Some recent studies propose to abandon using any form of syntactic subtrees as features and leverage hidden outputs of a neural syntax parser as implicit features instead [29], [31].

Besides various syntactic features, one can use lexical features, semantic similarities of verbs and nouns [6] in different DUs, tokens and POS tags at the beginning and end of each DU and whether the both of them are in the same sentence [14], bag of words along with the appearing of any possible word pair from both DUs [30]. In [7], neural tensor network with interactive attention was applied to capture the most important word pairs. Authors use them as additional features to word embeddings. In [17], researchers suggest to use some entity-related features to extract implicit discourse relations between sentences of one paragraph, such as whether entities in the current DU were used in previous sentences or not. Authors claim it could be useful for detection of “Expansion”-type relations (e.g., “Restatement”), or occurrence of a topic indication, which is frequent for “Comparison” (e.g., “Contrast”, “Concession”) and “Temporal” relations. Other representative semantic properties were discovered in [13] for three relation types from Penn Discourse Treebank: “Comparison”, “Contingency” (e.g., “Cause”, “Condition”), “Expansion”. Authors find that “Comparison” relations are usually expressed by negation in one of the two arguments; “Contingency” relation can be discovered if one of the DUs is a subjective judgement, e.g., it can be manifested in the lexical choice of the main verb. “Expansion” relations, being general-specific,

can be encoded with pronouns tagging and named entity recognition in a Narrowing Entity Continuity feature by indefinite pronouns detection in DU1 and named entities extraction in DU2 and in a Parallel Entity Continuity feature by comparison of type of named entities in both DUs and detecting any continuity form in the predicate.

Recently, deep learning models that use low-level features were adopted for discourse parsing. In [10], authors propose a transition-based discourse parser that makes use of memory networks to take discourse cohesion into account and benefit discourse parsing, including cases of long span scenarios. Experiments were based on RST Discourse Treebank for English[1]. Several discourse parsing models were created for Chinese. In [15], a framework based on recursive neural network is proposed, it jointly models the subtasks of EDU segmentation, tree structure construction, center labeling, and sense labeling. In [7], researchers use word pairs from two discourse arguments to model pair specific clues, and integrate them as interactive attention into argument representations produced by the bidirectional long short-term memory network (Bi-LSTM). Pair patterns improve recognition of discourse relations. In [28], a text matching network is presented. It encodes the discourse units and the paragraphs by combining Bi-LSTM and CNN to capture both global dependency information and local n-gram information.

In this paper, we primarily rely on feature-engineering approach rather than on deep models for several reasons. The purpose of this work is to set a baseline for the discourse parsing of texts in Russian and investigate importance of various language factors rather to push the performance of the parser to the limit. Although deep models can perform better, they are not transparent enough for feature investigation. We also note that we are still lacking of training data for leveraging deep models. Commonly, these models have a lot of parameters (starting from hundreds of thousands) and tend to overfit on small datasets.

3. Annotated corpus

This study is based on Ru-RSTreebank[2] – first open discourse corpus for Russian [20], [27]. We use an updated version of Ru-RSTreebank that is currently freely available on demand. Currently, it consists of 179 texts, including news, news analytics, popular science, and research articles about linguistics and computer science (203,287 tokens in total). The set of rhetorical relations was customized to make it more suitable for Russian. The corpus was annotated with an open-source tool called rstWeb[3]. As to inter-annotator agreement, Krippendorff's unitized alpha is 81%.

The corpus contains the following types of annotations: segmentation of EDUs (mostly clauses), nuclearity of discourse units, types of discourse relations, rhetorical tree structures. In addition to ordinary multi-nuclear relation types, there is a relation type "Same-unit", which is used for annotations of cases when one discourse unit is interrupted by another one. A rhetorical tree fragment example is presented in **Figure 1**.

---

1 https://catalog.ldc.upenn.edu/LDC2002T07

2 http://rstreebank.ru/

3 https://corpling.uis.georgetown.edu/rstweb/info/

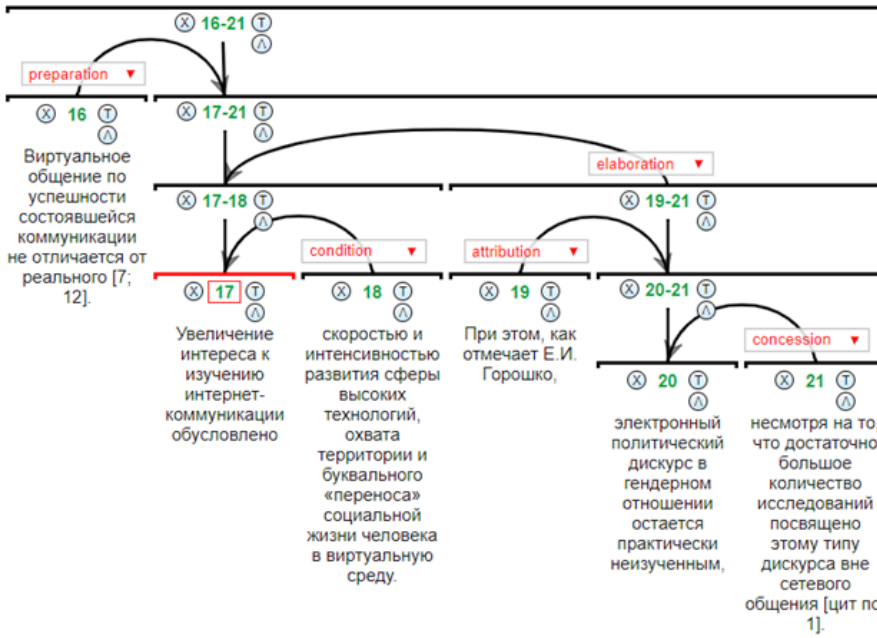


Figure 1: Visualization in rstWeb of an annotated fragment of Ru-RSTreebank

4. Features and Models for Discourse Parsing

In this work, we focus on two multiclass classification tasks. The objects for classification are pairs of DUs, that are given in the corpus. The first task is classification of DU pairs into 11 rhetorical labels. The second task is nuclearity relationship classification between DUs; there are three types of nuclearity in RST: “Satellite-Nucleus” (SN), “Nucleus-Satellite” (NS), “Nucleus-Nucleus” (NN).

4.1. Features

For both tasks, we consider combinations of various lexical, morphological, and semantic features. As lexical features, we use the list of marker phrases (or discourse connectives), nearly 450 items. It was manually composed on the basis of three sources: expressions extracted by experts from the annotated texts, the conjunctions used in complex sentences in Russian described in RusGram[4] and the list of functional MWUs suggested in the Russian National Corpus[5].

The set of features contains various numerical features:

[4] http://rusgram.ru

[5] http://ruscorpora.ru/obgrams.html

- Number of words.
- Average word length.
- Number of completely uppercase words.
- Number of words starting with an uppercase letter.
- Number of various morphological features. For instance, verbs have person and number.
- Part of speech tags for the first and the last word pairs of each DU.
- Features indicating the similarity between morphological features vectors of both DUs using various similarity measures namely Cosine, Hamming, Canberra, similarity measure for binarized vectors.
- Number of occurrences of stop words.
- Number of occurrences of each marker phrase.
- Occurence of each cue phrase at the beginning and the end of each DU.
- TF-IDF [23] of each DU.
- Cosine similarity between TF-IDFs.
- Jaccard index between lemmatized DUs.
- BLEU similarity measure.
- Averaged word embeddings of each DU. Embedding models were trained using word2vec [19].
- Sample of non-top11 classes examples along with the features described above were supplied to train a regressor, which predicts the probability of appearance of a mononuclear relation between DUs. This prediction is also used as a feature in the relation labeling.

4.2. Classification and Feature Selection Methods

We compared the effectiveness of various widely used supervised learning algorithms, namely, logistic regression, feedforward neural network (NN), support vector machine (SVM) with various kernels [2], and gradient boosting on decision trees (GBT) implemented in LightGBM [12] and CatBoost [4] packages. Feedforward neural network is a 2-layer perceptron regularized with dropout. The first layer activation function is ReLU. The outputs of the first layer are passed through the batch normalization. The activation on the output layer is softmax. As data imbalance highly affect the performance of neural network model, the SMOTE technique [3] is incorporated to oversample all classes but the majority class. We also experimented with ensembles by combining several models with soft voting.

The number of features in the original feature space is 3,273. Since only some features are informative, we perform feature selection: in some experiments we pick a strong subset of features selected by L1-regularized logistic regression model. A parameter of regularization is C. The higher C means the lower regularization strength. The best C for feature selector was found using grid search on 5-fold cross validation.

We also build ensembles of classifiers using soft voting. During the preliminary experiments, we found that ensembles of gradient boosting models with feature selection and linear SVM classifiers achieve the best performance.

5. Experiments

5.1. Dataset and Evaluation Procedure

The distribution of the classes in the original Ru-RSTreebank corpus is skewed. For experiments, we excluded "Elaboration" and "Joint" relations, since they are not very informative, although they are the most common. We also excluded "Same-unit" since it is has an utility function. Finally, we took the first 11 most representative classes, for which the dataset contains at least 320 examples. Therefore, we selected 8 mononuclear relations (these relations are marked with postfix "_r") and 3 multinuclear relations (they are marked with postfix "_m"). The result dataset for experimental evaluation contains 6,790 examples, the distribution of the classes is depicted in **Figure 2**.

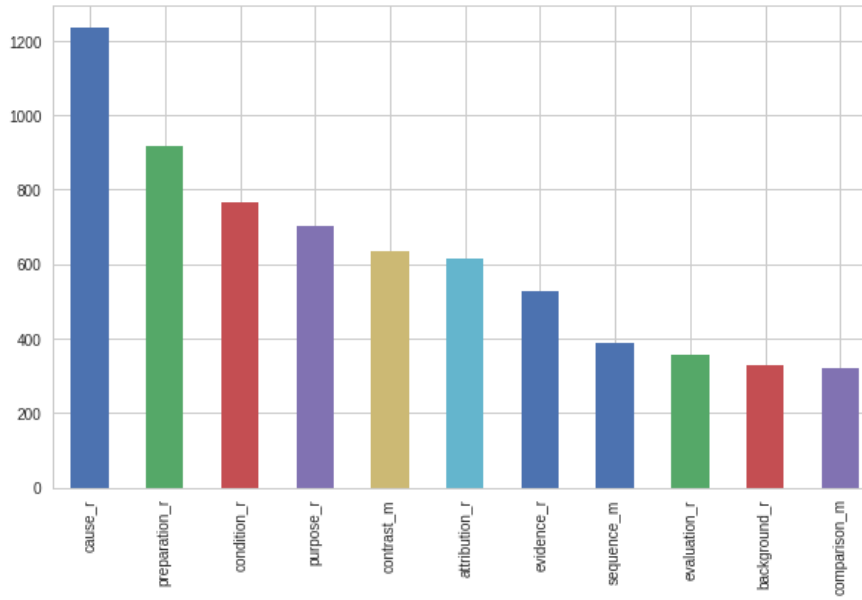


Figure 2: Distribution of rhetorical relation classes in the result dataset

Prior to feature extraction, the following text preprocessing steps were taken: tokenization, lemmatization, part-of-speech tagging, and morphological analysis using MyStem [24]. The pipeline was implemented via IsaNLP[6] Python library.

For evaluation, we used the standard metrics: precision, recall, and $_{F1}$. Macro-averages were employed as our main measurements, and accuracy was omitted, since the distributions of classes are unbalanced. We perform all our experiments using 5-fold cross validation with stratified randomized split of the dataset into 90% for training and 10% for testing.

[6] https://github.com/IINemo/isanlp

A randomized grid search algorithm was used to find the optimum logistic regression and SVM parameters: C and type of penalty (L_1, L_2), and neural network parameters: number of units for each layer, activation function for each layer, dropout rate. Randomized grid search was used for selecting the best hyperparameters for gradient boosting models: number of trees, number of leaves, learning rate, feature sampling ratio, and regularization coefficients. For selection of optimal number of iterations in a CatBoost model, we used its built-in overfitting detector.

After hyperparameter tuning, we get the following best parameters. Logistic regression: inverse regularization strength: 0.001 and L_2 penalty. SVM: inverse regularization strength: 0.0001 and L_2 penalty, kernel: linear. LightGBM: number of leaves: 36, number of iterations: 1,000, bagging fraction: 0.9, learning rate: 0.1. CatBoost: number of iterations: 2,000, learning rate: 0.1. NN: size of hidden layer: 100, dropout: 0.5, optimization algorithm: Adam, learning rate: 0.01, batch size: 128, number of epochs: 7.

5.2. Main Results

Table 1 summarizes the results of experiments with models for rhetorical relation classification. The results show that gradient boosting models outperform other models. Ensemble of CatBoost model with selected features and a linear SVM model owns the best score.

Table 1: Results of rhetorical relation classification models, %

Classifier	Macro F_1		Micro F_1	
	mean	std	mean	std
NN	49.43	1.52	55.78	1.16
Logistic Regression	50.81	1.06	53.81	1.84
LGBM	51.39	2.18	59.91	1.32
Linear SVM	51.63	1.95	56.61	1.54
L_1 Feature selection + LGBM	51.64	2.22	60.29	1.74
CatBoost	53.32	0.96	60.71	0.81
L_1 Feature selection + CatBoost	53.45	2.19	61.09	1.96
voting((L_1 Feature selection + LGBM), Linear SVM)	54.67	1.80	62.39	1.51
voting((L_1 Feature selection + CatBoost), Linear SVM)	54.67	0.38	62.32	0.41

We evaluated the importance of features related to the word order in the document. There are two types of discourse markers in the feature set: positional, i.e. whether a cue is found at the beginning or at the end of DUs and quantitative, i.e. a number of a cue in each DU. In Table 2, we see a performance drop when removing positional features. At the same time, we can observe that quantitative features do not significantly affect the F_1 score.

The results for distinguishing "Satellite-Nucleus", "Nucleus-Satellite", and "Nucleus-Nucleus" types of relations are presented in Table 3. We used the full set of features described in subsection 4.1. The experiment shows that the gradient boosting models strongly outperform feedforward neural network, SVM and logistic regression classifiers.

Table 2: F_1, % for rhetorical relation classification task with different feature sets

Feature set	Macro F_1		
	Logistic Regression	Linear SVM	CatBoost
All features	51.5	50.6	52.4
w/o quantitative features	−0.3	+0.1	−0.1
w/o positional features	−4.0	−4.0	−2.8

Table 3: F_1 for the nuclearity recognition models, %

Classifier	Macro F_1		Micro F_1	
	mean	std	mean	std
Linear SVM	63.01	0.58	64.20	0.52
NN	63.32	0.88	64.59	0.75
Logistic Regression	63.66	0.37	65.02	0.26
L1 Feature selection + LGBM	67.82	0.86	69.17	0.73
CatBoost	68.03	0.45	69.37	0.36
LGBM	68.81	0.77	70.17	0.67
L1 Feature selection + CatBoost	68.82	0.84	70.31	0.76

From the whole set of features (3,624 features), CatBoost model for rhetorical type relation classification selected 2,014 as important features. Analysis of this features is presented in Table 4. We can see that the most important features for this model are related to discourse markers. Table 4 also shows the performance drop when removing features from this model. As we can see, after removing the information about 1,887 features related to discourse markers, this model loses 2.49% of macro F_1.

Table 4: Important features selected by CatBoost model per feature type

Type	Features	Number	% in selected	Performance drop, %
Lexical	4 elements of TF-IDF vectors for the first DU; 4 elements of TF-IDF vectors for the second DU;	8	0.4	0.11
Morpho-syntactic	Combinations of punctuation, nouns, verbs, adverbs, conjunctions, adjectives, prepositions, pronouns, numerals, particles at the beginning of a first DU; Combinations of punctuation, verbs, adverbs, nouns, pronouns, adjectives, conjunctions, prepositions, particles, numerals at the end of a first DU; Number of nouns in instrument case, pronouns, adverbs in a first DU; Various combinations of verbs, pronouns, nouns, adverbs, conjunctions, punctuation, particles at the beginning of a second DU; Various combinations of punctuation, nouns, verbs, pronouns, adverbs, adjectives, prepositions, conjunctions, particles at the end of a second DU; Number of occurrences of conjunctions, adverbs, adjectives, pronouns, adpositions in a second DU; Number of passive verbs, gerunds and infinitives in a second DU; Correlation between morphological features vectors of DUs.	119	5.9	0.45
Textual	Number of occurrences of 355 markers in a first DU (18%) Number of occurrences of 331 markers in a second DU (17%) Occurrences of 298 markers at the beginning of X (16%) Occurrences of 326 markers at the end of X (17%) Occurrences of 335 markers at the beginning of Y (19%) Occurrences of 242 markers at the end of Y (13%)	1,887	93.69	2.49

5.3. Error Analysis

The classification report of the best performed model using a variety of measures is presented in Table 5. In Figure 3, we also provide the confusion matrix generated by this model. Asymmetric relations labeling has relatively better performance, we achieved 74.36% F_1 score for "Attribution" relation.

The worst performance, under 50% F_1 score, was obtained with 4 classes that have least number of training instances: "Comparison" (320 samples), "Evidence" (529 samples), "Evaluation" (356 samples), and "Background" (328 samples). For example, "Evidence", "Evaluation", and "Background" are often recognized as "Cause", the most represented class (1235 samples). The model has a very low recall score on "Background" relation, often labeling it as "Preparation". Macro averaged F_1 score for the classification on the top 7 relations is 72.34±1.37%.

Table 5: Relation labeling performance for each class, %

Class	Precision	Recall	F_1-score
attribution	73.11	75.77	74.36
purpose	71.87	73.71	72.70
condition	73.60	65.75	69.36
preparation	57.82	81.09	67.49
cause	51.73	69.96	59.46
contrast	68.43	56.69	56.69
sequence	54.46	54.55	54.22
evidence	44.75	34.53	38.95
comparison	50.43	31.25	38.49
evaluation	31.89	17.46	22.56
background	24.09	5.15	8.41

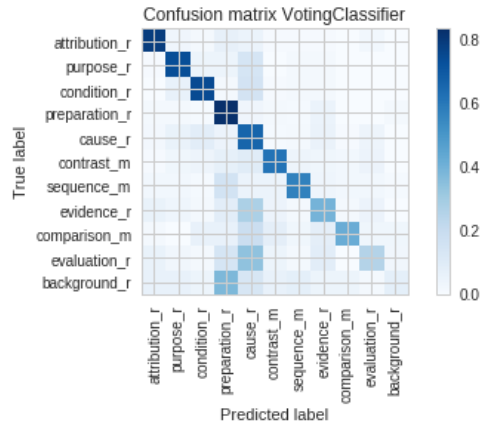


Figure 3: Confusion matrix for the best model

Errors with relation labeling partly occur when there is semantic similarity between true type and predicted type, such as in pairs “Preparation”—“Background”, “Comparison”-“Contrast”, “Cause”—“Evidence”, “Purpose”—“Cause”, “Preparation”—“Attribution”, “Preparation”-“Sequence”. In other cases, such as “Cause”—“Preparation” or “Preparation”—“Attribution”, errors can be caused by stylistic difference in news texts/scientific texts that are included in corpus. There are also cases when relation types are not semantically close to each other, these ones need more thorough investigation. For example, if “Cause” is predicted instead of “Contrast”, the error can be explained by occurrences of possible cause markers in nucleus or satellite, and corresponding punctuation marks: ‘[В _основе_ фразеологического сочетания лежат две заимствованные из турецкого языка лексемы_:_] [а сама идиома является точной калькой турецкого выражения.]’ (‘[Two lexemes borrowed from Turkish are at the heart of the phraseological unit_:_] [and the idiom itself is a calque of the Turkish expression.]’), ‘[Текст перевода при этом не является копией или подобием

исходного текста.] [Он _порождается_ путем воплощения на языке перевода указанной концептуальной структуры.]' ('[The translated text is not a full copy or a semblance of the original text.] [It is created by emphasizing a specified conceptual structure in the language of translation.]').

6. Conclusion and Future Work

We investigated the performance of different algorithms and features for discourse relations labeling and nuclearity type classification. We found that textual, morpho-syntactic, and lexical features are equally important in the relation labeling; both positional and quantitative textual features improve the quality of classification. Source code of the experiments is available online[7].

In our future work we are going to implement the complete pipeline for discourse parsing of Russian texts including segmentation and discourse tree construction. We also looking forward to employ state-of-the art deep learning techniques and pretrained language models for relation classification.

7. Acknowledgments

This paper is partially supported by Russian Foundation for Basic Research (project No. 17-29-07033, 17-07-01477).

We would like to express our gratitude to the corpus annotators T. Davydova, A. Tugutova, M. Vasilyeva and Y. Petukhova.

References

1. *Afantenos, S. et al.:* Discourse parsing for multi-party chat dialogues. In: Proceedings of the 2015 conference on empirical methods in natural language processing. (2015).
2. *Boser, B. et al.:* A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on computational learning theory. pp. 144–152 ACM (1992).
3. *Chawla, N. V. et al.:* SMOTE: Synthetic minority over-sampling technique. Journal of artificial intelligence research. 16, 321–357 (2002).
4. *Dorogush, A. V. et al.:* CatBoost: Gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363. (2018).
5. *Feng, V. W., Hirst, G.:* A linear-time bottom-up discourse parser with constraints and post-editing. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers). pp. 511–521 (2014).
6. *Feng, V. W., Hirst, G.:* Text-level discourse parsing with rich linguistic features. In: Proceedings of the 50th annual meeting of the association for computational linguistics (volume 1: Long papers). pp. 60–68 (2012).

[7] http://nlp.isa.ru/paper_dialog2019/

7. *Guo, F. et al.:* Implicit discourse relation recognition using neural tensor network with interactive attention and sparse learning. In: Proceedings of the 27th international conference on computational linguistics. pp. 547–558 (2018).
8. *Hernault, H. et al.:* HILDA: A discourse parser using support vector machine classification. Dialogue & Discourse. 1, 3, (2010).
9. *Hirao, T. et al.:* Single-document summarization as a tree knapsack problem. In: Proceedings of the 2013 conference on empirical methods in natural language processing. pp. 1515–1520 (2013).
10. *Jia, Y. et al.:* Modeling discourse cohesion for discourse parsing via memory network. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: Short papers). pp. 438–443 (2018).
11. *J.Y., C., R., J.:* Discourse structure for context question answering. In: Proceedings of the workshop on pragmatics of question answering at hlt-naacl. (2004).
12. *Ke, G. et al.:* Lightgbm: A highly efficient gradient boosting decision tree. In: Advances in neural information processing systems. pp. 3146–3154 (2017).
13. *Lei, W. et al.:* Linguistic properties matter for implicit discourse relation recognition: Combining semantic interaction, topic continuity and attribution. In: Thirty-second aaai conference on artificial intelligence. (2018).
14. *Li, J. et al.:* Recursive deep models for discourse parsing. In: Proceedings of the 2014 conference on empirical methods in natural language processing. pp. 2061–2069 (2014).
15. *Lin, C. A. et al.:* A unified RvNN framework for end-to-end Chinese discourse parsing. In: Proceedings of the 27th international conference on computational linguistics: System demonstrations. pp. 73–77 (2018).
16. *Lin, Z. et al.:* Recognizing implicit discourse relations in the penn discourse treebank. In: Proceedings of the 2009 conference on empirical methods in natural language processing. pp. 343–351 (2009).
17. *Louis, A. et al.:* Using entity features to classify implicit discourse relations. In: Proceedings of the 11th annual meeting of the special interest group on discourse and dialogue. pp. 59–62 (2010).
18. *Mann, W., Thompson, S.:* Rhetorical structure theory: Toward a functional theory of text organization. Text-Interdisciplinary Journal for the Study of Discourse. 8, 3, 243–281 (1988).
19. *Mikolov, T. et al.:* Efficient estimation of word representations in vector space. In: ICLR workshop. (2013).
20. *Pisarevskaya, D. et al.:* Towards building a discourse-annotated corpus of Russian. In: Computational linguistics and intellectual technologies. Proceedings of the international conference dialogue. p. 23 (2017).
21. *Qin, B. et al.:* A planning based framework for essay generation. arXiv preprint arXiv:1512.05919. (2015).
22. *Sagae, K.:* Analysis of discourse structure with syntactic dependencies and data-driven shift-reduce parsing. In: Proceedings of the 11th international conference on parsing technologies. pp. 81–84 (2009).
23. *Salton, G., Buckley, C.:* Term-weighting approaches in automatic text retrieval. Information processing & management. 24, 5, 513–523 (1988).

24. *Segalovich, I.:* A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In: MLMTA. pp. 273–280 (2003).
25. *Somasundaran, S.:* Discourse-level relations for opinion analysis. University of Pittsburgh (2010).
26. *Soricut, R., Marcu, D.:* Sentence level discourse parsing using syntactic and lexical information. In: Proceedings of the 2003 conference of the north american chapter of the association for computational linguistics on human language technology-volume 1. pp. 149–156 (2003).
27. *Toldova, S. et al.:* Automatic mining of discourse connectives for russian. In: Conference on artificial intelligence and natural language. pp. 79–87 (2018).
28. *Xu, S. et al.:* Employing text matching network to recognise nuclearity in chinese discourse. In: Proceedings of the 27th international conference on computational linguistics. pp. 525–535 (2018).
29. *Yu, N. et al.:* Transition-based neural RST parsing with implicit syntax features. In: Proceedings of the 27th international conference on computational linguistics. pp. 559–570 (2018).
30. *Zhang, B. et al.:* Shallow convolutional neural network for implicit discourse relation recognition. In: Proceedings of the 2015 conference on empirical methods in natural language processing. pp. 2230–2235 (2015).
31. *Zhang, M. et al.:* End-to-end neural relation extraction with global optimization. In: Proceedings of the 2017 conference on empirical methods in natural language processing. pp. 1730–1740 (2017).

SIMULATION OF BACKGROUND KNOWLEDGE AND BRIDGING IN RUSSIAN

Dikonov V. G. (dikonov@iitp.ru)
IITP RAS, Moscow, Russia

This paper introduces a knowledge-based semantic approach towards bridging annotation of Russian texts. Our method simulates human background knowledge by using compact domain descriptions based on an extended version of SUMO ontology and lexical-semantic data from the "Universal Dictionary of Concepts". Our approach supports a wide and extensible range of bridging relations. The tagger that implements it can build complex bridges with multiple arcs, supports making assumptions and can be adapted to annotate other languages supported by the underlying dictionary of concepts.

Keywords: computational linguistics, bridging, reference, anaphora, linguistic ontology

МОДЕЛИРОВАНИЕ ФОНОВЫХ ЗНАНИЙ И ПОИСК АССОЦИАТИВНЫХ АНАФОР В ТЕКСТАХ НА РУССКОМ ЯЗЫКЕ

Диконов В. Г. (dikonov@iitp.ru)
ИППИ РАН, Москва, Россия

1. Bridging

The notion of bridging, also known as indirect or associative anaphora, introduced by Clark [2] captures an essential mechanism of text interpretation inside a human mind. Every comprehensible piece of text contains some identifiable entities and statements about them. As we read or listen, we encounter new entities (new information) and refer them with previously mentioned ones (given information in Clark's terms). This is a way to construct a mental model of the reality described in the text. Pairs of related entities are viewed as a kind of anaphora where the previously given entity becomes the antecedent of the new. Unlike the common definition of anaphora, the relation between the entities in bridging is not limited to identity. It can be of any semantic type meaningful to the Listener. Building references requires deep background

knowledge of the relevant domain, especially when the Speaker skips "redundant" details to improve the speed of communication. When the Listener fails to find a directly related antecedent of some new entity, he is forced to insert a suitable intermediate concept. This leads to creation of "bridges" with multiple "arcs".

(1) На станции метро «Владыкино» в Москве найдено *взрывное устройство*. Найденный предмет *обследовали* с использованием служебных **собак**.
(An *explosive device* was found at the Moscow underground station Vladykino. The object was *examined* with service **dogs**)

Here the Listener assumes the existence of policemen, who were not mentioned directly, and constructs the following possible bridge: *explosive device* ^{isObjectOf} *examine* ^{hasAgent} *policemen* ^{isUserOf} *dog*. This assumption is based on common knowledge about police work and terrorism, implanted by television newscasts. It is possible to build arbitrarily long bridges by adding new assumptions.

Introduction of new concepts associated with the given information from background knowledge is also a productive mechanism of creativity. Its proper modeling combined with good plausibility filters might give AI the ability to invent.

2. Overview of the approach

Existing works in the field of bridging fall into two groups: semantic approaches and syntactic ones. Syntactic approaches choose particular syntactic patterns, usually definite NPs, and treat the ability of certain words to fill such patterns as a criterion of a non-typed bridging relation. Later research by Hou [5] departs from a single pattern restriction, but still lacks the ability to explicitly represent the meaning of the detected bridging relations. The first published paper on bridging in Russian [9] follows the same path and uses Russian genitive NPs as the clue pattern.

A semantic approach always ascribes a semantic type to the discovered relations. The authors of such approaches often impose restrictions on the types of bridges they detect in order to accommodate to their resources and relation search methods. Papers by Poesio [7], Lassale [6] concentrate only on part-whole relations. Recasens [8], Zikánová [10], etc. add set-subset, cohyponymy, predicate-argument and symptom relations. Many studies rely on Princeton Wordnet as the source of lexical data and a knowledge base to estimate semantic relatedness of words. Unfortunately, English Wordnet provides only part of the information needed to simulate the mental mechanism of bridging. It offers good lexical coverage, usable (though poorly organized) taxonomy, but is very limited in the field of semantic relations other than part-whole. In particular, it lacks cause-result and predicate-role relations. Roitberg et al. [9] wrote that absence of a (large scale) Russian Wordnet prevents the use of semantic methods on Russian material. We would answer that there are alternative resources for Russian and they have some advantages over the English Wordnet. One of them is briefly described in **section 2.1**.

We take a semantic approach based on a rich background knowledge base (KB). The target language is Russian, but our KB is a language-neutral semantic resource. As a result, our bridging tool can be adapted to work with other languages supported

by the underlying semantic dictionary UNLDC [3] (English, Hindi, French etc). Our project bears resemblance with the work by Fan [4], yet it is different in some key points. Both projects use a knowledge base encoded as a semantic graph and support simple taxonomy based inference. However, the structure and contents of the KBs are different. The set of relations in our study is wider. Our tool supports making assumptions and builds complex chain relations with intermediate concepts like the relation between the bomb and dogs in **example 1**.

2.1. Resources

Our relation search engine operates with ontology concepts instead of words. We use a modified version of SUMO ontology with greatly extended taxonomy (extended ontology). This extension exists in the framework of developing the "Universal Dictionary of Concepts" (UNLDC) [3]. The extended ontology is an experimental resource and is different from the internal ontology of the linguistic processor ETAP[1] (ETAP ontology), which is also based on SUMO and mentioned further in this paper.

UNLDC translates the concepts of the extended ontology into Russian and several other languages. The Russian lexicon used in this project contains 42,973 Russian words and multi-word expressions with 66,896 senses total. These senses are linked with 48,883 concepts of the extended ontology (both original SUMO concepts and the added ones). UNLDC also has a growing semantic network that includes many relation types not available in the Wordnet, including the cause-result and argument ones. The types of semantic relations supported in this project are described in **section 3.2**. UNLDC is an open public resource. Its core parts are available for download at GitHub2. The extended ontology is a supplement to UNLDC.

3. Knowledge base

Modeling the mechanism of human association reference requires an imitation of human knowledge about the subject domain of the text, which consists of:

a) set of concepts relevant to the domain,
b) semantic relations that hold between such concepts.

It also needs imitation of the relevant subset of human linguistic ability sufficient for transition from an NL text to a set of concepts. The latter includes at least chunking and morphology engines to identify sentences and lemmatize words, a semantic lexicon linking the words with concepts and some kind of lexical disambiguation.

---

[1] ETAP is a multipurpose linguistic processor developed by the laboratory of computer linguistics at the Institute of Information Transmission Problems (IITP) in Moscow. It supports robust syntactic parsing, English ↔ Russian machine translation, paraphrasing, semantic parsing using two different frameworks, question answering and more.

[2] https://github.com/dikonov/Universal-Dictionary-of-Concepts

3.1. Concept inventory

Using ontology concepts to abstract away from lexical variation and peculiarities of different natural languages always poses the problem of choosing the right degree of abstraction or "semantic grain" for the task. Consider the following example:

(2) Во Владимирской области произошло столкновение товарного *поезда* с застрявшим на переезде *грузовиком*. **Водитель** успел выскочить из кабины. **Машинист** получил травмы.
(A freight *train* hit a *truck* stuck at a crossing in the Vladimir region. The **driver** managed to jump out of the cabin. The **train driver** was injured.)

Bridging is expected to establish relations of association between *машинист (train driver)* and *поезд (train)*, *водитель (driver)* and *грузовик (truck)* based on the fact that each type of driver controls a particular type of vehicle. This information is embedded in definitions of Russian words.

Initially we had three different sets of concepts offered by SUMO, ETAP Ontology and UNLDC to choose from. Straight ontology rendering of this example would use the same class label "SocialRole" (SUMO) / "DriverRole" (Etap Ontology) for the truck and the train drivers. This would not allow the bridging process to see the difference between the two driver entities and link them with the Train and Automobile concepts correctly.

On the other hand, UNLDC uses a very fine-grained set of concepts, that correspond to word senses from several natural languages. In particular, it includes most of the English Wordnet senses. UNLDC concepts can reflect even stylistic distinctions between members of the same Wordnet synset. Semantic classes roughly parallel to NL POS categories are imposed on top. This level of detail is an overkill for most text processing tasks except translation.

The extended ontology offers a fourth option—an optimized set of concepts, more general than lexical senses and more specific than most SUMO/Etap Ontology concepts. It is produced by an automatic procedure. We a) merge into one concept all synonymous senses regardless of the POS class of the source words, e.g. *катание (act of rolling as a ball)* gets merged with *катить (cause to move by turning like a ball)* and all their synonyms b) merge pairs of predicates like *катить (cause to move by turning)* and *катиться (move by turning)*, which differ only by the regular transformation of their argument frames. Each new concept receives a unique OWL-compatible name and a link to an upper SUMO class or another new concept. The new concepts inherit semantic relations from UNLDC semantic network, including *is_a* and *instance_of*, which create subtrees of new concepts within SUMO classes, and other types translated into the bridging relation set, e.g. *катание (roll—act of rolling as a ball)* ^{subProcess} *боулинг (bowling game) = "rolling (balls) is part of playing bowling"*.

3.2. Relations

The relation types supported by our bridging tool are listed in Table 1. This set of relations can be extended through editing of the knowledge base. All relations are directed and have corresponding reverse relation types. Type labels are taken from the Etap Ontology or follow the same style.

Table 1: Bridging relation types

Group	Relation / Reverse relation	Examples (X—Y)	Comment
Function	hasFunction / isFunctionOf	restaurant—serve meals baker—to bake	Y is what X does or is for.
	hasRoleAt / isRoleAt	company—accountant tourists—guide cathedral—priest	Y is a function in respect to the group or object X. There may be multiple persons/ objects with the same function.
	hasChief / isChiefOf	team—trainer company—director country—president	The leader of a group
Part ↔ whole	hasPart / isPartOf	room—wall	Parts that are always present
	hasOptionalPart / isOptionalPartOf	room—chandelier	Parts that may be absent
	hasDetachablePart / isDetachablePartOf	lock—key violin—bow	Required accessories that are not physically attached
	hasMember / isMemberOf	parliament—MP government—minister	All members of the group X are Y-s.
	hasSubEvent / isSubEventOf	eat—swallow	
Object ↔ matter	hasIngredient / isIngredientOf	tea—water water—oxygen	Y is one of the raw materials used and irrevocably changed or chemically bound in making X.
	hasSubstance / isSubstanceOf	table—wood ocean—water	X is a mass of pure Y. There may be parts made of other substances.

Group	Relation / Reverse relation	Examples (X—Y)	Comment
Event ↔ role	hasAgent / isAgentOf	buy—buyer fly—airplane	
	hasAgent2 / isAgent2Of	buy—seller	
	hasObject / isObjectOf	write—letter	
	hasInstrument / isInstrumentOf	eat—spoon	
	hasLocation / isLocationOf	study—school	
	hasStartingPlace Point / isStarting PlacePoint	delivery—warehouse (as an order in a webshop)	
	hasTerminalPlace Point / isTerminal PlacePoint	delivery—home (as an order in a webshop)	
	hasRecipient / isRecipientOf	delivery—customer (as an order in a webshop)	
	hasBeneficiary / isBeneficiaryOf	sing—audience	X has an object or message delivered to Y
	hasSource / isSourceOf	passport—Russia	
Cause ↔ result	hasResult / isResultOf	murder—death	
	newstatus-agent	compete—winner compete—looser	Y is a new social role of the agent of the event X
	newstatus-object	matriculation—student	Y is a new social role of the object of the event X

Group	Relation / Reverse relation	Examples (X—Y)	Comment
Temporal	before / after	grab—arrest—jail	Relative position at the timeline. Used in describing typical sequences of events concurrence.
	concurrence		Events that occur at the same time but neither is a subEvent of the other.
	hasTime / isTimeOf	breakfast—morning	Customary period
Misc. association	hasResident / isResidentOf	Berlin—Berliner	Resident of a place
	hasBeliever / isBeliefOf	Pope—Christianity socialist—socialism	Supporter and teaching supported
	hasAuthor / isAuthorOf	writer—book	Y is an object designed by X.
	hasMaker / isMakerOf	blacksmith—horseshoe	Y is one of many manufactured objects
	hasFrame / isFrameOf	clash—public protest study—university study—seminar	A typical scene (event, institution, proposition) associated with event X and forming its background.
	isUserOf / isUsedBy	woodcutter—ax pilot—airplane	Y is a default instrument of X e.g an attribute of profession
	hasOwner / isOwnerOf	cop—uniform	X typically possesses Y
	hasAttribute / isAttributeOf	exam—passing grade	
Cohyponyms	cohyponym	hands—legs mother—son	Only usable at low taxonomy levels.
Equivalence	SameAs	projector—apparatus shopper—client	Y is another name for X in the given domain

3.3. Domain descriptions

Relations and concepts are used to make semantic graphs containing generalized descriptions of different subject domains. Together such domain descriptions and the extended ontology constitute our knowledge base for bridging.

The graphs consist of triplets, where the relation labels take the place of predicates. A domain description can be saved as an RDF document. Each triplet has an additional annotation field, containing a list of domain names. Domain annotation is used to limit the scope of statements applicable only to certain parts of actual reality.

This kind of data can be imported from various domain ontologies that a) cover the domains relevant to the text to be processed, b) provide non-taxonomic relations used to construct bridges, c) have their concepts linked with the dictionary used to lemmatize/disambiguate the text. The extra fourth field (domain annotation) can be filled with the ontology's declared domain.

Our goal is to interpret texts dealing with everyday life and typical news topics: shopping, medical care, education, traffic, crime and police, sport, banking, politics etc. We model a very basic level of common background knowledge of Russian people, essential to understand contemporary Russian texts, reflecting the reality of Russia and late USSR. We did not have a suitable ontology to fill the knowledge base. The domain descriptions used in our experiments are written manually and later augmented with data from the UNLDC semantic network. We found that the amount of labor needed to describe a single domain is agreeable.

We start by enumerating a few key concepts of the domain (not including any individual persons and institutions). At the next step we link them to each other using the relations from Table 1. Later we enumerate key events concerning the domain and corresponding predicates, e.g. *matriculation, studying, reading, writing, answering, evaluation, passing exams, graduation*, etc. in the educational domain, list their default argument slot fillers, e.g. *student, professor, textbook*, etc, and specify typical temporal and causal relations between the events, e.g. *studying* ^{before} *passing exams*. Everything is done ad-hoc to replicate human background knowledge.

The main reason to do it is to capture typical domain-bound sequences of events (scripts) that people follow in their life and work. Scripts are presented as chains of predicate concepts placed along the abstract timeline and connected by the temporal and/or causal relations. A typical scripted activity is fishing where an angler has to *attach (a fly to the hook)* ^{before} *throw (the line into the river)* ^{before} *wait* ^{concurrence} *watch (the cork)* ^{before} *strike (fish)* ^{hasResult} *pull (the line)*, etc. This information is not present in Wordnet or general purpose ontologies, but it turned out to be very useful for bridging. It can explain the relations between participants of the events e.g. *fish* and *cork* by connecting the events they take part in *fish* ^{isAgentOf} *strike* ^{hasResult} *bob* ^{hasAgent} *cork*. UNLDC does provide some cause-result links, but they are limited to universal connections between concepts, embedded in their definitions, e.g. *to grow (vegetables)* ^{hasResult} *growth (of the plants)*.

Another reason is that manually formulated domain descriptions help to identify most important keywords making up the lexical footprint of the domain. We take premade concepts from UNLDC, which already have associated Russian words. Consequently, the domain description graphs are accompanied by a cloud of keywords that help to identify domain texts.

The resulting sketch description is immediately useful and can be tested with our bridging tool. We make a test run and check, if there are any important keywords/concepts missing from the domain description. This final step can be repeated many times to improve the recall of bridging relations. The typical size of a domain description is 100–1000 triplets. A fragment is shown in **Figure 1**.

Domain	Subdomains		Triplet		
EducationalProcess	SchoolEducationalInstitution		SchoolEducationalInstitution	hasChief	Headmaster
EducationalProcess	SchoolEducationalInstitution	Matriculation	OrderRequest	hasAgent	ParentGenitor
EducationalProcess	SchoolEducationalInstitution	Matriculation	OrderRequest	hasRecipient	SchoolEducationalInstitution
EducationalProcess	SchoolEducationalInstitution	Matriculation	OrderRequest	hasRecipient	Headmaster
EducationalProcess	SchoolEducationalInstitution	Matriculation	OrderRequest	hasTopic	ChildJuvenile
EducationalProcess	SchoolEducationalInstitution	Matriculation	OrderRequest	hasResult	Matriculation
EducationalProcess	SchoolEducationalInstitution	Matriculation	Matriculation	hasAgent	ChildJuvenile
EducationalProcess	SchoolEducationalInstitution	Matriculation	Matriculation	hasTerminalPoint	SchoolEducationalInstitution
EducationalProcess	SchoolEducationalInstitution	Matriculation	Matriculation	hasResult	EnrollRegister
EducationalProcess	SchoolEducationalInstitution	Matriculation	EnrollRegister	hasAgent	SchoolEducationalInstitution
EducationalProcess	SchoolEducationalInstitution	Matriculation	EnrollRegister	hasObject	ChildJuvenile
EducationalProcess	SchoolEducationalInstitution	Matriculation	EnrollRegister	Result-newstatus	Schoolchild

Figure 1: A few lines of a domain description showing the process of enrolling a child in a school. "The parents make an application to the school. The child gets enrolled and becomes a pupil"

The domains have their own taxonomy. Statements made in the general domains, such as *Education* and *Shopping* apply together with all statements from more specific domains, such as *University* and *Supermarket*.

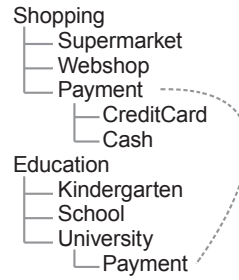


Figure 2: A fragment of the taxonomy of domains

The KB describes a default general state of affairs. The statements in the domain descriptions are just "usually true". No claim for universal truth can be made here. Actual truth in the real world or a fictional reality described in some concrete text has to be determined during understanding of the text or a situation in the real world. For example, the *TerminalPlacePoint* argument slot of the concept *Carrying* is always filled by some *Region*. The statement *Carrying* ^{hasTerminalPlacePoint} *Region* is universally true. However, in the domain of supermarkets shoppers usually carry goods to the checkout counter. Therefore, the description of the supermarket domain contains the statement *Carrying* ^{hasTerminalPlacePoint} *Checkout*, which is expected to be true in the domain. It makes the content of the domain descriptions unfit for a general purpose ontology, where all statements must be universally true. Instead each sub-domain section of a domain description could be viewed as a small domain ontology.

4. Bridging annotation

Our bridging annotation tool has two major functions:

1) search through a corpus and detect fragments of text that match known domains,
2) generate a set of potential bridging relations for the fragments found.

We use a corpus of newspaper texts as a source of examples. It consists of automatically parsed news feeds and full articles in the ETAP TGT format. The current version of the program uses only lemmatization tags. Syntax relations are used only to detect multiword expressions. It can use ETAP combinatorial dictionary entriy tags for disambiguation and falls back to lemmas if they are not available. A simple TF-IDF ranked keyword search is used to extract fragments that contain higher than average density of keywords linked with available domain descriptions. The length of the fragments is not set and usually falls between 3 and 15 sentences. Each fragment gets tagged with the applicable domains with weight numbers. There is a weigh threshold which can be adjusted to tune the output between better domain detection and quantity of examples.

The bridging annotation option works as follows: an example text is scanned for any nouns, verbs and multi-word expressions (MWEs), e.g. *банк России (bank of Russia), барная стойка (bar stand),* present in UNLDC. The words/expressions whose lexical senses match the domains ascribed to the example text form a set of possible reference words and antecedents. The set contains all words suitable for bridging in the whole text. MWEs are represented by their head words that carry the lexical meaning of the corresponding expression.

The words are taken one by one in the linear order of the text and paired with every preceding word of the same set within a rolling window of configurable number of sentences. This creates candidate pairs of words which are turned into two sets of concepts, associated with different senses of both words/MWEs. The concepts linked with the possible reference word and mentioned in the background knowledge base are paired with all concepts linked with the possible antecedent.

Resulting pairs of concepts are fed to a search function which returns all possible bridges between them, if any. The bridges may consist of either a single semantic relation or a chain of 1–2 intermediate concepts with relations between them. Since the extended ontology has taxonomic relations between its extra concepts within SUMO/Etap Ontology classes, the search function can use subclass_superclass_sibling criterion [4] to improve recall and relate antecedent concepts not mentioned in the domain description.

The resulting bridges are filtered by applying such criteria as number of intermediate concepts (“bridge arcs”), number of assumed extra concepts, distance between the reference and antecedent, saliency, etc. Each confirmed antecedent word receives a list of discovered reference words and clusters of bridge links are formed.

An interesting feature of our tool is building of a possible associations list. Intermediate concepts, which are not linked with any words in the text but occur in complex bridge relation, e.g. *Policeman* in **Example 1**, are remembered. Most frequent associations are returned together with the list of discovered bridging pairs.

5. Problems

Like every other ontology based system, our approach falls prey to the expert knowledge input bottleneck. The amount of background knowledge provided by domain descriptions is never enough (just like with us humans) but extending them manually is a labor intensive process that gets harder with more elaborate descriptions.

The tool demonstrates domain bias. Lack of a relevant domain description provokes our tool to switch to other domains which have partially similar lexical footprint. As a result, news reports about politics and wars, for instance, get interpreted in terms of crimes and terrorism. Sport events can get mixed with theater performances because both actors and athletes play and win contests and those domains share a certain amount of keywords. This problem can be mitigated by making brief descriptions of interfering domains that cover problematic keywords.

Use of very general ontology classes in domain descriptions creates spurious assumptions, yet it is hard to avoid. For example, the domain of police work includes the concept of arresting some *Human*. It makes the system assume that every entity of a *Policeman* arrests every entity of a *Human* mentioned in the text. It is impossible to enumerate all possible objects of arresting. A text-wide resolution of identity anaphora and semantic parsing is needed to filter out bad bridges and select correct ones.

The system can make bridges that are irrelevant or redundant from a human point of view. For example, it can link words *зачетка (student's grade book)* and *дверь (door)*: *grade book* ^{hasOwner} *student* ^{isAgentOf} *opening* ^{hasObject} *door*. It is hard to make a filter that would prevent such cases. Such filter must introduce the notion of the reader's intention, i.e. what we want to learn from the text.

There is no good stopping rule in assumption generation, except to ban all concepts not explicitly mentioned in the text. In a story about hijacking of a car that results in a chase, crash and explosion, the computer will happily (mis-)assume an existence of a bomb and some terrorists, because the bag of concepts (*Automobile, Impacting, Explosion, Policeman, Criminal*) has enough similarity with the domain of terrorism. This problem can also affect humans when there is no sufficient context to rule out wrong assumptions.

6. Evaluation

We used two different methods to assess the performance of the bridging tagger. The standard approach, which relies on precision/recall measurement against a manually tagged test corpus, hit its limits and proved to be impractical for our project. It happened because the very nature of the modeled process implies high variability and individual bias.

The second evaluation, described in **section 6.2**, was based on manual expert assessment of the tagger output without a reference corpus. This method is better suited for evaluation of highly variable results, such as translation, which also records a particular interpretation of a text.

6.1. Standard approach

A pilot sample of a test corpus was made and tagged by 6 annotators. The sample consists of two short texts, 2,627 words in total, from the domains of shopping and cinema. We tried to follow formal rules similar to the ones implemented in the software, but inter-annotator agreement was so bad that we rejected the idea of making a larger corpus following the same procedure. Every annotator seemed to have a different set of associations. Out of 197 unique pairs of reference+antecedent words in the test material only 1 pair was universally accepted by all annotators and 148 pairs (75.1%) were chosen by only one person. Table 2 provides an overview.

Table 2: Percentage of detected bridges vs number of annotators sharing them

Annotators	1	2	3	4	5	6
Pairs %	75.1%	13.2%	7.1%	3%	1%	0.5%

In most cases when several annotators selected the same bridging pair with a semantically complex relation, they interpreted it differently. For example, three annotators expressed the relation between *покупатель (buyer)* and *магазин (store)* in the following three different ways: *1) buyer* ^{hasLocation} *store*, 2) *buyer* ^{isAgentOf} *buying* ^{hasFrame} *store*, 3) *buyer* ^{isRecipientOf} *retailing* ^{isFunctionOf} *store*. All three variants are correct and acceptable.

Identification of the words that represented referring and antecedent entities in the texts was much more uniform. 59% of the words were chosen by at least 3 annotators and 44% were chosen by more than 3 people.

This situation is well aligned with the theory of bridging explained in section 1. The Listener produces associations based on his unique background knowledge, prior information and current goals. Every instance of understanding, even by the same person, may follow a different path of associations. It is unrealistic to expect that several people will produce identical sets of bridging links.

6.1.1. Tagger perfomance

Given the same test data our bridging tool produced 532 candidate bridges. Comparison between the collective of human annotators and the program shows that the computer was able to tag 78 (39.5%) out of 197 referent+antecedent word pairs tagged by at least one human annotator. It compares favorably against the numbers of bridges found by each single human. Table 3 shows individual recall of human annotators and the computer.

Table 3: Number of detected bridges per annotator out of the total pool of 197 relevant pairs

Annotators	A	B	C	D	E	F	Computer
Bridges	22	23	35	46	72	84	**78**
%	11.1%	11.6%	17.7%	23.3%	36.5%	42.6%	**39.5%**

Only one of the annotators managed to find more bridges than the system. All semantic types ascribed by the computer to the 78 bridging relations it detected were correct.

The remaining 454 links that were not confirmed by human annotators still contained some valid bridges that were overlooked by all six annotators and a lot of plausible but wrong assumptions, i.e. relations that are "usually true" but in the given context they became false.

Since people could not collectively exhaust all possible ways to interpret the test texts using an open set of bridging relations and build a gold standard corpus, we decided to analyze output of the tagger as is.

6.2. Expert evaluation

The second evaluation used a new set of 10 short texts, covering the domains of banking, crime/terrorism and education, 3,649 words together. They were processed automatically with different settings of the tagger and the output was manually assessed by an expert in two rounds.

Round 1 was used to evaluate the plausibility of associations produced by computer. Each bridging link was marked "good" or "bad" without reading the text itself and considering only a pair of words/expressions linked and semantic type of the proposed bridging relation. The "bad" mark was given to bridges that contradicted the expert's knowledge of the world. This procedure evaluates the system's ability to make good assumptions and uncovers eventual defects of the knowledge base. Here are some common cases:

- Overly general classification of some lexical senses makes some relations look improbable. For example, the text about evacuation of Russian specialists from Iraq yielded the following bad assumption: *граждане SubjectPerson* ^{isAgentOf} *DepartureByAircraft* ^{hasInstrument} *Airplane* ^{hasOwner} *Human пекарь (baker).* It is correct to assume that a person may own an airplane, but it is highly unlikely that a common baker would be that person. Such situations are caused by the problem of general class labels in the domain descriptions, as mentioned in section 5. On the other hand, it is hard to justify existence of e.g. a special class of "people that are rich enough to own a plane" in a general purpose ontology.
- Lack of validation with complex reasoning while building multi-arc bridges. The bridge *президент (president) Human* ^{isRecipientOf} *Payment* ^{hasTerminalPlacePoint} *BudgetFund бюджет* is wrong, even though both its parts are feasible. The knowledge that a head of state does not personally receive payments to the state budget is not available. This is mitigated by the existence of another possible relation between the same words: *президент President* ^{isChiefOf} *Nation* ^{isOwnerOf} *BudgetFund бюджет*.

Despite such problems, the system demonstrated overall high quality of generated assumptions, no less than 85% and reaching 96%, depending on the settings of the bridging tagger (see 6.2.1).

Round 2 evaluated the same bridging relations again, but this time they were put in context. The expert had to carefully read the source text and decide, whether the assumptions were correct or contradicted the text contents. Each bridging link received a second "good" or "bad" mark. False bridges were further tagged according to the nature of the error. There are two important problems:

- Lexical disambiguation errors. One or both bridged words may be labeled with wrong lexical senses, which results in false assumptions. For example, the word "life" in

  Виктор Д. перепробовал в своей жизни много профессий.
  Victor D. tried many occupations in his life.

  got wrongly interpreted as *Human (as in "saved many lives").* This caused a false bridge **Human** ^{isOwnerOf} *PrivilegeAdvantage льготами (social benefits).* The correct sense label here is *Life* and it does not take the relation of ownership.
- Reference errors. Many texts contain several entities of the same semantic type and the tagger can falsely relate what is said about one of them with another. For example,

  В сельской школе N4 Аксайского района ... ребята осваивают компьютер просто играючи. В селе Покровском Неклиновского района школьники уже и сами разрабатывают учебные программы.
  Children easily acquire skills while working with computers … in the village school N.4 of Aksaisky district. In the settlement Pokrovskoe of Neklinovsky district pupils started to develop their own educational software.

- It is evident from the text that the computers used by the pupils of the two schools are different. Therefore, the bridging relation between pupils from Pokrovskoe and the computers from the school N.4 is false.
- Another possible case is a bridge between two members of the same corefence group, e.g. *Директор* ^{isChiefOf} *предприятие* ^{hasMember} *руководитель (Director* ^{isChiefOf} *enterprize* ^{hasMember} *manager),* where director and manager are the same person.
- Reference errors constitute 15–20% of all bridges deemed to be good assumptions at Round 1. A recent paper [11] by Pagel and Rösiger applies a partially similar rule-based approach to German and confirms positive effect of using coreference resolution. They report a 3.3% improvement of the F1 measure.

The first evaluation demonstrated that the recall of perfectly true bridges produced by the computer was on par with humans or better and the problem lies in its precision. Full evaluation of the tagger's output during round 2 ensures quality sufficient for a "gold standard" annotation. The resulting set of texts with labeled bridging relations following our approach and manual coreference annotation can be released as a small corpus, if there is some public interest.

6.2.1. Tuning parameters

The tagger has several tuning parameters that influence its performance. It is possible to adjust the length of the look-back window, which tells how many preceding sentences are taken into consideration while searching for antecedents. Another parameter is the maximum number of arcs, i.e. number of intermediate concepts used to explain the semantic relation between the reference word and its antecedent.

The tested window size range is 1–5. Narrowing the window improves the ratio of perfectly good bridges to failed hypotheses and lowers the number of found antecedents. **Figure 3** shows that windows of 4 and 5 sentences applied to our test set show almost equal number of detected bridges. Smaller windows of 3 and 2 sentences caused steep decline of that number.

The bridge length could be set to 1–3 arcs. "One" means that all referent-antecedent pairs must be connected by a single semantic relation. "Two" allows a single intermediate concept. "Three" provided two intermediate concepts, one of which may be an associated entity not actually mentioned in the text. Greater bridge length brings more freedom in constructing complex bridging relations, e.g. *actor* ^{isObjectOf} *makeup* ^{hasLocation} *dressing room* ^{isLocationOf} *mirror,* but increases the number of false bridges.

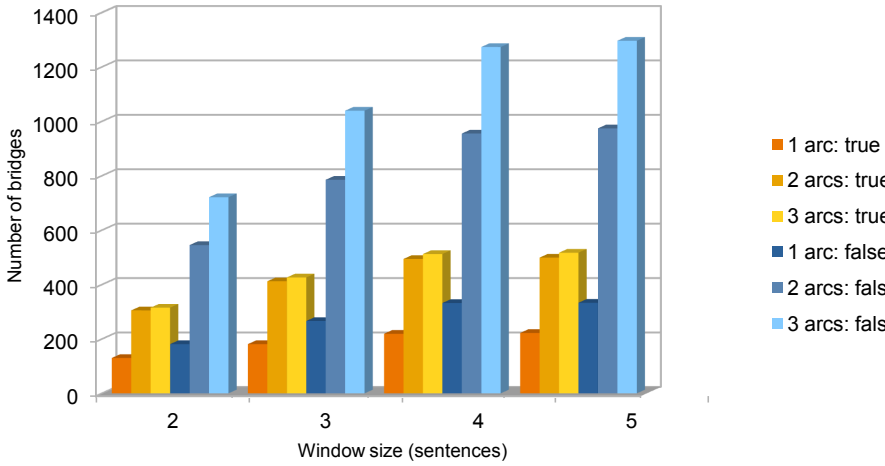


Figure 3: Number of true and false bridges in relation to 1) number of arcs and 2) antecedent search window size

The optimal balanced combination for a regular text seems to be window size 4 with 2 arcs. Three arcs may help when no bridges are found in the regular way. This can happen with a text that tries to say more with fewer words and leaves out more information than average.

6.2.2. Results

There are several combinations of settings that provide best results in one or another aspect. Table 4 sums them up. The F1 score calculation bears a special note because, as explained in section 6.1, we do not know the total number of all conceivable true bridges (relevant samples). The expert evaluation procedure does not consider any bridges except those generated by the tagger itself. Therefore the number of relevant samples used to determine the recall is always equal to the total amount of true bridges marked during round 2 (see section 6.2).

Comparing our results with other studies cannot be straightforward because of differences in methods and bridge identification criteria. In particular, paper [9] by Roitberg and Khachko reports F1 measure 0.65 for Russian with a completely different approach based on syntactic criteria that does not explain the semantics of the relation between reference words and antecedents and covers only nouns, while we also include verbs. Pagel and Rösiger apply a closer approach to German and report F1 measure of 11.1% (no coreference) to 14.1% (with corefence).

Table 4: Tagger performance using different settings

	Arcs	Window	% plausible bridges (round 1)	% true bridges (round 2)	F1	# true bridges	# false bridges
Highest recall	3	5	85.4	22.04	0.361	**801**	2561
Balanced	2	4	90.47	30.5	**0.44**	633	1390
Best precision	1	2	96.57	**37.32**	0.255	156	258

7. Conclusion

We develop an extensible semantic knowledge base geared towards bridging resolution. It opens up a possibility to explore semantic approaches in Russian and use richer background information than previous studies. All established bridging relations receive a semantic interpretation, which is not limited by a fixed set of predefined labels. Flexibility granted by combining multiple relations and intermediate concepts in "multi-arc" bridges allows to represent complex associations but creates problems mentioned in sections 5 and 6. Support for complex semantic relations is an important feature of our bridging tagger.

Most authors in the field narrow down the problem by imposing artificial constraints on the types of bridges they consider. For example, papers [6], [7] limit the relation types to part-whole. Paper [9] ignores semantic types but imposes a syntactic limitation. It greatly simplifies formal evaluation but results in ignoring most of the possible bridges in any text. Such works fail to cover the full scope of the studied phenomenon. We prefer to look at the problem in a more general and holistic way and build a model which covers wider range of possible implicit relations than previous studies following the semantic approach.

The mental process explained in **section 1** is always subjective and implies great variability. Different people see different relations because 1) they have different background knowledge (it includes education, prior experience, cultural bias, etc.), 2) different intentions and 3) the space of possible implicit relations is so vast that no one can exhaust it. In our study we cap variability which stems from factors 1 and 2. We look for relations that are based on an explicitly formulated knowledge base (KB) and follow explicitly defined rules. However, experiments showed that even this controlled space of possible relations is bigger than six expert annotators could collectively cover during the first evaluation. It is very difficult to make a "gold standard" corpus with a rich set of semantic relation types and complex "muti-arc" bridges, because annotators naturally produce different interpretations of the same text. There is no way to be certain that the reference tagging is complete and any other bridges in the same corpus will be wrong.

Output of our bridging tool is hard to rate using the traditional method which requires comparison with a reference corpus because bridging belongs to the class of problems that allow many alternative solutions, just like translation. This is why the alternative method of expert evaluation is more feasible.

We can confirm that the list of most useful features for bridging in paper [11] is true. It includes 1) semantic connectivity and 2) distance between reference word and its antecedent. It is also worth to explore the possibilities of populating the domain descriptions and ranking plausibility of different alternative antecedents by ML.

8. Acknowledgments

This study was supported by the Russian Science Foundation (grant No. 16-18-10422)

References

1. *Boguslavsky I. M., Dikonov V. G., Frolova T. I., Iomdin L. L., Lazurski F. V., Rygaev I. P., Timoshenko S. P.* (2016) Plausible Expectations-Based Inference for Semantic Analysis. Proceedings of the 2016 International Conference on Artificial Intelligence (ICAI"2016). USA: CSREA Press, 2016. pp. 477–483. ISBN: 1-60132-438-3.
2. *Clark, H. H.* (1975) Bridging. In Proceedings of the 1975 workshop on Theoretical issues in natural language processing, Association for Computational Linguistics, pp. 169–174.
3. *Dikonov V. G.* (2013) Development of lexical basis for the Universal Dictionary of UNL Concepts; Proceedings of the International Conference "Dialogue". Issue 12(19), Moscow, RGGU Publishers. P 212–221.
4. *Fan, J., Barker, K., & Porter, B.* (2005) Indirect anaphora resolution as semantic path search. In Proceedings of the 3rd international conference on Knowledge capture ACM, pp. 153–160.
5. *Hou, Y., Markert, K., & Strube, M.* (2013) Global Inference for Bridging Anaphora Resolution. In HLT-NAACL pp. 907–917.
6. *Lassalle, E., & Denis, P.* (2011) Leveraging different meronym discovery methods for bridging resolution in French. In Discourse Anaphora and Anaphora Resolution Colloquium, Springer Berlin Heidelberg, pp. 35–46.
7. *Poesio, M., Mehta, R., Maroudas, A., & Hitzeman, J.* (2004-B) Learning to resolve bridging references. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, p. 143.
8. *Recasens M., Martı M. A., Taulé M.* (2007) Text as scene: Discourse deixis and bridging relations. Procesamiento del lenguaje natural, 39:205–212.
9. *Roitberg A. M., Khachko D. V.* (2017) Bridging Anaphora Resolution for the Russian Language. Proceedings of the International Conference "Dialogue 2017", Moscow, May 31—June 3, 2017.
10. *Zikánová Š., Hajičová E., Hladká B., Jínová P., Mírovský J., Nedoluzhko A., Poláková L., Rysová K., Rysová M., Václ J.* (2015) Discourse and Coherence. From the Sentence Structure to Relations in Text, volume 14 of Studies in Computational and Theoretical Linguistics. Charles University in Prague, Praha, Czechia.
11. *Pagel J., Rösiger I.* (2018) Towards Bridging Resolution in German: Data Analysis and Rule-based Experiments. In Proceedings of the Workshop on Computational Models of Reference, Anaphora, and Coreference (CRAC), NAACL, New Orleans, US.

AN APPROACH TO CUSTOMIZATION OF PRE-TRAINED NEURAL NETWORK LANGUAGE MODEL TO SPECIFIC DOMAIN

Dudarin P. V. (p.dudarin@ulstu.ru),
Tronin V. G. (v.tronin@ulstu.ru),
Svyatov K. V. (k.svyatov@ulstu.ru)
Ulyanovsk State Technical University, Ulyanovsk, Russia

Nowadays the majority of tasks in NLP field are solved by means of neural network language models. These models already have shown state-of-the-art results in classification, translation, named entity recognition and so on. Pre-trained models are accessible in the internet, but the real life problem's domain could differ from the origin domain which the network was learned. In this paper an approach to vocabulary expansion for neural network language model by means of hierarchical clustering is presented. This technique allows to adopt pre-trained language model to a different domain. In the experimental part the proposed approach is demonstrated on specific domain of textual artifacts of software development process. This field is actively studied this days due the expensiveness of the process and its impact on the modern world and society.

Key words: NLP, language model, neural network, RNN, ULMFiT, transfer learning, clustering, fuzzy graph clustering, word-to-vec

ПОДХОД К АДАПТАЦИИ ПРЕДОБУЧЕННОЙ ЛИНГВИСТИЧЕСКОЙ МОДЕЛИ НА БАЗЕ НЕЙРОННОЙ СЕТИ К СПЕЦИФИЧЕСКОЙ ПРЕДМЕТНОЙ ОБЛАСТИ

Дударин П. В. (p.dudarin@ulstu.ru),
Тронин В. Г. (v.tronin@ulstu.ru),
Святов К. В. (k.svyatov@ulstu.ru)
Ульяновский Государственный Технический Университет, Ульяновск, Россия

В современном мире большинство задач обработки текстов (NLP) решаются при помощи лингвистических моделей на базе нейронных сетей. Эти модели уже показатели выдающиеся результаты в классификации, машинном переводе, извлечении именованных сущностей

и многих других. Предобученные модели доступны в сети интернет, но в практических задачах связанных с конкретными предметными областями используемый словарь может сильно отличаться от словаря на котором обучалась нейронная сеть. В данной работе представлен подход к адаптации предобученной лингвистической модели на базе нейронной сети к специфической предметной области. Помимо собственно алгоритма, представлен эксперимент демонстрирующий применимость предложенного подхода на примере предметной области процесса разработки программного обеспечения. Эта предметная область активно изучается ввиду высокой стоимости данного процесса и широкого влияния на современный мир и общество.

Ключевые слова: NLP, лингвистические модели на базе нейронной сети, нейронные сети, RNN, ULMFiT, перенос знаний, кластеризация нечетких графов, word-to-vec

1. Introduction

Code production is a complex and expensive process. Many resources are spent to get instruments of monitoring, control and prediction software development results. Many papers are dedicated to this theme for example in [15] an approach to project architecture analysis is proposed. Besides the code itself there are a lot of information produced during the software development process. For example, tasks are tracked in a task tracking system, where each issue could be commented and discussed by team members. During the code review phase the commit content is discussed by developers. All this information is textual, thus NLP methods are required to analyze it and extract knowledge about the effectiveness of communication among team members, about emotional condition of the team, specific relations between colleagues. This knowledge could be used to monitor software development process, predict quality and timing, reveal conflicts on the early stages.

Traditional methods in information retrieval [13] work well with large texts and text corpuses. But the most information generated during software process is presented as short sentences and short dialogues. Not only in software processing but also in many other domains, the short text processing is becoming a new trend [6]. Since the short text has rarely external information, it is more challenging than document [18]. To coupe with this task different clustering techniques are used [3], [20]. Each clustering procedure needs a similarity measure, like the one used in Serelex system [16] published in 2013. In 2015 word2vec as the most used technique to obtain this measure in NLP tasks was introduced [14].

Although the word embedding approach has shown good efficiency that is shown in [2], lately an approach of construction neural network language models get a leading position in NLP benchmarks [19], almost every state-of-the-art results are obtaining by means of neural networks. But the process of neural network learning is quite long and computationally expensive.

Besides there are a lot of task in specific domains where there is no opportunity to train special neural network. In this case the idea of transfer learning [10] looks very promising. Authors of ULMFiT propose using their universal architecture to train

language model and then to tune them for specific NLP tasks. But in ULMFit the tokens list is limited, authors recommend using up to 60,000 tokens. And as long as different word forms are treated as different tokens, ULMFiTs vocabulary is even more limited. On the contrary, modern word embedding models [11] have 250–400 thousand of lemmas. Word embedding technique being combined with thesaurus could demonstrate even higher performance[12]. In case of Russian language with its huge possible word forms language model approach allows to construct general purpose neural networks like casual phases generator only. And does not allow to include specific terms, neologism, swear words, rare used words and so on. In ELMo [4] and BERT [5] words are split into parts and then fed to neural network. But these models take a lot of calculation resources and could be afforded by huge corporations like Google. There are some multilingual pre-trained ELMo [9] and BERT models. But as for now they demonstrate very poor performance for Russian language. For example 'happy birthday to' really common phrase without double meaning could not be continued correctly by available models.

In this paper an approach to customization of pre-trained neural network language model to specific domain is proposed. This technique allows to process word outside the tokens list and thus to get benefits from transfer learning.

The rest of this paper is organized as follows. In **section 2** the detailed technique description is presented. **Section 3** shows an experimental results. And section 4 concludes the paper.

2. Language Model Customization

General idea of proposed approach is to add an extra layer of words pre-processing before the neural network language model (**Figure 1**).

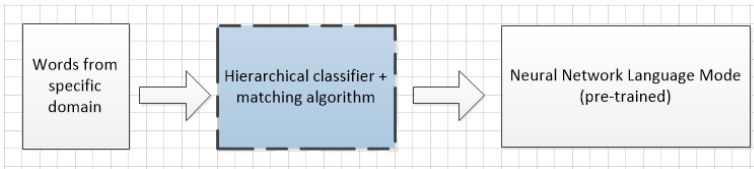


Figure 1: Additional pre-processing layer

This layer consists of two part: hierarchical classifier that groups words from neural network vocabulary and matching algorithm that matches new words to linear combination of words from vocabulary $new\_word = weight_1 * word_1 + ... + weight_N * word_N$.

2.1. Tokens Hierarchical Clustering

The first layer of neural network language model is an embedding layer which transforms one-hot encoded vectors into n-dimensional vectors of the embedding vectors space. Each coordinate of one-hot vector references to a word in a vocabulary of language model.

Lets define W_{lm}—a set of words included in tokens list of neural network. The task is to organize words from tokens list into a tree, where leaf nodes contain single word $w_i \in W_{lm}$, and other nodes are clusters that include all the words below in the hierarchy $w_{kj} \in C_k \subset W_{lm}$. $|W_{lm}| = N$.

This task could be completed by performing procedure which is a hierarchical modification [8] of $\epsilon - clustering$ [1], [17]. This procedure needs to be provided with a similarity measure for objects, let denote it as μ. There are a lot of pre-trained word embedding models for each language. This model provide a vector for each word and than the Euclidean or Manhattan distance could be calculated. In this paper the '*ruwikiruscorpora_upos_skipgram*_300_2_2019'[1] model was used.

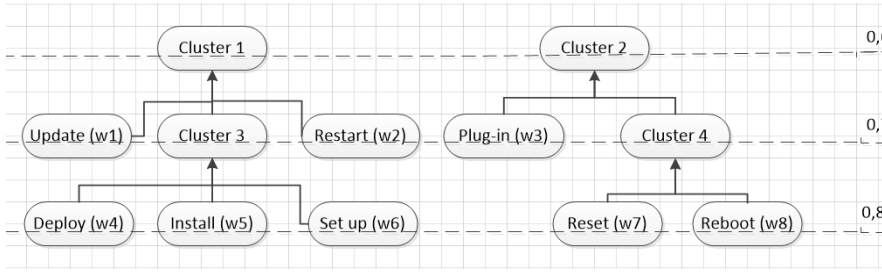


Figure 2: Hierarchical clustering sub-tree sample

One of the main advantages of graph based approach is its ability to be interpreted by human. The classifier could be easily modified by experts to add information domain specifics [7]. At least all the words that are not from domain vocabulary could be cut of the classifier. On the **Figure 2** a part of sample classifier is shown. This sub-tree consist of two main branches dedicated to software and hardware installation process. Each level has a number (ϵ) that indicates the step of hierarchical clustering procedure when these level was obtained and it means that all the branches on this level has mutual similarity less than ϵ.

Thus a hierarchical classifier wit additional layer information could be obtained.

2.2. Specific Domain Words Matching

The task of the matching step is to construct vectors for words from specific domain in order they could be processed by pre-trained neural network. These vectors should have N components, where N equals to amount of inputs of neural network $N = |W_{lm}|$.

For each word w there are two possible cases. The word is already included into language model tokens list $w = w_i \in W_{lm}$ and in this case corresponding vector $v = (0, 0, ..., 1, 0, ..., 0)$, where component with 1 has i index. Another case when word $w \notin W_{lm}$. In this case there are some possible strategies to get a vector form. The first one, and the most evident, is to replace a given word with the most similar one

[1] The model was downloaded from open resource https://rusvectores.org/ru/models.

according to similarity measure μ. Means, to choose $i, \mu(w, w_i) = max(w, w_j) \forall j \in [1, N]$. This strategy does not require any classifier, but it is not efficient when there are some equidistant words in the tokens list, especially when they are significantly differ in their semantic meaning. In order to have an alternative way of matching, in the experimental part the first strategy also included.

In general case proposed technique is following:

1. If $max(w, w_j) = \mu(w, w_i) > 0,9$[2] $\forall j \in [1, N]$ then $v = (0, 0, ..., 1, 0, ..., 0)$, with 1 on the i-th place.
2. Start with $\epsilon = 0,9$ and find all the words $W_{nn} = \{w_j | \mu(w, w_j) >= \epsilon, j \in [1, N]\}$. If $|W_{nn}| = 0$ then set $\epsilon = \epsilon - \delta\epsilon$. In this paper $\delta\epsilon = 0,05$, according to hierarchical clustering procedure specifics.
3. Get all the clusters $C_{nn} = \{c_j | \exists i\, w_i \in W_{nn}\}$ i.e. all the parent nodes in classifier for leaf nodes in W_{nn}.
4. Start with layer $l = 0,9$ and get all nodes from this layer $L_l = \{c_j | c_j \in C_{nn} \& layer(c_j) = l\}$. If $|L_l| > 2$[3] then change $l = l + \delta_l$ and move to the previous step. In this paper δ_l has been chosen as 0,05, according to hierarchical clustering procedure specifics.
5. For each node (cluster) define a weight according the distance to the cluster center. $weight = \mu(w, cluster\ center_i) / |\sum_j \in_{Ll} \mu(w, cluster\ center_j)|$
6. For each child node define weight the same as at the previous step and multiply to parent's weight $weight = parent\ weight * children\ weight$.
7. Stop when all the leaf nodes get weights. All the other weights are set to 0 $weight_i = 0 \forall i \notin W_{nn}$ As a result $v = (weight_1, weight_2, weight_3, ..., weight_N)$

This algorithm is illustrated on **Figure 3**. Firstly the similarity of word 'mount' to other words is calculated. The most similar words 'install', 'set up' and 'plug-in' were detected. Then layer by layer from the bottom to the top parent nodes are detected, until only 2 nodes left. Next, top-to-bottom process starts. Based on the distance to the cluster centers (0.8 and 0.71), node weights are calculated 0.53 and 0.47 respectively. And finally, weights for children nodes of 'cluster 3' are calculated. Thus a vector for word 'mount' will be (0, 0, 0.53, 0, 0.24, 0.23, 0, 0, ...).

Thus each word from the domain is converted into vectors in N-dimensional vector space.

Besides even words that are present in language model token list could be rematched to another words or set of words. This could be useful in case of word's meaning is changed significantly in the domain. For example: 'mount', 'branch', 'bug' in software development domain.

---

[2] The value $\epsilon = 0,9$ was obtained by experimental way and could differ from model to model.

[3] this threshold is heuristic and need to be surveyed more thoroughly in future studies

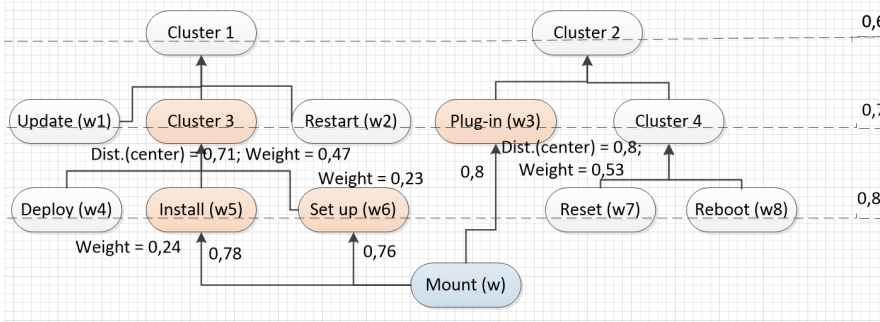


Figure 3: Matching process sample for word 'mount'

3. Experiment results

For experimental purposes pre-trained neural network language model for Russian language with architecture ULMFiT has been chosen.[4] This model has been trained on news portal (lenta.ru) and has perplexity 36,23.

All the most popular neural network language models take as an input sequence of words, to be more specific - sequence of words indexes in tokens list. This make difficult to use custom input vectors with pre-trained neural network. In this paper hard code solution was used: the 'fastai' library has been modified to change not used input components into hard coded vectors. More thorough and wide experiments will need to make changes into the core of neural network framework where the embedding function is located.

To show the technique some common phrases from developers chats is used (all the experiments were with Russian words, but translation is provided next to the phrases):

1. 'кто может **смонтировать** новый жесткий диск?' (translation: 'who can **mount** a new hard drive?')
2. 'тут есть **баг** и тебе нужно исправить его ' (translation: 'this part has a **bug** you need to fix it')
3. 'этот класс не может быть унаследован от этого **интерфейса**' (translation: 'this class could not be inherited from this **interface**')

The chosen language model does include 'mount' in common meaning and does not include words 'bug', 'interface' in its tokens list. The aim is to be able to proceed this sentences with pre-trained neural network.

The first step is to construct a hierarchical classifier. The input layer of the current network has 60,000 neurons. The resulting hierarchy has about 80,000 nodes, 60 levels. The part of hierarchy is shown on Figure 2.

The second step is to construct vectors for words that are absent in tokens list. Word 'смонтировать' ('mount') is related to words 'инициировать','установить' and

[4] https://github.com/ppleskov/Russian-Language-Model

‘подключить’ (’install’, ’set up’ and ’plug-in’). This case is shown on Figure 3. For the other two words:

1. ‘баг’ (’bug’): ‘неисправность’, ’ошибка’, ‘недочет’ (’failure’, ’error’, ’lack’)
2. ‘интерфейс’ (’interface’): ‘структура’,’правило’,’протокол’ (’structure’, ’rule’, ’protocol’)

Then the sentences could be processed by neural network language model. The first 3–5 generated words has been taken as an output result:

1. Input: ’who can **mount** a new hard drive?’. Output of 15 words contains: ‘сервер’, ‘процессор’, (’server’, ‘processor core’)
2. Input: ’this part has a **bug** you need to fix it’. Output of 15 words contains: ‘приложение’, ‘исправление’ (‘application’, ’patch’).
3. Input: ’this class could not be inherited from this **interface**’. Output of 15 words contains: ‘протокол’, ‘нарушение’ (’protocol’, ‘violation’)

Get some experiments with NN and compare results of two ways. Access perplexity of the resulting NN.

The results below were generated when one the most similar word has been used instead of vector calculation.

1. Input: ’who can **mount** a new hard drive?’. Output of 15 words contains: ‘монтаж’, ‘ремонт’ (‘installation’,’repair’)
2. Input: ’this part has a **bug** you need to fix it’. Output of 15 words does not contains any words related to software domain.
3. Input: ’this class could not be inherited from this **interface**’. Output of 15 words contains: ‘родственники’, ‘матери’ (‘relatives’, ’mothers’)

Neural network output in the first case uses the common word meanings and produces wrong context. In the second and third cases some words were just ignored and the context produced was based on insignificant word.

4. Conclusion

In this paper an attemt to apply transfer learning technique to special domains was made. The proposed approach allows to use not learned words with pre-trained neural network language model. It is important in domains with insufficient amount of texts to train custom language model or when the calculation resources are limited. Also this technique could be used to prototype and check ideas (hypothesis) before starting to teach custom language model.

The results have shown an application capability of proposed approach. But more thorough and wide experiments need to be done. These experiments will need to make changes into the core of neural network framework where the embedding function is located. This will allow to compute perplexity measure of proposed approach and comparison with the other approaches. Further studies will involve comparison of different neural network architectures within proposed approach, searchin a way of fine tuning the language model and comparison of effectiveness in different

NLP benchmarks. Besides it is important to develop extension to existing neural network frameworks to support not only a custom head but custom tails also.

5. Acknowledgements

The reported study was funded by RFBR and the government of Ulyanovsk region according to the research project № 18-47-732005 and by RFBR and the government of Ulyanovsk region according to the research project № 18-47-732004.

References

1. *Rosenfeld A,:* Fuzzy graphs. Fuzzy Sets and Their Applications to Cognitive and Decision Processes. Academic Press, New York. pp. 77–95. (1975).
2. *Arefyev, N. et al.:* How much does a word weigh? Weighting word embeddings for word sense induction. CoRR. abs/1805.09209, (2018).
3. *Avendaño, D. P. et al.:* Clustering abstracts of scientific texts using the transition point technique. In: Computational linguistics and intelligent text processing, 7th international conference, cicling 2006, mexico city, mexico, february 19-25, 2006, proceedings. pp. 536–546 (2006).
4. *Che, W. et al.:* Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In: Proceedings of the CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. pp. 55–64 Association for Computational Linguistics, Brussels, Belgium (2018).
5. *Devlin, J. et al.:* BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. (2018).
6. *Dudarin, P. et al.:* Methodology and the algorithm for clustering economic analytics object. Automation of Control Processes. In: Automation of Control Processes, Vol. 47, № 1. pp. 591–604 RGGU, Ulyanovsk, Russia (2017).
7. *Dudarin, P. et al.:* An approach to feature space construction from clustering feature tree. In: Kuznetsov, S. O. et al. (eds.) Artificial intelligence. pp. 176–189 Springer International Publishing, Cham (2018).
8. *Dudarin, P. V., Yarushkina, N. G.:* An approach to fuzzy hierarchical clustering of short text fragments based on fuzzy graph clustering. In: Abraham, A. et al. (eds.) Proceedings of the second international scientific conference “intelligent information technologies for industry” (iiti’17). pp. 295–304 Springer International Publishing, Cham (2018).
9. *Fares, M. et al.:* Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In: Proceedings of the 21st nordic conference on computational linguistics. pp. 271–276 Association for Computational Linguistics, Gothenburg, Sweden (2017).
10. *Howard, J., Ruder, S.:* Universal language model fine-tuning for text classification. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers). pp. 328–339 Association for Computational Linguistics, Melbourne, Australia (2018).

11. *Kutuzov, A., Kuzmenko, E.:* WebVectors: A toolkit for building web interfaces for vector semantic models. In: Ignatov, D. I. et al. (eds.) Analysis of images, social networks and texts: 5th international conference, aist 2016, yekaterinburg, russia, april 7–9, 2016, revised selected papers. pp. 155–161 Springer International Publishing, Cham (2017).
12. *Loukachevitch, N., Parkhomenko, E.:* Recognition of multiword expressions using word embeddings. In: Kuznetsov, S. O. et al. (eds.) Artificial intelligence. pp. 112–124 Springer International Publishing, Cham (2018).
13. *Manning, C. D. et al.:* Introduction to information retrieval. Cambridge University Press, New York, NY, USA (2008).
14. *Mikolov, T. et al.:* Efficient estimation of word representations in vector space. CoRR. abs/1301.3781, (2013).
15. *Nadezhda, Y. et al.:* An approach to similar software projects searching and architecture analysis based on artificial intelligence methods. In: Abraham, A. et al. (eds.) Proceedings of the third international scientific conference "intelligent information technologies for industry" (iiti'18). pp. 341–352 Springer International Publishing, Cham (2019).
16. *Panchenko, A. et al.:* Serelex: Search and visualization of semantically related words. In: Serdyukov, P. et al. (eds.) Advances in information retrieval. pp. 837–840 Springer Berlin Heidelberg, Berlin, Heidelberg (2013).
17. *Raymond, T. Y., Bang, S.:* Fuzzy relation, fuzzy graphs and their applications to clustering analysis. Fuzzy Sets and their Applications to Cognitive and Decision Pro-cesses. Academic Press. Pages 125-149. (1975).
18. *Tang, J. et al.:* Enriching short text representation in microblog for clustering. Frontiers of Computer Science. 6, 1, 88–101 (2012).
19. *Xu, J. et al.:* Self-taught convolutional neural networks for short text clustering. Neural networks : the official journal of the International Neural Network Society. 88, 22–31 (2017).
20. *Zhao, Q. et al.:* P.: keyword clustering for automatic categorization. In: In: 2012 21st international conference on pattern recognition (icpr). IEEE (2012. (2012).

GAPPING PARSING USING PRETRAINED EMBEDDINGS, ATTENTION MECHANISM AND NCRF

Emelyanov A. A. (login-const@mail.ru),
Artemova E. L. (echernyak@hse.ru)
Moscow Institute of Physics and Technology; Nationa Research University Higher School of Economics, Moscow, Russia

The article is devoted to the problem of automatic gapping resolution for the Russian language. We use BERT Language Model as embeddings with bidirectional recurrent network, attention, and NCRF on the top. Unlike other models these are using BERT, we apply BERT only as embedder without any fine-tuning. As a result, our implementation took second place in the AGRR-2019 competition.

Key words: gapping resolution, BERT embeddings, attention, NCRF

РАЗРЕШЕНИЕ ГЭППИНГА С ИСПОЛЬЗОВАНИЕМ ПРЕДОБУЧЕННЫХ ЭМБЕДДИНГОВ, МЕХАНИЗМА ВНИМАНИЯ И NCRF

Емельянов А. А. (login-const@mail.ru),
Артемова Е. Л. (echernyak@hse.ru)
МФТИ, Москва, Россия

Статья посвящена проблеме автоматического разрешения гэппинга для русского языка. Мы используем языковую модель BERT в качестве эмбеддингов с двунаправленной рекуррентной нейронной сетью, механизмом внимания и NCRF на верхнем слое сети. В отличие от других моделей, в которых используется BERT, мы применяем BERT только в качестве эмбеддингов без какого-либо дообучения. В результате наша реализация заняла второе место в конкурсе AGRR-2019.

Ключевые слова: разрешение гэппинга, BERT эмбеддинги, механизм внимания, NCRF

1. Introduction

In natural language along with the surface level (the material that we read or hear), there is deep structure. The deep structure can differ in many aspects from the surface. One of the cases is omition of repeating elements. If the elements can be unambiguously restored from the previous linguistic context, such procedure is called ellipsis. Our work touches upon one type of ellipsis, namely gapping. Gapping is omition of repeating predicate in coordinate (and probably subordinate) structure while its arguments remain expressed. Consider the example 'John likes tea, and Mary coffee', where the second clause lacks the predicate 'likes', but has two of its arguments ("Mary" and "coffee") [Schuster et al. 2017].

While having been studying theoretically for decades [Ross 1970], [Hankamer 1979], [Coppock 2001] the phenomenon still has been not illustrated with sufficient corpora which is a prerequisite for developing of automatic systems. In the framework of the Automatic Gapping Resolution for Russian competition (AGRR-2019) such corpus for the Russian language was presented.

The data consists labeled text sequences (see **Section 2.1** for detailes). So we decided to address this problem as sequence labelling task and predict gap label for each token in the input sentence. For the purpose of this paper, we consider neural network solution for automatic gapping resolution for Russian language in proceedings of AGGR-2019 challenge [Ponomareva et al. 2019]. Our solution based on BERT language model [Devlin et al. 2018], use bidirectional LSTM [Hochreiter, Schmidhuber 1997], Multi-Head Attention [Vaswani et al. 2017], NCRFpp [Yang et al. 2018] (being neural network version of CRF++framework for sequence labelling) and Pooling Classifier (for classification) on the top.

2. Task description

2.1. Data format

Input data consists of sentences without any additional markup (raw texts). For each sentence output should contain 7 columns. First column should have 0 or 1 in it, depending on presence of gapping construction in the sentence. Other output cells separated with tab symbol correspond gapping element names (cV, cR1, cR2, V, R1, R2) and should contain char offsets (first symbol in each sentence has offset 0 1) for annotation borders (two numbers separated by colon (:) symbol) for each gapping element. If the provided sentence lacks certain gapping element, the corresponding cell should not contain any symbols. For example: "Аналогичным образом, среднегодовой прирост ВВП на душу населения, который в странах, расположенных к югу от Сахары, составлял в период с 1965 по 1973 год 3 процента, cV [упал cV] cR1 [с 1980 до 1986 года cR1] cR2 [на 2,8 процента cR2], R1 [в 1987 году R1] — V[] R2 [на 4,4 процента R2] и R1 [в 1989 году R1] — V[] R2 [на 0,5 процента R2]". For the binary presence-absence classification for each sentence all the output cells except the first one are ignored. For gap resolution task cells in columns cR1, cR2, R1, R2 are ignored. For the full annotation task all output cells are evaluated [Ponomareva et al. 2019].

2.2. Tasks

The task contains three parts [Ponomareva et al. 2019]:

1. Binary presence-absence classification. For each sentence decide if there is a gapping construction in it.
2. Gap resolution. Predict the position of the elided predicate and the correspondent predicate in the antecedent clause.
3. Full annotation. In the clause with the gap predict the linear position of the elided predicate and annotate its remnants. In the antecedent, clause finds the constituents that correspond to the remnants and the predicate that corresponds the gap.

For more details about data and task see description on github[1].

3. System description

We propose modeling the task as both sequence labeling and classification jointly with a neural architecture.[2] The system's architecture is shown in **Figure 1** and consists of seven parts:

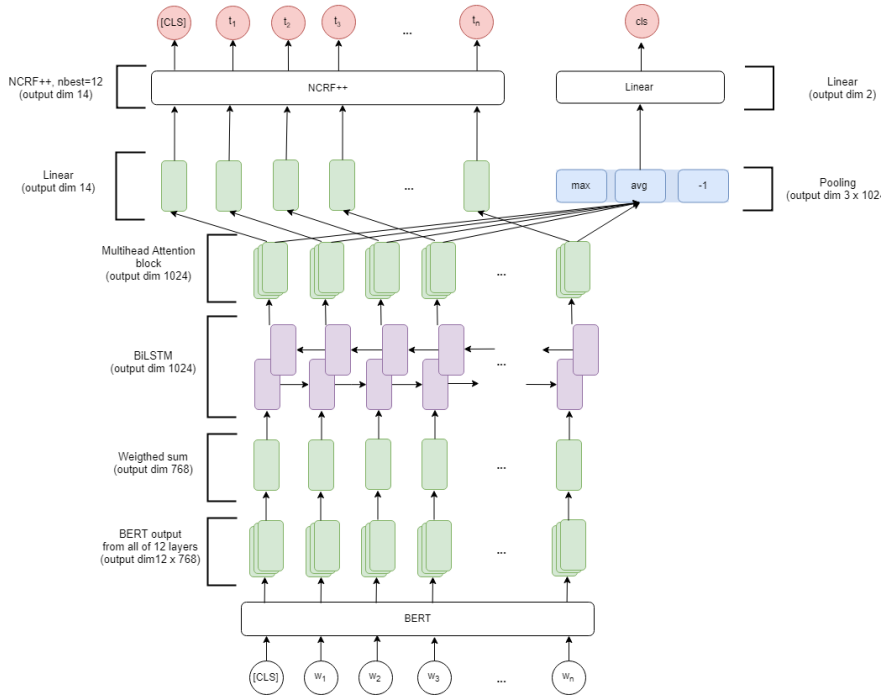


Figure 1: The system architecture

[1] Full AGRR-2019 description available at https://github.com/dialogue-evaluation/AGRR-2019.

[2] Code is available at https://github.com/king-menin/AGRR-2019.

1. BERT Embedder;
2. Weighted aggregation of BERT output;
3. Recurrent BiLSTM layer;
4. Multi-Head Attention;
5. linear layer;
6. NCRF++ inference layer for sequence labelling;
7. Concatenation operation of Max Pooling, Average Pooling and last output of Multi-Head Attention layer, later passed to linear layer for classification.

3.1. Neural network architecture

3.1.1. BERT Embedder

The BERT embeddings layer contains Google's original implementation of BERT language model. Each sentence is preprocessed as described in BERT paper [Devlin et al. 2018]:

1. Process input text sequence to WordPiece embeddings [Wu et al. 2016] with a 30,000 token vocabulary and pad to 512 tokens.
2. Add first special BERT token marked "[CLS]".
3. Mark all tokens as members of part "A" of the input sequence.

Entities labels converted to format IOBX format. For example:

JimHen##sonwasapuppet##eer

B-PERI-PERXOOOX

But instead of BERT's original paper [Devlin et al. 2018] we keep "B" ("Begin") prefix for labels and do a prediction for "X" labels on training stage. BERT neural network is used only to embed input text and don't fine-tune on the training stage. We freeze all layers except dropout here, that decreases overfitting.

We take hidden outputs from all BERT layers as the output of this part of the neural network and pass to the next level of the neural network. So the shape of output is 12×768 for each token of 512 length's padded input sequence.

3.1.2. BERT weighting

Here we sum all of BERT hidden outputs from previous part:

$$o_i = \gamma \times \sum_{i=0}^{m-1} b_i s_i \tag{1}$$

where

- o_i is output vector of size 768;
- $m = 12$ is the number hidden layers in BERT;
- b_i is output from i BERT hidden layer;
- γ and s_i is trainable task specific parameters.

Because we do not fine-tune BERT, we should adapt it outputs for our specific sequence labeling task. The suggested weighting approach is similar to ELMo [Hochreiter, Schmidhuber 1997], but with a lower number of weighting vectors parameters s_i.

3.1.3. Recurrent part

This part contains two LSTM networks for forward and backward passes with 512 hidden units so that the output representation dim is 1024 for each token. We use a recurrent layer for learning long time dependencies in an input sequence [Vaswani et al. 2017].

3.1.4. Multi-Head Attention

After applying the recurrent layer, we should learn any other dependencies in a sequence for each token. We can achieve this result with Self-Attention (**Figure 2**).

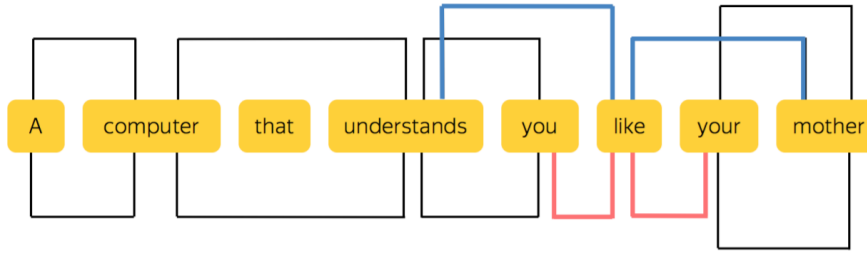


Figure 2: David Talbot's example of Self-Attention

That can be formulated as $D(d_h|S)$, where D is some hidden dependency; d_h is the h head of attention, and S is all sequence. In our architecture, we use Multihead-Attention block as proposed in the paper "Attention is all you need" [Vaswani et al. 2017]. We took 3 heads and value and key dim 64.

3.1.5. Inference for sequence task

After the input sequence was encoded, we gave the final representation of each token in a sequence. This representation is passed to Linear layer with *tanh* activation function and gets a vector with 14 dim, that equals to the to the number of entities labels (include supporting labels "pad" and "[CLS]"). The inference layer takes the extracted token sequence representations as features and assigns labels to the token sequence. As the inference layer, we use Neural CRF++ layer instead of vanilla CRF. That captures label dependencies by adding transition scores between neighboring labels. NCRF++ supports CRF trained with the sentence-level maximum log-likelihood loss. During the decoding process, the Viterbi algorithm is used to search the label sequence with the highest probability. But also, NCRF++ extends the decoding algorithm with the support of *nbest* output [Yang et al. 2018].

3.1.6. Inference for classification task

For the classification inference, we use Pooling Linear Classifier block as proposed in ULMFiT paper [Howard, Ruder 2018]. We pass output sequence representation H from Multihead-Attention part to different Poolings and concat (as shown in **Figure 1**):

$$h_c = [h_T, maxpool(H), meanpool(H)] \quad (2)$$

where [ ] is concatenation; h_T is last output significant vector of Multihead-Attention part (which does not have "pad" label).

The result of concat Pooling (3×1024) is passed to Linear layer, and that predicts binary classification.

3.2. Postprocessing prediction

After getting labels for the sequence of WordPiece tokens, we should convert prediction to word level labels in sequence labeling task. Each WordPiece token in the word is matched with neural network label prediction. We use ensemble classifier on labels by count all predicted labels for one word except "X" and select label for a word with the higher number of votes.

For the final prediction we have two strategies of making full gapping annotation:

1. Before the submission deadline: in this submission we don't use classification result for sequence label prediction, but train joint model and select classification result by the following rule: if any word's predicted labels of all sequence contains any label except "O", "[CLS]" and "pad" mark sample as 1 (has gapping), 0—otherwise. Sequence labels are not changed.

2. After the submission deadline: first, classification result is taken into account and all words in sequence are marked as "O" , next if binary classification prediction is not 0 (gapping present) and predicted sequence labeling are returned.

For the Full annotation task, we should predict "V" label. We use the following rule because in data all labels "V" have zero length representation in text: mark word with "V" label if the word was marked by "R2". If the word wasn't marked, mark word with "V" label if the word was marked by "R1".

4. Training the system

4.1. Data conversion

Because train, dev and test datasets contain symbolwise markup, but BERT take words sequence as input we convert datasets to word level IOB markup [Ramshaw, Marcus 1995]. After that, each word was tokenized by WordPiece tokenizer and word label matched with IOBX labels.

On the prediction stage result, labels were received by voice classifier as described in **section 2.2**. After this, we transform word predictions to symbolwise markup.

4.2. Training Procedure

The proposed neural network was trained with joint loss:

$$L = L_{SL} + L_{clf} \tag{3}$$

where L_{SL} is maximum log-likelihood loss [Yang et al. 2018] for the sequence labeling task and L_{clf} is Binary Cross Entropy Loss for the classification task.

We use Adam with a learning rate of $1e-4$, $\beta_1 = 0.8$, $\beta_2 = 0.9$, $L2$ weight decay of 0.01, learning rate warmup, and linear decay of the learning rate. Also, gradient clipping was applied for weights with $clip = 5.0$.

Training of proposed neural network architecture was performed on one GPU with the batch size equal to 16, the number of epochs equal to 100, which required only around 5 GB of memory instead of fine-tuning all BERT model, which would have required more than 8 GB GPU memory. All training procedure lasted around five hours on one GPU with the evaluation of dev set on each epoch.

The final model was trained on train and dev datasets.

5. Results and discussion

5.1. Evaluation results

The evaluation of the training stage was produced on dev dataset. **Table 1** shows word-level metrics precision, recall, and f1-measure. The evaluation metric of AGGR-2019 [Ponomareva et al. 2019] competition is symbolwise and measured by organizations evaluation script. For dev set, we obtained the following scores: Binary classification quality (f1-score): 0.958 and Gapping resolution quality (symbol-wise f-measure): 0.958. This difference in word and symbolwise is because words prediction isn't used classification results.

Table 1: Dev dataset evaluation metrics

label	precision	recall	f1-score	support
B_O	0.98	0.98	0.98	74,268
B_{R1}	0.95	0.92	0.94	1,500
I_{R1}	0.93	0.92	0.93	1,769
B_{R2}	0.94	0.95	0.94	1,473
I_{R2}	0.94	0.91	0.93	3,255
B_{cR1}	0.86	0.85	0.86	1,382
I_{cR1}	0.86	0.86	0.86	2,022
B_{cR2}	0.89	0.90	0.89	1,355
I_{cR2}	0.87	0.92	0.89	2,395
B_{cV}	0.96	0.94	0.95	1,382
I_{cV}	0.00	0.00	0.00	1
avg / total	0.832	0.831	0.833	90,802

The evaluation of test dataset presented in Table 2. In this submission, we do not use classification result. After deadline submission takes into account classification result and marks all words in sequence as "O" than Binary Classification prediction is 0 (no gapping) and select predicted word labels otherwise. This difference in evaluation metrics means that single neural network architecture (for Gapping Resolution only) is overfitted on "O" label.

Table 2: Test dataset evaluation metrics

	binary			gap resolution	full
Team	precision	recall	f-measure	f-measure	f-measure
fit$_p$redict	0.969	0.950	0.959	0.900	0.892
EXO (**ours**) after deadline	0.946	0.946	0.946	0.859	0.836
EXO (**ours**) before deadline	0.899	0.964	0.931	0.815	0.786
Koziev Ilya	0.774	0.903	0.834	0.677	0.647
Derise	0.801	0.906	0.850	0.665	0.622
Meanotek	0.891	0.781	0.832	0.635	0.514
МГУ-DeepPavlov	0.934	0.644	0.762	0.600	0.588
Vlad	0.778	0.915	0.841	0.574	
MorphoBabushka	0.763	0.619	0.683	0.466	0.440
nsu-ai	0.485	0.123	0.196	0.037	0.036

5.2. Error analysis

First of all, we have some errors with converting from origin data format (symbolwise markup) to word markup and back to origin after prediction. For example with extra spaces, bad Unicode symbols and there are some symbols, which are absent in WordPiece vocabulary.

Another error connected with neural network prediction mistakes. Even though we use classification results in after deadline submission network was overfitted on label "O" and there are many false positives in prediction as shown in Section 4.

The last kind of errors connected with the bad learned structure of gapping.

6. Related work

The related work has several parts: first, the phenomena of gapping has received some attention of the NLP community recently, see [Schuster et al. 2017], [Park 2016]. Secondly, our work follows the recent trend of using trained neural languages models, such as [Devlin et al. 2018], [Peters et al. 2017], [Howard, Ruder 2018]. Thirdly we model the task of gapping resolution as a joint sequence labeling and classification task following other joint architectures [Liu, Ian 2016], [Nguyen et el. 2016].

7. Conclusion and future work

We have proposed neural network architecture that solves three tasks (described in **Section 1**) without any additional data. However, it heavily exploits GPU memory cost on train and prediction steps. Our method took second place on AGGR-2019 competition [Ponomareva et al. 2019]. This neural network architecture can be used for other tasks, that can be reformulated as a sequence labeling task for Russian or any other language (listed in BERT documentation [Devlin et al. 2018]).

As improvements of the system, we can fine-tune BERT embeddings and try to do different layers after BERT or pass other modern language models as an input.

Acknowledgements

We are grateful to Sesame Street for their fruitful inspiration.

References

1. *Elizabeth Coppock* (2001). Gapping: In deffence of deletion.
2. *Jorge Hankamer* (1979). Deletion in coordinate structuresn.
3. *John Robert Ross* (1970). Gapping and the order of constituents. In Manfred Bierwisch and Karl Erich Heidolph, editors, Progress in Linguistics. De Gruyter, The Hague.
4. *Sebastian Schuster, Matthew Lamm, Christopher D. Manning* (2017). Gapping Constructions in Universal Dependencies v2. Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017), pages 123–132, Gothenburg, Sweden.
5. *Ponomareva M. Smurov I., Shavrina T. O., Droganova K., Bogdanov A.* (2019) AGRR: Automatic Gapping Resolution for Russian. https://github.com/dialogue-evaluation/AGRR-2019.
6. *Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova* (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Google AI Language.
7. *Hochreiter, S. and Schmidhuber, J.* (1997). LSTM can solve hard long time lag problems. In Advances in Neural Information Processing Systems 9. MIT Press, Cambridge MA. Presented at NIPS 96.
8. *Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin* (2017). Attention Is All You Need. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.
9. *Jie Yang, Shuailong Liang, Yue Zhang* (2018). Design Challenges and Misconceptions in Neural Sequence Labeling. Proceedings of COLING 2018, Santa Fe, New Mexico, USA.
10. *Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc VLe, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al.* (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

11. *Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer* (2017). Deep contextualized word representations. arXiv:1802.05365.
12. *Jeremy Howard, Sebastian Ruder* (2018). Universal Language Model Fine-tuning for Text Classification. arXiv:1801.06146v5.
13. *Lance A. Ramshaw, Mitchell P. Marcus* (1995). Text Chunking using Transformation-Based Learning. arXiv:cmp-lg/9505040v1.
14. *Sang-Hee Park* (2016). Towards a QUD-Based Analysis of Gapping Constructions. Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers.
15. *Bing Liu, Lane Ian* (2016). Attention-based recurrent neural network models for joint intent detection and slot filling. arXiv:1609.01454.
16. *Thien Huu Nguyen, Kyunghyun Cho, Ralph Grishman* (2016). Joint event extraction via recurrent neural networks. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

TRACING CULTURAL DIACHRONIC SEMANTIC SHIFTS IN RUSSIAN USING WORD EMBEDDINGS: TEST SETS AND BASELINES

Fomin V. (wadimiusz@gmail.com),
Bakshandaeva D. (dbakshandaeva@gmail.com),
Rodina Ju. (julia.rodina97@gmail.com)
National Research University Higher School of Economics, Moscow, Russia

Kutuzov A. (andreku@ifi.uio.no)
University of Oslo, Oslo, Norway

The paper introduces manually annotated test sets for the task of tracing diachronic (temporal) semantic shifts in Russian. The two test sets are complementary in that the first one covers comparatively strong semantic changes occurring to nouns and adjectives from pre-Soviet to Soviet times, while the second one covers comparatively subtle socially and culturally determined shifts occurring in years from 2000 to 2014. Additionally, the second test set offers more granular classification of shifts degree, but is limited to only adjectives.

The introduction of the test sets allowed us to evaluate several well-established algorithms of semantic shifts detection (posing this as a classification problem), most of which have never been tested on Russian material. All of these algorithms use distributional word embedding models trained on the corresponding in-domain corpora. The resulting scores provide solid comparison baselines for future studies tackling similar tasks. We publish the datasets, code and the trained models in order to facilitate further research in automatically detecting temporal semantic shifts for Russian words, with time periods of different granularities.

Key words: word embeddings, diachronic semantic shifts, temporal language change

КУЛЬТУРНО ОБУСЛОВЛЕННЫЕ ДИАХРОНИЧЕСКИЕ СДВИГИ ЗНАЧЕНИЙ СЛОВ В РУССКОМ ЯЗЫКЕ: ТЕСТОВЫЕ СЕТЫ И БАЗОВЫЕ ДИСТРИБУТИВНЫЕ АЛГОРИТМЫ

Фомин В. (wadimiusz@gmail.com),
Бакшандаева Д. (dbakshandaeva@gmail.com),
Родина Ю. (julia.rodina97@gmail.com)
НИУ Высшая школа экономики, Москва, Россия

Кутузов А. (andreku@ifi.uio.no)
Университет Осло, Осло, Норвегия

В статье представлены размеченные вручную датасеты для тестирования систем автоматического отслеживания изменений значений слов в русском языке. Датасеты взаимодополняемы: первый из них посвящён семантическим изменениям прилагательных и существительных в русском языке советского периода в сравнении с периодом до 1918 года, тогда как второй связан с тонкими семантическими изменениями, происходившими между годами с 2000 по 2014. Кроме того, второй датасет предлагает гранулярную классификацию степени изменения, однако содержит только прилагательные.

Создание этих датасетов позволило оценить качество работы нескольких распространённых алгоритмов определения семантических сдвигов, многие из которых ранее не применялись к русскому языковому материалу. Все использованные алгоритмы основаны на дистрибутивных моделях, обученных на соответствующих корпусах, саму задачу оценки алгоритмов можно отнести к задачам классификации. Итоговые оценки могут считаться базовым уровнем (baseline) для сравнения с будущими подходами. Датасеты, программные реализации использованных алгоритмов и обученные дистрибутивные модели выложены в открытом доступе. Мы надеемся, что это поможет будущим исследованиям в области автоматического отслеживания семантических сдвигов в русском языке с различной временной гранулярностью.

Ключевые слова: дистрибутивная семантика, диахронические семантические сдвиги, изменения языка во времени, датасеты

1. Introduction

In any language, words undergo semantic shifts over time, that is, they acquire new meanings and lose the old ones. The sources and nature of these shifts can vary. For example, words can completely change what they mean (cf. the Russian adjective 'красный' shifting from '*beautiful*' to '*red*'[1]). Sometimes, these shifts are linked to global processes of language development, and the granularity of the involved time spans is often very large-scale: we are talking about centuries or more.

Another type of diachronic semantic shifts is related to changes which occur on comparatively small time spans (decades or even years). The traditional classification in [25] assigns them to the category of *substitutions*: changes that have non-linguistic causes, for example that of technological progress. Such shifts often do not radically change the core meaning of a word, but instead significantly restructure the sets of typical associations which the word triggers in the speakers' minds. Cf., for example, the metonymical changes in the semantic structure of the Russian word 'Болотная [площадь]' '*bolotnaya [square]*', when after the 2012 mass protests, its meaning has widened to include not only the toponym in Moscow, but also the protest movement itself. These 'cultural shifts' are sometimes defined as socially and culturally determined changes in the people's associations for a given word [8].

This definition plays well with the so called distributional hypothesis [6]. In the present paper, we treat word meaning (or lexical semantics) as a **function of words' contexts in natural texts**. It is important that there exist many other linguistic theories and definitions of what meaning is: for example, one can postulate that the word meaning equals to the definition of this word in a well-established dictionary. We do not claim superiority of any of these theories; however, in our work, we stick to the distributional one. Thus, for us semantic shifts are primarily changes in the word's typical contexts and associations. The task of automatic discovery of diachronic semantic shifts using data-driven methods (especially distributional semantic models) is becoming increasingly popular in contemporary computational linguistics: see [18] and [26] for only a few of the recent surveys. However, to the best of our knowledge, there is only one publication applying these methods to Russian material [16]; it is behind the paywall, and thus not publicly accessible.

The presented paper aims to establish at least some foundation for further studies in automatically tracing temporal semantic shifts (of different granularities) for Russian. Correspondingly, our main **contributions** are as follows:

1. we release a manually annotated dataset of socially and culturally determined semantic shifts in Russian adjectives, based on news texts published from 2000 to 2014, covering many important social and political events;
2. we re-package in a more machine-readable and coherent form the dataset of Russian nouns and adjectives which shifted their meaning in the Soviet period [16];
3. several well-established algorithms of tracing diachronic semantic shifts using word embeddings are tested on the aforementioned datasets; this provides publicly available baseline scores for this task, when applied to Russian.

[1] Vasmer's Etymological Dictionary of the Russian Language, 1986

The rest of the paper is structured as follows. In **Section 2** we put our work in the context of the related research, both globally and in Russia. **Section 3** describes the process of annotating the presented gold standard datasets. **Section 4** introduces the corpora we employed to test our hypothesis and to train distributional semantic models, in its turn described in **Section 5**. We evaluate the algorithms of semantic shifts detection in **Section 6**, and in **Section 7** we discuss the results and conclude.

2. Previous work

Linguistics has a long story of studying diachronic semantic shifts. Just like words change, methods and sources used to trace changes in word's meaning and goals set in such research also evolve. The hand-picking approach [27] was historically the first, but then, with the expansion of corpus and computational linguistics, distributional semantics in particular, the data-driven approach was brought to the foreground. Time spans used to compare word's usage were initially as large as centuries [24], but then have begun to shrink to decades [7] and even years [14, 19]. Word embedding models, which appeared to be the most efficient tool in the field, also vary. For example, there are models that are trained separately on the corresponding time spans and aligned using specific methods, such as 'local anchors' [30] or orthogonal Procrustes transformation [9]. In other cases, word embeddings are trained jointly across all time periods, like in dynamic models of [1] and [28]. In yet other approaches, the models are trained incrementally, so that the model for every next time period is initialised with the embeddings from the previous period [14]. As for the data, in the beginning extensive and inclusive corpora were used, which primarily contain book texts [7, 12], but later researchers have taken interest in more specific corpora, for instance, of newspapers [3, 10, 28].

Unfortunately, in most cases only British and American English material is analysed, while there are much less studies dedicated to other languages [26]. As postulated in [18], the lack of studies in which existing methods are applied to other languages apart from English is among top priority open challenges in the field of detecting diachronic semantic shifts.

For temporal semantic shifts studies in Russian, [4] provides an inspiring linguistic foundation. Examining the corpus[2] (exploring the contexts in which the particular word appears and counting the number of word's uses for each period) gave insights about transformations of meaning which words undergo. Unfortunately, the words were once again hand-picked and analysed manually, and thus the whole book covers only 20 cases. However, this research served as a starting point for the only (known to us) publication which tests word embedding models on Russian data to detect language change [16]. In it, gold datasets are constructed in two different ways (by applying 'pseudowords' technique and deriving changed words from [4] and [21]), and then used to evaluate some basic algorithms that had been tested before for English. [16] conclude that Kendall's τ and Jaccard distance 'remain the most relevant for detecting semantic shifts'. In the presented work, we continue experiments with word

[2] In this case the Russian National Corpus (RNC), texts from the XIX and the XX centuries.

embedding models. Our results, however, are somewhat different; see **Section 6** for details. In fact, the evaluation setup in [16] seems rather convoluted; we propose more straight-forward and practically applicable approach. Additionally, we take into account the intensity of temporal semantic shifts.

It is important to draw the distinction between two types of semantic change: cultural shifts and linguistic drift [8]. They are linked in complex ways to two different families of data-driven methods to measure semantic change: 'global' (identifying the movement of word's vector in the whole semantic space while comparing two time periods) and 'local' (detecting variations in comparatively short lists of word's nearest semantic neighbours). After considering the decades from 1800 to 1990, [8] come to the conclusion that 'global' measures are appropriate when dealing with linguistically driven changes, whereas 'local' measures are better choice for capturing social and cultural shifts. This distinction further raises the question of how it is related to the granularity of time periods under analysis. [18] states that 'corpora with smaller time spans are useful for analysing socio-cultural semantic shifts', whereas the study of linguistically motivated semantic changes requires bigger time bins. We have used both global and local measures to compare the semantics of words as used in Russian corpora from different time periods.

3. Datasets

We present two gold datasets:

1. **Macro**: a list of manually chosen Russian words that have undergone semantic shifts from the pre-Soviet times through the Soviet times; this dataset is borrowed from [16].
2. **Micro**: a newly created manually annotated dataset of Russian adjectives that have undergone cultural semantic shifts in the years from 2000 to 2014.

The **Macro** dataset is a list of 215 words. 43 of them (38 nouns and 5 adjectives) are 'target' and labelled as having changed their meaning from pre-Soviet times through Soviet times, based on the linguistic research from [21] and [4]. Additionally, there are 4 fillers (words belonging to the same part of speech and the same frequency tier, randomly sampled from the Russian National Corpus) per each target word ($38 \times 4 = 152$ nouns and $5 \times 4 = 20$ adjectives). The words that have undergone semantic shifts are tagged with the class label 1, and the fillers are labelled with 0.

The **Micro** dataset (manually annotated in this work) also contains human judgements about how much the meaning of a word has shifted over a given time. However, it is limited to adjectives only[3]. To create it, we iterated over consequent pairs of yearly news texts corpora from 2000 to 2014 (2000 — 2001, ..., 2013 — 2014), and the word embedding models trained on them (see more in Section 5). To 'seed' the dataset with possible candidates for diachronic semantic shifts, for each pair of years we produced

[3] Adjectives are less studied in diachronic research, but at the same time present an interesting case of words less susceptible to noisy fluctuations of meaning. Note that the corresponding datasets for nouns or verbs can be easily constructed following the same workflow. We leave this for future work.

10 adjectives which shifted most according to the Global Anchors algorithm [29] (see more on it in **Section 6**). This gave us 140 adjectives with supposedly high ratio of real shifts. Then, for each of these 'seed samples', one randomly sampled adjective from the same corpus and the same frequency tier was added to the dataset as a filler. Additionally, about 20 of the seed samples were manually replaced by random fillers, mostly because of PoS tagging errors.

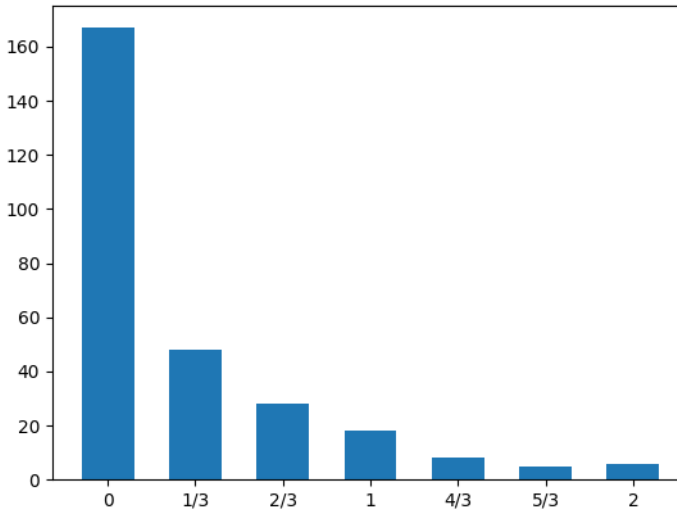


Figure 1: Distribution of average scores in the '**Micro**' dataset

3 human annotators independently labelled the resulting dataset of 280 adjectives from 14 year pairs. To help making the decision, each annotator was provided with a random sample of sentences containing the given adjective in the news texts from both years in the current pair. Note that during the annotation, there was no difference between the seed words and the fillers: annotators made their decision without knowing the source of adjectives. Each adjective was annotated with one of three labels:

- 0 (meaning not changed)
- 1 (meaning somewhat changed)
- 2 (meaning significantly changed)

After the first annotation round, there was a brief reconciliation, where all 3 annotators discussed the strongest disagreement cases (about 15 words out of total 280). After this reconciliation, the inter-rater agreement as measured by Krippendorff's Alpha [15] was 0.62. Considering the complexity and ambiguity of the task at hand, we believe this score to be quite high[4].

[4] The dataset is available online at https://github.com/wadimiusz/diachrony_for_russian

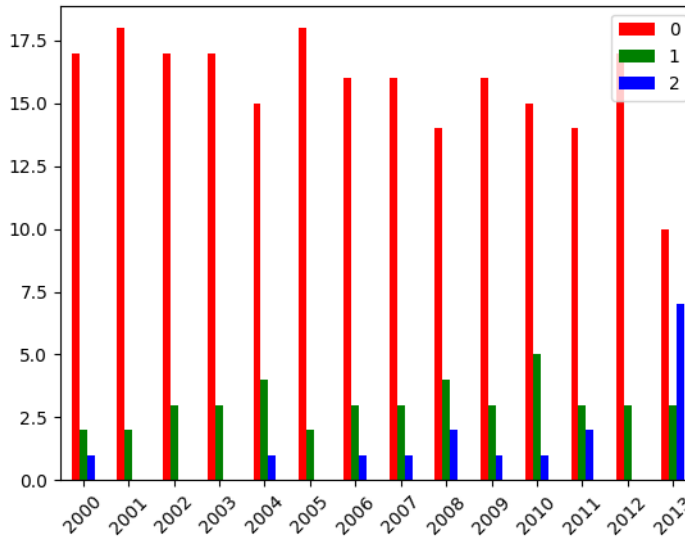


Figure 2: Distribution of class labels in the '**Micro**' dataset by years

The average score for each sample (over 3 annotators) could be 0, $\frac{1}{3}$, ..., $\frac{5}{3}$, 2, seven options overall. The resulting quantized class of each word was the arithmetic mean of the three scores rounded to the nearest integer. The distribution of the mean scores over the whole dataset (before quantization into 3 classes) is shown in Figure 1, and the yearly distribution of class labels is shown in Figure 2.

Table 1: Socio-cultural semantic shifts in adjectives in 2014, as compared to 2013 (excerpts)

Class	Adjective	English translation
2	крымский	*'Crimean'*
2	приёмный	*'1) adopted; 2) something receiving'*
2	луганский	*'of Luhansk'*
1	правый	*'1) right; 2) right-wing'*
1	кипрский	*'Cyprian, Cypriot'*
0	серый	*'gray'*
0	балетный	*'of ballet'*

Table 1 contains examples of the adjectives from the 2013–2014 year pair (including two adjectives with stable meaning: 'серый' and 'балетный'). Many of the shifted adjectives are related to the Ukrainian events of 2014 that had an enormous effect on Russian news texts. For example, in 2014, the 'крымский' and 'луганский' toponyms (previously mostly associated with quiet provincial regions or seaside resorts) are almost exclusively mentioned in the context of the Russian-Ukrainian

military conflict, particularly related to the annexation of the Crimea and the appearance of the self-proclaimed Luhansk People's Republic.[5]

Although geographically motivated adjectives, like the ones from the previous paragraph, are quite specific in nature (their usage can easily change due to events occurring in the corresponding locations), it still seems reasonable to include them in the dataset, since, as mentioned above, we take into account the extra-linguistic causes of semantic shifts. Moreover, the dataset contains adjectives of many different types anyway, and we believe that the toponymic ones fit well into this diversity.

The changes are not limited to this: in 2013, the word 'приёмный' was most frequently used in the sense of '*adopted*', because of very active public discussion about the law prohibiting American citizens from adopting Russian orphan children (it was enacted on January 1, 2013). However, in 2014 this topic almost disappeared from the news discourse, and the usage of 'приёмный' adjective returned to its sense of '*something receiving*' (there is also a related noun 'приёмная' *'reception office'*). Both senses had existed long before: in this case, the diachronic semantic shift consists of significant changes in the balance between these senses (a secondary sense stepping forward for social and cultural reasons, and then 'moving backwards' again).

A few adjectives were annotated as belonging to the 1 class, since the assessors believed that some changes did occur to the cultural context around these words, but they were not as significant as those described in the previous paragraph. One example is 'правый' which in 2014 acquired heavy '*right-wing*' associations with the Ukrainian nationalistic movement 'Правый сектор' (literally, '*Right sector*'), but not enough to seriously decrease its usage in the primary '*right*' sense.

Another example is 'кипрский'. In 2014 it became associated with the 'offshore' concept slightly stronger than earlier in Russian media space. This was due to the number of corruption investigations involving high-ranking Russian officials and deoffshorization initiatives that were proposed by the Russian Ministry of Finance. In 2014, 'кипрский' was often used as a generalised image, an epitome of offshore company (when no particular company is implied), as in this example:

'На заседании в Госдуме директор по маркетингу Games Operation Division (Mail. Ru Group) Михаил Кочергин назвал World of Tanks «кипрской, офшорной» игрой.' (*At a meeting in the State Duma Mikhail Kochergin, head of marketing, Games Operation Division (Mail.Ru Group), called World of Tanks a 'Cyprian, offshore' game.*)

However, it was not an entirely new sense, and that was the reason for labelling this adjective as 1. It is also worth mentioning that the inter-rater agreement was quite low for this particular adjective (the three scores were '0', '1' and '2'). The **Micro** dataset preserves the scores assigned by each assessor, so that it is possible, if desired, to use only adjectives with high inter-rater agreement.

[5] Of course, the dictionary definitions for these adjectives ('*related to the Crimea*' and '*related to Luhansk*' correspondingly) still remained the same. But as already noted before, distributionally, contexts actually form the meaning (including typical connotations). From this point of view, these words have certainly significantly shifted their positions in the imaginary semantic space of Russian.

4. Employed corpora

In accordance with the datasets described in **Section 3**, we employ two sets of Russian corpora:

1. '**Macro**': less granular corpora, extracted from the main body of the Russian National Corpus (RNC),[6] 3 time bins overall:
   - texts produced before 1917 (**pre-Soviet**, 75 million tokens),
   - texts produced in 1918–1990 (**Soviet**, 96 million tokens),
   - texts produced after 1991 (**post-Soviet**, 85 million tokens).
2. '**Micro**': more granular corpora extracted from the RNC newspaper subcorpus and the *lenta.ru* dataset (14 time bins overall):
   - news texts produced in 2000,
   - news texts produced in 2001,
   - ...
   - news texts produced in 2014.

The '**Macro**' set of corpora (first used in [16]) is aimed at tracing the 'long-term' diachronic semantic shifts, while the '**Micro**' set of corpora is supposed to illustrate cases of 'short-term' cultural shifts heavily influenced by the current social and political events.

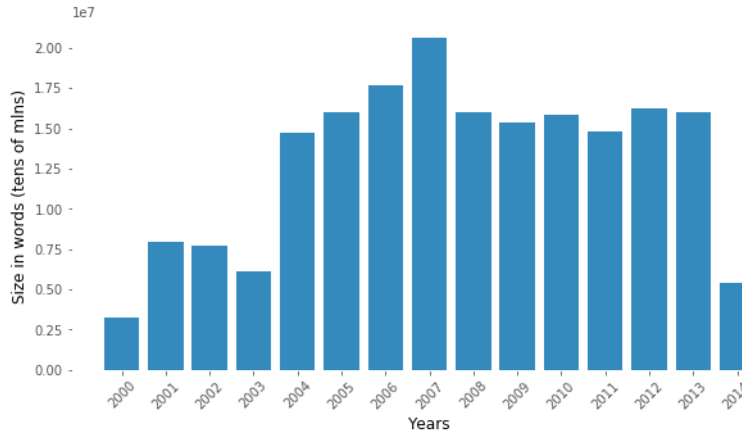


Figure 3: '**Micro**' corpora sizes per year

This second corpus collection (used here for the first time) features Russian news texts and comprises the whole RNC newspaper corpus[7] and the *lenta.ru* dataset texts[8] from the year 2000 up to 2014. The *lenta.ru* dataset additionally contains texts

6 http://ruscorpora.ru/en/

7 http://ruscorpora.ru/search-paper.html

8 https://github.com/yutkin/lenta.ru-news-dataset

produced in the years up to 2018, but we decided to stick to the 2000–2014 time span covered by the RNC newspaper corpus, to preserve the diversity.

The RNC newspaper corpus includes articles from the mass media of the 2000s; there are seven newspapers represented in approximately equal shares: 'Izvestia', 'Trud', 'Komsomolskaya Pravda', 'Sovetskiy Sport', 'RBC', 'RIA News', 'New Region'. The *lenta.ru* part consists of news articles crawled from one of the largest Russian news web sites. There are almost 194 million words in the whole '**Micro**' corpus, with about 12 million words per a yearly corpus on average. **Figure 3** shows yearly corpora sizes.

All the texts in the aforementioned corpora were lemmatized and tagged using a *UDPipe 1.2* model[9] trained on the Russian Universal Dependencies SynTagRus treebank [5]. Sequences of proper names immediately following each other were merged together if agreeing in case and number ('Александр_PROPN Сергеевич_PROPN' was transformed into 'Александр::Сергеевич_PROPN'). After that, lemmas were lowercased and words belonging to the functional parts of speech removed, to make it easier to train the word embedding models, described in the next section.

5. Word embedding models

All the approaches to trace semantic shifts employed in **Section 6** use distributional word embeddings [2, 6]. Word embeddings represent the meaning of words as dense vectors learned from lexical co-occurrences in the training corpora. In particular, we trained CBOW models [20] on all 14 + 3 = 17 tagged and lemmatized corpora described in **Section 4**, in two variations:

1. **static** models: trained separately on the corresponding time bins (years in the '**Micro**' corpora and larger time periods in the '**Macro**' corpora);
2. **incremental** models: the model M_{n+1} for the year $n + 1$ is initialised with the vectors trained from the previous year model M_n (essentially, this means simply continuing the training of the very first time bin model: the word vectors are updated using the co-occurrence signal from the new textual data).

Incremental models were shown to perform better for tracing temporal semantic shifts in some setups for English [17], and that was the reason we decided to test this mode for Russian data as well. Note that it also presupposes incrementally updating the vocabulary of the models.

All the models were trained with the vector dimensionality 300, context window size of 5 words to the left and 5 words to the right, for 10 epochs, with no downsampling. For the '**Macro**' models we ignored the words with corpus frequency less than 10, and for the '**Micro**' models, this parameter was set to 5, due to the corpora being significantly smaller.

[9] Note that we re-tagged the '**Macro**' corpora, instead of using them as they were presented in [16] (tagged with Mystem).

6. Experiments

In this section, we evaluate several existing approaches to semantic shifts detection on the Russian data presented above. Note that these experiments do not attempt to completely solve the problem at hand: the algorithms are very basic and intended only to provide solid baselines for further research in the area. In essence, all of the evaluated systems are implemented as logistic regression classifiers, taking as an input two word embedding models and one of the features enumerated below, and returning the semantic shift class of a word:

1. cosine distance between second order word similarity vectors in two models, using the **Global Anchors** algorithm [29];
2. cosine distance between embeddings of the word after aligning two models using the **orthogonal Procrustes transformation** [9];
3. **Jaccard distance** between top n neighbour lists of one and the same word in two models [11];
4. **Kendall's** τ between top n neighbour lists of one and the same word in two models [13];
5. **combined**: all the aforementioned metrics used as input features of a logistic regression.

Procrustes alignment is an SVD-based orthogonal transformation used to as closely as possible 'align' one embedding space to another (for example, a model trained on the 2011 corpus and the model trained on the 2012 corpus). After this, one can calculate cosine distances between word embeddings from different models, as if they were trained together. The **Global Anchors** algorithm measures how much the lists of word cosine similarities to other 'anchor words' are different between two models. In the case of *global* anchors, these lists are simply full intersection of two models' vocabularies, and thus it allows to analyse how the word position has changed related to all other words in the model.

Jaccard distance is a metrics used to measure the diversity between samples (here, samples are lists of a word's nearest neighbours in a given model by cosine similarity). It is defined as the difference of the samples' union and intersection sizes divided by the size of the union of two samples. **Kendall's** τ coefficient measures the rank correlation of intersections of two words' neighbours' lists. Its main difference in comparison to **Jaccard distance** when applied to our task, is that it pays attention to the relative order of n nearest neighbours in two models, not only the size of their intersection.

In the terms presented in [8], we can classify Jaccard distance and Kendall's τ metrics as *local* neighbourhood measures, while the orthogonal Procrustes transformation and Global Anchors are *global* measures. The **combined** approach merges them all to find out if there are any complementary aspects in which these metrics can help each other. Because of space limitations, we do not describe the employed algorithms themselves in details. We refer the interested reader to the corresponding sources or to the survey in [18].

The 5 methods enumerated above were evaluated on both '**Micro**' and '**Macro**' datasets and the corresponding sets of diachronic word embedding models. For each

method, the evaluation workflow was as follows. For each word in the current dataset, we calculated the numerical degree of its semantic shift between two models according to the current method. In the '**Macro**' dataset, these two models were always pre-Soviet and Soviet. In the '**Micro**' dataset, each word belongs to a particular sequential pair of years (e.g., 2005 and 2006), so the corresponding embedding models were used. Logistic regression classifiers were trained, using the annotated classes in the datasets as gold labels, and the calculated semantic shift degrees as features (thus, the classifiers were trained on 4 features with the **combined** method and on only 1 feature in all other cases). The '**Macro**' dataset presents a binary classification problem (shift or not shift), while the '**Micro**' dataset is a more granular ternary classification task (not shifted, somewhat shifted or significantly shifted).

These classifiers were then evaluated using stratified 9-fold cross-validation[10]. We report macro-averaged F1 score, since our class distribution is very imbalanced, but all classes are equally important. All the methods were tested both on static and incremental word embedding models, as described in Section 5.

Tables 2 and 3 present the results of the evaluation experiments with both datasets. As expected, the binary task of the '**Macro**' dataset turned out to be consistently easier for the systems under evaluation. Among the tested methods, the 'global' ones (**Global Anchors** and **Procrustes** alignment) always outperform the 'local' approaches (**Jaccard distance** and **Kendall's** τ). **Kendall's** τ performed worst of all, often ending up lower than random choice. The advantage of the global methods is most clear in the '**Micro**' dataset. This emphasises the importance of taking into account the word 'trajectory' in relation to the whole vector space, when trying to detect slight social and cultural changes (while for more profound semantic shifts, it is often enough to compare the nearest neighbours' lists). Also, in the more vague ternary problem of the '**Micro**' dataset, it pays off to take into account signals from several methods: the **combined** approach yields the best performance overall. At the same time, when making strict choice between two well-defined classes in the '**Macro**' dataset, single cosine distance between word vectors after **Procrustes alignment** seems to predict semantic shifts pretty well without any other features.

Table 2: Macro F1 scores, 'Macro' dataset

Model set	Global Anchors	Procrustes	Kendall	Jaccard	combined
Static	0.675	**0.767**	0.504	0.646	0.722
Incremental	0.598	0.681	0.475	0.576	0.617
Random choice					
≈ 0.5					

Using embedding models trained incrementally in our case did not yield strong improvements like those reported in [16] (it should be noted that our evaluation setup

[10] The number of folds was motivated by the fact that there are 18 words labelled as class 2 ('significant shifts') in the '**Micro**' dataset, so 9 folds allowed us to equally distribute these samples: 2 per each fold.

is significantly different). In the '**Micro**' dataset, incremental models were indeed marginally better than the static ones for all methods except **Procrustes** and **Combined**—but these two turned out to be the best, so overall the static models definitely win, being at the same time computationally simpler. It seems that employing global methods can in many cases compensate for inherent randomness in the initial states of word embedding models.

Table 3: Macro F1 scores, 'Micro' dataset

Model set	Global Anchors	Procrustes	Kendall	Jaccard	combined
Static	0.453	0.468	0.136	0.301	**0.503**
Incremental	0.462	0.459	0.194	0.326	0.442
Random choice					
≈ 0.33					

In general, the achieved scores are significantly better than random, and prove that distributional word embedding models can be successfully used to trace semantic shifts of different kinds and with time periods of different granularities in Russian language material. This provides a strong baseline for further studies employing more complex approaches. At the same time, the scores show the challenging nature of the presented datasets, which we hope will cause the NLP practitioners and researchers to dig deeper into the problem of detecting temporal semantic shifts. Fully implemented code for the evaluated algorithms is published at https://github.com/wadimiusz/diachrony_for_russian/, together with the word embedding models we used.

7. Conclusion

We presented test sets and evaluated baseline approaches for the task of tracing diachronic (temporal) semantic shifts in Russian. We are publishing two manually annotated datasets for this task (one entirely new and another from prior work, but repackaged and re-structured in more coherent form). The datasets are complementary in that the first one is focused on socially and culturally determined semantic shifts occurring within time spans of years, while the second one allows to test the ability of systems to trace global changes in words' meaning occurring in the scale of centuries (in this case, comparing pre-Soviet and Soviet time periods).

Further on, we used these datasets and the word embedding models trained on the corresponding corpora to rigorously evaluate 4 well-established algorithms of tracing diachronic semantic shifts (2 of which have to the best of our knowledge never been tested on Russian), as well as the system using the output of all these methods. Overall, we found that the methods using global information outperform the ones based on local data (like comparing nearest neighbours' lists). However, it should be noted that these experiments are preliminary and are intended to only provide a solid baseline for further studies in diachronic evolution of Russian lexical semantics.

For example, one obvious direction for future work is using more features describing the words (their corpus frequencies, parts of speech, semantic classes, etc). We also plan to test more advanced distributional models like those employing

continuous time variables [23] or contextualised word embeddings [22]. Finally, there exists a more difficult but more fascinating task of sense-aware semantic shifts detection (tracing particular types of shifts, like narrowing, widening, or splitting the meaning into two). We believe the presented paper provides a good starting point for such kind of research in application to Russian texts.

References

1. *Bamler, R., Mandt, S.:* Dynamic word embeddings. In: Proceedings of the international conference on machine learning. pp. 380–389, Sydney, Australia (2017).
2. *Baroni, M. et al.:* Don't count, predict! A systematic comparison of context-counting vs. Context-predicting semantic vectors. In: Proceedings of the 52nd annual meeting of the association for computational linguistics. pp. 238–247, Baltimore, USA (2014).
3. *Boussidan, A., Ploux, S.:* Using topic salience and connotational drifts to detect candidates to semantic change. In: Proceedings of the ninth international conference on computational semantics. pp. 315–319, Oxford, United Kingdom (2011).
4. *Daniel, M., Dobrushina, N.:* Two centuries in twenty words (in Russian). NRU HSE (2016).
5. *Droganova, K. et al.:* Data conversion and consistency of monolingual corpora: Russian UD treebanks. In: Proceedings of the 17th international workshop on treebanks and linguistic theories (tlt 2018), december 13–14, 2018, oslo university, norway. pp. 52–65 Charles University, Faculty of Mathematics; Physics, Prague, Czeck Republic; Linköping University Electronic Press, Linköpings universitet (2018).
6. *Firth, J.:* A synopsis of linguistic theory, 1930-1955. Blackwell (1957).
7. *Gulordava, K., Baroni, M.:* A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In: Proceedings of the gems 2011 workshop on Geometrical Models of Natural Language Semantics. pp. 67–71, Edinburgh, UK (2011).
8. *Hamilton, W.L. et al.:* Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In: Proceedings of the conference on empirical methods in natural language processing. pp. 2116–2121, Austin, Texas (2016).
9. *Hamilton, W.L. et al.:* Diachronic word embeddings reveal statistical laws of semantic change. In: Proceedings of the 54th annual meeting of the association for computational linguistics. pp. 1489–1501, Berlin, Germany (2016).
10. *Hilpert, M., Gries, S.T.:* Assessing frequency changes in multi-stage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. Literary and Linguistic Computing. 24, 4, 385–401 (2008).
11. *Jaccard, P.:* Distribution de la flore alpine: Dans le bassin des dranses et dans quelques régions voisines. Rouge (1901).
12. *Jean-Baptiste Michel, A.P.A., Yuan Kui Shen, others:* Quantitative analysis of culture using millions of digitized books. Science. 331(6014), 176–182 (2011).
13. *Kendall, M.G.:* Rank correlation methods. Griffin (1948).
14. *Kim, Y. et al.:* Temporal analysis of language through neural language models. In: Proceedings of the 52nd annual meeting of the association for computational linguistics. pp. 61–65, Baltimore, USA (2014).

15. *Krippendorff, K.:* Content analysis: An introduction to its methodology. Sage (2012).
16. *Kutuzov, A., Kuzmenko, E.:* Two centuries in two thousand words: Neural embedding models in detecting diachronic lexical changes. In: Quantitative approaches to the russian language. pp. 95–112 Routledge (2018).
17. *Kutuzov, A. et al.:* Tracing armed conflicts with diachronic word embedding models. In: Proceedings of the events and stories in the news workshop at acl 2017. pp. 31–36, Vancouver, Canada (2017).
18. *Kutuzov, A. et al.:* Diachronic word embeddings and semantic shifts: A survey. In: Proceedings of the 27th international conference on computational linguistics. pp. 1384–1397 Association for Computational Linguistics, Santa Fe, New Mexico, USA (2018).
19. *Liao, X., Cheng, G.:* Analysing the semantic change based on word embedding. In: Natural language understanding and intelligent applications. pp. 213–223 Springer International Publishing (2016).
20. *Mikolov, T. et al.:* Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems. 26, 3111–3119 (2013).
21. *Ozhegov, S.:* About lexical change in Russian language during the soviet era (in Russian). Voprosy Yazykoznania. 2, 70–85 (1953).
22. *Peters, M. et al.:* Deep contextualized word representations. In: Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers). pp. 2227–2237 Association for Computational Linguistics, New Orleans, Louisiana (2018).
23. *Rosenfeld, A., Erk, K.:* Deep neural models of semantic shift. In: Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies. pp. 474–484, New Orleans, Louisiana, USA (2018).
24. *Sagi, E. et al.:* Tracing semantic change with latent semantic analysis. Current methods in historical semantics. 161–183 (2011).
25. *Stern, G.:* Meaning and change of meaning; with special reference to the English language. Wettergren & Kerbers (1931).
26. *Tang, X.:* A state-of-the-art of semantic change computation. Natural Language Engineering. 24, 5, 649–676 (2018).
27. *Traugott, E.C., Dasher, R.B.:* Regularity in semantic change. Cambridge University Press (2001).
28. *Yao, Z. et al.:* Dynamic word embeddings for evolving semantic discovery. In: Proceedings of the eleventh acm international conference on web search and data mining. pp. 673–681, Marina Del Rey, CA, USA (2018).
29. *Yin, Z. et al.:* The global anchor method for quantifying linguistic shifts and domain adaptation. In: Advances in neural information processing systems. pp. 9433–9444 (2018).
30. *Zhang, Y. et al.:* The past is not a foreign country: Detecting semantically similar terms across time. IEEE Transactions on Knowledge and Data Engineering. 28, 10, 2793–2807 (2016).

IMPORTANCE OF COPYING MECHANISM FOR NEWS HEADLINE GENERATION

Gusev I. O. (ilya.gusev@phystech.edu)
MIPT, Moscow, Russia

News headline generation is an essential problem of text summarization because it is constrained, well-defined, and is still hard to solve. Models with a limited vocabulary can not solve it well, as new named entities can appear regularly in the news and these entities often should be in the headline. News articles in morphologically rich languages such as Russian require model modifications due to a large number of possible word forms. This study aims to validate that models with a possibility of copying words from the original article performs better than models without such an option. The proposed model achieves a mean ROUGE score of 23 on the provided test dataset, which is 8 points greater than the result of a similar model without a copying mechanism. Moreover, the resulting model performs better than any known model on the new dataset of Russian news.

Key words: text summarization, headline generation, Russian language, deep learning, seq2seq, copynet, BPE

ВАЖНОСТЬ МЕХАНИЗМА КОПИРОВАНИЯ ДЛЯ ГЕНЕРАЦИИ НОВОСТНЫХ ЗАГОЛОВКОВ

Гусев И. О. (ilya.gusev@phystech.edu)
МФТИ, Москва, Россия

Генерация заголовков новостей — существенная проблема в области суммаризации (автореферирования) текстов, так как она довольно ограничена, по сравнению с другими типами суммаризации, но всё ещё сложна. Модели с ограниченным словарём будут плохо справляться с такой задачей, потому что новые именованные сущности могут регулярно появляться в новостях, и зачастую они должны быть в заголовках. Для генерации заголовков для новостей на морфологически богатых языках, таких как русский, необходимы модификации моделей из-за обилия возможных словоформ. Цель этой работы — показать, что модели, которые могут копировать слова из оригинальной новости, справляются с задачей генерации заголовков лучше, чем модели без такой возможности. Модель предложенной архитектуры на тестовом наборе данных имеет средний ROUGE, равный 23, что на 8 баллов больше аналогичной модели без возможности копирования. Более

того, и на предоставленном тестовом наборе данных, и на наборе данных РИА наша модель показывает результаты лучше, чем какая-либо из известных моделей.

Ключевые слова: автореферирование текстов, суммаризация текстов, генерация заголовков, глубокое обучение, seq2seq, copynet, BPE

1. Introduction

Summarization systems are attracting more and more interest due to the increasing number of available texts. One of the main areas of application of these systems is news summarization and headline generation. Good headlines are crucial for news agencies and useful for readers. On the one hand, they should be informative enough and easy to read. On the other hand, they should encourage users to open full articles and should not contain complete information. Different news agencies have a different balance between these two extremes.

The structure of many news articles and especially the ones published online can be described as the inverted pyramid, which means that the first sentence contains critical information and answers basic questions: who, what, where, when, why and how. The rest of the first paragraph includes some important details while the subsequent sections provide more details, background information, and citations. Thereby the first sentence usually contains enough information to be a headline, and it can be a hard baseline to outperform.

Headline generation can be seen as a particular type of abstractive text summarization. Unlike extractive summarization, models of abstractive summarization can generate new words that were not used in the original text. The primary purpose of such models is to capture main information from the original text and produce a shorter version of it.

This work was done within the framework of the headline generation competition of the “Dialogue” conference. Various systems of headline generation were planned to compare within this track. The main training dataset consists of Russian news articles from the “RIA Novosti” website[1] [5]. Another test dataset was closed, but the organizers set up the framework for model evaluation based on Docker.

In addition to these datasets, we used a corpus of news articles of Lenta.Ru for evaluation purposes.

In this paper, we present an approach to headline generation based on two key features over the standard seq2seq model with attention [2]. Byte-pair encoding [9] is the first improvement, and the second is CopyNet [6], which either generates tokens from the vocabulary (as other networks do) or optionally copies a token from the source text. CopyNet model was applied to the task of headline generation in Ayana et al. [1]. We also present the test results of our model on the RIA corpus, Lenta.ru corpus and unknown test dataset from the organizers of the track. These results validate that models with a possibility of copying words from the original article outperform models without such an option.

[1] https://ria.ru

2. System description

The most simple architecture used was encoder-decoder with attention [2] over word-level tokens. One can utilize recurrent neural networks [2] or transformers [13] as encoder and decoder. We tried only LSTM models for encoding and decoding, whereas transformers should perform even better. There are several reasons we did not try to use them. First, there is no default implementation of them as a decoder in the framework we used. Second, they require much time to train on one GPU. Third, Gavrilov et al. [5] utilized Universal Transformer architecture in their work, but we achieved better scores even with LSTM encoder.

The first improvement over simple architecture was byte-pair encoding. It is crucial for many natural language processing tasks in the Russian language as it enables the use of rich morphology and decreases the number of unknown tokens. It often detaches word endings as each word form is less frequent than its stem. Moreover, many words in the Russian language share the same ending, thereby endings of many words can be encoded with the same token. The encoding was trained on the same datasets model trained using sentencepiece [9] library.

The second improvement was a copying mechanism as described in Gu et al.[6]. Many persons, organizations and locations are typically mentioned in the news. Furthermore, news documents contain unique numbers and dates. Some of these objects should be in the headlines. However, we can not cover all these elements using the fixed size vocabulary. A subword vocabulary can help us in some cases, but it is not enough to deal with this problem entirely. One of the possible solutions to this problem is using the copying mechanism that enables our model to copy tokens from the source text. The most popular solutions are CopyNet [6] and pointer-generator networks [12]. In our system, we used CopyNet primarily because it has an implementation as a part of AllenNLP framework [4].

We used the AllenNLP framework mainly because of its configuration system. Besides, it has a set of necessary modules that can be combined to build a flexible working model. Our code is available online as well as trained models and dataset splits.[2]

3. Data

3.1. RIA dataset

The organizers of the track provided a new dataset of news documents in Russian [5]. This dataset contains texts and titles of around 1 million news document published on the website “RIA Novosti (RIA news)” from January 2010 to December 2014. “Rossiya Segodnya” (Russia Today) news agency runs this website. Texts were lowercased before publishing. We split the dataset into the train, validation and test parts in a proportion of 90:5:5.

[2] https://github.com/IlyaGusev/summarus

3.2. Lenta.ru dataset

Another news dataset available online is the Lenta.ru dataset. It consists of about 800 thousand texts and titles from 1999 to 2018. We utilized this dataset to measure the performance of models trained on the RIA dataset to see how well these models can deal with texts of other style and period.

3.3. Secret test dataset

The test dataset of the conference track was available only for evaluation through a Docker container sent to the system. The public leaderboard was available for all participants. The first sentence baseline was hard to beat for this dataset.

We have some assumptions about the structure of this dataset. To begin with, the opening sentences of texts are meaningful. However, that is not true for almost every document in the RIA dataset. First sentences in the RIA dataset contain location, date, and author of the text. Thus either test dataset was sampled from different distribution or was carefully cleaned.

The second observation is linked with evaluation scores obtained. Scores of the models trained on the RIA dataset and evaluated on the Lenta dataset are almost similar with the scores achieved on the secret dataset thus we can consider the hidden dataset to be sampled from a different news agency than RT. The secret dataset could also be from a different time period.

4. Experiments

We utilized two architectures, namely standard encoder-decoder model with attention and CopyNet model. Models can have different token types; they can operate on word-level or use byte-pair encoding. We used only LSTM encoder-decoder for reasons described earlier. Moreover, we tried models of different size varying from 5 to 43 million of trainable weights. All models were trained on a single GeForce GTX 1080 with batch sizes depending on the size of the model. We used the loss evaluation on the validation dataset to determine when to stop training. The biggest model was trained for five days.

Byte-pair encoding model was trained on the train part of the RIA dataset. All models have a vocabulary size of 50000 tokens, though there were some experiments with an extended vocabulary. Models that used BPE do not require any additional preprocessing, whereas the texts for word-level models should be tokenized for better performance and the generated titles should be detokenized. The length of the source document was limited to 400 tokens for word-level models and 800 tokens for subword-level models.

No pretrained word or subword embeddings were used in the final models. However, we tried to utilize fastText [3] embeddings for Russian language and pretrained subword embeddings [7], but the performance of the resulting models was not better than the performance of fully trained models.

We did all the training with the Adam optimizer with learning rate 0.001. In most cases, we set a size of the beam search to 10.

The evaluation of models on the secret dataset was done using Docker. Participants were not able to see any error logs of the models. The time limit was set to 30 minutes, so big CopyNet models with beam search could not fit this requirement. We did some batching tricks that enable balancing between time and memory consumption. Eventually, our biggest model can meet the requirements with a beam-search size of 2 and no more.

5. Results

We present evaluation results on the RIA datasets in **Tab. 1**, the secret dataset results in **Tab. 2**, and the scores of the model trained on the RIA dataset and evaluated on the Lenta dataset in Tab.3. We evaluated our models with the standard ROUGE metric [10], reporting the F1 and recall scores for ROUGE-1, ROUGE-2, and ROUGE-L. ROUGE was measured with a Python package[3]. We used the mean of ROUGE-1-F, ROUGE-2-F, and ROUGE-L-F as the primary metric (R-mean-f). Also, we utilized BLEU [11] as an internal metric as it was easier to measure before we found a decent Python package for measuring ROUGE.

Table 1: RIA dataset evaluation

Model	R-1-f	R-1-r	R-2-f	R-2-r	R-L-f	R-L-r	R-mean-f	BLEU
seq2seq-bpe-5m	38.78	36.91	21.87	20.90	35.96	35.24	32.20	49.77
copynet-words-10m	39.48	38.39	22.57	22.05	36.95	36.69	33.00	51.99
copynet-bpe-10m	40.03	38.68	23.25	22.50	37.44	37.04	33.57	52.57
seq2seq-words-25m	36.96	35.19	19.68	19.02	34.30	33.60	30.31	44.69
seq2seq-bpe-25m	40.30	38.83	22.94	22.18	37.50	37.01	33.58	51.66
copynet-words-25m	40.38	39.46	23.26	22.83	37.80	37.70	33.81	52.99
copynet-bpe-43m	**41.61**	40.33	**24.46**	**23.76**	**38.85**	38.51	**34.97**	**53.80**
First Sentence [5]	24.08	**45.58**	10.57	21.30	16.70	**41.67**	17.12	—
UTransformer [5]	39.75	37.62	22.15	21.04	36.81	35.91	32.90	—

The baseline provided by the organizers was reasonably straightforward. They split a text into sentences and used the first one as the title. We slightly modified this baseline to achieve a better score. We removed full stops and constrained the number of words used as the title to 25. These steps seem reasonable as full stops are rarely used in news titles and titles should not be too long.

One can see that models with subword tokens perform better than the ones with word tokens of the same trainable weights count and vocabulary size. For example, R-mean-f for the word-level encoder-decoder model with 25 million weights is 30.31, which is significantly lower than R-mean-f for the subword-level model of the same weights count.

[3] https://github.com/bheinzerling/pyrouge

Table 2: Secret dataset evaluation. The score of * was without detokenization

Model	R-mean-f
seq2seq-bpe-5m	14.85
seq2seq-bpe-25m	15.40
copynet-words-10m	20.49*
copynet-bpe-10m	21.69
copynet-bpe-43m	**23.00**
First Sentence	19.50
First Sentence (modified)	19.89

The other observation is CopyNet having scores higher than standard encoder-decoder models with attention. That is especially true for the secret dataset where there were no successful tries to beat the first sentence baseline without the copying mechanism. Moreover, as we know, no other participants succeeded to outperform this baseline.

We did an ablation study for BPE and copying mechanism with models of 25 million trainable weights. Copying mechanism has a more significant impact than BPE, improving R-mean by 3.42 R-mean-f points on the RIA dataset, whereas BPE adds 3.27 R-mean-f points.

Table 3: Lenta dataset evaluation with a model trained on RIA dataset

Model	R-1-f	R-1-r	R-2-f	R-2-r	R-L-f	R-L-r	R-mean-f	BLEU
seq2seq-bpe-5m	19.38	17.35	8.27	7.43	16.94	16.55	14.86	25.14
seq2seq-words-25m	18.29	17.11	7.21	6.96	16.23	16.13	13.91	23.35
seq2seq-bpe-25m	20.75	19.06	8.77	8.11	18.15	17.97	15.89	28.21
copynet-words-25m	28.24	27.51	13.67	13.51	25.67	25.91	22.53	40.13
copynet-words-10m	26.37	26.38	12.67	12.74	24.04	25.06	21.02	38.36
copynet-bpe-10m	25.60	24.57	12.33	11.84	23.03	23.33	20.32	36.13
copynet-bpe-43m	**28.27**	27.61	**13.95**	13.63	**25.77**	26.19	**22.66**	**40.44**
First sentence	25.45	**40.52**	11.16	**18.63**	19.17	**37.80**	18.59	25.45

We also measured the scores of models trained on the RIA dataset applied to the Lenta dataset. News agencies have their own writing style of headlines, so we need to validate that models capture the essence of an article and not that style. Moreover, texts of RIA and Lenta articles also differ, so it should be harder for models to condition on Lenta articles. These scores confirm that copying mechanism is essential when transferring models between different news agencies. Remarkably, in some cases, BPE worsens models with copying mechanism. One of these cases was on the Lenta dataset with two models of 10 million weights.

Table 4: R-mean-f scores: Lenta dataset vs Secret dataset

Model	Lenta	Secret
seq2seq-bpe-5m	14.86	14.85
seq2seq-bpe-25m	15.89	15.40
copynet-bpe-10m	20.32	21.69
copynet-bpe-43m	22.66	23.00
First sentence	18.59	19.50

In Tab. 4 we provide a comparison between scores achieved on the secret dataset and scores achieved on the Lenta dataset. They are almost similar. We suppose that the organizers of the competition used some part of the Lenta dataset for evaluation as the secret dataset. We can explain slight differences in scores with another dataset split. The Lenta dataset contains period including one from 2010 to 2014 (period of the RIA dataset), so different dataset splits can contain this interval or not, and it can dramatically influence scores.

6. System and error analysis

We provide three examples of bad cases for most models. We do not state texts of articles here, but they are available in the RIA dataset. However, we have a reference title to catch what is this article about.

Almost in all severe cases, the model without both BPE or copying mechanism produces UNK tokens. Headlines with this token will be considered wrong by users. These type of errors are worse than wrong word forms sometimes produced by subword models.

Table 5: Beltukov example

Reference title	дело в отношении бельтюкова не скажется на “сколково” — вексельберг
seq2seq-words-25m	“сколково” UNK возбуждение дела против UNK
seq2seq-bpe-5m	“сколково” обеспокоен возбуждением дела против экс-главы фонда
seq2seq-bpe-25m	“сколково” считает возбуждение дела против вице-президента
copynet-words-10m	ситуация против бельтюкова не скажется на воплощении проекта “сколково”
copynet-bpe-10m	“сколково” озабочено возбуждение дела против бельтюкова
copynet-bpe-43m	руководство “сколково” озабочено возбуждение дела против бельтюкова

We have a person named Beltukov (бельтюков) in the example in Tab. 5. In the case of the word-level model without copying mechanism, he appears in the headline as UNK. He can appear in the headline as Beltukov only with the usage of the source tokens.

Table 6: Sevilla example

Reference title	**“бенфика” и “севилья” сыграют в турине в финале футбольной лиги европы**
seq2seq-words-25m	“бенфика” и “UNK” сыграют в финале лиги европы
seq2seq-bpe-5m	“бенфика” и “севильи” сыграют в финале лиги европы
seq2seq-bpe-25m	футболисты “бенфики” и “севильи” сыграют в финале лиги европы
copynet-words-10m	“бенфика” и “севилья” сыграют в финале лиги европы
copynet-bpe-10m	“бенфика” и “севилья” сыграют в финале лиги европы
copynet-bpe-43m	“бенфика” и “севилья” сыграют в финале лиги европы

There are two football clubs in the example in Tab. 6: Benfica (бенфика) and Sevilla (севилья). Benfica is in the vocabulary of the word-level model, whereas Sevilla is not. Small subword-level model without copying mechanism tried to reconstruct the name of the club but made an error (“севильи” instead of “севилья”). All the CopyNet models generated decent headlines.

Table 7: Mark Deutch example

Reference title	**лучшей смерти он себе и не желал — вдова журналиста марка дейча**
seq2seq-words-25m	UNK
seq2seq-bpe-5m	день на бали
seq2seq-bpe-25m	марк дейча
copynet-words-10m	“за заслуги перед отечеством”: “за заслуги перед отечеством”
copynet-bpe-10m	друзья и родные проводили в последний путь журналиста марка дейча
copynet-bpe-43m	марк дейч утонул на острове бали

The example in Tab. 7 was tough for some models. The death of Mark Deutch (“марк дейч”) was described in this article. Word-level models failed to generate a headline for this example. Subword word-level models caught whether the place or the person. CopyNet models with subwords succeeded to generate a meaningful headline.

7. Conclusion and future work

It was surprising not to see any other participants successfully using the CopyNet or pointer-generator networks, as these techniques are commonly used for other tasks of the text summarization.

In conclusion, our results validate that the copying mechanism applies to the task of headline generation well. Moreover, in most cases, it is required to use this mechanism due to a variety of named entities, numbers and dates in the news.

As for future work, we should try the transformer as a decoder, make an evaluation of pointer-generator networks for this task, and utilize reinforcement learning

methods described in Keneshloo et al. [8]. Besides, it would be nice to make a human evaluation and extend these models to other languages.

8. Acknowledgements

Authors are thankful to Kozlovskiy Borislav, Lyubanenko Vadim and Smirnova Elizaveta for proofreading and to Smurov Ivan for useful advises.

References

1. *Ayana, Shen S., Lin Y. et al.* (2017), Recent advances on neural headline generation, Journal of computer science and technology. pp. 768–784.
2. *Bahdanau D., Cho K., Bengio Y.* (2014), Neural Machine Translation by Jointly Learning to Align and Translate, arXiv:1409.0473.
3. *Bojanowski P., Grave E., Joulin A., Mikolov T.* (2017), Enriching Word Vectors with Subword Information, Transactions of the Association for Computational Linguistics, pp. 135–146.
4. *Gardner M., Grus J. et al.* (2017), AllenNLP: A Deep Semantic Natural Language Processing Platform, Proceedings of Workshop for NLP Open Source Software. pp. 1–6.
5. *Gavrilov D., Kalaidin P., Malykh V.* (2019), Self-Attentive Model for Headline Generation, arXiv:1901.07786.
6. *Gu J., Lu Z., Li H., Li V.* (2016), Incorporating Copying Mechanism in Sequence-to-Sequence Learning, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1631–1640.
7. *Heinzerling B., Strube M.* (2018), BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages, Proceedings of the Eleventh International Conference on Language Resources and Evaluation.
8. *Keneshloo Y., Ramakrishnan N., Reddy C.* (2018), Deep Transfer Reinforcement Learning for Text Summarization, arXiv:1810.06667.
9. *Kudo T., Richardson J.* (2018), SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 66–71.
10. *Lin C.Y.* (2004), ROUGE: A package for automatic evaluation of summaries, Text summarization branches out: ACL workshop.
11. *Papineni, K., Roukos, S., Ward, T., Zhu, W. J.* (2002), BLEU: a method for automatic evaluation of machine translation, 40th Annual meeting of the Association for Computational Linguistics, pp. 311–318.
12. *See A., Liu J. P., Manning C.* (2017), Get To The Point: Summarization with Pointer-Generator Networks, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1073–1083.
13. *Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I.* (2017), Attention Is All You Need, Advances in Neural Information Processing Systems, pp. 5998–6008.

АННОТИРОВАНИЕ ПАРАЛЛЕЛЬНЫХ ТЕКСТОВ: ПОНЯТИЕ «ДИВЕРГЕНТНЫЙ ПЕРЕВОД»

Инькова О. Ю. (Olga.Inkova@unige.ch)
ИПИ ФИЦ ИУ РАН, Москва, Россия; Женевский университет, Женева, Швейцария

Ключевые слова: база данных коннекторов, русский язык, корпусная лингвистика, лингвистическое аннотирование, параллельные корпуса, семантика коннекторов, статистические данные

ANNOTATION OF PARALLEL TEXTS: THE CONCEPT OF DIVERGENT TRANSLATION

Inkova Olga Yu. (Olga.Inkova@unige.ch)
Institute of Informatics Problems, FRC CSC RAS, Moscow, Russia; University of Geneva, Geneva, Switzerland

The annotation of parallel corpora, as well as building of supracorpora databases, challenges linguists with the question of how to define a functional equivalent of the linguistic units that serve as an object of a given study. The paper discusses the concept of divergent translation and whether it is theoretically important for the analysis of logical-semantic relations (LSR). It is shown that relations between states of things can be expressed not only by connectives but also by lexical means (referred to as “alternative lexicalizations” in the works of the Penn Discourse Treebank group) and grammatical tools (syntactic constructions and morphological forms), and by marks of punctuation. While the two latter ways are mentioned in grammars, they are usually not taken into account when the alternative ways of tagging LSR are described, nor are they annotated in corpora or databases. The supracorpora database of connectives, built on the basis of the French and Italian parallel subcorpora of the Russian National Corpus, introduces new functional capabilities. It stores a representative array of annotations tagged as “divergent translation” (more than 1,250, i. e. 7.7 per cent of the total number), which allows users to collect various statistical data. With these data, one could establish: (1) which LSR tend to be expressed by alternative means and how often they occur compared to connectives, (2) what these alternative means are, (3) which divergent translations may be used to render a given marker of LSR and how often each of them is used, (4) which alternative markers of LSR are specifically employed to convey one or another relation and which of them are able to express several LSR. The conclusive

part of the paper suggests that, for the analysis of divergent equivalents, it is central that one and the same alternative means is used by different translators when translating one and the same textual fragment into one and the same language as well as into several languages, which speaks for its productivity. The further development of multi-language and polyvariant parallel corpora and databases would let us find out to what extent the means conveying LSR differ in various languages.

Keywords: corpus linguistics, linguistic annotation, parallel corpora, semantics of connectives, Russian, statistical data, supracorpora database of connectives

1. Введение

Аннотирование параллельных корпусов и создание надкорпусных баз данных требует разработки не только новой информационной технологии (см., в частности [Бунтман и др. 2014]), но и системы терминов [Зализняк и др. 2017]. В данном исследовании проблема разработки терминологии рассмотрена на примере надкорпусной базы данных (НБД) коннекторов, в которую загружены параллельные тексты французского и итальянского подкорпусов Национального корпуса русского языка. Общий объем параллельных текстов составляет примерно 9 млн словоупотреблений. Представительный фрагмент НБД коннекторов доступен по адресу: http://a179.frccsc.ru/RSCF10004/main.aspx.

Информация о семантике и функционировании коннекторов хранится в НБД в виде двуязычных аннотаций, *моноэквиваленций* (МЭ), включающих фрагмент текста оригинала с коннектором (столбец 1 на **рис. 1**) и соответствующий ему фрагмент текста перевода (столбец 3).

Контекст коннектора	Коннектор и его признаки	Контекст эквивалента	Эквивалент и его признаки	Признаки МЭ
Голос превосходный, но ведь голос, в конце концов, можно и так слушать, не вступая в брак, не правда ли… **Впрочем**, неважно.	**впрочем** <"вопреки ожиданиям"> <иллокутивные> <повествовательное> <начальная > <p CNT q> <CNT>	Sa voix est splendide, mais enfin, une voix, on peut l'écouter sans l'épouser pour autant, pas vrai?... **Du reste**, c'est sans importance.	**du reste** <"вопреки ожиданиям"> <иллокутивные> <повествовательное> <начальная> <p CNT q> <CNT>	• Cngrn • Dvrg

Рисунок 1. Пример двуязычной аннотации из НБД коннекторов

В процессе аннотирования (подробнее см. [Зацман и др. 2016]; [Inkova, Popkova 2017]) коннектору (столбец 2 на **рис. 1**) и его функциональному эквиваленту (столбец 4) приписываются значения признаков, характеризующих их функционирование в данном контексте. На заключительном этапе заполняется пятый столбец, характеризующий саму МЭ, а именно тип межъязыковой эквивалентности: метки Cngrn и Dvrg предназначены для обозначения типа

перевода, который считается конгруэнтным (Cngrn), если коннектор переведен коннектором, как на **рис. 1**, и дивергентным (Dvrg), если коннектор переведен лексической единицей, не являющейся коннектором, морфологической формой или синтаксической конструкцией. Ср. (1), где *и|при этом* переведен в **а.** деепричастием, а в **б.** — относительным придаточным. В таких случаях проставлена метка Dvrg.

(1) — Да три полсотни с лишком надо будет приложить, — сказал Петрович *и* сжал *при этом* значительно губы. [Н. В. Гоголь. Шинель (1842)]
а. Trois billets de cinquante roubles au bas mot», dit Pétrovitch *en se pinçant* les lèvres. [Trad. H. Mongault]
б. Il faudra y mettre trois billets de cinquante roubles au moins, dit Pétrovitch *qui serra les lèvres avec des sous-entendus*. [Trad. B. Kreise]

По мере наполнения НБД появилась необходимость уточнить понятие «дивергентный» перевод, чему посвящен **раздел 2** настоящего исследования. Представительный массив аннотаций с пометой Dvrg (на 31.01.2019 в НБД сформировано 1263 таких МЭ, т. е. 7,7 % от общего количества МЭ) позволяет, в свою очередь, сформулировать некоторые гипотезы относительно способов выражения логико-семантических отношений (ЛСО). Эти гипотезы рассматриваются в **разделе 3**. Там же приводятся некоторые статистические данные, а также иллюстрируются возможности, предоставляемые НБД для контрастивного исследования способов выражения ЛСО.

2. Понятие дивергентного перевода

Термин «дивергентный перевод» заимствован нами из работы [Johansson 2007], посвященной использованию многоязычных корпусов в контрастивных исследованиях. Тип межъязыкового соответствия характеризуется по трем параметрам: тип текста (оригинал vs. перевод), эксплицитный или имплицитный характер соответствия (наличие эквивалента vs. нулевой перевод) и синтаксическая природа переводного соответствия (конгруэнтный vs. дивергентный эквивалент). Например, соответствие норвеж. *nok* → англ. *probably* охарактеризовано как «перевод, эксплицитный, конгруэнтный», а норвеж. *nok* → англ. *I suppose* как «перевод, эксплицитный, дивергентный» [Там же: 25]. Однако понятие «дивергентный перевод» определено в работе без учета особенностей семантического или функционального класса, к которому принадлежит переводимая языковая единица (ЯЕ).

В случае коннекторов определение соответствия как конгруэнтного или дивергентного вызывает немалые трудности при аннотировании корпусов параллельных текстов. Но проблема определения, каким языковым средством выражено то или иное логико-семантическое отношение (ЛСО), возникает и при аннотировании одноязычных корпусов, см., например [Taboada 2009]. Первая трудность связана с тем, что функциональный класс коннекторов является морфологически неоднородным и открытым [Инькова-Манзотти 2001: 20–23]. Помимо сочинительных (*а, но* и др.) и подчинительных (*потому что,*

если и др.) союзов, а также их «аналогов» (*однако, тем не менее, например*), к этому классу относятся некоторые предлоги, способные сочетаться с именами с пропозитивным содержанием (*после, вместо* и др.), и предикативные структуры, характеризующиеся высокой степенью устойчивости, например, *дело (заключается/состоит) в том, что* и др.

Однако хорошо известно, что отношения между ситуациями могут выражаться ЯЕ, не принадлежащими к классу коннекторов. В системе аннотирования, разработанной в рамках Penn Discourse Treebank (PDTB), некоторые из них квалифицируются как «альтернативные лексикализации» [Prasad et al. 2010]. В PDTB зафиксировано 624 таких случая. Этот класс включает *i*) лексически и синтаксически свободные конструкции (например, *The increase was due mainly to…, These measurements can indicate temperature changes, That is why*, выражающие причинные или следственные отношения[1]); *ii*) конструкции, синтаксически свободные, но с фиксированным лексическим составом (*What's more, To begin with*), *iii*) некоторые устойчивые наречные выражения, синтаксический статус которых позволяет их отнести к классу коннекторов, но эта функция не является у них основной, а возникает в определенном контексте (*for one thing, as well, too, even, especially*). В НБД коннекторов в тех случаях, когда коннектор переведен одной из перечисленных ЯЕ (2) или, наоборот, такая ЯЕ переводится коннектором (3), межъязыковое соответствие квалифицируется как дивергентное.

(2) Вот что было-с. *Да еще* слухи о земельной реформе, которую намеревался произвести пан гетман. [Михаил Булгаков. Белая гвардия (1924)]
Voilà ce qui existait. *Il fallait ajouter à cela* les bruits qui couraient, selon lesquels le seigneur hetman avait l'intention de procéder à une réforme agraire. [Trad. C. Ligny]

(3) Да приложи к письму какой-нибудь петербургский гостинец… сигар, *что ли*. [И. А. Гончаров. Обломов (1848–1859)]
Ajoute à ta lettre quelque cadeau pétersbourgeois… Des cigares, *par exemple*. [Trad. L. Jurgenson]

[1] В [Toldova et al. 2018] речь уже идет не об альтернативных способах выражения причинных отношений, а о коннекторах; ср. приводимые на с. 6 в качестве «discourse connectives for causal relations» глаголы *порождать, позволять, объяснять, давать, изменять, вызывать, приводить к*) и пример: [Неудачно остановившаяся машина стала помехой для быстрых кругов многих гонщиков, включая Фернандо Алонсо,] [и **это вызвало** расследование FIA.] Такая позиция представляется теоретически уязвимой. Эти глаголы, безусловно выражая семантику причинности, участвуют в описании ситуации, входя в пропозициональное содержание высказывания. Включение в состав «коннектора» демонстратива *это*, осуществляющего резумптивную анафору, не спасает дела. Если следовать логике исследователей, то коннектором следует признать и предикативные структуры с подлежащим, осуществляющим резумптивную анафору, типа *эта ситуация, это положение дел* и др. Ср. модификацию приведенного примера: [Неудачно остановившаяся машина стала помехой для быстрых кругов многих гонщиков, включая Фернандо Алонсо,] [и **эта ситуация вызвала** расследование FIA.] Если говорить о коннекторе в исходном примере, то им является только союз *и* (ср. Миша много гуляет, *и* это очень хорошо).

В (2) семантика *да еще*, выражающего аддитивные ЛСО, передается безличной конструкцией франц. *Il fallait ajouter à cela* ‘нужно добавить к этому’. В (3) частица *что ли* употреблена в контексте ЛСО спецификации: первый фрагмент текста содержит наименование множества (*какой-нибудь петербургский гостинец*), а во фрагменте текста со *что ли*, — элемент этого множества (*сигары*). Однако вряд ли можно считать *что ли* выразителем этого ЛСО. Эта частица сохраняет свое модальное значение предположения и может употребляться и в других контекстах, далеких от семантики спецификации; ср. (4), где можно говорить скорее об ЛСО переформулирования, поиска более адекватного наименования, между *спокойный* и *традиционный*.

(4) И хочется чего-то более спокойного, традиционного, *что ли*. [Вероника Стрельникова. Опять акробатика, милый? // «Даша», 2004]

Словари справедливо не указывают для *что ли* значения показателя спецификации. Во французском переводе, напротив, использован специализированный показатель этого ЛСО — *par exemple*.

Помимо ЯЕ, относящихся к «альтернативным лексикализациям», в НБД зафиксированы и другие функциональные эквиваленты коннекторов. Их можно разделить на две группы: грамматические (синтаксические конструкции, морфологические формы), примеры которых мы видели в (1), и лексические. Так, отношение результативной аналогии, при которой длящееся в момент наблюдения действие или состояние субъекта, описанное в главной части, непосредственно продолжает то же действие или состояние субъекта, описанное в придаточной части, или является его результатом, часто выражается лексемами с семантикой неизменности [Кобозева, Инькова 2018: 206–210]: в (5)– итал. *rimanere immobile* ‘остаться неподвижным’ и франц. *s’immobiliser* ‘застыть в неподвижности’:

(5) Катерина Ивановна *как* стояла на месте, *так и* осталась, точно громом пораженная. [Ф. М. Достоевский. Преступление и наказание (1866)]
а. Katerìna Ivànovna *rimase lì immobile*, come colpita da un fulmine. [Trad. G. Kraiski]
б. Catherine Ivanovna *s’immobilisa* sur place comme frappée par la foudre. [Trad. E. Guertik]

Наконец, семантика коннектора может быть выражена семантически насыщенным знаком препинания: двоеточием или тире, причем как в оригинале (6), так и переводе (7).

(6) Конечно, от оперуполномоченного многое зависит, и таинственны, туманны дорожки к высотам жизни — зав. баней, хлеборез. [Василий Гроссман. Жизнь и судьба (1959)]
Bien sûr, bien des choses dépendaient de l’oper, bien sûr, les voies qui mènent vers les sommets de la vie, *par exemple* être responsable du bain ou des rations de pain, sont sombres et mystérieuses. [Trad. A. Berelowitch]

(7) Подлинная многопланность разрушила бы драму, *ибо* драматическое действие, опирающееся на единство мира, не могло бы уже связать и разрешить ее. [М. М. Бахтин. Проблемы поэтики Достоевского (1963)]
Une véritable multiplicité de plans serait préjudiciable à la pièce**:** l'action dramatique qui s'appuie normalement sur l'unité de l'univers représenté serait alors incapable de servir de lien et d'apporter des solutions. [Trad. I. Kolitcheff]

Эти знаки препинания, особенно двоеточие, совместимы с ограниченным числом ЛСО (как правило, разного рода причинными ЛСО и ЛСО спецификации), поэтому, в отличие от других знаков препинания (точка, точка с запятой и запятая), они могут служить сигналом того, что между двумя фрагментами текста существует некоторое ЛСО.

Во всех перечисленных выше случаях в НБД для МЭ проставляется признак Dvrg.

В процессе аннотирования были выявлены, однако, такие случаи перевода, когда двух меток — Cngrn и Dvrg — оказалось недостаточно для характеристики переводного соответствия, поскольку речевая реализация (РР), т. е. та форма, в которой коннектор встретился в конкретном контексте и которая является единицей аннотирования в НБД [Инькова 2018а], может быть переведена сочетанием коннектора и ЯЕ, не являющейся коннектором. Это могут быть как РР, выражающие одно ЛСО, так РР, представляющие собой сочетание коннекторов, выражающих разные ЛСО:

(8) Звали его полностью Лазарь Рувимович Цехновицер, он был худой, длинноносый, курчавый, *а также* учился играть на скрипке.
[Сергей Довлатов. Иностранка (1986)]
Son nom exact était Lazare Rouvimovitch Tsekhnovitser, il était maigre, avait le cheveu frisé, le nez long *et* apprenait *même* le violon.
[Trad. J. Michaut-Paternò]

(9) Соня дала свой адрес *и при этом* покраснела.
[Ф. М. Достоевский. Преступление и наказание (1866)]
Sonia donna son adresse *et* rougit *en le faisant*. [Trad. É. Guertik]

В (8) показатель аддитивных ЛСО *а также* переведен сочетанием союза *et* ‘и’ и градационной частицы *même* ‘даже’; в (9) РР *и при этом*, сочетающая показатель соединительных ЛСО и ЛСО сопутствования, переведена союзом *et* и деепричастием глагола *faire* ‘делать’ с анафорическим местоимением *le* ‘это’, передающими одновременность двух действий.

Для таких случаев в новой версии НБД введена метка Cngrn + Dvrg. На 01.02.2019 эта метка проставлена для 89 МЭ (учитываются все направления перевода).

3. О чем нам говорит дивергентный перевод?

В работах, выполненных в рамках проекта РНФ «Логическая структура текста: контрастивный анализ показателей логико-семантических отношений в русском, французском и итальянском языках», наличие большого количества дивергентных эквивалентов или стимулов перевода используется как один из параметров лингвоспецифичности коннектора [Инькова 2018б]. Однако функциональные возможности НБД позволяют получить и другую статистику по признаку Dvrg. Эта статистика позволяет, например, увидеть, для каких ЛСО дивергентный перевод показателя является распространенным явлением, а для каких — скорее редкость[2].

Таблица1. Доля дивергентных переводов для ЛСО (выборочно) в направлении русский-французский

ЛСО	Dvrg	Всего	%
сопутствование	62	348	**17,8%**
спецификация	86	726	**11,8%**
сравнительные	29	288	**10,1%**
аналогия	10	102	9,8%
переформулирование	81	864	9,4%
сопоставительные	24	282	8,5%
причина	32	398	8%
условные	39	580	6,7%
аддитивные	43	683	6,3%
уступительные	43	730	5,9%
альтернатива	9	420	*2,1%*
отрицательная альтернатива	3	170	*1,7%*
соединительные	1	263	*0,4%*
«вопреки ожиданиям»	40	1346	*0,3%*

В НБД самая высокая доля дивергентных переводов — у показателей отношения сопутствования: почти 18%, а наименьшая (приближающаяся к нулю) — у показателей соединительных ЛСО и ЛСО «вопреки ожиданиям». У ЛСО сравнения, аналогии, причины и переформулирования доля дивергентных переводов составляет около 10%. Чуть больше их у ЛСО спецификации (почти 12%).

Важно также узнать, какой из показателей того или иного ЛСО отличается наибольшей долей дивергентных переводов и каких именно. В Таблице 2 приведены данные для ЛСО переформулирования и причины в направлении перевода русский — французский.

---

[2] Подчеркнем, что статистику, приводимую в работе, можно получить непосредственно в НБД, задав соответствующий запрос. Кроме того, абсолютные цифры в таблицах являются в НБД гиперссылками, отсылающими к аннотациям, отвечающим критериям запроса, что позволяет пользователю сразу получить доступ к корпусу примеров и проводить их дальнейший семантический анализ.

Таблица 2. Распределение дивергентных переводов по показателям ЛСО в НБД

Переформулирование	Всего МЭ	Dvrg, %
то есть	633	11,22
одним словом	68	5,88
словом	84	0

Причина	Всего МЭ	Dvrg, %
потому что	180	12,22
ибо	50	10
так как||*то*	31	3,23

Мы видим, что, если у обоих ЛСО доля дивергентных переводов сопоставима (9,4% для переформулирования и 8% для причины), распределение этих переводов по показателям каждого из ЛСО существенно разнится. В группе показателей переформулирования лидирует *то есть* (11,22%), тогда как у *словом* пока не зафиксировано ни одного дивергентного перевода. В группе причинных ЛСО лидирует *потому что* с 12,22%, чуть меньше таких переводов у *ибо* (10%).

НБД позволяет также узнать, каковы эти дивергентные переводы. Например, для *потому что* наиболее частотным из 7 зафиксированных в НБД дивергентных переводных эквивалентов является двоеточие (40,1%), как в (10), где два из четырех переводчиков прибегают к такому решению; за ним следуют различные причастные конструкции в прошедшем времени, со вспомогательным глаголом (18,2%) и без него (4,6%), и в настоящем времени (18,2%), а затем еще один знак препинания: тире (9,1%), как в (11), где оба переводчика выбрали это решение.

(10) <…> коллежский асессор Ковалев не мог слышать запаха, *потому что* закрылся платком *и потому что* самый нос его находился бог знает в каких местах. [Н. В. Гоголь. Нос (1832–1833)]
<…> l'assesseur de collège Kovaliov ne pouvait pas s'en rendre compte**:** il avait caché son visage sous un mouchoir, et d'ailleurs son nez se trouvait en cet instant Dieu sait où. [Trad. B. de Schloezer]
<…> le major Kovaliov ne s'en trouvait point incommodé: il tenait son mouchoir sur son visage, et d'ailleurs son nez se promenait… Dieu sait où. [Trad. H. Mongault]

(11) Он по необходимости сидел в классе прямо, слушал, что говорили учителя, *потому что* другого ничего делать было нельзя. [И. А. Гончаров. Обломов (1848–1859)]
Il lui fallut se bien tenir en classe, écouter ce que disaient les professeurs — il y était bien obligé. [Trad. A. Adamov]
Se pliant à la nécessité, il se tenait droit en classe, écoutait ce que disaient les professeurs — on ne pouvait rien faire d'autre [Trad. L. Jurgenson]

Два других дивергентных эквивалента: конструкция *Adj+que+fin_être* (12) и выделительная конструкция *c'est N qui/que* зарегистрированы с единичными употреблениями.

(12) Ничто не нарушало однообразия этой жизни, и сами обломовцы не тяготились ею, *потому что* и не представляли себе другого житья-бытья. [И. А. Гончаров. Обломов (1848–1859)]
Rien ne troublait la monotonie de cette vie, qui ne pesait point aux habitants d'Oblomovka, *incapables qu'ils étaient* même de s'imaginer un train-train différent. [Trad. L. Jurgenson]

У *ибо* зафиксировано пока три дивергентных эквивалента: двоеточие (7) и причастие прошедшего времени без вспомогательного глагола (по 40%), которые мы уже видели для *потому что*, и относительное придаточное (20%). Заметим, что его выбирают не только французские (13а), но и 3 из 15 итальянских переводчиков этого контекста; один из таких переводов приведен в (13б.)

(13) <...> вскрикнул, может быть, в первый раз от роду, *ибо* отличался всегда тихостью голоса. [Н. В. Гоголь. Шинель (1842)]
а. <...> s'exclama, pour la première fois de sa vie, sans doute, le malheureux Akaki Akakiévitch, *qui d'ordinaire parlait à voix très basse*. [Trad. H. Mongault]
б. <...> e fu forse questo il primo urlo che avesse emesso dacchè era nato, *lui che parlava sempre sottovoce*. [Trad. Duchessa D'Andria]

Иная семантика дивергентных переводов у *то есть*. Хотя здесь тоже встречаются двоеточие и тире, но на их долю приходится в совокупности всего 5,6%, а самые частотные дивергентные переводы — глагольное выражение *vouloir dire* 'хотеть сказать' (39,5%), как в (14), и глагол *entendre* 'понимать что-то под чем-л.' (15,5%). Используются также глаголы *signifier* 'означать', *dire* 'сказать', *s'expliquer* 'объясниться', *appeler* 'называть'.

(14) Между нами уже было не так, как раньше, *то есть* мы дружили, но все больше я на нее покрикивала [Светлана Алексиевич. Время секонд хэнд (2013)]
Ce n'était déjà plus comme avant entre nous, *je veux dire*, on s'entendait bien, mais je lui criais de plus en plus souvent dessus. [Trad. S. Benech]

Заметим, что если некоторые переводы, и не только дивергентные, являются контекстуально обусловленными, что часто приводит к значительной перестройке контекста оригинала (см. на примере *хотя* [Нуриев 2018]), и зафиксированы с единичными вхождениями, то другие могут рассматриваться как регулярные модели перевода. Для сопоставительного *а* — это местоименный повтор (15) [Зализняк, Микаэлян 2018], для *при этом* — деепричастие настоящего времени (1)[3].

[3] Подробный семантический анализ дивергентных переводных эквивалентов не входит в задачу настоящего исследования, да и невозможен в силу ограниченности его объема. Примеры такого анализа см. в посвященной этому вопросу монографии *Семантика коннекторов: контрастивное исследование* под ред. О. Ю. Иньковой.

(15) Обломов задумался, *а* Алексеев барабанил пальцами по столу.
[И. А. Гончаров. Обломов (1848–1859)]
Oblomov devint pensif, *Alexéev, lui*, pianotait sur la table. [Trad. L. Jurgenson]

Кроме того, как альтернативный способ выражения ЛСО могут использоваться ЯЕ, имеющиеся также в языке оригинала, как в приведенных выше примерах, или же не существующие в системе языка оригинала, как в (16), где для перевода условных отношений, выраженных в оригинале коннектором, используется отсутствующая в русском языке синтаксическая конструкция с двумя предикатами в ирреалисе (и с обязательной инверсией в первом), соединенными асемантичным союзом *que* ‘что’.

(16) <…> и не представляли себе другого житья-бытья; а *если б и* смогли представить, *то* с ужасом отвернулись *бы* от него [И. А. Гончаров. Обломов (1848–1859)]
<…> incapables qu’ils étaient même de s’imaginer un train-train différent; *en auraient-ils été* capables *qu’ils l’auraient condamné* avec horreur.
[Trad. L. Jurgenson]

Наконец, мы можем узнать, является ли дивергентный перевод «специализированным» средством выражения некоторого ЛСО или может обслуживать несколько ЛСО. К первым можно отнести безличный глагол франц. *il s’ensuit* ‘из этого следует’, который может служить эквивалентом только для коннекторов ЛСО следствия. Ко вторым — относительное придаточное, использующееся при переводе коннекторов сопоставления (20 %), спецификации (17,14 %), аддитивности (14,29 %), переформулирования (11,43 %), сопутствования (11,43 %), причины (8,58 %), уступительных (5,72 %) и др.

4. Заключительные замечания

Теоретический интерес понятия «дивергентный перевод» для описания средств выражения ЛСО не вызывает, на наш взгляд, сомнения. Основное преимущество параллельного корпуса состоит в том, что он служит средством выявления различий в системах языков, особенно в тех случаях, когда они не фиксируются словарями и грамматиками. Использование НБД коннекторов, созданной на основе параллельных текстов предоставляет исследователям принципиально новые возможности и позволяет установить: 1) для каких ЛСО характерны альтернативные способы выражения и какова их частотность по сравнению с коннекторами; 2) каковы эти альтернативные способы выражения; 3) какие дивергентные переводы может иметь тот или иной показатель ЛСО и какова частотность каждого из них; 4) какие альтернативные средства выражения ЛСО являются специфическими для того или иного отношения, а какие могут обслуживать несколько ЛСО.

Важным при таком анализе представляется тот факт, что одни и те же модели дивергентного перевода зафиксированы в НБД для показателей одного и того же ЛСО в разных языках, в разных направлениях перевода и у разных

переводчиков одного и того же контекста оригинала. Это позволяет сделать обобщения относительно регулярности того или иного альтернативного способа выражения ЛСО и установить, каков набор этих способов в каждом из сопоставляемых языков, а также чем этот набор отличается в разных языках.

Литература

1. *Buntman N. V., Zalizniak Anna A., Zatsman I. M., Kruzhkov M. G., Loshchilova E. Yu., Sitchinava D. V.* (2014) Informatsionnye tekhnologii korpusnykh issledovanij: printsipy postroeniya krosslingvistitcheskikh baz dannkh [Information technologies for corpus studies: underpinnings for cross-linguistic database creation]. Informatics and applications. 2014. Vol. 8. No. 2. Pp. 98–110.
2. *Inkova-Manzotti O. Yu.* (2001) Konnektory protivopostavleniya vo frantsuzskom i russkom yazykakh: Sopostavitel'noe issledovanie [Connectors of opposition in French and Russian: A comparative study]. Moscow: Informelektro. 434 p.
3. *Inkova O., Popkova N.* (2017). Statistical data as information source for linguistic analysis of Russian connectors. Informatics and applications. 2017. Vol. 11, No. 3. Pp. 123–131.
4. *Johansson S.* (2007) Seeing through multilingual corpora: On the use of corpora in contrastive studies. Amsterdam: John Benjamins. 377 p.
5. *Inkova O. Yu.* (2018a) Nadkorpusnaja basa dannykh kak instrument izutcheniya formal'noj variativnosti konnektorov [Supracorpora database as an instrument of the study of the formal variability of connectives]. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018". Moscow, May 30–June 2, 2018. http://www.dialog-21.ru/media/4299/inkovaoyu.pdf.
6. *Inkova O. Yu.* (2018б) Lingvospetsifitchnost' konnektorov: metody i parametry opisaniya [The language-specificity of connectives: methods and parameters of description]. Semantika konnektorov: kontrastivnoe issledovanie [Semantics of connectives: a contrastive study], O. Inkova (ed.). Moscow: TORUS PRESS. Pp. 5–23.
7. *Inkova O. Yu. ed* (2018в) Semantika konnektorov: kontrastivnoe issledovanie [Semantics of connectives: a contrastive study]. Moscow: TORUS PRESS.
8. *Kobozeva I. M., Inkova O. Yu.* (2018) Kak i ego dvukhmestnye varianty [Kak and its bi-places variantes]. Semantika konnektorov: kontrastivnoe issledovanie [Semantics of connectives: a contrastive study], O. Inkova (ed.). Moscow: TORUS PRESS. Pp. 168–239.
9. *Nuriev V. A.* (2018) Khotya. Semantika konnektorov: kontrastivnoe issledovanie [Semantics of connectives: a contrastive study], O. Inkova (ed.). Moscow: TORUS PRESS. Pp. 269–300.
10. *Prasad R., Joshi A., Webber B.* (2010) Realization of Discourse Relations by Other Means: Alternative Lexicalizations. Proceedings of the 23rd International Conference on Computational Linguistics (Beijing, China — August 23–27, 2010): Posters. Pp. 1023–1031.
11. *Taboada M.* (2009) Implicit and explicit coherence relations. Discourse, of Course. An overview of research in discourse studies, J. Renkema (ed.). Amsterdam/Philadelphia: John Benjamins. Pp. 127–140.

12. *Toldova S., Pisarevskaya D., Kobozeva M., Vasilyeva M.* (2018) The cues for rhetorical relations in russian: “cause–effect” relation in Russian rhetorical structure treebank. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2018”. Moscow, May 30–June 2, 2018. http://www.dialog-21.ru/media/4338/toldovas.pdf.
13. *Zalizniak Anna A., Mikaelyan I. L.* (2018) Soyuz a [La conjunction a]. Semantika konnektorov: kontrastivnoe issledovanie [Semantics of connectives: a contrastive study], O. Inkova (ed.). Moscow: TORUS PRESS. Pp. 24–79.
14. *Zalizniak Anna A., Zatsman I. M., Inkova O. Yu.* (2017) Nadkorpusnaja basa dannykh konnektorov: postroenie systemy terminov [Supracoprora database on connectives: term system development]. Informatics and applications. 2017. Vol. 11, No. 1. Pp. 101–109.
15. *Zatsman I. M., Inkova O. Yu., Kruzhkov M. G., Popkova N. A.* (2016) Predstavlenie krosslingvistitcheskikh znanij o konnektorakh v nadkorpusnykh basakh dannykh [Representation of cross-lingual knowledge about connectors in supracorpora databases]. Informatics and applications. 2016. V. 10. No. 1. Pp. 106–118.

AN ANAPHORA RESOLUTION SYSTEM FOR RUSSIAN BASED ON ETAP-4 LINGUISTIC PROCESSOR[1]

Inshakova E. S. (e.s.inshakova@gmail.com)
Laboratory of Computational Linguistics,
A. A. Kharkevich Institute for Information Transmission Problems, Moscow, Russia

The paper presents a rule-based system of automated anaphora resolution for Russian. The system is based on the resources of ETAP-4 linguistic processor: the Russian combinatorial dictionary (RCD), the ETAP parser, and the ontology OntoEtap. In this paper, I describe the ordered algorithms for resolution of different pronouns and provide the results of their evaluation.

Keywords: anaphora resolution, ETAP-4 linguistic processor

СИСТЕМА РАЗРЕШЕНИЯ АНАФОРЫ ДЛЯ РУССКОГО ЯЗЫКА НА БАЗЕ ЛИНГВИСТИЧЕСКОГО ПРОЦЕССОРА ЭТАП-4

Иншакова Е. С. (e.s.inshakova@gmail.com)
Лаборатория компьютерной лингвистики,
Институт проблем передачи информации им. А. А. Харкевича РАН, Москва, Россия

[1] This work was supported by the RSF grant 16-18-10422.

1. Introduction

In this paper, I present a system of automated anaphora resolution, which is a module of the ETAP linguistic processor [Boguslavsky et al., 2008][2]. This system is basically a system of deterministic rules written in the FORET formal language [Cinman 1995], with access to the Russian combinatorial dictionary (RCD) and the OntoEtap ontology [Boguslavsky et al. 2013]. It is domain and genre unspecific, meant to provide an antecedent for every pronoun in texts of any type; thus, it makes an important part of SemEtap—a system for deep semantic analysis and question answering being currently developed in IITP.

The paper has the following structure. In **section 2**, I describe the anaphora resolution module of ETAP: in **subsection 2.1**, I give its general outline, in **subsections 2.2**, **2.3**, **2.4** and **2.5** I discuss in more detail the resolution of reflexives, relative pronouns, the *tot* pronoun and 3rd person pronouns respectively. The subject of **section 3** is evaluation of the system's performance. **Section 4** contains conclusions and directions for further work.

2. Describing the system

2.1. A general overview

ETAP-4 is a linguistic processor whose main options are machine translation between Russian and English and semantic analysis of Russian texts; ETAP was also used to create SynTagRus[3], a Russian dependency treebank [Boguslavsky et al. 2014].

The stages of text processing in ETAP are as follows:

0. Splitting a text into sentences.
1. Morphological analysis.
2. Syntactic analysis.
3. Anaphora resolution.
4. Different options of translation from Russian (into semantic structures / into English).

The anaphora resolution module comes into action after the input text is fully parsed (its input can be a separate sentence / group of sentences or a corpus of parsed sentences in the .tgt format). It augments the parse trees with two types of non-tree links: 1) 'quasi-syntactic' links for zero anaphora that connect the syntactic host and the controller of zero pronouns; 2) coreference links between pronouns and the heads of their linear closest non-pronominal antecedents. All the mentions that form a coreferential chain are merged into coreference groups. Each coreference group automatically gets a unique number, which serves as a coreferential index.

[2] ETAP-4 is downloadable at http://proling.iitp.ru/etap4download.

[3] http://ruscorpora.ru/search-syntax.html.

Our algorithm is an ordered set of rules (each rule uses the results of earlier ones' work—coreference links or absence thereof). The (groups of) rules are listed here in the order they are applied:

1. A rule for various types of zero anaphora[4].
2. Rules for reflexive pronoun *sebja* '-self', reflexive possessive *svoj* '-self-s' and reciprocal *drug druga* 'each other'.
3. Rules for relative pronouns *kotoryj* 'which', *kto* 'who', *chto* 'that' and *chej* 'whose'.
4. A rule for the switch-reference pronoun *tot*.
5. Rules for 3rd person anaphora.

Like in many other rule-based or hybrid anaphora resolution systems (e.g. [Mitkov 2002], [Haghighi and Klein 2010], [O'Connor and Heilman 2013], [Lee et al. 2011]), the process of resolving 3rd person pronouns in ETAP falls into three stages:

a) filtering out the nouns that cannot control 3rd person pronouns at all;
b) generating a set of potential antecedents for every pronoun;
c) passing them through the ordered sieves of restrictional rules to select the right candidate:
    i. the ontological sieve;
    ii. the syntactic sieve that favors anaphoric pairs in certain syntactic configurations and filters out the other ones;
    iii. the discourse sieve that sorts out candidate antecedents on the basis of their relative discourse prominence.

2.2. Reflexive and reciprocal pronouns

The algorithm for reflexive (*sebja* 'oneself', *svoj* '-self's') and reciprocal (*drug druga* 'each other') pronouns includes the following stages:

1. Searching for a predicate Z—the head of the local finite or non-finite clause (including nominalizations) that contains the pronoun. That means finding a predicate (a verb, event noun, adjective or adverb) that governs the pronoun directly or via a sequential chain of words that are not finite verbs and don't have their own subjects.

2. Finding the subject of this clause (the 'short-distance' antecedent). It is the noun that depends on the predicate Z via PREDIC, or DAT-SUBJECT, or QUASI-AGENT / AGENT[5] (if Z is a noun) relation. For the reciprocal pronoun *drug druga*, the antecedent can be only short-distanced (within the same minimal clause).

---

[4] I won't discuss here the process and examples of resolution of this type of anaphora for space reasons and because zero anaphora annotation standard in ETAP has not yet been developed, so the quality of its resolution cannot be evaluated.

[5] See a comprehensive list of Russian SSRs in Mel'čuk 1974 or on the SynTagRus website, and of English SSRs in [Mel'čuk 1988].

Our rules handle reflexive cataphora: *Svoj*$_i$ *put' v zhurnalistiku on*$_i$ *nachal barabanschikom v Nju-Jorke*—lit. 'His$_i$ road to journalism, he$_i$ began as a drummer in New York'). They contain subrules for detecting oblique antecedents [Paducheva 1983], e.g. *Mne*$_i$ *ne hvataet vremeni dlja sebja*$_i$ 'I$_{DATi}$ am short of time for myself$_i$', *Kazhdomu avtoru*$_{DATi}$ *svoj*$_i$ *podhod kazhetsja bolee jestestvennym*—lit. 'To each author$_i$, their$_i$ own approach seems more commonsensical'.

3. If the predicate Z is a nominalization or an infinitive governed by a predicate with the feature OB-INF (and the pronoun is reflexive), the algorithm goes up the parse tree to the predicate that governs Z and selects its subject as the 'distant' antecedent (in Russian, the binding domain for reflexive pronouns is a finite clause: see Rappaport 1986). In most cases, the next rule selects the most recent antecedent, and only in the contexts like *zastavljat' sebja* + INF—the 'distant' one (...*ona predlozhila sozdat' jedinuju platformu*$_i$, *kotoraja*$_i$... *zastavit vlast' sebja*$_i$ *slushat'* '...she suggested creating a united platform$_i$ that$_i$... would make the government listen to it$_i$').

4. Ruling out idiomatic expressions where reflexive pronouns are used non-referentially: *vesti sebja* 'to behave', *ne po sebe* '[to feel] uneasy', *svoego roda* 'a sort of' and many others[6].

2.3. Relative pronouns

The rule for this type of anaphora first finds a noun (Z) that is the syntactic host of a relative clause, i.e. governs its main predicate (Q) via the RELAT relation. Then it proceeds to the predicate Q and goes down the tree to the relative pronoun (it finds pronouns that depend on Q immediately or via a preposition, a noun, a noun and a preposition). If the pronoun fills a valency slot of some predicate, the rule checks whether the noun Z fits its selectional restrictions.

(1) *protestantskaja etika*$_i$ (Z), *znachenie kotoroj*$_i$ *raskryl* (Q) *Maks Veber*...
'protestant ethics$_i$ (Z), the meaning of which$_i$ was (Q) revealed by Max Weber...'

If the noun Z depends on another noun (but not via APPOS, COORD or EXPLIC), which can also depend on a noun, the rule the rule checks the selectional restrictions for all those nouns to decide which of them is the antecedent.

However, many cases of such syntactic ambiguity require world knowledge, for which the state-of the-art ETAP does not yet provide any ready resources.

2.4. The pronoun "tot"

The rule for this pronoun is mainly based on the paper [Kreydlin, Chekhov 1988]. It also resolves the pronoun *tot* in '...*da i tot*'-constructions, which were not analysed in that paper. The rule deals with 3 types of contexts:

[6] See a full list of such expressions in [Inshakova 2016].

1. The antecedent of *tot* is an argument (but not a subject) in the clause that precedes the clause containing the pronoun *tot*. The predicate that governs its antecedent also has a topically more prominent argument (it is usually a subject):

(2) *Bauer$_j$ pytalsja privlech k sudu i avtora «Njurnbergskix zakonov» Globke$_i$, no tot$_i$ okazalsja jemu$_j$ ne po zubam…*
'Bauer$_j$ tried to bring Globke$_i$, the author of the 'Nuremberg laws', to trial, but he$_i$ turned out to be a hard nut to crack for him$_j$.'

2. The antecedent of *tot* is a nominal subject, and there is another nominal subject between the pronoun *tot* and its antecedent.

The rule checks number and gender agreement between the pronoun and its antecedent as well as selectional restrictions of the predicate that governs the pronoun. It prefers candidate antecedents from the class 'PhysicalObject' to other nouns.

2.5. 3rd person anaphora

2.5.1. Filtering out 'non-antecedents'

At the preliminary stage of 3rd person pronoun resolving, the system discards nouns that cannot be antecedents of such pronouns at all (by means of assigning them the NON-ANTEC feature). The list of 'non-antecedents' includes many groups of expressions; here I will give only four examples (others are listed in [Inshakova 2016]):

1. **Nominal predicates**:

(3) *On skazal, chto Efimova chestnyj, kvalifitsirovannyj rabotnik.*
'He said that Efimova was an honest, competent employee.'

(4) *Glavnyj element$_j$ konstruktsii vertoleta—rama$_j$.*
'The main element$_j$ of the helicopter structure is a frame$_i$.'

2. **Nouns** from the class **GradableParameter** or with the feature **CHARACTROD** 'genitive of characterization', depending by **ATTRIB**: *problema pervostepennoj vazhnosti* 'an issue of primary importance'; *kategorija eksponatov povyshennogo riska* 'the category of higher risk exhibit items'.

3. Nouns that depend by the **COMPL-APPOS** relation (directly or via a preposition): *dom nomer vosem'* 'house number eight'; *ves v sto tonn* 'a hundred-ton weight'; by **LOCUT**: *drug druga* 'each other'; *kuram na smex* ≈'enough to make a cat laugh'.

4. **Nouns in idiomatic expressions**, e.g. *slava bogu* 'thanks God', *chto tolku…* 'what's the use of…', *ne mozhet byt' i rechi o…*'[It's] out of the question', *delo ne v etom* 'it is not the point'.

2.5.2. Creating the set of candidate antecedents

The next rule picks out possible antecedents (within the search scope of 3 sentences—the current one plus two preceding sentences). It sets the following restrictions:

1. The antecedent's **location and POS feature**: to the left of the pronoun (our system does not resolve 3rd person cataphora yet). It can be a noun, an adjective or a participle (*mitingujuschie* 'protesters').

2. **Gender, number and person features:** agreement or disagreement, e.g.—*Ja*$_i$ *zhit' hochu!*—*zakrichal on*$_i$ '— I$_i$ want to live!—he$_i$ cried'; *redaktor Znamenskaja* 'editor-M Znamenskaja-F'; *gosudarstvo*$_i$-N *v litse prinadlezhaschix jemu*$_i$-M *monopolij* 'the state$_i$ in the person of monopolies that belong to it$_i$'. For unknown words with missing morphological features, the rule checks the gender and number of their predicates or adjectival modifiers (if there are any).

3. **Selectional restrictions** for the candidate antecedents (if the pronoun fills a valency of some word). Candidates that don't belong to any of the ontological concepts listed in the corresponding column of the predicate word's GP (if there are any) are filtered out.

4. **Syntactic restrictions**. For 3rd person pronouns, the best known ones are **binding principles B and C** (Chomsky 1981) or **co-dependency** for dependency trees [Paducheva 1977]. The rule forbids contexts like *Masha*$_i$ *znala bol'she nejo*$_{*i}$ 'Masha$_i$ knew more than her$_{*i}$'; *Petja*$_i$ *vidit jego*$_{*i}$ *dom* 'Petja$_i$ sees his$_{*i}$ house'; *Kak najti Petju*$_i$, *on*$_{*i}$ *ne znaet*—lit.'How to find Petja$_i$, he$_{*i}$ does not know'. There is also a counterpart of the i-within-i restriction, (*problemy*$_i$ *ix*$_{*i}$ *povtornoj restavratsii* 'issues$_i$ of their$_{*i}$ repeated restoring').

5. The rule forbids anaphoric links to conjuncts, except for cases like *arxitektor i sotsial'nyj reformator Vjacheslav Glazychev* 'architect and social reformer Vjacheslav Glazychev'.

6. The rule also forbids anaphoric links from subjects or addressees of speech verbs to nouns within the direct speech governed by these verbs:—*Skoro pridet Petja*$_i$,—*skazal on*$_{*i}$ *jemu*$_{*i}$ '— Petja$_i$ will come soon,—he$_{*i}$ told him$_{*i}$'.

2.5.3. Sorting out incorrect candidates

The architecture of ETAP processor does not allow the system to rank a set of candidate antecedents, which can consist of >15 words for each pronoun. Instead, the anaphoric rules are 'eliminating': they either discard the candidate antecedents that don't meet certain ontological or syntactic constraints, or take every two nouns/adjectives Q and Z from the set of possible antecedents, compare the relative prominence of Q and Z and delete the coreferential link with the less prominent candidate.

The **ontological sieve** deals with such selectional restrictions that are not specified in the government patterns in RCD, but can be extracted from the dictionary zone SEM-ONTO of the RCD lexemes, which contains their ontological correlates from the ontology OntoEtap.

Table 1. Ontological restrictions stipulated in the ontology-based rule

Ontological correlate of the word W that governs the pronoun	Ontological correlate the antecedent should have	Syntactic relation between W and the pronoun	Examples
1. IntentionalProcess, SocialRole, SocialRelation, BiologicalAttribute, IntentionalRelation, NormalBiologicalEvent, PathologicProcess, PhysiologicProcess, EmotionalBehavioralProcess, StateOfMind, Human, Proposition, Model, PropositionalAttitude, TraitAttribute, Intelligence	Agent	PREDIC / QUASI-AGENT / AGENT / DAT-SUBJECT	*On vidit* 'he/it sees'; *oni studenty* 'they are students'; *jego teorija* 'his theory'
2. Human&SocialRelation	Human	1-COMPL	*Jejo drug* 'her friend'
3. Creation, Manufacturing	Artifact	1-COMPL	*Izgotavlivat' ix* 'to produce them'
4. SocialInteraction	Human, Organization	2-COMPL + preposition *s* 'with'	*Videt'sja s nim* 'to meet him'
5. AnimalAnatomicalStructure	Animal	QUASI-AGENT	*Jego plecho* 'his shoulder'
6. ChemicalProcess, NaturalProcess, PhysicalAttribute, StateChange, SurfaceChange	PhysicalObject	PREDIC / QUASI-AGENT / AGENT	*On plavitsja* 'it melts'
7. Motion, ShapeChange, StateChange, SurfaceChange	PhysicalObject	1-COMPL	*Chistit' jejo* 'to clean it/her'
8. Artifact, GeopoliticalArea, Human, Animal	Human	ATTRIB	*Ix gorod / sobaka / mashina* 'their town / dog / car'
9. Event	Event	PREDIC	*On byl obrjadom* 'it was a ceremony'
10. Location, BodyPosition + certain locative prepositions	PhysicalObject	1-COMPL / ADVERB	*Ivan zhil cherez tri doma ot nejo* 'Ivan lived three houses away from her/it'

Candidates that don't belong to the needed ontological classes are filtered out, as the bolded noun in (5):

(5) *Nel'zja cheloveka*$_i$ *zastavit' idti na* **miting**$_{*i}$. *Jemu*$_i$ *mozhno rekomendovat', jego*$_i$ *mozhno prizvat'.*
'One cannot force people$_i$ to go to **rallies**$_{*i}$. They$_i$ can be recommended to, they$_i$ can be encouraged.'

The **lexical functional rule** is applied to pronouns that depend on verbs/nouns that are values of lexical functions, if some of the candidate antecedents belong to the set of arguments of the given LF, and some do not:

(6) *...vpolne jestestvennymi byli dva proekta, kotorye Xruschev predlozhil* **nomenklature**$_{*i}$ *... Vo-pervyx, snjat' s nejo otvetstvennost'*$_i$ *za terror, perelozhiv jejo*$_i$ *na Stalina...*
'...two projects Khrushchev offered to the **Nomenklatura**$_{*i}$ were quite natural... Firstly, to exonerate it from responsibility$_i$ for the terror and to shift it$_i$ on Stalin...'

The **syntactic sieve** up to date deals with the following types of constructions:

1. Coordinate chains where the antecedent is an n^{th} conjunct noun and the pronoun directly or indirectly depends on the n+1th conjunct noun:

(7) *Vo vremja Olimpiady politsejskim*$_{*i}$ *dano ukazanie: vystupat' protiv bezdomnyx*$_i$ *i zaschischajuschix ix*$_i$ *aktivistov...*
'During the Olympics, policemen$_{*i}$ were instructed to force against the homeless$_i$ and the activists who defended them$_i$ ...'

2. Specifying constructions where the minimal clause that contains the pronoun depends on its antecedent (or its syntactic host) via EXPLIC or JUXTAPOSE relation:

(8) *Ljudi*$_{*i}$ *zdes' xodjat s telezhkami*$_i$ *iz supermarketov: v nix*$_i$ *udobno skladyvat' banki i butylki.*
'People$_{*i}$ here walk with shopping carts$_i$: they$_i$ are convenient to collect cans and bottles.'

3. Some types of syntactic parallelism. These are present in the following constructions:

1) In a coordinate chain, the antecedent depends on the n^{th} conjunct via a syntactic relation R, and the pronoun depends on the n+1th conjunct via the same relation R:

(9) *...ljudej*$_i$*, ix*$_i$ *otnoshenija*$_{*i}$ *s Bogom, ix*$_i$ *put', prisutstvie Xrista v nix*$_i$.
'...people$_i$, their$_i$ relationship with God, their$_i$ [spiritual] path, presence of Christ in them$_i$.'

2) The antecedent depends on a word that belongs to a lexeme L via a syntactic relation R, and the pronoun depends on another word that is also a form of the lexeme L, via the same relation R:

(10) *...po isku*$_{*i}$ *...on priznal ispol'zovanie logotipa*$_i$ *nezakonnym i nalozhil zapret na jego*$_i$ *ispol'zovanie.*
'...on the suit$_{*i}$ ... he declared the use of the logo$_i$ to be illegal and imposed a ban on its use$_i$.'

Discourse rules compare candidates by pairs. They deselect a candidate antecedent:

1) if it is an adjunct (i.e. depends via ADVERB, or ATTRIB, or 4/5-COMPL relation) located 1 or 2 sentences away from the pronoun or it is inside an adjunct clause, and the other candidate holds an argument position in the current sentence.
2) if it is a possessor of another candidate noun, which belongs to the ontological class PhysicalObject.
3) if it is not not the previous finite clause's 'forward-looking center' in terms of the Centering theory [Grosz et al., 1995]:

(11) *Kogda krestjanin*$_i$ *zhenil starshego* **syna**$_{*i}$, *on*$_i$ *schital sebja*$_i$ *«starikom» i otdeljalsja so «staruxoj» v otdel'noe pomeschenie.*
'When a peasant$_i$ married off [his] eldest **son**$_{*i}$, he$_i$ considered himself$_i$ an "old man" and moved with his$_i$ "old-wife" to a separate room.'

If, after all those sieves are applied, there still remain several possible antecedents for a pronoun, the last rule selects the most recent candidate.

3. The testing corpus, evaluation and error analysis

Our system participated in the AnCor evaluation campaign (2019), where only 3rd person pronouns were considered. It was evaluated on the test corpus of texts collected from the OpenCorpora online corpus (opencorpora.org), untagged but split into sentences. Its performance on this corpus and for this class of pronouns turned out to higher in precision but poorer in recall (micro-averaged) than the results shown on ETAP team's own test set. In this paper, I will also present the results of evaluation on our own testing corpus[7].

This corpus is an anaphorically annotated subset of SynTagRus in the .tgt format, collected in 2017–2019. It comprises 43 texts (mainly newspaper articles and fiction, a small amount of news texts and interviews), and in total is 6,315 sentences long. It contains 3,621 pronoun—antecedent pair.

When annotating the testing corpus I kept to the 'soft' criterion of annotation, i.e. agreed with the tagger's choice of non-closest and/or pronominal antecedents, if the resulting coreference groups were correct. Anaphora to coordination chains, comitative constructions and split antecedents was annotated as several anaphoric links from a pronoun to each element of its disjoint antecedent.

The SynTagRus-based corpus comes in two versions: 1) with 'gold', manually corrected syntactic structures and morphological tags, and 2) with 'raw', uncorrected structures and morphology. In this version, unknown words (= absent in RCD) were often unidentified by ETAP parser and lack morphological and POS tags. In this way, the impact of syntactic and morphological correctness on the system's performance can be estimated (**Table 2**).

[7] Texts from the RuCor corpus ([Toldova et al., 2014]; downloadable at http://rucoref.maimbava.net), parsed by ETAP, comprised the development corpus.

Table 2. The results of the system's evaluation

Corpus	Pronouns	Syntactic structures	Criteria	Precision	Recall	Pairwise F1
AnCor	3rd person	'Raw'	macro, soft	69.90	53.80	59.30
			micro, soft	78.70	52.40	62.90
			macro, strong	58.10	45.00	49.40
			micro, strong	58.70	39.10	46.90
SynTagRus	3rd person	'Raw'	macro, soft	68.00	63.78	65.82
			micro, soft	66.81	62.46	64.56
		Gold	macro, soft	72.51	68.72	70.56
			micro, soft	75.40	71.05	73.16
	Reflexive and reciprocal	'Raw'	macro, soft	88.31	84.14	86.17
			micro, soft	84.76	76.25	80.28
		Gold	macro, soft	91.78	89.75	90.75
			micro, soft	90.94	87.26	89.06
	All	'Raw'	macro, soft	76.96	71.23	73.98
			micro, soft	73.78	68.04	70.79
		Gold	macro, soft	82.33	78.76	80.50
			micro, soft	81.99	78.12	80.01

Because the ETAP-based system extensively employs syntactic information, it is quite predictable that its performance depends on the quality of parsing to some extent. However, this dependence is not as strong as might be expected. The difference in performance on SynTagRus-based corpora with 'gold' and 'raw' syntactic structures for 3rd person pronouns is 4.74 F1 points / 6.72% (macro), or 8.60 F1 points / 11.76% (micro). For reflexive and reciprocal pronouns the difference is 4.58 F1 points / 5.05% (macro), or 8.78 F1 points / 9.86% (micro). For all types of pronouns it is 6.52 F1 points / 8.10% (macro), or 9.22 F1 points / 11.52% (micro). The fact that correct parse trees turn out to have a lower impact on resolution of reflexives and reciprocals than on resolution of pronominals is also somehow against expectations (because the correctness of structure seems to be crucial for syntactic anaphora resolution).

Error analysis. The comparison of scores for pronominals vs. reflexives and reciprocals supports the conclusion from [Toldova et al. 2016] that systems that handle syntactic anaphora quite well tend to have more mistakes in discourse anaphora resolution. **Table 3** shows the actual role of salience-related errors (in the AnCor test set, the ETAP-based system made 250 errors in total).

In fact, the top three causes of errors have something to do with discourse issues. Such factors as spelling and punctuation errors (present in the AnCor corpus but not in the SynTagRus-based corpora) and unknown words, mostly proper names, on the contrary, don't decrease the score as dramatically as discourse rules do, although ETAP was designed to parse error-free texts and fails to process such words and sentence structures.

Table 3. Distribution of errors

Error type	Percentage
1. Discourse salience rules discard the correct candidate	19.2
2. Parsing errors and incorrect morphological tags that result in impossible syntactic configurations for pronouns and their antecedents (e.g. i-within-i, incorrect binding, attaching pronouns to argument positions with such selectional restrictions that disallow the antecedent, etc.). Some of the discourse-related errors may also be caused by incorrect parsing	15.2
3. Wrong choice of the linear closest candidate (i.e. the rules' failure to choose the most prominent candidate)	12.8
4. Antecedent beyond the search scope of 2 sentences	10.0
5. Too restrictive ontological rule that can deselect the correct candidates because they 'don't satisfy' the ontological constraints of words that govern the pronouns	8.0
6. Too restrictive rule that creates anaphorical links	6.8
7. The antecedent is an unknown lexeme	4.8
8. Pronoun—antecedent number disagreement (e.g. *Zimbabwe—they*)	4.8
9. Imperfections in RCD entries (e.g. lack of selectional restrictions or certain lexical functions)	3.6
10. The parser selects a wrong homonym (e.g. *predpolagat'1* 'suppose' instead of *predpolagat'3* 'presuppose')	2.4
11. Misprints and punctuation errors	2.4
12. Incorrectly assigned or lacking NON-ANTEC feature (for non-referential expressions)	2.4
13. Incorrect choice of a non-closest candidate in the case of JUXTAPOSE construction	1.6
14. Malfunction of the LF-based rule	1.6
15. Other types of errors (a single error for each type)	4.0

4. Conclusion and further work

The first item on our laboratory's agenda now is developing non-pronominal coreference resolution algorithms. Secondly, the anaphora resolution system has to access sources of world/encyclopaedic knowledge and to fully use the potential of the ontology OntoEtap (which is still rather limitedly used in anaphora resolution) and its inference engine. This will enable it to resolve more complicated cases of anaphora, like those in the Winograd Schema Challenge, at the stage of semantic analysis (see an account of the initial stage of work on this issue in [Boguslavsky et al. 2019]). Thirdly, the anaphorically annotated part of the SynTagRus corpus has to be made publically available.

References

1. *Apresyan Yu. D. et al.* (2007) Lexical Functions in Actual NLP-Applications. In: Wanner L. (ed.): Selected Lexical and Grammatical Issues in the Meaning–Text Theory. In honour of Igor Mel'čuk. Studies in Language Companion. Series 84. Amsterdam: Benjamins Academic Publishers. P. 199–230.
2. *Boguslavsky I. M.* (2014) SynTagRus—a deeply annotated corpus of Russian. In: Blumenthal P., Novakova I., Siepmann D. (eds): Nouvelles perspectives en sémantique lexicale et en organisation du discours. Frankfurt am Main. P. 367–379.
3. *Boguslavsky I. M. et al.* (2008) Parser of the ETAP system and its evaluation with the aid of a deeply annotated corpus of Russian texts [Sintaksicheskij analizator sistemy ETAP i jego otsenka s pomoschju gluboko razmechennogo korpusa russkix tekstov]. In: Proceedings of the international conference 'Corpus linguistics—2008' (St. Petersburg). P. 56–74.
4. *Boguslavsky I. M. et al.* (2013). Semantic representation for NL understanding. Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2013), p. 132–144.
5. *Boguslavsky I. M. et al.* (2019) Knowledge-based approach to Winograd Schema Challenge. In: current volume.
6. *Chomsky N.* (1981) Lectures on Government and Binding. Foris.
7. *Grosz, B., Aravind J. and Weinstein, S.* (1995). Centering: a framework for modelling the local coherence of discourse. Computational Linguistics, 21 (2), 203–225.
8. *Haghighi A. and Klein D.* (2010) Coreference resolution in a modular, entity-centered model. In Proc. of HLT-NAACL. P. 385–393.
9. *Inshakova E. S.* (2016) Resolution of syntactic pronominal anaphora in the ETAP-3 system [Razreshenie sintaksicheskoj mestoimennoj anafory v sisteme ETAP-3]. In: Information Technology and Systems (ITaS'16). Proceedings of the 40th Interdisciplinary Conference and School (St. Petersburg, 2016).
10. *Kreydlin G. E., Chekhov A. S.* (1988) Interrelation of semantics, information structure and pragmatics in lexicographic description of anaphoric pronouns (the case of pronouns of the TOT group) [Sootnoshenie semantiki, aktual'nogo chlenenija i pragmatiki v leksikograficheskom opisanii anaforicheskix mestoimenij (na materiale mestoimenija gruppy TOT)] Institute of Russian Language of AS USSR. The experimental and applied linguistics task group [IRYa AN SSSR. Problemnaja gruppa po eksperimental'noj i prikladnoj lingvistike]. Preprints. Iss. 178. Moscow.
11. *Lee H. et al.* (2011) Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In Proceedings of the CoNLL-2011 Shared Task.
12. *Mel'čuk I.* (1974) Toward a theory of Meaning ⇔ Text linguistic models [Opyt teorii lingvisticheskix modelej "Smysl ⇔ Text"]. Moscow.
13. *Mel'čuk I.* (1988) Dependency syntax: Theory and practice. Albany, NY.
14. *Mitkov R.* (2002) Anaphora resolution. Longman.
15. *O'Connor B., Heilman M.* (2013) Arkref: A rule-based coreference resolution system. arXiv preprintarXiv:13101975.

16. *Paducheva E. V.* (1977) On the semantics of syntax. Materials toward the transformational grammar of Russian [O semantike sintaksisa. Materialy k transformatsionnoj grammatike russkogo jazyka]. Moscow.
17. *Paducheva E. V.* (1983) Reflexive pronoun with an oblique antecedent and the semantics of reflexivity [Vozvratnoe mestoimenie s kosvennym antetsedentom i semantika refleksivnosti]. Semiotics and Informatics [Semiotika i informatika]. Iss. 21. P. 3–33.
18. *Rappaport G. C.* (1986) On anaphor binding in Russian. In: Natural Language & Linguistic Theory, 4(1):97–120.
19. *Toldova S. Ju., Roytberg A., Nedoluzhko A., Kurzukov M., Ladygina A., Vasilyeva M., Azerkovich I., Grishina Y., Sim G., Ivanova A., Gorshkov D.* (2014) Evaluating Anaphora and Coreference Resolution for Russian. Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue". Issue. 13(20). P. 681–695.
20. *Toldova S., Roytberg A., Ladygina A., Azerkovich I., Vasilyeva M. D.* (2016) Error analysis for anaphora resolution in Russian: new challenging issues for anaphora resolution task in a morphologically rich language, in: Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016), co-located with NAACL 2016, San Diego, California, June 16, 2016. Stroudsburg, PA: Association for Computational Linguistics. P. 74–83.

В КОПИЛКУ МИКРОСИНТАКСИЧЕСКИХ НЕОЖИДАННОСТЕЙ: ДВЕ РУССКИЕ АНТОНИМИЧНЫЕ СИНТАКСИЧЕСКИЕ ФРАЗЕМЫ С КОМПАРАТИВАМИ[1]

Иомдин Л. Л. (iomdin@iitp.ru)
Институт проблем передачи информации
им. А. А. Харкевича РАН, Москва, Россия

Данная статья продолжает серию исследований микросинтаксиса русского языка. Подробно рассматриваются две конструкции, которые достаточно близки по синтаксическому устройству и по семантике: это единицы типа *как можно лучше* и *как нельзя лучше*, в которых первые два словесных элемента фиксированы лексически, а третий фиксирован грамматически, поскольку заполняется формой компаратива. Показано, что между этими единицами есть существенные семантические различия; в частности, первая из них ориентирована проспективно (*сыграй как можно лучше* vs. ?*сыграл как можно лучше*), а вторая — скорее ретроспективно (*все сложилось как нельзя лучше* vs. ?*Реши эту задачу как нельзя лучше, чтобы сдать экзамен*). Рассматриваемый материал используется также для уточнения некоторых общих представлений о русском компаративе.

Ключевые слова: микросинтаксис, синтаксические фраземы, компаратив, Микросинтаксический словарь русского языка, корпус с микросинтаксической разметкой

ADDING TO THE TREASURY OF RUSSIAN MICROSYNTACTIC CURIOSITIES: TWO ANTONYMIC SYNTACTIC IDIOMS WITH COMPARATIVES

Iomdin L. L. (iomdin@iitp.ru)
Institute for Information Transmission Problems (Kharkevich Institute), Russian Academy of Sciences, Moscow, Russia

[1] Работа выполнена при поддержке Российского фонда фундаментальных научных исследований (грант № 19-07-00842). Автор выражает фонду искреннюю признательность.

The paper continues a series of research studies into the microsyntax of Russian. Two constructions that are sufficiently close to each other in syntactic structure and semantics are considered in detail: these are linguistic units of the type *kak možno lučše* ≈ 'in the best way possible' and *kak nelzja lučše* '≈ 'it can never be better'. In both constructions, the first two elements are determined lexically while the third one is fixed grammatically since it can be instantiated by (almost) any comparative form. It is demonstrated that the two units possess substantial semantic differences; in particular, the former unit is oriented prospectively (cf. *sygraj kak možno lučše* 'play as well as you possibly can' but hardly ?*sygral kak možno lučše* ≈ 'he has played as well as he possibly could') while the latter unit is, rather, oriented respectively (cf. *vse složilos' kak nel'zja lučše* ≈ 'everything turned out in a way that could never be better' but hardly ?*Reši etu zadaču kak nel'zja lučše, čtoby sdat' ekzamen* ≈ 'solve this problem in a way that could never be better, to pass the exam'. The material under consideration is also used to discuss certain general subtleties of the Russian comparative.

Keywords: Microsyntax, syntactic idioms, comparative, Microsyntactic dictionary of Russian, corpus annotated with microsyntactic elements

1. Вводные замечания

Как показывает опыт составления Микросинтаксического словаря русского языка, [Иомдин 2018], работа над которым ведется в рамках микросинтаксических исследований автора (см., в частности, [Иомдин 2017], с дальнейшей библиографией) значительная часть содержащихся в нем элементов приходится на долю адвербиалов, т.е. таких микросинтаксических лексических единиц, поведение которых близко к поведению тех или иных классов наречий. Например, адвербиальное выражение *все равно1* 'независимо ни от чего' в конструкциях типа

(1) ***Все равно*** *ты не слышишь,* ***все равно*** *не услышишь ни слова, //****все равно*** *я пишу, но как странно писать тебе снова* (И. Бродский)[2],

ведет себя как сентенциальное наречие типа *наверняка, несомненно, обязательно,* а выражения *все равно2* 'безразлично', как в ...

(2) *Нам все равно — листы ли, листья — Как называется предмет,// Каким — не только для лингвистов — Дышать осмелился поэт* (В. Т. Шаламов).

и *все равно3* 'равносильно', как в

(3) *Наблюдать умиранье ремесел — // Все равно что себя хоронить* (А. А. Тарковский)

ведут себя как предикативные наречия (у которых могут быть собственные актанты, как субъект состояния, выраженный словом *нам* в примере (2)).

[2] Большинство приводимых в статье примеров заимствовано из основного и поэтического подкорпуса НКРЯ или глубоко аннотированного корпуса СинТагРус.

Адвербиальные выражения *между прочим* и *между тем* — тоже элементы микросинтаксического словаря — ведут себя как вводные наречия типа *например,* а выражение *не у дел* по синтаксическому поведению похоже на локативное наречие типа *здесь* или *вокруг*: оно легко сочетается с глаголом-связкой, как в

(4) *При мутации этого белка возбудитель СПИДа оказывается* ***не у дел*** (А. Волков).

Среди адвербиальных микросинтаксических единиц заметное место занимают словосочетания типа «предлог + существительное», такие как *без спросу, в частности, в меру* (ср. *все хорошо в меру), до отказа, за рубежом, на вкус* (ср. *приятный на вкус*), *под силу, про себя* (ср. *читайте про себя, а не вслух*), значение которых подчас очень далеко от значения мотивирующего существительного, что и обусловливает целесообразность постулирования таких выражений как самостоятельных лексических сущностей.

Рассматриваемые здесь единицы типа *как можно лучше* и *как нельзя лучше* в своих основных значениях тоже являются адвербиалами. Чаще всего они выступают как наречия образа действия:

(5) *Его взыскательные уши // Еще упрашивали мглу, // И лед, и лужи на полу // Безмолвствовать* ***как можно суше*** (Б. Л. Пастернак) (≈ '… безмолвствовать очень сухо');

(6) *В Европе холодно. В Италии темно. // Власть отвратительна, как руки брадобрея. // О, если б распахнуть, да как нельзя скорее, // На Адриатику широкое окно* (О. Э. Мандельштам) (≈ '… распахнуть очень скоро …').

Ниже эти микросинтаксические единицы будут рассмотрены подробно. Изложение материала строится по следующему плану. В **разделе 2** дается описание синтаксических и сочетаемостных особенностей обеих единиц и некоторых их аналогов. **Раздел 3** посвящен нетривиальным семантическим свойствам, характеризующим этот класс выражений. В **разделе 4** приводятся некоторые замечания, касающиеся семантики русского компаратива в целом и вытекающие из наблюдений над рассмотренными единицами. В заключение (**раздел 5**) обосновывается трактовка двух рассматриваемых единиц как антонимичных, несмотря на существование высказываний, содержащих эти единицы и являющихся ситуативно равнозначными.

2. Синтаксис и сочетаемость

2.1. Синтаксическая структура: синхронный срез

Оба рассматриваемых типа фразем — *как можно* + COMP и *как нельзя* + COMP синтаксически организованы единообразно. Каждая из них состоит из трех словесных элементов. Порядок этих элементов жестко фиксирован: на первом месте стоит элемент *как*, на втором — предикативное наречие (*можно* или *нельзя)* и на третьем — форма компаратива (наречного или адъективного).

Синтаксической вершиной всего выражения является именно форма компаратива. В любых конструкциях, где встречается такое выражение, можно удалить первые два элемента, и грамматическая правильность сохранится: ср.

(7) … *Беспокойство своё я постарался спрятать как можно глубже и ничем его не проявлять* (М. А. Булгаков) vs. *Беспокойство своё я постарался спрятать глубже*;

(8) *Имя у него было как нельзя более подходящее — псевдонима не требовалось* (В. Белоусова) vs. *Имя у него было более подходящее*.

Обратное, разумеется, неверно: отнюдь не к каждому компаративу можно добавить элементы *как можно* или *как нельзя*, чтобы конструкция осталась правильной:

(9) *Он живет гораздо ближе,* но не **Он живет гораздо как можно ближе,*

(10) *Пушкин родился раньше Лермонтова,* но не **Пушкин родился как нельзя раньше Лермонтова*.

Причины такой несимметричности мы рассмотрим ниже, в **разделе 3**.

Компаратив, занимающий третью позицию конструкции, может быть наречным, как в (7), (9) и (10), или адъективным, как в (8). Он может быть как синтетическим (типа *интереснее*), так и аналитическим (типа *более интересно*); кроме того, он может выражаться и сочетанием, в котором (качественное) наречие или прилагательное подчиняется слову *менее*:

(11) *Он старался быть как можно менее заметным*.

В подавляющем большинстве случаев рассматриваемые конструкции состоят ровно из трех слов: *как, можно/нельзя,* компаратив. Бывает, однако, что между *как* и предикативом появляется частица, зависящая от этого предикатива:

(12) *Ради бога, Макар Алексеевич, как только можно скорее займите сколько-нибудь денег* (Ф. М. Достоевский);

(13) *Каждый из вас должен настрелять как только можно больше зайцев* (И. Грекова).

До середины XX века в ходу был и вариант конструкции, где перед предикативом *можно* стояла частица *ни* (по-видимому, совершенно синонимичный современной трехсловной конструкции):

(14) *Насчет Батурина один исход вижу: продать как ни можно скорей, пока с молотка не продали.* (И. А. Бунин).

Изредка в первой из рассматриваемых фразем слово *как* заменяется количественным наречием *сколько, насколько* и *сколь*:

(15) *В возможно скором времени вы наберете и оттиснете сколько можно более экземпляров, и затем всю зиму разбрасывать* (Ф. М. Достоевский).[3]

(16) *Раковский привез нынче в 6 ч. вечера требование сколь можно скорее оставить Одессу* (И. А. Бунин).

(17) *— Это по пути, — насколько можно терпеливей объяснил я.* (Е. Евтушенко).

Что касается формального синтаксического представления наших конструкций, то представляется вполне допустимым считать, что на синхронном уровне компаративу подчиняется предикатив *можно/нельзя,* а последнему подчиняется слово *как* (которое в таком случае следует признать местоименным наречием). В терминах синтаксических зависимостей, принятых в теории «Смысл — Текст» И. А. Мельчука, разумно использовать для обеих внутриконструкционных связей одно и то же вспомогательное синтаксическое отношение, применяемое для описания синтаксических фразем:

как ←вспом— *можно* ←вспом— *лучше*.

2.2. Синтаксическая структура в диахронии

Если же обратиться к истории развития синтаксиса наших конструкций, то тут дело обстоит не столь очевидным образом. До сих пор достаточно часто встречаются случаи, когда вместо предикатива *можно* используются личные глаголы с модальным значением (в первую очередь *мочь* и *суметь*), ср.

(18) *В командировки она теперь рвалась как могла чаще* (Ю. Трифонов);

(19) *Я, как только мог небрежнее, спросил у Евгении Петровны, есть ли письма от Вас* (И. А. Гончаров);

(20) *— Арсентий, вырежь доски для печати, и как можешь скорей.* (Ю. Арбат).

(21) *На генералке в БДТ я позволил себе как мог деликатней дать понять, что исполнение одной из женских ролей в Москве Нифонтовой кажется мне предпочтительней* (С. Алешин);

(22) *— Зато ты по-прежнему стройная, — сказал, как сумел, добродушней* (Н. В. Кожевникова).

На наш взгляд, считать части конструкции типа *как могла* в (18), *как только мог* в (19) и т. д. синтаксически зависящими от компаратива было бы большой натяжкой: нетрудно видеть, что форма модального глагола (число, род, даже лицо в (20)) определяется тут полнозначным глаголом, при котором компаратив выступает в роли обстоятельства (*рвалась, спросил, вырежь* и т. д.),

[3] Обращает на себя внимание и сочетание *возможно скором* в начале предложения: это вариант нашей конструкции, при котором компаратив заменяется положительной степенью прилагательного; *в возможно скором времени = в как можно более скором времени.*

при этом сам этот глагол может стоять как в постпозиции к модальному в (18), так и в препозиции к нему в (19).

В примере (21) дело обстоит еще сложнее: обстоятельство *деликатней* характеризует выражение *дать понять,* а модальный глагол согласуется с вершинным глаголом *позволил.* Тем самым предложения типа (18)–(22) содержат по существу два глагольных центра, соотношение между которыми требует отдельного исследования, которое невозможно предпринять в рамках короткой статьи. Мы ограничимся указанием на то, что в некоторых случаях, возможно, в качестве одного из этих центров выступает вводная конструкция (*как только мог небрежнее* в предложении (19)), а также добавим, что и фразы с конструкциями, составляющими центральный предмет нашего рассмотрения, скорее всего не укладываются в рамки простого предложения, представляются скорее сложносоставными (при всей неопределенности этого термина) и требуют дальнейшего исторически ориентированного изучения. Приведем лишь один исторический пример, из которого вытекает, что и частеречный статус первого слова конструкции — *как* — нельзя определить однозначно как наречие: в предложении из НКРЯ

(23) *Позади же сего видно, как сей высокомысл сам во всех ищет*; *а об нем все думают так мало, как не можно меньше* (Н. И. Новиков, 1769),

последняя часть которого очевидно совпадает с нашей конструкцией, элемент *как* явно представляет сравнительный союз.

2.3. Синтаксические роли конструкций

Набор синтаксических ролей, которые наши конструкции могут выполнять в предложении, в общем совпадает с теми ролями, которые играют изолированные компаративы. Так, в примере (7) конструкция *как можно глубже* является обстоятельством при глаголе *спрятать*, в примере (8) имеет место аналитический компаратив от прилагательного *подходящий,* а вся конструкция *как нельзя более подходящее* представляет собой именную часть сказуемого. В предложении

(24) *Но он всегда старался производить на халдеев как можно более двусмысленное впечатление…* (В. Пелевин)

конструкция *как можно более двусмысленное* является определением к существительному *впечатлению,* а в предложении

(25) *Старайся сделать «твою» жизнь как можно интенсивнее, остальное приложится* (В. Г. Короленко)

конструкция *как можно интенсивнее* играет роль второго дополнения при глаголе *сделать.* Наконец, в предложении

(26) *Надо, чтобы пришло как можно больше народу*

конструкция *как можно больше* (вместе со словом *народу* — заполнителем валентности *больше* играет при глаголе *пришло* роль подлежащего.

2.4. Отсутствие второго компарата

Интересно, что наши конструкции, несмотря на наличие в них компаратива, не могут присоединять к себе второго компарата — ни в родительном падеже, ни с помощью союза *чем*: выражения типа **работать как можно лучше Пети, *работать как можно лучше всех, *работать как можно лучше, чем Петя* находятся за пределами грамматической нормы. Единичные примеры типа

(27) … *конкурс, в течение которого надо пройти в решении поставленной задачи как можно лучше других участников,*

встречающиеся в небрежных текстах, представляются неграмматичными, да и семантически избыточными.

2.5. Неприемлемость аттенуатива

В принципе, наши конструкции не принимают смягченной формы компаратива на *по-*. Выражения типа ?*как можно получше,* ?*как нельзя покрупнее* представляются весьма странными, хотя первое из них изредка встречается. Показательными являются соответствующие цифры встречаемости конструкции типа *как можно получше* в НКРЯ: если таких (строго трехсловных) конструкций с нормальным компаративом в основном корпусе НКРЯ нашлось около 10 тысяч, то конструкций со смягченным компаративом — около сотни (т. е. приблизительно в сто раз меньше), причем из них на современные тексты (начиная с середины XX века) приходится всего порядка 15 вхождений; ср.

(28) … *Я колесил по городу без цели и при разминках с «Волгами», окрашенными в голубой цвет, стремился прижаться к ним как можно поближе.* (К. Воробьев).

По мнению автора, замена стандартного компаратива смягченным, даже если признать ее грамматичной, ни в одном случае не является семантически оправданной, поскольку не приводит ни к какому изменению смысла. Это наблюдение, на наш взгляд, согласуется со сделанным ранее наблюдением о редкой встречающимися сейчас сочетаниями аттенутивного компаратива с наречиями высокой степени типа ?*гораздо побольше,* ?*намного поинтереснее* и т. п. [Богуславский-Иомдин 2009].[4]

Добавим, что вторая синтаксическая фразема с *как нельзя,* в отличие от *как можно* + COMP, допускает опущение компаратива, вместо которого могут фигурировать некоторые наречия и адвербиалы:

(29) *Поэтому почтовый грузовик с брезентовым тентом над кузовом пришелся как нельзя кстати* (В. Солоухин) (= *как нельзя более кстати*);

[4] Заметим, что близкое по форме ходовое выражение *куда подальше* на самом деле семантически отличается от намного подальше: в этой конструкции *куда* скорее всего, используется вместо неопределенного местоименного наречия *куда-либо*.

(30) *Приключение их закончилось как нельзя вовремя: к посту возвращались Антон и сталкер, и с ними шел кто-то еще* (Д. Глуховский);

(31) *Здесь как нельзя к месту вспоминается, что писал Гоголь своему «близорукому приятелю»* (И. Липовецкая) (= *как нельзя более к месту).*

2.6. Аналоги

У первой из рассматриваемых нами конструкций — *как можно* + COMP — есть два синонимичных варианта, которые, хотя и не могут тягаться с ней по употребительности, тем не менее представлены в текстах достаточно широко.

Первый из таких вариантов — микросинтаксическая конструкция *возможно* + *COMP,* ср.

(32) *В августе И. Е. Репин, с которым я виделся почти ежедневно, попросил меня передать Владимиру Галактионовичу его горячую просьбу — посетить возможно скорее «Пенаты».* (К. И. Чуковский).

Иногда эта двухсловная конструкция дополняется в препозиции наречием *сколь, сколько, насколько* и даже *как*, без какого-либо изменения в значении:

(33) *Особенно много опытов было сделано на обезьянах, причем хотели насколько возможно детальнее изучить вопрос о работе больших полушарий* (И. П. Павлов).

(34) *А случается и так, что регулировщики бывают сознательными организаторами аварий, потому что потом зарабатывают на ремонте, растягивая его насколько возможно подольше и делая его насколько можно хуже* (Е. Евтушенко).[5]

(35) *Ей показалось, что и она и все они притворяются, и ей стало так скучно и неловко в этом обществе, что она сколько возможно менее ездила к графине Лидии Ивановне* (Л. Н. Толстой).

(36) *За порубку лесов надо было взыскивать сколь возможно строже, но за загнанную скотину нельзя было брать штрафов* (Л. Н. Толстой).

(37) *Имеем в виду одно обстоятельство: чтобы для начальства как возможно меньше беспокойства было* (М. Е. Салтыков-Щедрин).

Второй вариант конструкции *как можно* + COMP носит разговорный характер и характеризуется существенно более ограниченной сочетаемостью: это микроконструкция *как* + COMP, фигурирующая в примерах типа.

(38) — *А зачем ты … крысу на поднос повесил? — Хотел как чуднее сделать* (А. Аверченко);

---

[5] Обратим внимание, что в одном и том же предложении фигурируют два разных варианта конструкции — *насколько возможно* + COMP и *насколько можно* + COMP.

(39) *Что ты на меня злишься? Мне хочется как лучше.* (Д. Гранин);

(40) *Осциллограф — это хорошо, но я хотел как проще* (Форум 2012 года, НКРЯ).

(41) *Может, в России только так и надо? Больше не желаем «как лучше»?* (М. Гиматов).

Эта конструкция встречается почти исключительно в контексте глаголов желания (практически всегда это *хотеть* или *хотеться*), почти исключительно с участием наречного компаратива и почти исключительно при опущении предиката, относительно которого уместно говорить о способе действия. В приведенных примерах этот предикат не опущен лишь однажды — в (38) ('сделать как можно чуднее'). Львиная доля примеров здесь приходится на сюжет *хотели как лучше* (с продолжением *...а получилось как всегда*), т. е. если использовать нашу базовую конструкцию, 'хотели сделать как можно лучше'): благодаря этому сюжету встречаемость данного варианта конструкции в текстах оказывается довольно заметной. Следует добавить, правда, что эту конструкцию очень трудно автоматически или хотя бы полуавтоматически детектировать в корпусах из-за весьма высокой доли false positives — при микросинтаксической разметке корпуса такими выражениями придется полагаться в основном на ручную обработку.

3. Семантика

Рассмотрим теперь основные семантические особенности наших синтаксических фразем. Этих особенностей три.

Во-первых, обе единицы — *как можно* + COMP и *как нельзя* + COMP **антропоцентричны**. В смысле любого высказывания, содержащего такие конструкции, обязательно присутствует человек, совершающий какое-либо действие. Эти конструкции не используются при описании каких-либо природных фактов, не предполагающих участия человека. Нельзя, например, сказать что-либо вроде

(42) **В словаре развитого языка как можно больше слов,*

(43) **Крупные метеориты падают на землю как нельзя реже* и т. д.

В тех случаях, когда такое высказывание все-таки используется, оно имплицитно характеризует фрагмент картины мира, в которой присутствует человек, обязанный учитывать соответствующий факт: скажем, (43) становится осмысленным в повествовании о геологе, тщетно ищущем возможности исследования крупных метеоритов, а фраза из лингвистической статьи «Парадоксы валентностей»

(44) *Подоплека такого принципа в том, что актанты — товар штучный, и их должно быть как можно меньше* (В. А. Плунгян, Е. В. Рахилина)

апеллирует к ученому, который должен строить свое описание так, чтобы в нем фигурировало мало актантов.

Более того, действие, предполагаемое к совершению в высказывании с нашими конструкциями, скорее всего, должно быть намеренным, сознательным и целесообразным: трудно представить себе высказывания типа

(45) ?*Он мог сверзиться как можно ниже*

или даже

(46) ?*Надо, чтобы младенец закричал как можно громче*:

чтобы предложение (46) стало осмысленным, необходимо, чтобы речь шла о младенце, который уже способен к сознательным действиям.

Это в особенности относится к первой из рассматриваемых фразем; конструкция *как нельзя* + COMP в некоторых случаях может характеризовать и непроизвольное действие или состояние, ср.

(47) *И вдруг все закончилось как нельзя лучше* (Ф. М. Достоевский).

Вторая семантическая особенность обеих конструкций состоит в том, что в их пресуппозицию существенным образом входит (а) **модальность возможности** и (б) **идея желательности**. Конкретнее говоря, в предложении

(48) *Гуляйте на свежем воздухе как можно чаще*

содержится рекомендация к собеседнику приложить максимально возможные усилия, чтобы добиться максимальной частоты его прогулок и утверждается, что такая частота прогулок является желательной. При этом совершенно неважно, для кого именно из участников ситуации последняя является желательной — для производителя действия, для говорящего (и то и другое допустимо в (48)) или даже для третьего лица, выступающего в качестве прямого или опосредованного каузатора ситуации, как в

(48') *Бабушка просила передать, чтобы ты гуляла на свежем воздухе как можно чаще.*

Третья семантическая особенность рассматриваемых конструкций состоит в том, что свойства, выражаемые компаративом в ее составе, **располагаются вблизи верхнего полюса соответствующей шкалы**.[6] Если мы говорим, что нечто следует *сделать как можно лучше* или что нечто работает *как нельзя лучше*, то мы имеем в виду, что нечто следует сделать *очень хорошо* или что нечто работает *очень хорошо*. Такие высказывания нельзя понимать лишь как требование приложить максимум усилий, чтобы повысить качество чего-либо и удовольствоваться тем уровнем качества, которого удалось достичь. В любом случае речь идет о желательности **приближении к пределу**.

Важное различие между первой и второй фраземами, по нашему убеждению, состоит в **направлении движения** к этому пределу. Когда говорят что-либо вроде

[6] Под шкалой мы здесь понимаем варьирование как «положительных» свойств (хороший, хорошо, большой, много, высокий, частый), так и «отрицательных (плохой, плохо, маленький, мало, низкий, редкий).

(49) *Нужно бросить камень как можно дальше*,

имеется в виду, что нужно произвести (пусть мысленный) **выбор** броска так, чтобы расстояние до камня оказалось максимально большим из тех, которые достижимы для бросающего: это **приближение** к верхнему полюсу шкалы дальности со стороны нижнего полюса.

Напротив, предложение

(50) *Он бросил камень как нельзя дальше*

означает, что говорящий, оценивая расстояние до камня, **отталкивается** от верхней шкалы дальности и убеждается, что это расстояние не могло оказаться еще ближе к пределу.

Не удивительно поэтому, что фразема *как можно* + COMP употребляется в первую очередь проспективно (оценивается или предписывается ожидаемая ситуация: *сыграй как можно лучше*), а фразема *как нельзя* + COMP обычно употребляется ретроспективно (оценивается уже существующая ситуация: *сыграно как нельзя лучше*).

Именно этими семантическими особенностями наших синтаксических фразем объясняется упомянутая в **разделе 2.1** несимметричность их поведения и поведения одиночных компаративов: для функционирования этих фразем необходимо соблюдение сложных семантических условий, перечисленных выше.

Добавим в заключение этого раздела, что в целом изложенные семантические свойства рассматриваемых единиц нельзя считать стопроцентно истинными. В таких ситуациях, характеризующих тонкую семантику и прагматику конкретных языковых единиц и, в значительной степени, требования здравого смысла, речь может идти скорее о тенденциях, чем об абсолютных, не знающих исключений, правилах.

4. Общие наблюдения над компаративом

В последнее время русский компаратив, серьезные научные наблюдения над которым осуществлялись еще в середине XX века (см. в особенности [Еськова 1955] и [Еськова 1963]), привлекает повышенное внимание исследователей, и в новейших работах (см, в частности, [Князев 2007], [Богуславский-Иомдин 2009], [Сичинава 2015]) приводятся достаточно нетривиальные факты, характеризующие синтаксическое поведение и семантику различных типов компаративов. В частности, коротко в работах Н. А. Еськовой и подробнее в двух последних из цитированных работ отмечается, что смягченный компаратив на *по-* (аттенуатив) выступает в двух принципиально разных ипостасях. А именно, в одних ситуациях он ведет себя практически так же, как обычный, несмягченный компаратив, сравнивающий по какому-то признаку два компарата и, возможно, добавляя количественную оценку различий между компаратами (аттенуатив сообщает, что она невелика; ср. *Картошка дороже апельсинов* и *Картошка подороже апельсинов*). В других же ситуациях аттенуатив сближается с превосходной степенью (суперлативом) прилагательного или

наречия, характеризуя приближение степени качества или свойства к верхнему полюсу шкалы, ср.

(51) *Принеси камень потяжелее*

'принеси камень из самых тяжелых, хотя, возможно, и не самый тяжелый из наличествующих'[7],

(52) *Ешьте побольше сырых овощей* ≈ 'ешьте очень много сырых овощей (но все-таки чуть-чуть оставьте несъеденными)'

и т. д. — тем самым речь может идти о смягченном суперлативе. Обратим внимание, что в таких ситуациях аттенуатив вообще не присоединяет второй компарат.

Следует сказать, что это далеко не единственный тип ситуаций, когда русский компаратив выступает в значении суперлатива. Другие случаи представлены 1) прилагательными *лучший* и *худший*, которые морфологически не однозначны и могут выступать как формы компаратива, так и формы суперлатива, 2) конструкциями типа *лучше всего* и *лучше всех*, *быстрее всего* и т. д., в которых формально местоименное слово в родительном падеже можно расценивать как выражение второго компарата, но содержательно все выражение эквивалентно суперлативу[8]. Легко обнаружить, что и наши синтаксические фраземы *как можно* + COMP и *как нельзя* + COMP, равно как и их аналоги, описанные выше в **разделе 2.6**, разделяют это свойство: они превращают сравнительную степень в суперлатив.

5. Оправдание заглавия

В заключение отметим следующее. Автор отдает себе отчет в том, что выражения типа *как можно лучше* и *как нельзя лучше* далеко не всякий лингвист будет готов признать антонимичными, резонно полагая, что перед ним скорее синонимы — ведь обе единицы, в общем, указывают на очень высокую степень «хорошести» какого-либо действия, состояния, факта или предмета. Мы тем не менее придерживаемся мнения, что важным обстоятельством здесь является бесспорная антонимичность опорных слов упомянутых единиц — *можно* и *нельзя*, смысл которых в их основных значениях различается на отрицание. Тот факт, что в определенных условиях высказывания с этими словами оказываются ситуационно близкими или даже тождественными (скажем, в случаях типа *Можно ли попросить соль?* vs. *Нельзя ли попросить соль*?; *А можно потише?* vs. *А потише нельзя?*) смысловой антонимичности самих единиц никак не отменяет — точно так же, как не исчезает смысл отрицания у частицы *не* из-за того, что некоторые типы высказывания с этой частицей и без нее оказываются ситуационно равнозначными, как в случаях типа *выпить ли чаю — не выпить ли чаю,*

---

[7] Следует оговориться, что (51) может пониматься и как смягченный компаратив с опущенным вторым компаратом: *Этот камень слишком легкий. Принеси камень потяжелее*. Автор благодарит анонимного рецензента за данное наблюдение.

[8] Заметим, что в таких конструкциях родительный падеж компарата нельзя заменить на выражение *чем+S*.

жду, пока он уснет — жду, пока он не уснет. Но это даже не главное. Главный аргумент в пользу антонимичности рассмотренных конструкций является тот факт, что семантические различия между нашими выражениями достаточно велики, даже если не учитывать факт присутствия в их толкованиях них антонимичных единиц *можно* и *нельзя*: Эти различия в первую очередь состоят в ориентации изменения степени признака относительно верхнего полюса шкалы, которой придерживается говорящий: у наших двух конструкций эта ориентация прямо противоположна — движение к верхнему полюсу (*как можно* COMP) vs. движение от верхнего полюса (как нельзя COMP). Именно этим и объясняется наша трактовка рассматриваемых единиц как антонимичных.

Литература

1. *Богуславский И. М., Иомдин Л. Л.* (2009). Семантика смягченной сравнительности: русские компаративы на по- // Von grammatischen Kategorien und sprachlichen Weltbildern — Die Slavia von der Sprachgeschichte bis zur Politsprache. Festschrift für Daniel Weiss zum 60. Geburtstag / Berger, Tilman; Giger, Markus; Kurt, Sybille; Mendoza, Imke (Hrsg.). Wiener Slawistischer Almanach. München-Wien. Bd. 1. S. 319–333.
2. *Еськова Н. А.* (1955). Степени сравнения в современном русском литературном языке. Московский городской педагогический институт им. В. П. Потемкина. Автореферат канд. дисс. М., 1955. 18 с.
3. *Еськова Н. А.* (1963). О некоторых формах сравнительной степени // Вопросы культуры речи. Вып. 3. М., 1963. С. 145–149.
4. *Иомдин Л. Л.*(2017). Между синтаксической фраземой и синтаксической конструкцией. Нетривиальные случаи микросинтаксической неоднозначности. SLAVIA, časopis pro slovanskou filologii, ročník 68, 2017, sešit 2–3, s. 230–243.
5. *Иомдин Л. Л.* (2018). Еще раз о микроконструкциях, сформированных служебными словами: То и дело. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Москва, 30 мая — 2 июня 2018 г.). М.: Изд-во РГГУ, 2018. Вып. 17 (24). С. 267–283.
6. *Князев Ю. П.* (2007). Грамматическая семантика. Русский язык в типологической перспективе. Языки славянских культур, — 704 с.
7. *Сичинава Д. В.* (2015) К описанию русского компаратива на *по-* на материале Национального корпуса русского языка. // Acta linguistica petropolitana. Труды Института лингвистических исследований РАН / Отв. ред. Н. Н. Казанский. Т. XI. Ч. 1. Категории имени и глагола в системе функциональной грамматики / Ред. М. Д. Воейкова, Е. Г. Сосновцева. СПб.: Наука. С. 701–718.
8. *Iomdin Leonid* (2016). Microsyntactic Phenomena as a Computational Linguistics Issue. // Grammar and Lexicon: Interactions and Interfaces. Proceedings of the Workshop. Osaka, Japan. 2016, pp. 8–18 (http://aclweb.org/anthology/W/W16/W16-3803.pdf), ISBN 978-4-87974-706-8.

References

1. *Boguslavsky I. M., Iomdin L. L.* (2009). The semantics of attenuated comparativeness: Russian comparatives with the po- prefix. [Semantika smjagchennoj sravnitel'nosti: russkie komparativy na po-]. // Von grammatischen Kategorien und sprachlichen Weltbildern — Die Slavia von der Sprachgeschichte bis zur Politsprache. Festschrift für Daniel Weiss zum 60. Geburtstag / Berger, Tilman; Giger, Markus; Kurt, Sybille; Mendoza, Imke (Hrsg.). Wiener Slawistischer Almanach. München-Wien, 2009. Bd. 1. S. 319–333 (In Russian).
2. *Es'kova N. A.* (1955). Degrees of Comparison in Modern Literary Russian [Stepeni sravnenija v sovremennom russkom jazyke]. // Moskovskij gorodskoj pedagogicheskij institut im. V. P. Potemkina. Ph.D. Thesis Abstract Moscow. 18 p. (In Russian).
3. *Es'kova N. A.* (1963). On certain forms of the comparative degree. [O nekotoryx formax sravnitel'noj stepeni]. // Voprosy kul'tury rechi. Issue 3. Moscow. P. 45–149 (In Russian).
4. *Iomdin Leonid* (2016). Microsyntactic Phenomena as a Computational Linguistics Issue. // Grammar and Lexicon: Interactions and Interfaces. Proceedings of the Workshop. Osaka, Japan. 2016, pp. 8–18 (http://aclweb.org/anthology/W/W16/W16-3803.pdf), ISBN 978-4-87974-706-8.
5. *Iomdin Leonid* (2017). Between the syntactic idiom and syntactic construction. Nontrivial cases of microsyntactic ambiguity. [Mezhdu sintaksicheskoj frazemoj i sintaksicheskoj konstruktsiej. Netrivial'nye sluchai mikrosintaksicheskoj neodnoznachnosti]. // SLAVIA, časopis pro slovanskou filologii, ročník 68, sešit 2–3, s. 230–243 (In Russian).
6. *Iomdin L. L.* (2018). Once again on microconstructions formed with functional words: to i delo'every now and then' [Eshche raz o mikrokonstruktsijax, sformirovannyx sluzhebnymi slovami: to i delo.] // Computational Linguistics and Intellectual Technologies. International Conference (Dialog'2018). Moscow: RGGU Publishers. Issue 17(24). P. 267–283 (In Russian).
7. *Knjazev Ju. P.* (2007). Grammar semantics. The Russian language in typological perspective. [Grammaticheskaja semantika. Russkij jazyk v tipologicheskoj perspective]. Jazyki slavjanskix kultur Publishers — 704 p.
8. *Sitchinava D. V.* (2015). On the description of the Russian comparative with po-prefix on the material of the Russian National Corpus. [K opisaniju russkogo komparativa na *po-* na material Natsional'nogo korpusa russkogo jazyka] // Acta linguistica petropolitana. Proceedings of the Institute of Linguistic Studies, RAS. / N. N. Kazansky, Ed. v. XI. part 1. St. Petersburg, Nauka, p. 701–718 (In Russian).

THE CORPUS OF CONTACT-INFLUENCED RUSSIAN OF NORTHERN SIBERIA AND THE RUSSIAN FAR EAST[1, 2]

Khomchenkova I. A. (irina.khomchenkova@yandex.ru)

Lomonosov Moscow State University; Vinogradov Russian Language Institute & Institute of Linguistics, RAS; Moscow, Russia

Pleshak P. S. (polinapleshak@yandex.ru)

Lomonosov Moscow State University; Institute of Linguistics, RAS; Moscow, Russia

Stoynova N. M. (stoynova@yandex.ru)

Vinogradov Russian Language Institute & Institute of Linguistics, RAS; NRU HSE; Moscow, Russia

The paper presents a spoken corpus of contact-influenced Russian, which consists of oral spontaneous Russian speech of bilingual speakers of indigenous languages of Northern Siberia and the Russian Far East (Samoyedic, Tungusic, Chukotko-Kamchatkan). The texts included in the corpus were transcribed in ELAN in Standard Russian orthography and provided with a special system of manual annotation of contact-induced features developed for the corpus. The paper focuses mainly on this system of annotation, which is relevant in a wider context of annotating any kind of speech with "deviations" from the standard language variety (bilinguals', learners', dialectal speech etc.). The annotation tags are grouped in several separate levels: contact-induced morphological, syntactic, phonetic, lexical features etc. The exact meanings for the annotation tags were proposed on empirical grounds. Transcribed and annotated texts gain morphological annotation and search implementation based on the Tsakorpus platform. The aim of the project is to provide a useful resource for linguistic studies on language contact.

Key words: corpus linguistics, spoken corpora, Russian, minor languages of Russia, language contact

[1] The research was conducted with support of RSF grant No. 17-18-01649 (Dynamics of language contact in the circumpolar region).

[2] Many thanks to our colleagues who granted us their field records to include in the corpus and to the anonymous reviewers of "Dialogue-2019".

КОРПУС КОНТАКТНО-ОБУСЛОВЛЕННОЙ РУССКОЙ РЕЧИ БИЛИНГВОВ-НОСИТЕЛЕЙ МАЛЫХ ЯЗЫКОВ СЕВЕРА СИБИРИ И ДАЛЬНЕГО ВОСТОКА

Плешак П. С. (polinapleshak@yandex.ru)
МГУ им. М. В. Ломоносова;
Институт языкознания РАН; Москва, Россия

Стойнова Н. М. (stoynova@yandex.ru)
ИРЯ им. В. В. Виноградова; Институт языкознания, РАН;
НИУ ВШЭ; Москва, Россия

Хомченкова И. А. (irina.khomchenkova@yandex.ru)
МГУ им. М. В. Ломоносова; ИРЯ им. В. В. Виноградова;
Институт языкознания, РАН; Москва, Россия

В статье описан создаваемый нами корпус контактно-обусловленной русской речи, который состоит из устных спонтанных текстов на русском языке, записанных от билингвов Севера Сибири и Дальнего Востока, носителей самодийских, тунгусских и чукотско-камчатских языков. Тексты расшифрованы в стандартной русской орфографии и снабжены специально разработанной ручной разметкой контактно-обусловленных грамматических особенностей в программе ELAN. Наиболее подробно в работе обсуждается опыт разметки, который может быть интересен в более широком контексте аннотирования речи, так или иначе отклоняющейся от литературной нормы (речи билингвов, изучающих иностранный язык, диалектной речи и т. д.). Разметка разделена на несколько уровней: контактно-обусловленные морфологические, синтаксические, лексические, фонетические особенности и т. д. Корпус частично доступен онлайн на платформе Tsakorpus с возможностью поиска по разработанной нами разметке контактно-обусловленных черт, морфологической разметке и метаданным. Цель проекта — создание удобного ресурса для исследований в области языковых контактов.

Ключевые слова: корпусная лингвистика, корпуса звучащей речи, русский язык, малые языки России, языковые контакты

1. Introduction

In the paper, we will present a new corpus of Russian spoken by bilinguals and discuss some problems of annotating "deviations" from the standard language variety, relevant for corpora of speech of bilinguals, learners, heritage speakers, people with speech disorders, as well as for child speech and dialectal corpora.

The corpus constitutes a transcribed and annotated collection of oral spontaneous Russian speech of bilingual speakers of indigenous languages of Northern Siberia and Russian Far East (Samoyedic, Tungusic and, to a smaller extent, Chukotko-Kamchatkan). The majority of the texts are short narratives.

The transcription is made in ELAN in standard Russian orthography with a simplified intonation marking and with the manual annotation of contact-induced features.

The text collection, which is planned to be included in the corpus, consists by the moment of ca. 100 hours of records. Ca. 29 hours of records have been already transcribed and annotated, these texts are available offline in the ELAN-format. A small test text sample was added to the online resource, which is being created for the corpus on the Tsakorpus platform: http://web-corpora.net/tsakorpus_russian_nonst/corpus.html. Transcribed and annotated texts gain morphological annotation and search implementation based on the platform.

The resource is aimed to be used by specialists on language contact to trace the influence of indigenous languages of the area on the Russian speech of their speakers. The corpus is the most convenient to study contact-induced morphosyntactic features. However, it also can be used in other studies on language contact, e. g. studies on lexicon and phonetics.

The paper is comprised of 6 parts. In **Section 2**, we discuss some corpus projects which are similar to ours. **Section 3** presents the text collection included in the corpus: the amount of data, types and genres of texts, the narrators and languages they speak. **Section 4** describes our conventions of transcription (**4.1**), the system of annotation of contact-induced grammatical features used in the corpus (**4.2**) and the online searching interface (**4.3**). In **Section 5**, we list some studies on language contact based on our corpus data. Section 6 contains brief concluding remarks and plans on further development and use of this corpus.

2. Similar projects on bilinguals' Russian

There are some parallel projects, devoted to other varieties of Russian, spoken by bilinguals or learners. For example, resources the most close to ours are corpora made by Linguistic Convergence Laboratory, HSE—the corpus of Daghestanian Russian (DagRus, http://www.parasolcorpus.org/dagrus/#, cf. [Daniel & Dobrushina 2013]) and the corpus of Chuvash Russian (ChuvashRus, http://www.parasolcorpus.org/chuvashrus/). These corpora consist of oral spontaneous texts collected in the form of sociolinguistic interviews. In contrast to our corpus, they do not include any special annotation of contact-induced features.

One more similar resource is Russian Learner Corpus created in Linguistic Laboratory for Corpus Studies, HSE (http://www.web-corpora.net/RLC/), cf. [Rakhilina 2016];

[Rakhilina et al. 2016]. It consists of texts of speakers who learn Russian as their second language and of heritage speakers of Russian. The texts are mostly written. They are provided with the annotation of non-standard grammatical features ("errors" in terms of its creators), similar to ours.

Texts of bilingual speakers were also included in the spoken subcorpus of Russian National Corpus (http://ruscorpora.ru/search-spoken.html), see [Savchuk 2018] for more detail. They are provided with standard grammatical annotation of Russian National Corpus. Unfortunately, by the moment, the user has no possibility to separate this text collection from oral texts of monolinguals.

3. Text collection

The text collection consists of spontaneous oral texts, mostly short narratives (folklore, biographies) and descriptions (ethnographic texts, recipes etc.); some texts are everyday dialogues with linguists. They were collected by us and by our colleagues as a "by-product" of current language documentation projects. For many of them we have also parallel (or near-parallel) versions in the indigenous language.

The overall text collection includes ca. 100 h. of records. Tungusic and Samoyedic varieties are the best represented by the moment. We also have modest collections from speakers of Chukchi, Yakut and Yukaghir.

By now, we have transcribed and annotated ca. 29 h. (out of 100 h.), which is approximately 117,000 words. The **Table 1** represents the total amount of textual data in hours and words.

Table 1. Text collection

	all in hours	annotated in hours	annotated in words
Enets (Forest and Tundra)	26.5	12.5	49,128
Nenets	9	1.5	9,292
Nganasan	10	6	19,072
Nanai	42	8	29,076
Ulch	8.5	1	10,564
Even	1	0	0
Chukchi	1.5	0	0
multilingual speakers from Lower Kolyma	2	0	0
total amount	100.5	29	117,132

The majority of these languages have a comparable sociolinguistic situation: they are endangered; the typical speaker acquired Russian at school age, but now they use actively both Russian and the indigenous language or almost only Russian.

For each text we also collected some metadata: 1) technical information on the record: file name, record date, record place, duration; 2) information on the text: type, genre, content, existence of a parallel version in the indigenous language; 3) information on the narrator: name, code, indigenous language(s) (s)he speaks; 4) information

on transcription and annotation: annotator, date, size (in clauses). For narrators, we also have separate more detailed metadata: name, birth place, birth date, place of residence, level of education, acquisition age (for Russian), indigenous language(s) (s)he speaks, and short sociolinguistic biography. Unfortunately, there remain a lot of gaps by the moment.

4. Transcription and annotation

4.1. Transcription in ELAN and the structure of tiers

The text for the corpus are transcribed in ELAN in standard Russian orthography with a simplified intonation marking (rising and falling tones indicated after words bearing phrasal accents) and a small number of special marks for the features of oral spontaneous speech (self-corrections, pauses, non-speech sounds, fragments in the indigenous language), see **Fig. 1**. The standard orthography was chosen for technical reasons, cf. [von Waldenfels et al. 2014] for the same decision and its reasons. Contact-induced or dialectal phonetic features are not reflected in the transcription, some of them are annotated in special tiers (see **Section 4.2**).

Fig. 1. Transcription and annotation in ELAN

Proper names are marked with square brackets (e. g. *M*[*аш*]*а*) to become automatically anonymized in the web version of the corpus[3]. Fragments in the indigenous language, which sometimes occur in our texts, are transcribed if the annotator is familiar with the language enough or remain untranscribed (in this case we use a special mark *CS*).

The texts are segmented into clauses, or intonation units, more or less corresponding to clauses in oral speech. Ideally, 1 ELAN-annotation ≈ 1 clause. In practice, we rely more on intonation and pauses than on syntactic structure. In case of discrepancy between clausal boundaries and pausation, annotation boundaries correspond to pauses.

[3] The corresponding audio-fragments have not been anonymized by the moment.

We have separate transcription tiers for each participant of the conversation. Besides the transcription tier, which is synchronized with the audio data, there is a word tier, 6 special tiers for the annotation of contact-induced features (see **Section 4.2**), having the word tier as a parent, and some technical tiers. The latter include tiers for comments and for the translation of code-switching fragments. See the structure of the tiers in **Fig. 2**.

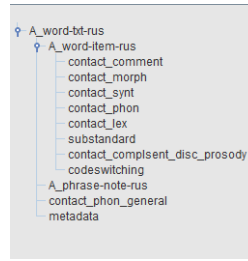


Fig. 2. Structure of tiers in ELAN

4.2. Annotation of contact-induced features

We use 5 ELAN-tiers for annotation of contact-induced features on different levels: phonetics, lexicon (loanwords and calques), morphology (including productive derivation, inflection and the use of grammatical categories), syntax (clause-level), one general tier is reserved for complex sentences, discourse and prosody.

One more tier ("substandard") is used for peculiarities that are presumably of non-contact nature (see some examples below and the discussion on the choice between particular tags).

To make the manual annotation of contact-induced features more structured and convenient for the search, we use Controlled Vocabulary incorporated into ELAN. For each tier (level) we have a set of tags among which the annotator can choose, see **Fig. 1**. The particular features and their values were chosen on empirical grounds. After a preliminary set had been proposed, it was used in the annotation during the testing period. Afterwards, the tags were discussed, some tags were added, which is considered to be enough for the text collection so far. Some tags are specific for texts produced by speakers of a concrete indigenous language. However, most of them are general enough to be used throughout the whole Northern Siberian corpus and even to be applied to other text collections. The tool is flexible and more tags can be added if needed.

The tags are ascribed to the words. There can be more than one tag ascribed to one word. Syntactic tags are ascribed to the word that manifests the syntactic relation. Usually, this is the dependent. For instance, the non-standard agreement tag is attached to adjectives, the non-standard argument encoding tag is attached to nouns etc. Intonation tags are attached to the accent-holder.

One of the problems with the Controlled Vocabulary is that it lacks hierarchical structure. So, in one tier, the annotator chooses among several possibilities without any further subdivision. We resolved this problem first, setting up separate tiers for

each level (phonetics, lexicon, morphology, syntax, complex sentences & discourse & prosody), and second, introducing complex names for tags: e. g. *agr_adj_gender*, *agr_adj_num*, *agr_adj_case*. All the three tags are used to indicate phenomena of disagreement but only for adjectives (in contrast to the verbal or anaphoric disagreement). Moreover, each tag is specified for the features that are involved. Therefore, we have a large set of disagreement tags within the syntactic tier, which are the following: *agr_adj_gender*, *agr_adj_num*, *agr_adj_case*; *agr_verb_gender*, *agr_verb_num*, *agr_verb_pers*; *gr_anaph_num*, *agr_anaph_gender*.

Such a fine-grained subdivision in the disagreement domain is due to the fact that it is one of the most frequent features in the non-standard speech. Having this powerful inventory, one can search choosing different sets of tags, in accordance to the purposes (see the description of a corpus-based study on gender disagreement in **Section 5**). The inventory of morphological features, which are more rare, is smaller.

By the moment, we use 73 tags in total. The level of morphology (including word-formation, inflection, use of grammatical categories) contains 10 tags[4]. The level of syntax (only within the clause) is the most elaborated and it contains 23 tags. The level of complex sentences, discourse and prosody contains 12 tags. The level of lexicon contains 3 simple tags: one for loanwords, one for calques, and one for non-evident cases. The level of phonetics includes 19 tags, almost all of them are very specific (the non-standard realization of a particular phoneme or a small group of phonemes) and the inventory of tags in use varies a lot across particular local varieties of bilingual Russian included in our sample. The phonetic and prosodic features, in contrast to morphological and syntactic ones, are marked with special tags not very consistently, since they are too frequent to mark them all and not clear enough for perception to mark them appropriately during the transcription without any additional instrumental analysis. So phonetic and prosodic tags are used only to mark striking clear cases just for an easy search of illustrative examples.

The level of "substandard" (non-contact) features contains 6 tags (one for each level: phonetics, morphology, syntax, lexicon etc.). These (dialectal, regional, register) features are not in our main focus, so they are not annotated very consistently either. The main reason to annotate them in our corpus is to make it possible for a user to differentiate between these features and contact-induced ones. In less clear cases, we use the corresponding "contact" tag, the "substandard" tags are reserved for more evident cases of non-contact features. However, since we cannot attribute all cases for

[4] Non-standard inflection and derivation patterns must be interpreted as under-acquisition of Russian rather than copying of the corresponding indigenous patterns (such as lexical calques or argument encoding patterns inherited from the indigenous language). In our annotation we do not differentiate between these two types of features, marking all of them as contact-induced. Another problem is to differentiate between contact-induced under-acquisition and non-standard inflectional and derivational patterns that can be produced also by monolinguals as occasional speech errors or as features of uneducated speech. There is no clear borderline between them. Our technical decision is to provide with tags as many cases as possible, ranking them according to the probability to be contact-induced. The annotator distributes them between the "contact" tier (= probable to be contact-induced, cf. the form *стаёт* 'becomes') and the "substandard" one (= less probable to be contact-induced, cf. *радовается* 'is glad'), basing on his/her intuition, see below on the substandard tier.

sure, we try to annotate everything that deviates from standard monolingual Russian[5] not to miss any relevant information. The aim of the annotator is not to make a right choice in all particular cases (it is generally impossible without a special investigation), but rather to rank the attested peculiarities roughly according the probability to be of a contact nature[6]. Therefore, our "contact" tags mark cases that are likely to be of contact nature and our "substandard" tags mark cases that have a chance to be interpreted as contact-induced. We leave the final decision to users of the corpus, giving them the access to both types of cases.

Table 2. Contact-induced features and tags

level (tier)	N of tags	examples (tags)
phonetics	19	***с***аръ 'king' (affr), ***т***арик 'oldman' (clust)
lexicon	3	*крупы* ***налила*** (calque)
morphology	10	*за неделю* ***пил*** (asp), *обитает****ся*** (refl)
syntax	23	*икра нету* (neg), *сетк****а*** *кинет* (gov_dom)
complex sentences & discourse & prosody	12	*попросили кто-нибудь увез* (subord_compl), *А чо* ***грит*** *мне от тебе надо\ Я\ ей говорю* (disc_word), *А там [****красивый****\ девушки]RHEME гоняют\ их* (pros_accent)
substandard	6	*у мен****е*** (morph), ***с*** *города* (synt), *балдой* (lex)

The full list of tags with short descriptions and illustrative examples is available at http://web-corpora.net/tsakorpus_russian_nonst/corpus.html.

4.3. Web-interface

The online interface for the corpus was implemented on the platform Tsakorpus (https://bitbucket.org/tsakorpus/tsakonian_corpus_platform), which had been developed by T. Arkhangelsky for small spoken corpora created in ELAN. The platform provides the possibility of search on grammatical features (the annotation mystem), search and filtration on metadata and search on any specific tag set used in a particular corpus (the annotation of contact-induced grammatical features in our case), see **Fig. 3**. Search results are given per clauses (with possibility of enlarging the context), both the transcribed fragment and the audio-fragment are available, see **Fig. 4**.

[5] The question arises, which monolingual variety must be chosen as tertium comparationis. The best option would be to use a text sample, the most comparable to ours: oral narratives produced by monolinguals of the same area and of the same sociolinguistic background as our narrators. Having no access to such texts, we rely on the intuition of the annotator, trying to mark with any tag as many "non-standard" features as possible.

[6] A more "honest" and simple way would be to mark on equal terms everything that the annotator assesses as non-standard, without any further differentiation during the annotation process. But in this case too much useless information would fall into the annotation.

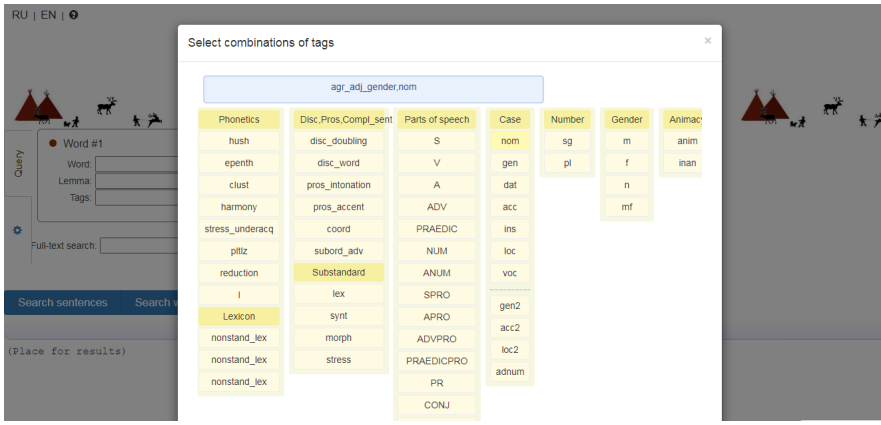


Fig. 3. Web-interface: search on grammatical features and contact-induced features

Fig. 4. Web-interface: search results

The user can find comparable samples of “standard” and “non-standard” uses, combining the search on grammatical tags provided with the platform and the search on our tags of contact-influenced features. For instance, one can find all occurrences of standard prepositional phrases (such as *в доме* ‘in the house’), using grammatical tags (the query “PR”), and then all non-standard occurrences with preposition drop (such as *доме* ‘(in) the house’), using tags of contact-induced features (the query “prep_drop”), see **Section 5** for the study based on these data.

At the moment, the online resource is working in a test mode at http://web-corpora.net/tsakorpus_russian_nonst/corpus.html. Only a small part of our transcribed and annotated text collection has been placed on the web. We are planning to enlarge the range of metadata types available for search, to make the search on contact-induced features more user-friendly and then to add the whole text collection. The next step is disambiguation of the grammatical annotation, which will make the search much more effective.

5. Using the corpus

The aim of the project is to provide a useful resource for linguistic studies of contact-induced language changes. In this section, we present studies conducted on the data of this corpus to illustrate possibilities of its application.

In some of them, the corpus served just as a source of examples, which were used to describe non-standard grammatical features attested in bilingual Russian in detail. Basing on the data of the Tungusic subcorpus, [Oskolskaya and Stoynova 2017] proposed a classification of uses of the construction *делал был*, *делал было* (V.PST + *be*.PST) in Nanai Russian and compared them to those of the similar construction in monolingual Russian and the pluperfect construction with the verb 'be' in Nanai.

One more way of using corpus data in the research of contact features in grammar is to calculate the frequency of "standard" (typical of monolingual Russian) and "non-standard" uses in the Russian speech of bilingual speakers and to reveal correlations with the grammatical context. In [Khomchenkova et al. 2018], gender disagreement in Russian speech of speakers of Southern Tungusic and Samoyedic languages of the elder generation was investigated (*бабка помер* 'old woman die.PST.MASC', *моя папка* 'my.FEM father'). The corpus data show that bilingual speakers are less likely to follow the standard agreement pattern for adjectives and more likely to choose the standard form of verbs and especially of anaphoric elements.

The data of the corpus can also be used to get a complex picture on some particular variety of bilingual Russian. For instance, in the grammatical description of Souther Tungusic Russian (Stoynova, to appear) the author gives some quantitative data on the relative frequency of different contact-induced grammatical features attested in this variety.

The list of some other studies on the data of this corpus is available at http://web-corpora.net/tsakorpus_russian_nonst/publ.html.

6. Conclusion

The present project has three main advances. First, it contributes to the overall collection of spoken corpora of Russian that are open source and can be used in linguistic studies. Second, it represents the speech of bilingual speakers and can serve as a representative data source for studies on language contact. Third, an important point, which was described in the paper in great detail, is a special system of annotation of contact-induced grammatical and lexical features created for the corpus. It reflects the peculiarities attested in particular varieties of Russian we deal with. However, it is quite flexible to be adapted for other contact-influenced varieties of Russian. The presence of such annotation gives the possibility to apply quantitative methods in studies of contact-induced features as these are difficult to search using only morphological tagging, concrete lemmas and regular expressions.

This experience also contributes to a more general problem, relevant for corpus linguistics, namely the problem of annotating any kind of speech, "deviating" anyhow from the standard language variety, including speech of learners, heritage speakers, children, people with speech disorders, as well as speech with regional and dialectal features.

We are planning to develop the project in the following directions. First, we will continue transcribing and annotating the existing text collection; the expansion to other bilingual varieties is in further plans as well. Second, we will continue the work on the online resource. The whole transcribed text collection will be placed on the web. The search interface will be improved—particularly, the search on different types of metadata will be added and the search on contact-induced features will become more user-friendly. One of our current plans is manual disambiguation of the automatic morphological annotation, which is used in the corpus.

References

1. *Bayda K., Kholodilova M., Kozhemjakina A., Romanova E., Remizova T., Storozheva A., Tarasova N., Zorina A., Morozova V., Panova A., Dobrushina N.* (2018) ChuvashRus Corpus, Moscow: Linguistic Convergence Laboratory, NRU HSE, available online at http://www.parasolcorpus.org/chuvashrus/.
2. *Daniel M. A., Dobrushina N. R.* (2013), A corpus of Russian as L2: the case of Daghestan [Russkij jazyk v Dagestane: problemy jazykovoj interferencii], Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue", 12(1), Moscow: RSUH, pp. 186–211.
3. *Dobrushina, N., Daniel M., von Waldenfels R., Maisak T., Panova A.* (2018), Corpus of Russian spoken in Daghestan, Moscow, Linguistic Convergence Laboratory, NRU HSE, available online at http://www.parasolcorpus.org/dagrus/.
4. *Khomchenkova I. A., Pleshak P. S., Stoynova N. M.* (2018), Gender disagreement in the contact-influenced Russian of Northern Siberia and the Russian Far East, presented at the conference "TheGen", Berlin, 14–15.06.2018.
5. *Oskolskaya S. A., Stoynova N. M.* (2017), Nanai verb categories in the Russian speech of Nanai speakers: 'be'-constructions [Nanajskije glagoljnyje kategorii v russkom jazyke nanajcev: konstrukcii tipa "byli delali"], presented at the conference "Russian grammar: describing, teaching, testing [Russkaja grammatika: opisanije, prepodavanije, testirovanije], Helsinki, June 7–9, 2017.
6. *Rakhilina E., Vyrenkova A. et al.* (2016), Russian Learner Corpus. Moscow: Linguistic Laboratory for Corpus Studies, NRU HSE, available online at http://www.web-corpora.net/RLC/.
7. *Rakhilina E., Vyrenkova A., Mustakimova E., Ladygina A., Smirnov I.* (2016), Building a learner corpus for Russian, Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, Umeå, 16th November 2016, pp. 66–75.
8. *Rakhilina E. V.* (2016), On a new instrumentary for describing Russian grammar: the corpus of errors [O novyx instrumentax opisanija russkoj grammatiki: korpus ošibok], Russkij jazyk za rubežom, 3, pp. 20–25.
9. *Russian National Corpus*, available online at http://www.ruscorpora.ru.
10. *Savchuk S. O.* (2018), Russian speech in polyethnic regions [Russkaja rečj v polietničeskix regionax], presented at the conference "Indigenous languages in contact with Russian: morphosyntactic and semantic interference", 30.11–01.12, 2018, Moscow: Russian Language Institute RAS, available at: https://drive.google.com/file/d/1GmJ4wxGM4DIWdSZftzrggcsnVvoWIyd3/view?usp=sharing.

11. *Stoynova N.* (to appear), Russian in contact with Southern Tungusic languages: evidence from Contact Russian Corpus of Northern Siberia and the Russian Far East, Slavica Helsingiensia, in print.
12. *von Waldenfels R., Daniel M., Dobrushina N.* (2014), Why standard orthography? Building the Ustya river basin corpus, an online corpus of a Russian dialect, in: Kompjuternaja lingvistika i intellektualjnye tekhnologii: Po materialam jezhegodnoj Mezhdunarodnoj konferentsii "Dialog" (Bekasovo, 4–8 June 2014) / V. Selegey. (ed.) № 13(20). M.: RSUH,. pp. 720–728.

ЕДИНАЯ МУЛЬТИКАНАЛЬНАЯ АННОТАЦИЯ КАК ИНСТРУМЕНТ АНАЛИЗА ЕСТЕСТВЕННОЙ КОММУНИКАЦИИ[1]

Кибрик А. А. (aakibrik@gmail.com)
Институт языкознания РАН,
МГУ им. М. В. Ломоносова, Москва, Россия

Коротаев Н. А. (n_korotaev@hotmail.com)
РГГУ, Институт языкознания РАН, Москва, Россия

Федорова О. В. (olga.fedorova@msu.ru)
МГУ имени М. В. Ломоносова,
Институт языкознания РАН, Москва, Россия

Евдокимова А. А. (arochka@gmail.com)
Институт языкознания РАН, Москва, Россия

Данная статья вносит вклад в область мультиканального анализа дискурса. Мультиканальный анализ дискурса исследует многочисленные каналы, задействованные в естественной коммуникации, такие как вербальная структура, просодия, жестикуляция, движения головы, взгляд, положения туловища и т. д., рассматривая их как части единого процесса. Для изучения того, как взаимодействуют между собой коммуниканты и как устроена координация между различными коммуникативными каналами, мы ввели понятие единой мультиканальной аннотации. Единая мультиканальная аннотация была реализована в среде ELAN. В частности, мы рассмотрели три конкретных примера: (1) временную координацию между речью коммуникантов и их мануальной жестикуляцией; (2) распределение зрительного внимания коммуникантов между речью и мануальной жестикуляцией их собеседников; (3) взаимодействие между положением тела и движениями головы коммуникантов.

Ключевые слова: мультиканальный дискурсивный анализ, единая аннотация, речь, мануальные жесты, движения головы, взгляд, положения туловища

[1] Работа выполнена при финансовой поддержке РФФИ, проект 19-012-00626.

UNIFIED MULTICHANNEL ANNOTATION: A TOOL FOR ANALYSING NATURAL COMMUNICATION

Kibrik A. A. (aakibrik@gmail.com)
Institute of Linguistics RAS, Lomonosov Moscow State University, Moscow, Russian Federation

Korotaev N. A. (n_korotaev@hotmail.com)
RSUH, Institute of Linguistics RAS, Moscow, Russian Federation

Fedorova O. V. (olga.fedorova@msu.ru)
Lomonosov Moscow State University, Institute of Linguistics RAS, Moscow, Russian Federation

Evdokimova A. A. (arochka@gmail.com)
Institute of Linguistics RAS, Moscow, Russian Federation

This paper contributes to the research field of multichannel discourse analysis. Multichannel discourse analysis explores numerous channels involved in natural communication, such as verbal structure, prosody, manual gesticulation, head movements, eye gaze, torso postures, etc., and treats them as parts of an integrated process. For the purposes of investigating the way participants interact with one another and the way different communication channel correlate, we introduce the notion of an integrated multichannel annotation created with ELAN software. In particular, we consider three topics: (1) temporal alignment between participants' speech and manual gesticulation; (2) distribution of participants' visual attention as they watch their interlocutors talking and gesticulating manually; (3) interrelationship between participants' torso postures and head movements.

Key words: multichannel discourse analysis, unified annotation, speech, manual gesticulation, head movements, eye gaze, torso postures

1. Введение. Мультиканальная лингвистика

Мультиканальная (мультимодальная) лингвистика изучает все реальное многообразие «живой» коммуникации между людьми: слова, интонацию, жестикуляцию, направление взгляда, мимику, см. **рис. 1**. Интерес к изучению мультиканальности возник еще в древности, см., напр., [Квинтилиан 1834: XI], однако современная лингвистика, берущая начало в первые десятилетия XX в., долгое время занималась исключительно письменными текстами, то есть вербальным каналом [Linell 1982]. В конце XX в. ситуация начала меняться в сторону изучения

устного дискурса, то есть к вербальному каналу прибавился просодический, см., в частности, работы У. Чейфа [Chafe ed. 1980]; [1994]. Наконец, в XXI в. на наших глазах формируется новый мультиканальный подход [Kress 2010]; [Кибрик 2010]; [Knight 2011]; [Adolphs, Carter 2013]; [Müller et al. eds. 2013–2014]; [Кибрик 2018], принимающий во внимание все каналы общения между людьми, в том числе кинетические [Efron 1941]; [Крейдлин 2002]; [Kendon 2004]; [Бутовская 2004]; [McNeill 2005]; [Гришина 2017].

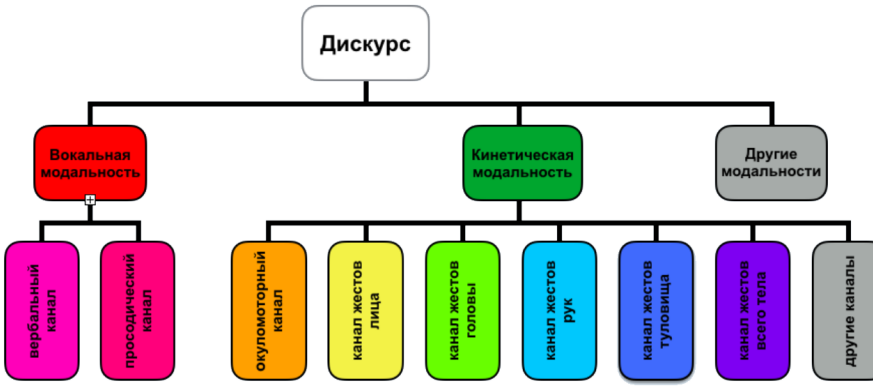


Рис. 1. Модель мультиканального дискурса

Одно из наиболее важных и перспективных направлений развития мультиканальной лингвистики — это разработка и создание мультиканальных ресурсов (корпусов). В отличие от моноканальных и мономодальных ресурсов, уже имеющих свою историю и традицию, мультиканальные ресурсы в настоящий момент находятся на этапе становления. Мультиканальный ресурс — это сочетание двух компонентов. Во-первых, это совокупность медиафайлов, фиксирующих коммуникацию между людьми. Во-вторых, это аннотация коммуникативных событий, содержащихся в медифайлах. Данная работа посвящена описанию принципов единой мультиканальной аннотации ресурса «Рассказы и разговоры о грушах», проект по разработке которого осуществляется в Институте языкознания РАН (сайт multidiscourse.ru). В **разделе 2** изложены основные характеристики собранного ресурса. **Раздел 3** содержит описание принципов мультиканальной аннотации, осуществляемой в рамках проекта. В разделе 4 предложены три конкретных примера применения разработанной аннотационной схемы на практике.

2. Ресурс «Рассказы и разговоры о грушах»

Ресурс состоит из отдельных коммуникативных эпизодов (так называемых записей), в каждом из которых участвуют четыре человека — Рассказчик, Комментатор, Пересказчик и Слушатель. Название корпуса обусловлено тем, что двое из участников вначале смотрят известный в лингвистике стимульный

материал — шестиминутный «Фильм о грушах», созданный в 1970-е гг. в Беркли исследовательской группой под рук. У. Чейфа (linguistics.ucsb.edu/faculty/chafe/pearfilm.htm). В этом фильме, не включающем никаких речевых действий персонажей, показана история о взаимодействии ряда лиц, в том числе садовника, собирающего груши, мальчика, крадущего корзину с грушами, и еще нескольких. Фильм представляет собой хорошо продуманную цепь физических и социальных событий и давно зарекомендовал себя как эффективный способ получать компактные и сравнимые между собой образцы устного дискурса.

Фильм просматривается участниками, которые в дальнейшем выполняют роли Рассказчика и Комментатора. На **рис. 2** можно видеть общий дизайн коммуникативной ситуации. Сначала Рассказчик в режиме монолога рассказывает Пересказчику содержание фильма (этап рассказа). Затем наступает интерактивный этап, в ходе которого Комментатор дополняет или уточняет рассказ, а Пересказчик задает вопросы обоим собеседникам, видевшим фильм (этап разговора). После этого появляется Слушатель, и Пересказчик пересказывает ему фильм, опять в режиме монолога (этап пересказа). В конце Слушатель письменно записывает еще один, вторичный, пересказ фильма. Два последних этапа необходимы для того, чтобы мотивировать участников к полноценной и осмысленной коммуникации.

Рис. 2. Общий дизайн коммуникативной ситуации

Ресурс «Рассказы и разговоры о грушах» состоит из двух частей. Первая часть была собрана летом 2015 г.; она включает 24 записи длительностью около 9 часов и объемом около 100 тыс. словоупотреблений; в записях приняли участие 96 человек от 18 до 36 лет, в том числе 34 мужчины и 62 женщины. Вторая часть корпуса была собрана летом 2017 г.; она включает 16 записей длительностью около 6 часов и объемом около 60 тыс. словоупотреблений; в записях приняли участие 64 человека от 18 до 36 лет, в том числе 16 мужчин и 48 женщин.

При записи корпуса были использованы оригинальные технические решения. *Аудиозапись* велась при помощи шестиканального рекордера ZOOM Handy Recorder, что обеспечивало автоматическую синхронизацию (параметры записи 96 kHz / 24 bit). Для *видеозаписи* были использованы три индивидуальные промышленные видеокамеры JAI GO-5000M с частотой 100 к/с и разрешением 1392×1000, которые записывали крупным планом каждого из трех основных коммуникантов. Для последующей аналитической работы важно, что эти камеры позволяют получить запись в формате mjpeg; данный формат выгодно отличается от остальных отсутствием межкадрового сжатия, что является необходимым условием для дальнейшего покадрового аннотирования. Частота записи 100 к/с позволяет проводить анализ собранного видеоматериала с точностью до 10 мс, что является показателем высокой точности аннотирования. Насколько нам известно, оборудование этого типа для лингвистических исследований было использовано впервые. Кроме того, общая видеокамера GoPro Hero с частотой 50 к/с и разрешением 2700 × 1500 (для записей 2015 г.) и частотой 120 к/с и разрешением 1920 × 1080 (для записей 2017 г.) записывала общий план. Для *регистрации движений глаз* были использованы две пары очков-айтрекеров фирмы Tobii Glasses II с частотой 50 Гц и разрешением видеокамеры 1920 × 1080. В очки-айтрекер вмонтирована миниатюрная видеокамера, регистрирующая движения глаз испытуемого, вторая видеокамера снимает окружающую обстановку.

3. Мультиканальная аннотация

При аннотировании корпуса мы различали два вида разметки: базовую и дополнительную. Базовая разметка включает в себя аннотирование вокального компонента (вербальный и просодический каналы), мануального (жестов рук), цефалического (жестов головы) и окуломоторного (направления взгляда) каналов. *Вокальная* аннотация, выполненная в программах Praat (fon.hum.uva.nl/praat) и MS Word, состоит в членении речевого потока на значимые фрагменты (элементарные дискурсивные единицы — ЭДЕ, слова, заполненные и незаполненные паузы, неречевые звуки), а также в приписывании свойств ЭДЕ и отдельным их частям [Кибрик, Подлесская ред. 2009]; [Kibrik et al. 2019]. Для *аннотирования мануального* канала применяется новая оригинальная методика, разработанная в среде ELAN (tla.mpi.nl/tools/tla-tools/elan); она основана на сегментировании потока мануального поведения на периоды неподвижности и отдельные движения, которые затем формируют функциональные единицы — жесты, адапторы и смены позы. Для этих единиц затем указываются их характеристики — рукость жестов, фазовая структура, функциональный тип и т.д., подробнее см. [Литвиненко и др. 2017]. Аннотирование *цефалического* канала основано на принципах, разработанных для аннотирования мануальных жестов. В ходе аннотирования *окуломоторного* канала производится экспорт данных айтрекинга на видеосцену, затем с помощью программы Tobii Pro Glasses Analyzer извлекаются данные о временной развертке всех фиксаций длительностью выше 100 мс, на которые потом в ручном режиме накладывается аннотационная схема с указанием направления взгляда, см. [Федорова 2017].

Дополнительная разметка включает в себя аннотирование *жестов тела*, *мимики*, а также *референциальную* аннотацию, которая содержит разметку языковых выражений с конкретной референцией. Описания принципов аннотирования коммуникативных каналов, а также образцы базовой и дополнительной разметок можно найти на сайте проекта multidiscourse.ru на вкладках «Корпус» и «Принципы аннотации».

Для изучения того, как взаимодействуют между собой участники коммуникации и как устроена координация между каналами, используется *единая мультиканальная аннотация*. В такой разметке сводятся результаты аннотаций, выполненных независимо для отдельных каналов. В базовой части единой аннотации учитываются вокальные, окуломоторные, цефалические и мануальные действия трех основных участников записи. Дополнительная часть единой аннотации включает остальные коммуникативные каналы. С технической точки зрения мультиканальная аннотация представляет собой файлы формата .eaf, используемые в программе ELAN; при этом вокальная и окуломоторная разметки, изначально реализованные в других программах, конвертируются в этот формат. Для унификации обозначений мы используем следующие соглашения.

(1) Мультиканальное поведение каждого участника фиксируется в отдельном наборе слоев аннотации. Порядок следования слоев и их названия идентичны в каждом наборе — с точностью до начальной литеры, указывающей на роль участника. Например, в слое N-mGesture размечаются мануальные жесты Рассказчика, а в слое C-mGesture — мануальные жесты Комментатора.

(2) Единицы мультиканального поведения (ЭДЕ, слова, мануальные и цефалические жесты, фиксации и проч.) фиксируются в виде непустых интервалов в слоях с независимой временно́й привязкой (в системе ELAN для этого используются т. н. «стереотипы» None и Included in). В качестве названия интервала используется стандартный идентификатор, включающий в себя указание на роль участника, кодовое обозначение канала и типа единицы, а также трех- или четырехзначный номер.

(3) Свойства единицы (напр., вербальное наполнение ЭДЕ, рукость жеста, направление взгляда) отмечаются в интервалах зависимых слоев (слоев со «стереотипом» Symbolic Association). Значения свойств либо выбираются из закрытого списка, либо вводятся вручную по специальным правилам.

Более подробное описание разработанной схемы базовой мультиканальной аннотации см. на сайте проекта multidiscourse.ru на вкладке «Принципы аннотации»; фрагмент аннотации, реализованной для записи #22, представлен в Приложении.

К настоящему моменту полностью аннотированы и выложены на сайт три записи #04, 22 и 23 длительностью около 1 часа, которые составляют эталонный подкорпус. *Эталонный подкорпус* — это экспериментальная площадка, на которой тестируются различного рода гипотезы, чтобы потом верифицировать

их на более обширном материале. Как можно видеть по **табл. 1**, несмотря на то что мы не устанавливали для испытуемых временных ограничений, все три записи похожи друг на друга по длительностям этапов — на рассказ приходится около 20% от времени записи, на разговор — примерно 50%, и на пересказ — 30%.

Табл. 1. Длительность записей эталонного подкорпуса и их этапов

#	общая длительность записи (мин:сек,мсек)	рассказ (мин:сек,мсек)	разговор (мин:сек,мсек)	пересказ (мин:сек,мсек)
04	24:36,240	05:22,640 (21,9%)	12:37,920 (51,3%)	06:35,680 (26,8%)
22	18:04,960	03:37,960 (20,1%)	08:48,280 (48,7%)	05:38,700 (31,2%)
23	16:26,520	03:52,400 (23,5%)	07:41,240 (46,8%)	04:52,880 (29,7%)

Ниже в **разделе 4** описаны три конкретных примера применения разработанной единой мультиканальной аннотационной схемы; исследования **4.1** и **4.3** выполнены на материале эталонного подкорпуса, исследование **4.2** — на материале записей #04, 06, 23.

4. Примеры использования единой мультиканальной аннотации

4.1. Взаимодействие вокальной модальности и мануального канала

В исследованиях временно́й координации вокальных и мануальных единиц часто утверждается, что жестикуляция имеет опережающий характер (напр., [Kendon 1980]; [McNeill 1992]; [Loehr 2012]; [Гришина 2017]). Для проверки этой гипотезы на нашем материале мы провели анализ координации на двух уровнях: верхнем и нижнем. Единицами верхнего уровня сегментации в мануальной жестикуляции являются жесты, в речи — ЭДЕ; единицами нижнего уровня — соответственно маховые фазы жестов (далее — «махи») и акцентированные словоформы (далее — «акценты»). Такие соответствия объясняются как содержательными причинами (жесты и ЭДЕ выступают базовыми единицами при анализе жестикуляционного и речевого потока и, предположительно, отражают центральные когнитивные феномены, отвечающие за процесс коммуникации; махи и акцентированные словоформы, в свою очередь, формируют содержательные центры внутри этих базовых единиц), так и сугубо формальными характеристиками.

Табл. 2. Длительность речевых и жестикуляционных единиц в эталонном подкорпусе

	Жесты	ЭДЕ	Махи	Акценты	Все словоформы
Количество	2714	2980	2713	3522	9587
Длительность (среднее), мс	846	1056	453	434	305
Длительность (медиана), мс	720	922	400	420	270

Как видно из табл. 2, внутри пар вида «жест — ЭДЕ» и «мах — акцент» наблюдается близость средних и медианных значений длительности. Это задает необходимую для дальнейшего анализа сопоставимость рассматриваемых единиц.

В ходе работы были разработаны и внедрены три критерия соответствия жестикуляционных единиц речевым: два вида т. н. относительной точности пересечения и гармоническое среднее. Все эти критерии основаны на формальной близости единиц на общей временной шкале — см. рис. 3; подробнее — [Федорова и др. 2016] и [Коротаев 2018]). Было обнаружено, что *степень* координации в парах «жест — ЭДЕ» в целом выше, чем в парах «мах — акцент». Для оценки *характера* координации каждая пара была отнесена, во-первых, к одному из четырех интегральных типов, выделяемых на основании соотношения левых и правых границ (внутренние, объемлющие, ранние и поздние, см. рис. 4), во-вторых, к одному из трех позиционных типов, выделяемых на основании соотношения только левых границ.

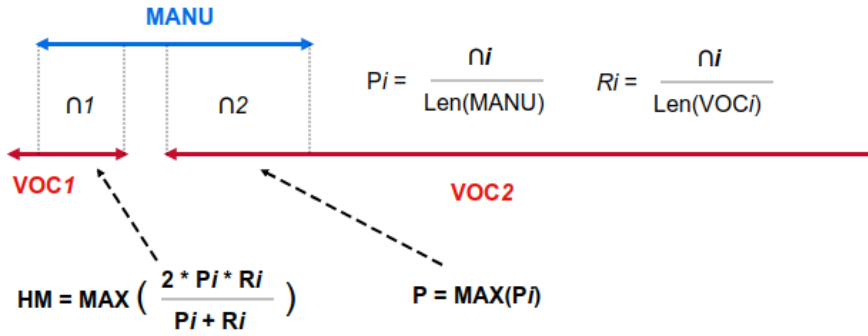


Рис. 3. Соотношение понятий точности пересечения (P) и гармонического среднего (HM; под используемым в формуле R понимается полнота пересечения)

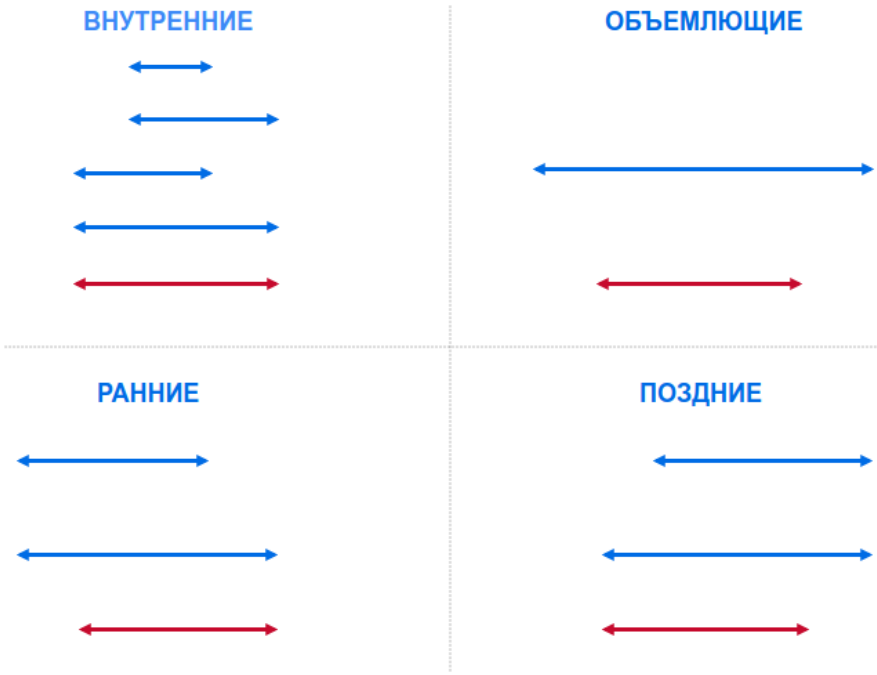


Рис. 4. Интегральные типы жестикуляционных единиц (обозначены синим цветом) с точки зрения соотношения их границ с границами соответствующих им вокальных единиц (обозначены красным цветом)

Исходная гипотеза была подтверждена только для первой оценки: и махи, и жесты демонстрируют тенденцию к интегральному опережению соответствующих им вокальных единиц. Преимущество ранних жестикуляционных единиц над поздними во всех случаях статистически значимо (биномиальный критерий, $p < 0{,}001$). Однако при опоре только на левые границы картина оказалась более сложной: если левые границы махов чаще опережают левые границы соответствующих акцентов, то для жестов, напротив, выявлена тенденция к запаздыванию левых границ относительно ЭДЕ. Так, на рис. 5 показаны количественные соотношения случаев опережения (A) и запаздывания (D) левых границ махов (круговая диаграмма слева) и жестов (диаграмма справа) относительно левых границ соответствующих им речевых единиц. Представлены данные о парах жестикуляционных и речевых единиц, выявленные по критерию максимального гармонического среднего; учитывались только такие случаи опережения и запаздывания, при которых расстояния между границами превышало «дельту» в 50 мс. Наблюдаемые различия в распределении махов и жестов по двумя позиционным типам статистически значимы («хи-квадрат», $p < 0{,}001$). При изменении критерия соответствия и / или «дельты» картина существенно не меняется.

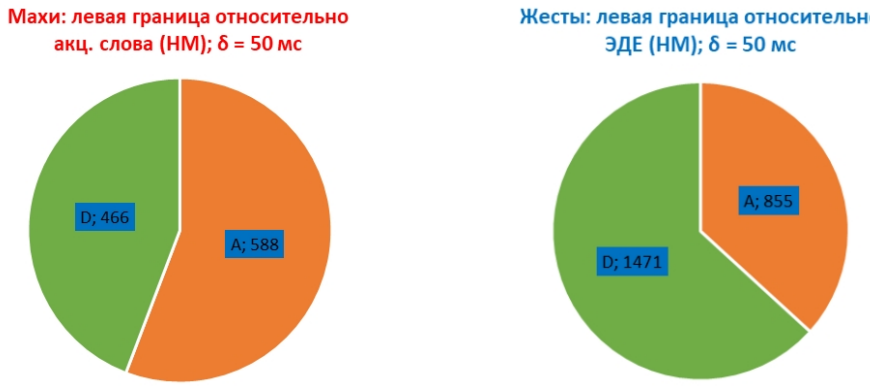


Рис. 5. Распределение махов и жестов по опережению (A) / запаздыванию (D) левых границ относительно левых границ речевых единиц, соответствующим им по критерию максимального гармонического среднего, при «дельте» в 50 мс

Кроме того, на характер распределений влияют и ингерентные свойства жестикуляционных единиц. Так, статистически значимое преимущество раннего интегрального типа над поздним наблюдается только для долгих и средних по длительности махов и средних жестов; для кратких махов, а также кратких и долгих жестов нет оснований отбрасывать нулевую гипотезу о случайности наблюдаемого распределения (биномиальный критерий, $p > 0{,}1$).

4.2. Взаимодействие мануального и окуломоторного каналов

Взаимодействие вокальной модальности и мануального канала, описанное в предыдущем разделе — наиболее изученная область мультиканальных исследований. Считается, что жестикуляция помогает общению коммуникантов; в частности, об этом свидетельствует мета-анализ в [Hostetter 2011]. Жестикуляция, с одной стороны, дублирует информацию, которая передается через речь, и, с другой стороны, передает дополнительную, а иногда и противоречащую, информацию. Таким образом, в процессе коммуникации собеседники принимают во внимание не только речь собеседника, но и жестикуляционный компонент. Однако так ли это на самом деле? Часто ли люди обращают внимание на жестикулирующие руки собеседника? Более точный ответ на этот вопрос дают современные исследования взаимодействия мануального и окуломоторного каналов[2].

На сегодняшний день проведено немного исследований, посвященных изучению зрительного внимания к жестикулирующим рукам собеседника. В пионерских работах [Gullberg, Holmqvist 1999]; [2002] было показано, что в естественноязыковой коммуникации 96% всего времени слушающий смотрит

[2] В данном разделе мы анализируем окуломоторный канал Рассказчиков и мануальный канал Пересказчиков и Комментаторов.

на лицо говорящего собеседника, в то время как на жестикулирующие руки он смотрит совсем редко, всего в 0,5% случаев, что покрывает только 7% жестикуляции собеседника. В работе Beattie et al. 2010 оказалось, что на лицо собеседника слушающий смотрит в 85% времени, а на руки приходится 2,1%. В разных исследованиях высказывались разные предположения о том, в каких именно случаях слушающий переводит взгляд с лица собеседника на его руки. Основное объяснение, предложенное в процитированных работах, связано с тем, что в большинстве случаев, когда испытуемый смотрит на лицо, периферическим зрением он видит и жестикулирующие руки — при стандартном расстоянии между собеседниками в 1,5–2 м и жестикуляции на уровне груди угол зрения оказывается около 8–10°, что в общем случае позволяет видеть движения рук без перевода взгляда, подробнее см. [Барабанщиков, Жегалло 2013].

При нашем дизайне Рассказчик также мог видеть руки Пересказчика периферическим зрением (так называемая ближняя периферия 8,5°), не переводя на них прямой взгляд с его лица. Поэтому, с одной стороны, айтрекинг не дает нам точного ответа на вопрос, сколько времени Рассказчик суммарно смотрел (фокальным или периферическим зрением) на жестикулирующие руки собеседников[3]. С другой стороны, однако, факультативность прямых взглядов на руки предоставляет нам возможность изучить индивидуальные различия между Рассказчиками — о чем говорят те случаи, когда Рассказчик мог бы и не переводить взгляд на руки Пересказчика, а тем не менее его перевел?

Были обнаружены значимые индивидуальные различия между Рассказчиками. Оказалось (см. **табл. 3**), что при одинаковых исходных условиях Рассказчик #23 минимально смотрит на жестикулирующие руки собеседников на протяжении всей коммуникации; внимание Рассказчика #04 привлечено к жестикулирующим рукам только на этапе пересказа, а внимание Рассказчика #06 привлечено к ним как на этапе разговора, так и на этапе пересказа. Данные различия не могут быть объяснены предлагаемыми ранее факторами, а являются следствием действия различных закономерностей распределения зрительного внимания участников коммуникации — общих, контекстно-(не) зависимых и индивидуальных, подробнее см. [Федорова, Жердев 2019].

Табл. 3. Доля зрительного внимания Рассказчика, направленного на жестикулирующие руки Пересказчика и Комментатора (по этапам, в % относительно суммарной жестикуляции)

#	рассказ	разговор		пересказ, Пересказчик
		Пересказчик	Комментатор	
04	0	3	0,8	25,5
06	0	25,6	28,4	19,7
23	0	0	1,2	8,1

[3] Что связано с отсутствием информации о нефокальных движениях взгляда.

4.3. Положение туловища и движения головы коммуникантов в зависимости от их коммуникативной роли

При аннотировании жестов туловища мы выделяем плечевую зону, зону спины и зону ног[4]. Положение туловища показывает степень заинтересованности собеседника и/или его вовлеченности в разговор, ср. [Kendon 1970]; [Scheflen 1973]; [Alibalia et al. 2001]. В частности, при анализе зоны спины мы рассматриваем движения корпусом, которые могут носить как коммуникативный (наклон вперед, когда коммуникант слушает, и отклонение на спинку стула, когда он сам говорит), так и выразительный характер (подача корпуса вперед или вперед-назад для усиления своей речи) (см. подробнее [Birdwhistell 1971]; [Bobick 1997]; [Frey, von Cranach 1973]; [Kendon 1973]; [Mehrabian 1968]).

Если рассматривать канал жестов туловища не изолированно, а во взаимодействии с другими каналами, можно выявить много важных закономерностей. Так, при взаимодействии с цефалическим каналом значимым оказывается расположение туловища относительно стула. Одни участники независимо от их роли в коммуникации опираются на спинку стула, другие опираются локтями на ноги или наклоняются вперед — иногда с дополнительной опорой на стул руками. Это различие влияет на количество так называемых «эховых» движений головы, т. е. движений, в том или ином смысле инициированных движениями в других каналах[5].

Как видно из **табл. 4**, в записи #22 наибольшее количество эховых движений головы зафиксировано у Пересказчика; при этом плечи и спина этого участника не опираются на спинку стула, а ноги убраны под стул. Меньше всего эховых движений у Комментатора, который опирается на спинку стула и спиной, и (время от времени) плечами. Наконец, Рассказчик опирается на спинку стула спиной, но не плечами — и частота эховых движений у этого участника ниже, чем у Пересказчика, но выше, чем у Комментатора. В записях #4 и #23 у всех участников наблюдаются другие закономерности в зонах туловища, что объясняется их индивидуальным стилем поведения.

В результате анализа эталонного подкорпуса в зависимости от ролевой характеристики и текущего статуса в коммуникации (говорящий или слушающий) выделяются следующие стратегии. Рассказчики (за исключением Рассказчика #22), предпочитают облокачиваться на спинку стула при слушании и не облокачиваться при говорении. Комментаторы чаще облокачиваются или имеют опору на руки и у них наблюдается наименьшее число эхо в цефалическом канале. Пересказчики независимо от текущего статуса чаще всего не облокачиваются на спинку стула, но имеют дополнительную опору на руки. Таким образом, у Рассказчиков и Пересказчиков не облокачивание на спинку стула является маркированным положением их текущего коммуникативного статуса.

---

[4] Таким образом, в данной работе мы понимаем туловище расширительно, ср. [Bull 1987].

[5] Принципы описания движений головы см. на сайте проекта multidiscourse.ru/annotation/.

Таблица 4. Положение туловища и эховые движения головы в эталонном подкорпусе. Условные обозначения: A — опирается на спинку стула; B — не опирается на спинку стула; a — опирается на локти[6]; b — опирается руками на стул; C — ноги скрещены и вытянуты; D — ноги скрещены под стулом; E — нога на ногу; F — ноги на ширине плеч; G — ноги вместе перед собой (ср. классификацию в Bull 1987: 186–187)

#	плечи	спина	ноги	эхо в цефал. канале[7]
Рассказчик				
04	Ba	A — когда слушает B — когда говорит	E	80 / мин.
22	B	A	C, E, D	48 / мин.
23	B	A — когда слушает B — когда говорит	D, C, G	80 / мин.
Комментатор				
04	Ba	B, A	F, D	48 / мин.
22	A, B	A	D	17 / мин.
23	Bb	B	C, D, F	17 / мин.
Пересказчик				
04	Ba	B	D — когда слушает C — когда говорит	48 / мин.
22	Bb	B	C, F, D	80 / мин.
23	Ba	A, B	F, G	17 / мин.

5. Заключение

В данной работе был предложен новый подход к анализу естественной мультиканальной коммуникации. Аннотации отдельных каналов, выполненные независимо друг от друга, были сведены в единую мультиканальную аннотацию, что позволило провести несколько конкретных исследований взаимодействия каналов. В настоящей статье было описано два примера взаимодействия базовых коммуникативных каналов (вокальной модальности vs. канала мануальных жестов; мануального vs. окуломоторного каналов), а также пример взаимодействия базового и дополнительного канала (цефалического vs. канала жестов туловища). Данная работа будет продолжена в сторону увеличения количества одновременно рассматриваемых каналов, вплоть до анализа канала жестов всего тела (см. **рис. 1**). Такой подход сделает возможным анализ таких исходно мультиканальных явлений, как, напр., смех [Chafe 2007].

---

[6] Локти могут находиться на бедрах или на коленях, на любой поверхности ног, на которую можно опереться или быть уперты в подвздошные кости.

[7] Среднее количество эховых движений в минуту.

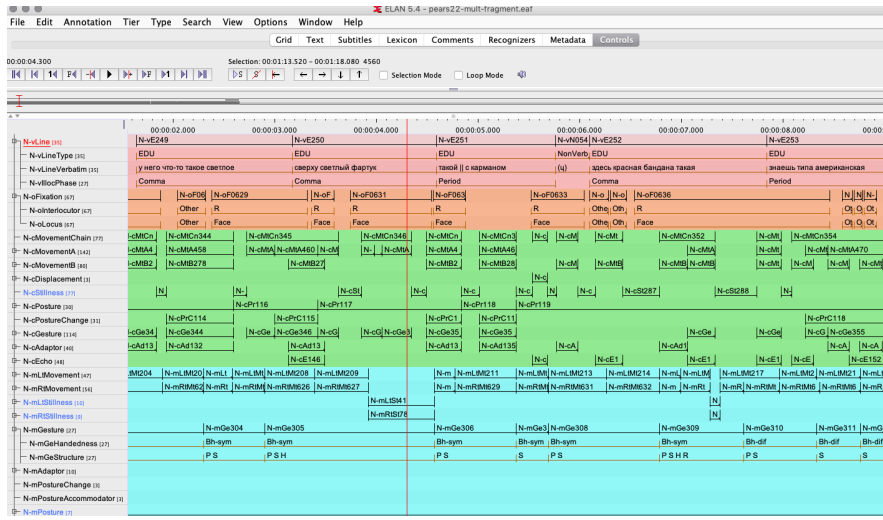
Литература

1. *Adolphs S., Carter R.* (2013), Spoken corpus linguistics: From monomodal to multimodal, N.-Y.: Routledge.
2. *Alibalia M. W., Heatha D. C., Myers H. J.* (2001), Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen, Journal of Memory and Language, Vol. 44 (2), pp. 169–188.
3. *Barabanshchikov V. A., Zhegallo A. V.* (2013), Registration and analysis of gazes [Registratsiya i analiz napravlennosti vzora cheloveka]. Moscow: Institut psikhologii RAN.
4. *Birdwhistell R. L.* (1971), Kinesics and context, London: Penguin Press.
5. *Bobick A.* (1997), Movement, activity, and action: the role of knowledge in the perception of motion, Royal Society Workshop on Knowledge- based Vision in Man and Machine, London.
6. *Bull P. E.* (1987), Posture and Gesture, New York: Pergamon Press.
7. *Butovskaya M. L.* (2004), Body language: nature and culture (evolutionary and cross-cultural foundations for non-verbal human communication) [Yazyk tela: priroda i kul'tura (evolyutsionnye i kross-kul'turnye osnovy neverbal'noy kommunikatsii cheloveka)], Moscow: Nauchny mir.
8. *Chafe W.* (1994), Discourse, consciousness, and time. The flow and displacement of conscious experience in speaking and writing, Chicago.
9. *Chafe W.* (2007). The importance of not being earnest: The feeling behind laughter and humor, Amsterdam: John Benjamins.
10. *Chafe W.* (ed.) (1980), The pear stories: Cognitive, cultural, and linguistic aspects of narrative production, Norwood: Ablex.
11. *Efron D.* (1941), Gesture and environment, Oxford: King's Crown Press.
12. *Fedorova O. V., Kibrik A. A., Korotaev N. A., Litvinenko A. O., Nikolaeva Ju. V.* (2016), Temporal coordination between gestural and speech units in multimodal communication [Vremennaya koordinatsiya mezhdu zhestovymi i rechevymi edinitsami v mul'timodal'noy kommunikatsii], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii], RGGU, Moscow, pp. 159–170.
13. *Fedorova O. V.* (2017), Distribution of the interlocutors' visual attention in natural communication: 50 years later [Raspredeleniye zritel'nogo vnimaniya sobesednikov v estestvennoy kommunikatsii: 50 let spustya], E. V. Pechenkova, M. V. Falikman (eds.) Cognitive science in Moscow: new research. Proceedings of the conference [Kognitivnaya nauka v Moskve: novye issledovaniya. Materialy konferentsii 15 iyunya 2017]. Moscow: BukiVedi, IPPiP, pp. 370–375.
14. *Fedorova O. V., Zherdev I. Ju.* (2019), Follow the hands of the interlocutor! (on strategies for the distribution of visual attention) [Sledi za rukami sobesednika (o strategiyakh raspredeleniya zritel'nogo vnimaniya], Experimental Psychology [Ekhsperimental'naya psikhologiya], Vol. 1.
15. *Frey S., von Cranach M.* (1973), A method for the assessment of body movement variability, M. von Cranach, I. Vine (eds.), Social communication and movement, pp. 389–418. London: Academic Press.

16. *Grishina E. A.* (2017), Russian gestures from a linguistic perspective [Russkaya zhestikulyatsiya s lingvisticheskoy tochki zreniya], Moscow: Jazyki slavyanskoy kul'tury.
17. *Gullberg M., Holmqvist K.* (1999), Keeping an eye on gestures: Visual perception of gestures in face-to-face communication, Pragmatics and Cognition, Vol. 7., pp. 35–63.
18. *Gullberg M., Holmqvist K.* (2002), Visual attention towards gestures in face-to-face interaction vs. on screen, I. Wachsmuth and T. Sowa (Eds.), Gesture and Sign Language in Human–Computer Interaction, Berlin Heidelberg: Springer, pp. 206–214.
19. *Hostetter A. B.* (2011), When do gestures communicate? A meta-analysis, Psychological Bulletin, Vol. 137(2), pp. 297–315.
20. *Kendon A.* (1970), Movement coordination in social interaction: Some examples described, Acta Psychologica, Vol. 32 (2), pp. 100–125.
21. *Kendon A.* (1973), The role of visible behaviour in the organization of social interaction, von Cranach, I. Vine (eds.), Social communication and movement: Studies of interaction and expression in man and chimpanzee. London: Academic Press, pp. 29–74.
22. *Kendon A.* (1980), Gesticulation and speech: Two aspects of the process of utterance, in M. R. Key (ed.), The relationship of verbal and nonverbal communication, pp. 207–227.
23. *Kendon A.* (2004), Gesture. Visible action as utterance, Cambridge.
24. *Kibrik A. A.* (2010), Multimodal linguistics [Mul'timodal'naya lingvistika], Yu. I. Aleksandrov, V. D. Solov'yev (eds.), Cognitive studies [Kognitivnyye issledovaniya], Vol. IV, Institute of psychology, Moscow, pp. 134–152.
25. *Kibrik A. A.* (2018), Russian multichannel discourse. Part I. Setting up the problem [Russkiy mul'tikanal'nyy diskurs. Chast' I. Postanovka problemy], Psikhologicheskiy zhurnal, Vol. 39 (1), pp. 70–80.
26. *Kibrik A. A., Korotaev N. A., Podlesskaya V. I.* (in press), Russian spoken discourse: Local structure and prosody, In Search for a Reference Unit of Spoken Language: A Corpus-Driven Approach.
27. *Kibrik A. A., Podlesskaja V. I.* (eds.), (2009), Night Dream Stories: A corpus study of spoken Russian discourse [Rasskazy o snovideniyakh: korpusnoye issledovaniye russkogo ustnogo diskursa]. Moscow: Jazyki slavyanskikh kul'tur.
28. *Knight D.* (2011), Multimodality and active listenership: A corpus approach, Bloomsbury, London.
29. *Korotaev N. A.* (2018), On temporal coordination between gestural and speech units in spontaneous spoken communication [O vremennoj koordinatsii zhestikulyatsionnyh i rechevyh edinits v nepodgotovlennoy ustnoy kommunikatsii], "Slovo i zhest" conference, Moscow, pp. 10–12.
30. *Kress G.* (2002), The multimodal landscape of communication, Medien Journal, Vol. 4, pp. 4–19.
31. *Kreydlin G. E.* (2002), Nonverbal semiotics [Neverbal'naya semiotika], New literary review, Moscow.
32. *Linell P.* (1982), The written language bias in linguistics, University of Linköping.

33. *Litvinenko A. O., Nikolaeva Ju. V., Kibrik A. A.* (2017), Annotation of Russian manual gestures: Theoretical and practical issues [Annotirovaniye russkikh manual'nykh zhestov: teoreticheskiye i prakticheskiye voprosy], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2017" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2017"], Moscow: RGGU, pp. 255–268.
34. *Loehr D.* (2012), Temporal, structural, and pragmatic synchrony between intonation and gesture, Laboratory Phonology, vol. 3 (1), pp. 71–89.
35. *McClave E. Z.* (2000), Linguistic functions of head movements in the context of speech, Journal of Pragmatics 32(7), pp. 855–878.
36. *McNeill D.* (1992), Hand and mind: What gestures reveal about thought, Chicago: University of Chicago Press.
37. *McNeill D.* (2005), Gesture and thought, Chicago.
38. *Mehrabian A.* (1968), Inference of attitude from the posture, orientation and distance of a communicator, Journal of Consulting and Clinical Psychology, Vol. 32, pp. 296–308.
39. *Müller C., Fricke E., Cienki A., McNeill D.* (eds.) (2013–2014), Body—Language—Communication: An international handbook on multimodality in human interaction, Berlin: Mouton de Gruyter. 2 vols.
40. *Scheflen A. E.* (1973), Communicational structure: analysis of a psychotherapy transaction, Bloomington: Indiana University Press.

Приложение. Фрагмент базовой мультиканальной аннотации записи #22



ЭВОЛЮЦИЯ ДИАЛЕКТНОЙ СИСТЕМЫ БЕЗУДАРНОГО ВОКАЛИЗМА В РЕЧИ ЖИТЕЛЕЙ МОСКВЫ: 4 ПОКОЛЕНИЯ

Князев С. В. (svknia@gmail.com)[1, 2],
Малыхина П. А. (malyhinapolina@rambler.ru)[2]

[1]Национальный исследовательский университет «Высшая школа экономики»

[2]Московский государственный университет им. М. В. Ломоносова, Москва, Россия

EVOLUTION OF DIALECTAL UNSTRESSED VOWELS' SYSTEM IN MOSCOW: 4 GENERATIONS

Knyazev S. V. (svknia@gmail.com)[1,2],
Malykhina P. A. (malyhinapolina@rambler.ru)[2]

[1]National Research University Higher School of Economics

[2]Moscow State Lomonosov University, Москва, Россия

The paper deals with evolution of one part of dialectal phonetic system (neutralization of non-high unstressed vowels' in different allophones as a function of stressed vowel's length or/and quality) over the course of three generations of speakers from one family, moved from a village to Moscow, Russian capital city. We discuss some methods of phonetic analysis that could be utilized in order to present sound changes observed and argue that the result obtained from a large data volume could be not so informative as compared to those, achieved from thorough analysis of every token. Our results show that the phonetic system starts to change immediately after the resettlement of a family: in the first generation of a family moved. The second and third generation displays yet more dramatic changes with only few markers of previous dialectal peculiarities remaining; along with this, the qualitative dissimilation survives somewhat longer than the quantitative one.

Key words: phonetics, dialects, unstressed vowels, evolution of vowel system

1. Введение

Вопросам утраты диалектных особенностей посвящена обширная литература. Общеизвестно, что разные диалектные черты утрачиваются с разной скоростью [Labov 1963, 1972, 2001], [Trudgill 1974], [Kochetov 2006] и др.; в работах [Labov, Rosenfelder & Fruehwald 2013], [Fruehwald 2017], [Daniel et all 2019] этот факт подтверждается статистически.

В процессе перехода носителей диалекта на литературный язык быстрее всего утрачиваются те звуковые особенности диалектной речи, которые:

1) осознаются и контролируются носителями,
2) существенно отличаются от литературных,
3) могут служить показателями диалектной основы литературного произношения.

Для современного русского литературного языка (далее — СРЛЯ) в качестве примера таких особенностей можно привести оканье, яканье, произношение [ɣ] на месте литературного [г]. Наоборот, те особенности, которые (почти) не осознаются и не контролируются говорящими (например, тайминг фразового акцента) или не являются / не воспринимаются говорящими как диалектные (например, произношение [ш'ч'] на месте [ш':]), сохраняются гораздо дольше.

Существуют и такие диалектные особенности, отнесение которых к той или иной группе явлений не столь очевидно. Среди них можно выделить диссимилятивное аканье (далее — ДА), представляющее собой качественную (исходно — количественную) зависимость реализации гласного предударного слога после исконно твердого согласного от подъема (исходно — длительности) ударного гласного. В системах с ДА в позиции перед открытым (долгим) ударным гласным [а] произносится краткий [ъ]-образный звук, а перед закрытыми (краткими) ударными [и], [ы], [у] — долгий [а] (например, *в*[ъ]*да*, но *в*[а]*ды*; *ст*[ъ]*ла*, но *ст*[а]*лу* при стандартном литературном *в*[а]*да*, *в*[а]*ды*; *ст*[а]*ла*, *ст*[а]*лу*); перед остальными ударными возможна вариативность в зависимости от типа ДА[1].

С одной стороны, ДА представляет собой модель, существенно отличную от той, что считается стандартом СРЛЯ (произношение [а] в первом предударном слоге после твердого согласного вне зависимости от подъема ударного гласного). С другой стороны, далеко не всегда этот тип вокализма осознаётся носителями как отступающий от литературного стандарта. Ситуация эта осложняется ещё и тем, что, например, в литературном произношении жителей Урала, Сибири, Дальнего Востока и ряда других регионов России гласный в первом предударном слоге после твердого согласного хоть и не зависит от подъема ударного гласного, но произносится менее открыто и более кратко, чем в московском или петербургском вариантах СРЛЯ.

[1] Качество согласных, отделяющих предударный гласный от ударного при этом роли не играет.

2. Цель исследования

Целью настоящего исследования была апробация методики анализа эволюции диалектной системы с ДА в речи 4-х поколений женщин одной семьи, представители второго поколения которой переехали из д.Алтухово Навлинского района Брянской области в Москву.

В качестве **информантов** в исследовании приняли участие:

1) Анна Павловна (прабабушка): 1924 года рождения, образование 10 классов, проживает в д. Алтухово Навлинского района Брянской области с рождения;
2) Людмила Михайловна (бабушка): 1953 года рождения, образование 10 классов, проживает в Москве с 18 лет;
3) Елена Владимировна (мама): 1975 года рождения, образование высшее, проживает в Москве с рождения;
4) Полина Альбертовна: 1998 года рождения, образование незаконченное высшее, проживает в Москве с рождения.

3. Материал для исследования

Материалом для исследования служили двух- и трехсложные слова с гласным на месте этимологических *о* и *а* в первом предударном слоге после твердого согласного и ударными гласными разного подъема (*кабак*, *капот*, *капут* и т. п.) в составе связного текста (автор — М. Кабанова[2]), который приводится ниже (тестовые слова выделены подчеркиванием):

> *У деревенского жителя в любое время много забот, но не в этом году. Покос, не успев начаться, уже был заброшен. На земле валялось несколько ржавых мотыг. Они были давно оставлены крестьянами и забыты. Уже стерлись следы от копыт — копыт лошади, одетой в тяжелую упряжку и пахавшей поле еще совсем недавно. Но и не шумят задорные трещотки, не пляшут гопак. Веселья — тоже не было. Кабак, в котором мы прятались, был заполнен темнотой, хоть мы и держали в руках свечки. Раздался стук. Дверь чуть не вылетела из пазов. Пазов, которые и так еле держались. Свечи задрожали, задув их, мы прижались к стенке.*
>
> *— Капут! — закричал я и ринулся к выходу, офицер побежал за мной.*
>
> *— Там машина!*
>
> *— У нее капот что ли открыт? — спросил офицер.*
>
> *— Да!*
>
> *— А это капот «Волги» или «Москвича»?*
>
> *— «Москвича», он здесь был забыт, когда я еще в школу ходил — отвечал я на бегу.*
>
> *В этот момент, офицер выстрелил, захлопнул капот и сел в салон. Попав туда, он пытался найти ключи, он наткнулся на иконки,*

[2] Данный текст не был создан специально для данного эксперимента, он используется для диагностики типов предударного вокализма.

фотографию, где пляшут гопак, одежду, моток, скрученный из проволоки , но ключей не было. Мне хотелось плакать. Этот позыв, который душил меня изнутри, становился все сильнее. Я знал: те люди убивали без разбору: попов, женщин, детей. Я пытался отвлечься: вспоминал резвый гопак отца, веселившегося на свадьбе брата, разглядывал салон, лики святых на картонке, упавший на пол моток. Это была машина попов. Пять лет назад они приезжали на выставку святынь, где теперь стоит тот самый кабак. Над ухом у меня просвистела пуля. Снова открывшись, загораживал дорогу капот. Шанса на спасение не было. Пули летали. Попав, одна из них убила офицера. Я выхватил из его руки пистолет, выскочил из машины. Закричал: «Гитлер, капут!», и ринулся на солдат. Боль поразила все тело, больше не было забот, мотыг, следов копыт, свечей. Гопак, отец, машина — все исчезло.

Для получения экспериментальных данных информанты 2–4 зачитывали приведенный выше текст вслух, для информанта 1 материал не мог быть получен тем же способом, поэтому данные отбирались из его спонтанной речи.

В работе были использованы экспериментально-фонетические **методы**: в ходе исследования анализировались длительность и значения первых двух формант (F-картина) ударных гласных и гласных 1-го предударного слога после (исконно) твердых согласных при помощи программы *Praat* и производилось их сопоставление для каждого из информантов. Значения формант (амплитудных максимумов в спектре) коррелируют с артикуляционными параметрами гласных: первая форманта (F1) — с их подъемом, положением языка на вертикальной оси (чем выше F1, тем более открытый гласный), а вторая (F2) — с рядом, то есть, положением языка на горизонтальной оси (чем выше F2, тем более передним является гласный); таким образом, по взаимному расположению формант можно судить о качестве гласного, в данном случае — чем больше разница между F2 и F1, тем больше гласный приближается к редуцированному (по сравнению с [а]).

4. Результаты исследования

4.1. Результаты 1

На начальном этапе исследования были проанализированы длительность и F-картина гласных ударного и первого предударного слога **в отдельных словах** с различными ударными гласными в произношении **информантов 2–4** с целью предварительного установления типа безударного вокализма в их речи (наличие ДА в речи информанта 1 не вызывало сомнений).

На **рисунках 1**, **2** и **3** приведены данные о длительности гласных в словах *попав*, *хотелось*, *назад*, *попов*, *машины*, *раздался* и *пазов*; на рисунке 4 — об F-картине гласных в словах *раздался* и *пазов*.

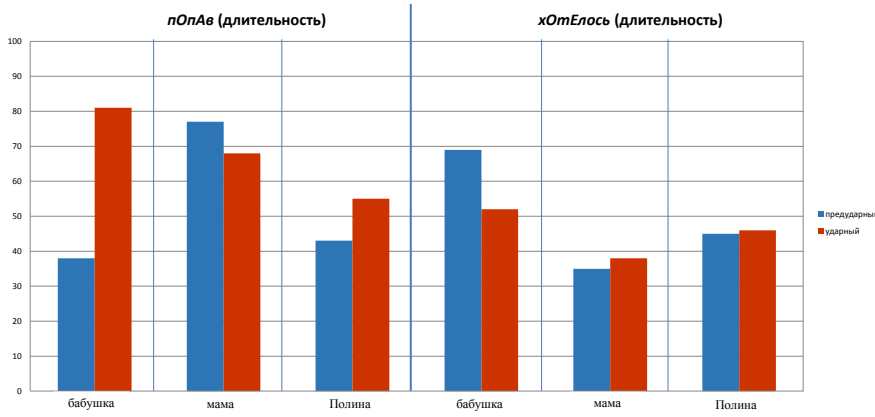


Рис. 1. Абсолютная длительность гласных (мс) в слове *попав* и *хотелось*, слева направо информанты 2, 3, 4. Столбец 1 — длительность предударного гласного, столбец 2 — длительность ударного гласного

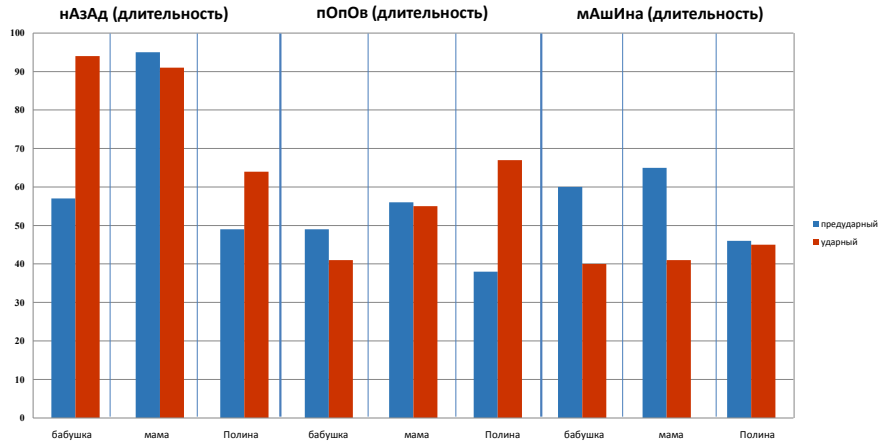


Рис. 2. Абсолютная длительность гласных (мс) в слове *назад*, *попов, машина*; слева направо информанты 2, 3, 4. Столбец 1 — длительность предударного гласного, столбец 2 — длительность ударного гласного

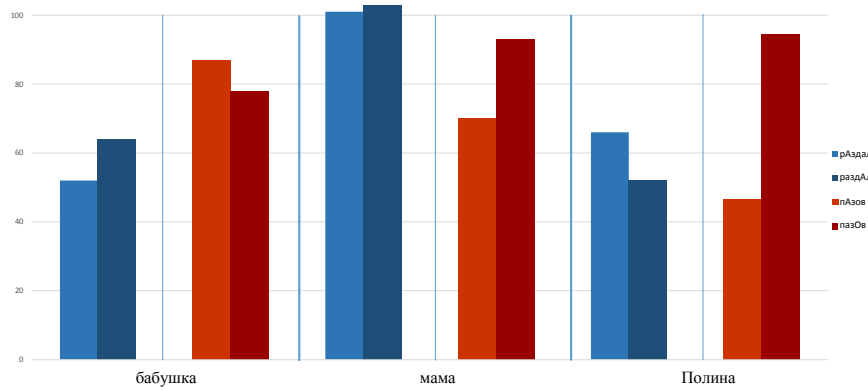


Рис. 3. Абсолютная длительность гласных (мс) в словах *раздался* и *пазов*, слева направо информанты 2, 3, 4. Столбец 1 — длительность предударного гласного в слове *раздался*, столбец 2 — длительность ударного гласного в слове *раздался*, столбец 3 — длительность предударного гласного в слове *пазов*, столбец 4 — длительность ударного гласного в слове *пазов*

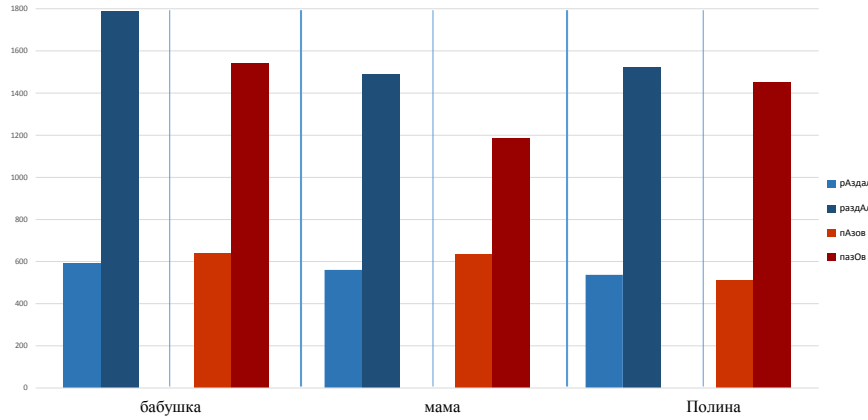


Рис. 4. F-картина предударных гласных в словах *раздался* и *пазов*, слева направо информанты 2, 3, 4. Столбец 1 — F1 предударного гласного в слове *раздался*, столбец 2 — F2 предударного гласного в слове *раздался*, столбец 3 — F1 предударного гласного в слове *пазов*, столбец 4 — F2 предударного гласного в слове *пазов*

4.2. Выводы 1

Данные, приведенные на рисунках 1–4, позволяют сформулировать предварительный вывод о том, что в произношении диктора 2 (бабушка) может быть зафиксирована зависимость длительности предударного гласного от подъема

ударного, в произношении дикторов 3 (мама) и 4 (Полина) этой зависимости не наблюдается; при этом качество предударного гласного зависит от подъема ударного у информантов 2 и 3 (что выражается в разнице значений F2 и F1, большей в слове *раздался*), но не у информанта 4.

У информантов 3 и 4 в отдельных словах фиксируются значения, на основании которых может создаваться впечатление о том, что либо влияние подъема ударного на длительность предударного есть — но обратное тому, которое наблюдается у информанта 2 (см. слово *машина* у информанта 3), либо имеется не связанная с подъемом зависимость между длительностью ударного и предударного гласных (см. слова *раздался* и *пазов* у информанта 4), однако анализ материала большего объема (см. ниже) позволяет утверждать, что эти предположения не соответствуют действительности.

4.3. Результаты 2

На втором этапе исследования были проанализированы длительность и F-картина гласных ударного и первого предударного слога на материале **всех тестовых слов** с различными ударными гласными в произношении **информантов 1–4**. Сначала были получены усредненные данные по всем произнесениям: для информанта 1 — всего 27 слов (7 перед ударным [а], 20 — перед остальными ударными гласными), для информантов 2–4 всего 26 слов (12 перед ударным [а], 14 — перед остальными ударными гласными). Полученные данные приведены на рис. 5 и 6.

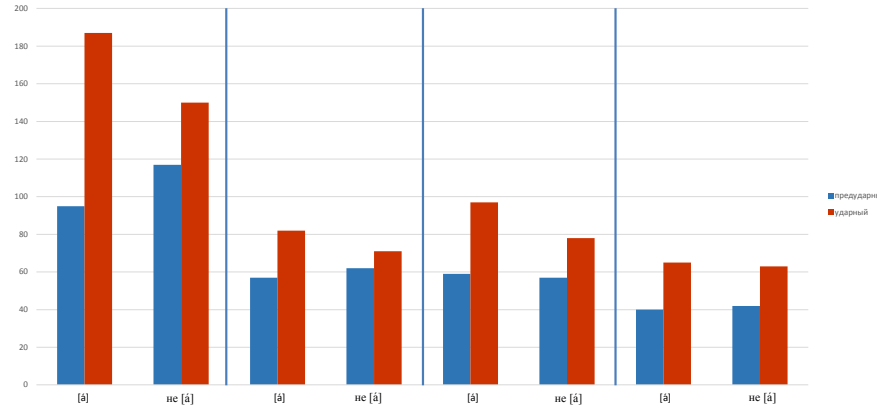


Рис. 5. Абсолютная длительность гласных (мс), усредненная по всем тестовым словам, слева направо информанты 1, 2, 3, 4. Столбец 1 — длительность предударного гласного, столбец 2 — длительность ударного гласного; слева в словах с ударным [а], справа — в остальных словах. Слева направо — прабабушка, бабушка, мама, Полина

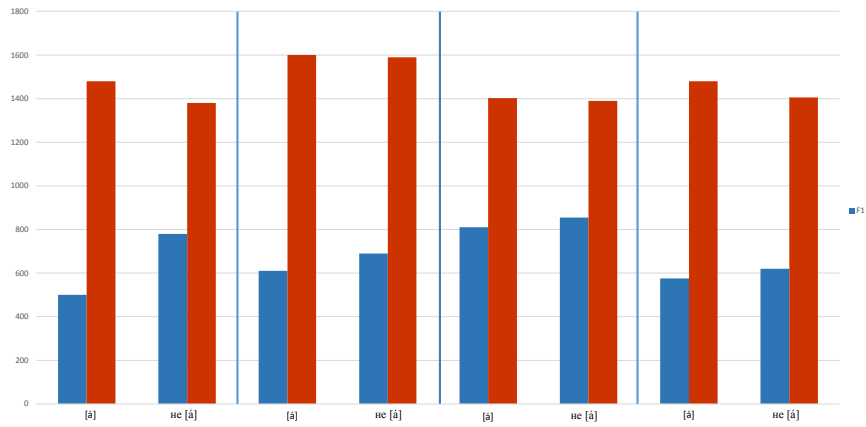


Рис. 6. F-картина гласных (F1 vs F2, гц), усредненная по всем тестовым словам, слева направо информанты 1, 2, 3, 4. Столбец 1 — F1 предударного гласного, столбец 2 — F2 предударного гласного; слева в словах с ударным [а], справа — в остальных словах

4.4. Выводы 2

На основании данных, представленных на рис. 5 и 6, можно утверждать, что в произношении дикторов 1 (прабабушка) и 2 (бабушка) имеет место как

- качественная зависимость предударного гласного от ударного (перед ударным [а] гласный предударного слога имеет более низкое значение F1, чем перед остальными ударными гласными, что свидетельствует о его более [ъ]-образной артикуляции), так и
- количественная (обратно пропорциональная): чем короче ударный гласный, тем больше длительность предударного.

При этом зависимость эта в произношении информанта 2 (бабушка) выражена в значительно меньшей степени, чем у информанта 1 (прабабушка).

В то же время в произношении дикторов 3 (мама) и 4 (Полина) этой зависимости не наблюдается, однако в целом гласный первого предударного слога у информанта 3 значительно (на 50%) дольше и в большей степени приближается к [а], чем у информанта 4, у которого этот гласный гораздо более централизованный (редуцированный, по положению F1), то есть, [ъ].

4.5. Результаты 3

Второй этап анализа позволил сделать вывод о наличии диссимилятивной зависимости между гласными ударного и первого предударного слогов в речи информантов 1 и 2 и её отсутствии в произношении информантов 3 и 4. Однако длительность гласных, особенно ударных, является величиной достаточно нестабильной вследствие ее чувствительности к целому ряду факторов, в том числе — фразовой

позиции слова (под фразовым акцентом или без); кроме того, разные гласные обладают и разной собственной длительностью (существенно увеличивающейся с понижением их подъема), поэтому один и тот же предударный гласный всегда характеризуется большей длительностью относительно узкого ударного гласного, чем относительно широкого. Поэтому для верификации данных, полученных в ходе второго этапа, был проведены измерения длительности гласного первого предударного слога относительно более постоянных величин:

1. общей длительности слова структуры CVCVC (с ударением на втором слоге),
2. длительности одного и того же сегмента (согласного [п] в словах структуры CVпVC с ударением на втором слоге)

на материале идентичных по количеству сегментов и ритмической структуре слов *гопак, попав, попов, капот, копыт, капут*, содержавшихся в экспериментальном тексте, прочитанном информантами 2–4.

Полученные данные приведены ниже в таблицах 1–4.

Таблица 1. Абсолютная длительность (мс) предударных гласных (слева) и длительность смычки последующего согласного [п] (справа), усреднено по словам *гопак, капот, копыт, попав, попов, капут*

	бабушка	мама	Полина
[а́]	66–78	50–95	43–75
[о́]	60–89	52–92	44–78
[у́], [ы́]	62–83	42–98	42–81

Таблица 2. Длительность предударных гласных, усредненная по тестовым словам *гопак, капот, копыт, попав, попов, капут* в процентах относительно длительности смычки последующего согласного [п]; слева — длительность предударного гласного, справа — длительность смычки [п]

	бабушка	мама	Полина
[а́]	84–100	52,8–100	57–100
[о́]	67,4–100	56,5–100	56,4–100
[у́], [ы́]	74,1–100	42,9–100	51,7–100

Таблица 3. Абсолютная длительность предударных гласных, усредненная по тестовым словам *гопак, капот, копыт, попав, попов, капут* в зависимости от общей длительности слова; слева — длительность предударного гласного, справа — общая длительность слова (мс)

	бабушка	мама	Полина
[а́]	66–369	50–356	43–291
[о́]	60–328	52–348	44–295
[у́], [ы́]	62–372	42–344	42–311

Таблица 4. Длительность предударных гласных, усредненная по тестовым словам *гопак, капот, копыт, попав, попов, капут* в процентах относительно общей длительности слова; слева — длительность предударного гласного, справа — общая длительность слова

	бабушка	мама	Полина
[а́]	17,8–100	14–100	14,8–100
[о́]	17,9–100	14,9–100	14,9–100
[у́], [ы́]	16,7–100	12,2–100	13,5–100

4.6. Выводы 3

Приведенные выше данные дают основания для следующих утверждений:

1. меньшая длительность предударного гласного у информанта 4 по сравнению с информантом 3 может быть обусловлена более высоким темпом его речи: средняя длительность одних и тех же слов у информанта 3 на 16,6% выше, чем у информанта 4;
2. при сопоставлении длительности предударного гласного с относительно постоянной величиной (флюктуации длительности смычки [п] в разных словах составляют всего около 5%) диссимилятивной зависимости не наблюдается не только у информантов 3 и 4, но и у информанта 2.

Последний факт может объясняться:

- либо тем, что при ДА расподобление по качеству гласного оказывается важнее, чем по количеству,
- либо тем, что оно представляет собой живую фонетическую зависимость: долгим является не любой гласный перед гласным определенного подъема, а только гласный перед тем гласным, который реально (фонетически) является кратким, таким образом поддерживается относительно постоянная суммарная длительность ударного и предударного слогов.

4.7. Результаты 4

Данные, усредненные на относительно большом количестве примеров, позволяют утверждать, что в произношении дикторов 1 (прабабушка) и 2 (бабушка) фиксируется система ДА (зависимость длительности и качества предударного гласного от подъема ударного), хоть и выраженная в разной степени, в то время как в произношении дикторов 3 (мама) и 4 (Полина) эта система уже разрушена. Тем не менее, перцептивный анализ произношения дикторов 3 и 4 свидетельствует о том, что и в их речи могут встречаться сверхкраткие редуцированные гласные в позиции перед [а́], отличающиеся от гласных в положении перед другими ударными гласными.

Для 4-го этапа анализа на этом основании были отобраны слова *начаться, валялось, пахавшей; забот, покос; году, копыт* с разными ударными гласными и **перцептивно различными** гласными первого предударного слога.

Результаты проведенных измерений длительности и F-картины ударных и предударных гласных в этих словах приведены в таблице 5 и на рисунках 7–9.

Таблица 5. Средняя длительность предударного гласного (слева) относительно средней длительности ударного (справа) и F1–F2 предударного гласного в словах *начаться, валялось, пахавшей; забот, покос; году, копыт*

	бабушка	**мама**	**Полина**
начаться	48–81 мс 400–? гц	59–73 мс 840–1763 гц	71–81,3 мс 314–1776 гц
валялось	33–64,7 мс 572–1828 гц	57,3–83,9 мс 902–1702 гц	48,6–76,4 мс ?
пахавшей	46,4–74,1 мс 641–1243 гц	40,1–64,7 мс 589–1191 гц	45,5–53,8 мс 262–1243 гц
покос	50,5–71,2 мс 697–1374 гц	60–59,6 мс 718–1251 гц	42,9–66 мс 533–1148 гц
забот	69,6–84,6 мс 615–1886 гц	61,4–77,8 мс 538–1656 гц	55–86,6 мс 512–1558 гц
году	91,2–34,4 мс 615–1804 гц	58–98,7 мс 595–1497 гц	54,2–47,6 мс 469–1518 гц
копыт	66,8–95,2 мс 710–1604 гц	49,2–71,1 мс 675–1518 гц	48,1–85 мс 624–1535 гц

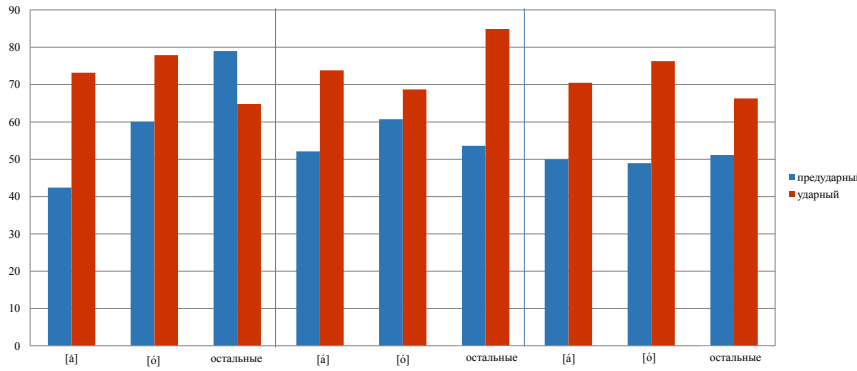


Рис. 7. Абсолютная длительность гласных (мс), усредненная по тестовым словам; слева направо информанты 2, 3, 4. Столбец 1 — длительность предударного гласного, столбец 2 — длительность ударного гласного; слева — в словах *начаться*, *валялось*, *пахавшей* (с ударным [а]), в центре — в словах *покос*, *забот* (перед ударным [о]), справа — в словах *копыт*, *году* (перед ударными [ы], [у])

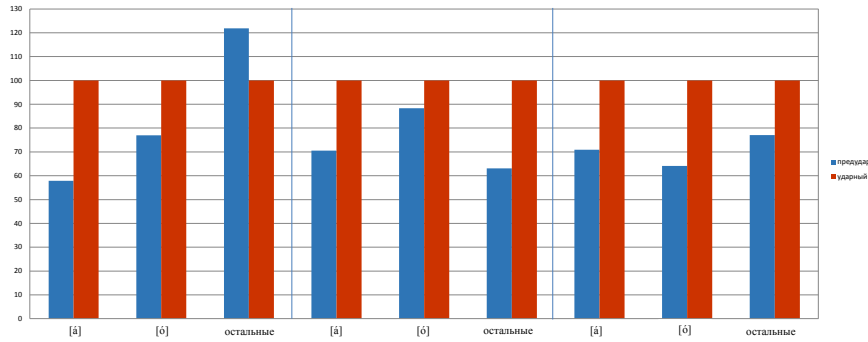


Рис. 8. Относительная длительность гласных предударных гласных (в % от длительности ударного), усредненная по тестовым словам; слева направо информанты 2, 3, 4. Столбец 1 — длительность предударного гласного, столбец 2 — длительность ударного гласного; слева — в словах *начаться*, *валялось*, *пахавшей* (с ударным [а]), в центре — в словах *покос*, *забот* (перед ударным [о]), справа — в словах *копыт*, *году* (перед ударными [ы], [у])

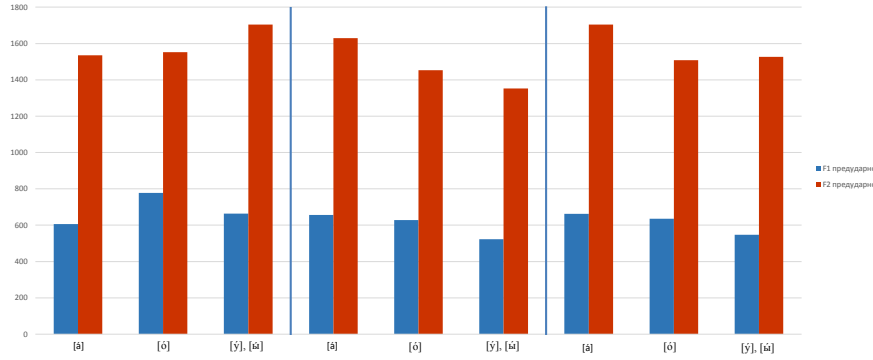


Рис. 9. F-картина гласных (F1 vs F2, гц), усредненная по всем тестовым словам, слева направо информанты 2, 3, 4. Столбец 1 — F1 предударного гласного, столбец 2 — F2 предударного гласного; слева — в словах *начаться*, *валялось*, *пахавшей* (с ударным [а]), в центре — в словах *покос*, *забот* (перед ударным [о]), справа — в словах *копыт*, *году* (перед ударными [ы], [у])

4.8. Выводы 4

Приведенные выше данные позволяют утверждать, что в исследованных отдельных словах у информанта 2 наблюдается отчетливая количественная диссимилятивная зависимость первого предударного гласного от подъема ударного (повторим, что, по данным 2-го этапа анализа на материале всех слов очевидно,

что существует и качественная), у информанта 3 количественная диссимиляция отсутствует, хотя присутствует слабо выраженная качественная (с повышением подъема ударного гласного разница между F2 и F1 уменьшается), в то время как у информанта 4 не наблюдается ни того, ни другого. Можно предположить, что перцептивный эффект ДА в речи информанта 4 является следствием произношение краткого (30–50 мс) гласного, отличающегося от стандартного литературного [а], в первом предударном слоге перед всеми ударными гласными.

5. Заключение

В статье описана апробация методики, применение которой к более обширному материалу может подтвердить или опровергнуть сформулированные ниже гипотезы.

Полученные в ходе настоящего исследования данные позволяют заключить, что при перемещении информанта из диалектной языковой среды в литературную система ДА начинает эволюционировать уже в первом поколении переселенцев: первоначально зависимость качества и длительности предударного гласного от качества и длительности ударного сохраняется, но степень контраста существенно уменьшается. Начиная со второго поколения переселенцев диалектная система претерпевает более значительные перемены: диссимилятивное аканье в целом сменяется недиссимилятивным; при этом диссимилятивная зависимость гласных по качеству сохраняется дольше, чем количественная. Прежняя система проявляется лишь в отдельных случаях: можно предположить, что это происходит в отдельных словах (в результате лексикализации), в позициях (например, под фразовым акцентом определенного типа) или эмоциональных состояниях.

Результаты проведенного исследования свидетельствуют также о том, что не всегда данные, усредненные на обширном материале, являются более точными, чем результат тщательного анализа конкретных языковых фактов существенно меньшего объема.

Особого внимания заслуживают, на наш взгляд, представляющиеся контринтутивными сведения о том, что результатом совпадения двух разных гласных (краткого редуцированного [ъ] и долгого полного [а]), позиционно распределенных при ДА, у представителей разных поколений одной семьи могут быть разные гласные: у более старшего информанта 3 это [а]-образный гласный длительностью около 60 мс, в то время как у самого молодого информанта 4 это [ъ]-образный гласный средней длительностью всего около 40 мс, то есть, редуцированный. На наш взгляд, одно из объяснений (кроме очевидного, связанного с более высоким темпом речи у информанта 4) может заключаться в том, что языковые привычки информанта 3 (1975 г. р.) формировались в конце 70-х годов, когда вариативность произносительных норм в московском регионе, особенно в общественном пространстве и в средствах массовой информации, была несравнимо меньше, чем в начале 2000-х, когда формировались языковые привычки информанта 4 (1998 г. р.), поэтому информант 4 мог слышать и усваивать не только произношение [а], но и [ъ] в первом предударном слоге при отсутствии в его речи живой тенденции к диссимиляции ударного и предударного гласных.

References

1. *Daniel et all* (2019). Dialect loss in the Russian North: modelling change across variables. To appear in: Language Variation and Change.
2. *Fruehwald, Joseph* (2017). Generations, lifespans, and the zeitgeist. Language Variation and Change, 29(1), 1–27.
3. *Kochetov, Alexei* (2006). The role of social factors in the dynamics of sound change: A case study of a Russian dialect. Language Variation and Change 18.01: 99–119.
4. *Labov, William* (1963). The social motivation of a sound change. Word 19. 273–309.
5. *Labov, William* (1972). Sociolinguistic patterns. Oxford: Blackwell.
6. *Labov, William* (2001). Principles of linguistic change. Vol. 2. Social factors. Language in Society. Oxford: Blackwell.
7. *Labov, William, Rosenfelder, Ingrid & Fruehwald, Josef* (2013). One Hundred Years of Sound Change in Philadelphia: Linear Incrementation, Reversal, and Reanalysis. Language, vol. 89, no. 1. 30–65.
8. *Trudgill, Peter* (1986). Dialects in contact. Blackwell.

ИНТРОСПЕКТИВНАЯ ПРОСОДИЧЕСКАЯ РАЗМЕТКА ПИСЬМЕННОГО ТЕКСТА И ЕГО РЕАЛЬНОЕ ОЗВУЧИВАНИЕ (сравнительный анализ на материале коллекции текстов Р. И. Аванесова)

Кривнова О. Ф. (okrivnova@mail.ru),
Смирнова О. С. (kisaolga@mail.ru)
Московский государственный университет
имени М. В. Ломоносова, Москва, Россия

Ключевые слова: устная речь, просодическое членение, словораздел, сегментирующая сила, просодический шов, просодическая разметка, интроспекция, озвучивание, перцептивный, статистический, синтаксический, инструментальный анализ

INTROSPECTIVE AND PERCEPTUAL LABELING OF PROSODIC PHRASING (a comparative analysis on the material of R. I. Avanesov texts collection)

Krivnova O. F. (okrivnova@mail.ru),
Smirnova O. S. (kisaolga@mail.ru)
Lomonosov Moscow State University, Moscow, Russia

This paper discusses the problems and results of a comparative analysis of two fundamentally different types of prosodic phrasing labeling realized for some literary Russian texts. The introduction examines the theoretical basis of the study and formulates specific tasks, the solution of which was necessary for comparative analysis and the achievement of the final goal of the study. The first section of the paper describes the experimental material, methods of research and the basic principles of experimental data processing. In the second, central section of the work, a detailed description of the parameters of comparative analysis of introspective labeling and perceptual one is given. The following parameters were taken into account in the comparative analysis: the general distribution of frequency of occurrence of text spaces with different indexes of word boundary strength; their contextual distribution with respective frequency data; relationship of prosodic breaks' strength with pauses. This section also contains many illustrations that demonstrate the main results of the comparative analysis of the target prosodic labeling of the experimental text material. Section 3 analyzes the relationship between the prosodic breaks' strength and pauses'

duration in both types of labeling analyzed. In conclusion results of the study are summarized and promising areas for further research on the relevant topics are noted.

Key words: phonetics, spoken language, prosodic phrasing, word boundary strength, prosodic breaks, labeling, introspection, insonification, perceptual, statistical, syntactic, instrumental analysis

Введение

Основная ***задача*** работы заключалась в том, чтобы провести сравнительный анализ двух типов разметки просодического членения (ПЧ) повествовательного художественного текста на русском языке. В рамках первого типа производится разметка ***письменного*** варианта текста на основе интроспекции автора разметки, когнитивного интерфейса «синтаксис-просодия», с использованием, возможно, т. н. внутренней речи для контроля адекватности разметки ПЧ с точки зрения смысла и выразительности предполагаемого звучания. Интроспективная разметка (далее И-разметка) часто используется при подготовке ответственных публичных мероприятий, создавая своеобразную партитуру будущего устного выступления. Данный метод в экспертном исполнении применялся на этапе лингвистической предобработки текста в синтезаторах речи типа «Текст-речь» [Dutoit 1997], на ранних этапах разработок, когда практически отсутствовали автоматические средства отображения ПЧ в синтезируемой речи.

Второй тип разметки основывается на перцептивным анализе ***озвученного*** текста обычными носителями русского языка, в ходе и результате которого они должны оценить сегментирующую силу каждого раздела в устном тексте с использованием определённой оценочной шкалы (далее П-разметка).

В теоретическом плане мы, как и ранее, см. [Кривнова, Смирнова 2018], исходим из признания слова основной рабочей единицей как при порождении, так и при восприятии любого текста, а также из того, что любая граница между словами (словораздел) в тексте имеет определенный сегментирующий ***потенциал***, который может реализоваться с разной вероятностью и с разной силой в зависимости от различных контролирующих факторов, в частности авторских предпочтений и фонетического опыта эксперта или аудитора. Сегментирующая сила словораздела фонетически реализуется разными просодическими средствами, что находит отражение в разной глубине просодических швов (ПШ) и просодического членения в целом на соответствующих словоразделах.

1. Экспериментальный материал и методика исследования

Много примеров с И-разметкой просодического членения содержится в приложении к книге Р. И. Аванесова «Русское литературное произношение» [Аванесов 1972]. Далее этот массив текстов обозначается как коллекция РИА.

В РИА используется пять особых маркеров, фиксирующих разную глубину ПЧ: -, |, /, //, /// в направлении возрастания плюс чистый пробел, т. е. автор фактически исходит из шестибалльной количественной шкалы для оценки сегментирующей силы словоразделов. Сам Р. И. пишет по поводу своей разметки следующее: «Членение речи на ритмико-интонационные и синтаксические группы самое приблизительное ввиду неразработанности вопроса… Минимальная пауза, или факультативная, потенциальная отмечается вертикальной пунктирной линией, небольшая пауза, отделяющая менее самостоятельные отрезки речи, обозначается одной вертикальной линией, более длительная пауза, отделяющая более самостоятельные отрезки речи, обозначается двумя вертикальными линиями. В некоторых случаях — для обозначения достаточно законченных отрезков речи — употребляется знак, состоящий их трех вертикальных линий… Нет сомнения в том, что за каждым из этих знаков в нижеприводимых текстах отделяются друг от друга отрезки текста, весьма различные в ритмико-интонационном и синтаксическом отношениях…» [Аванесов 1972: 215].

Для проведения дальнейшего сравнительного анализа исходные маркеры РИА были преобразованы в количественные показатели глубины ПШ в соответствии со следующим правилом:

- Словораздел с дефисом → **<–1>**
- Словораздел без маркера (чистый пробел) → **<0>**
- ¦ → **<1>**
- / → **<2>**
- // → **<3>**
- /// → **<4>**

Приведем в качестве примера интроспективной разметки РИА небольшой фрагмент из текста «Пушкин о родине»: **У <–1> нас <0> есть <0> благо <2> залог <0> всех <0> других <2> у <–1> нас <0> есть <0> надежда <0> и <–1> мысль <2> о <–1> великом <0> назначении <2> нашего <0> отечества <3>**[1].

Следует отметить, что к ссылкам Аванесова на паузы нужно относиться с определенной осторожностью, так как о фонетических средствах ПЧ, отличных от пауз, к тому времени практически ничего не было известно, да и сами паузы как маркер ПЧ были изучены весьма недостаточно даже в естественной звучащей речи, не говоря уже об их отражении в речи внутренней. К сожалению, тексты РИА, как нам известно, реально самим Аванесовым озвучены не были.

В прозаической части РИА собраны тексты весьма разнообразные по авторам, по размеру и времени создания, см. ниже таблицу 1. Сравнительный анализ И-разметки РИА в разных текстах тоже представляет определенный интерес, хотя размер текстов в целом небольшой.

---

[1] Несмотря на определенную архаичность лексики и синтаксиса, данное предложение в контексте всего текстового фрагмента А. С. Пушкина было озвучено диктором и размечено аудиторами без каких-либо трудностей. В то же время как пример для иллюстрации разметки это предложение удобно, так как при своей краткости содержит завершенную мысль и при адекватном интонировании легко понимается любым носителем русского языка.

Таблица 1. Краткая характеристика текстов РИА

Автор текста	Название текста	Количество слов
А. С. Пушкин	«О родине»	118
И. С. Тургенев	Отрывок из рассказа «Лес и степь»	311
Л. Н. Толстой	Отрывок из повести «Хаджи-Мурат»	541
М. Горький	Отрывок из статьи «О языке»	159
М. Пришвин	Рассказ «Говорящий грач»	200
А. Макаренко	Отрывок из «Книги для родителей»	199
К. Федин	Отрывок из романа «Необыкновенное лето»	296
К. Симонов	Отрывок из романа «Дни и ночи»	202
Всего	8	2026

Для получения перцептивной разметки тексты РИА были озвучены непрофессиональным диктором женщиной без опыта чтения перед микрофоном, без подготовки, в рамках одного сеанса записи при естественном для диктора темпе и громкости чтения, с небольшими паузами (5 сек) между отдельными текстами коллекции[2]. Запись производилась на качественной цифровой аппаратуре в студийных условиях.

При П-разметке озвученного текста каждому словоразделу должен быть поставлен в соответствие количественный ***субъективный*** показатель его сегментирующей силы, или иначе, глубины ПШ. В настоящем исследовании разметка производилась с использованием 5-ти балльной шкалы, согласованной с результатами перцептивных экспериментов, проведенных нами ранее с привлечением фонетистов-экспертов и обычных носителей русского языка в качестве аудиторов, подробнее см. об этом [Смирнова 2017]. В настоящей работе в качестве аудиторов выступали студенты филологического факультета МГУ, 8 человек (3 женщины и 5 мужчин), все носители современного русского литературного языка. Их задача состояла в том, чтобы прослушать звучащий текст (в домашних условиях) и в тех местах, где ими ощущались какие-либо фонетические границы/разрывы слитности произнесения, проставить в соответствующем графическом пробеле текста, напечатанном без знаков препинания и заглавных букв цифровой показатель от **1** до **5** в соответствии со степенью выраженности границы (**1** — минимальная выраженность границы, **5** — максимальная). В целях эксперимента озвученный текст был разрезан вручную, с помощью звукового редактора, на осмысленные фрагменты длиной 4–5 полнозначных слов, при этом общее время прослушивания и количество прослушиваний каждого фрагмента никак не регламентировалось. Для стандартизации заполнения перцептивных протоколов и минимизации доли ручного труда при их обработке нами использовалась электронная форма протокола, допускающая ввод только разрешенных значений оценочной шкалы и только в специальные поля

[2] Конечно, особый интерес представляет анализ И-разметки текста и его последующего озвучивания одним и тем же человеком. К сожалению, этой возможности в отношении Р. И. Аванесова у нас не было.

на словоразделах. Словоразделам без показателей наличия границы, т. е. чистым пробелам, в протоколах автоматически приписывался нулевой показатель, т. е. фактически в П-разметке использовались оценки от **0** до **5**. Кроме этого, была также произведена полная ***паузальная*** разметка озвученных текстов РИА с одновременным измерением длительности пауз на словоразделах.

Как видно из сказанного выше, оценочные шкалы рассматриваемых разметок совпадают не полностью: в них представлены только 5 формально одинаковых показателей глубины ПШ, а именно **0, 1, 2, 3, 4** и заметна разница в степени детальности краевых (полюсовых) участков оценочных шкал: в И-разметке **0**-показателю предшествует словораздел «-1» (исходно с дефисом), а в П-разметке показатель максимальной глубины равен **5**, а не **4**. Для проведения дальнейшего анализа необходимо было привести оценочные шкалы к какому-то единому «знаменателю». Исходя из общих фонетических соображений, мы приняли следующее рабочее решение: в И-разметке показатель «**-1**» был объединен с **0**-м, а в П-разметке были объединены показатели **4** и **5**. Действительно, предварительный анализ разметок показал, что словораздел с дефисом в И-разметке встречается как правило после простых предлогов и союзов, т. е. внутри традиционного фонетического слова, и в два раза реже **0**-го словораздела на стыке полнозначных слов. Фонетические основания для такого разграничения даже в настоящее время нельзя считать в общем случае убедительными. При этом в П-разметке словоразделы с исходной оценкой «**-1**» регулярно маркируются аудиторами как **0**-е, а словоразделы с **0**-м И-показателем, как правило, оцениваются аудиторами также как **0**-е, очень редко как ПШ малой глубины. Что касается максимальных показателей **4** и **5** в П-разметке, то аудиторы редко пользуются оценкой **5** и обычно на границе между текстами или самостоятельными предложениями, причем вариативно с оценкой **4**.

Кроме проблемы согласования краевых показателей шкалы, нужно учитывать также внутренние локальные смещения соседних оценок, с точки зрения покрытия ими конкретных словоразделов и их просодического маркирования, в том числе паузального. Подобные различия в оценочных шкалах и разметках на их основе могут быть выявлены только в результате сравнительного анализа с учетом разных параметров: частотного и контекстного распределения словоразделов с разными оценками, а также их фонетической реализации. С учетом сказанного выше, в обоих типах разметки ПЧ, которые будут анализироваться дальше, используются только показатели **0, 1, 2, 3, 4**.

Наконец, последний вопрос, который требуется обсудить в методическом разделе, связан с обобщением перцептивных оценок разных аудиторов для каждого отдельного словораздела в П-разметке. При проведении любого перцептивного исследования даже при поверхностном рассмотрении результатов обычно обнаруживаются испытуемые, реакции которых явно выбиваются из среднестатистической тенденции по разным причинам: из-за отсутствия фонетического слуха, соответствующего опыта и даже необходимой добросовестности. В нашем случае таких оказалось двое, их протоколы в последующем анализе не учитывались. Что касается протоколов остальных 6 аудиторов, то они использовались для получения ***медианного*** значения оценок на каждом словоразделе, с округлением

в большую сторону, поскольку при четном числе аудиторов могли получиться дробные значения. На основе медианных значений был построен окончательный вариант перцептивной разметки РИА, именуемый далее МЕД6+, который затем сравнивался с интроспективной разметкой коллекции. Ниже приводится пример фрагмента текста «Пушкин о родине» с его перцептивной разметкой.

***П-разметка*: У<0> нас <0> есть <0> благо <3> залог <0> всех <0> других <3> у <0> нас <0> есть <0> надежда <0> и мысль <2> о <0> великом <0> назначении <0> нашего <0> отечества <4>.**

Из данного примера видно, что в целом текстовая локализация маркированных словоразделов совпадает с И-разметкой данного примера, приведенной выше, но сила словоразделов в П-разметке имеет более высокие показатели. Одна из причин этого рассогласования может заключаться в том, что Аванесов ориентировался в своей разметке в основном на паузы, а аудиторы учитывали и другие просодические маркеры и изначально пользовались шкалой с бóльшим диапазоном[3].

Коэффициент ранговой корреляции Спирмена составляет для И- и П-разметок 0,86, что говорит о хорошем согласовании динамики изменения (возрастания-убывания) глубины ПШ в обеих разметках. Кроме данного интегрального вывода, безусловный интерес представляют более детальные сведения о соотнесенности И-показателя словораздела с его П-показателем для каждого словораздела с учетом его текстовой позиции, т. е. порядкового номера в РИА, см. **рис. 1**.

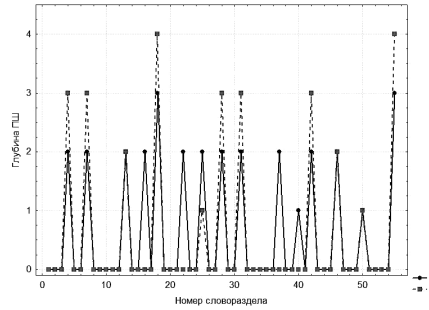


Рис. 1. Сравнительная динамика показателей сегментирующей силы для каждого словораздела в разных типах разметки (на примере фрагмента из текста Пушкина)

[3] Из фонетических исследований хорошо известно, что маркерами просодического членения могут быть разные просодические средства, см. об этом [Кривнова 2015]. В настоящей работе мы ограничились анализом влияния пауз на восприятие просодических швов прежде всего потому, что Аванесов в своей разметке ориентировался именно на этот просодический маркер, см. **раздел 3**. Зависимость восприятия просодических швов от способа их просодического маркирования, безусловно, представляет большой интерес и должна быть предметом самостоятельного изучения.

2. Сравнительный анализ интроспективной разметки коллекции РИА и перцептивной разметки ее озвученного варианта

С целью проведения сравнительного анализа обе разметки были введены базу дискурсивных признаков словораздела, формат и содержание которой подробно описаны в [Кривнова, Смирнова 2018]. Напомним, что указанная база реализована в двух форматах — как электронная таблица EXCEL и как таблица статистического пакета STATISTICA. В пакете STATISTICA производилась статистическая обработка данных для всех параметров сравнительного анализа. При анализе учитывались следующие параметры:

1. общее распределение частоты встречаемости словоразделов с разными показателями сегментирующей силы (**раздел 2.1**);
2. их контекстное распределение с частотными данными (**раздел 2.2**);
3. связь с физическими паузами на словоразделе.

2.1. Частотное распределение словоразделов с разными показателями сегментирующей силы И- и П-разметках РИА

На **рис. 2** приведены сравнительные гистограммы частот словоразделов с разными оценками сегментирующей силы для обоих целевых типов разметки РИА.

Из **рис. 2** видно, что в обоих случаях в РИА заметно и с практически равной частотой преобладает **0**-ой показатель. Это ожидаемый результат, так как фонетическая интеграция слов в просодические составляющие фразового уровня является необходимым условием ПЧ, наряду с локальным маркированием границ. Примерно одинаковая картина наблюдается и для показателя **3**, довольно надежно обеспеченного паузальным маркером, виртуальным во внутренней речи и реальным в звучащей, см. ниже **раздел 3**. Особый случай — показатель **4**, который в П-разметке по частоте больше, чем в И-разметке, что является техническим следствием объединения показателей **4** и **5**, которое было произведено нами для выравнивания оценочных шкал обоих разметок выравнивания шкал.

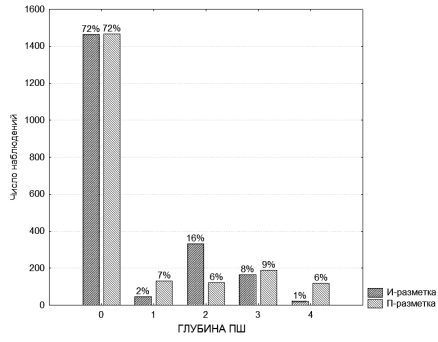


Рис 2. Сравнительная гистограмма частоты глубины ПШ в разных типах разметки РИА[4]

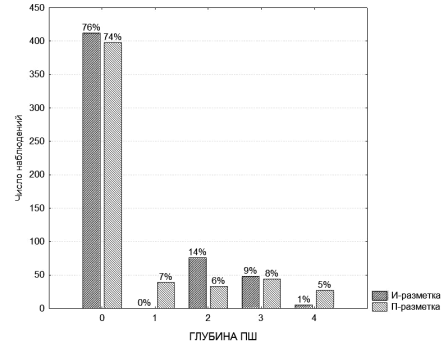


Рис. 2Т. Сравнительная гистограмма частот глубины ПШ для текста Толстого

[4] Здесь и далее %-значения частоты показателей на гистограммах даются относительно общего числа словоразделов в экспериментальном материале.

Аналогичная картина наблюдается и для текстов отдельных авторов. В качестве иллюстрации на **рис. 2Т** приведена сравнительная гистограмма частот словоразделов с разными показателями сегментирующей силы для текста «Толстой. Хаджи-Мурат», который в РИА является самым большим по размеру.

2.2. Контекстное распределение словоразделов с разной сегментирующей силой в И- и П-разметках РИА

2.2.1. При проведении этого этапа анализа мы рассматривали 3 класса контекстов: ***частеречный*** (ЧР), ***синтаксический*** (СК) и ***знаки препинания*** (ЗП). Что касается ЧР, то проведенное нами ранее исследование по влиянию частеречных признаков слов на глубину ПШ в соседнем словоразделе показало, что значимое влияние на этот показатель членения оказывает левый контекст, т. е. ЧР слова перед словоразделом, и прежде всего наличие существительного[5].

Рис. 3. подтверждает преобладающее влияние существительного на силу последующего словораздела, в отличие от всех прочих ЧР вместе взятых. Хотя в обоих случаях при всех типах разметки преобладают **0**-е показатели, после существительного их заметно меньше, а более высоких показателей, напротив, больше. В И-разметке достаточно активно используются также показатели **1** и **2**, с преобладанием последнего, а в П-разметке при наличии тех же показателей преобладает показатель **1**. Более высокий показатель **3** встречается после существительного редко и только в П-разметке[6].

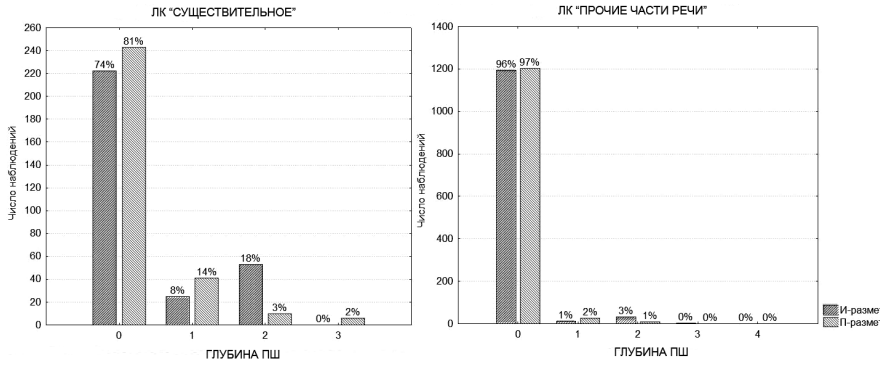


Рис 3. Сравнительные гистограммы частот глубины просодических швов в зависимости от ЧР левого контекста (ЛК) словораздела без знака препинания: слева — существительное, справа — прочие ЧР

[5] В указанном исследовании учитывалось 16 ЧР-категорий, различаемых в [НКРЯ 2005:122–123].

[6] Здесь стоит отметить, что ЧР как фактор глубины ПШ взаимодействует с другими факторами и его автономное рассмотрение не создает полной картины влияния ЧР-признака на сегментирующую силу последующего словораздела. Аналогичное замечание относится и ко всем другим контекстным факторам.

2.2.2. При анализе распределения показателей словоразделов в зависимости от синтаксической границы на словоразделе различались следующие типы синтаксических контекстов:

- **В** — внутри элементарной клаузы
- **К** — конец клаузы внутри самостоятельного предложения
- **КК** — конец самостоятельного предложения внутри абзаца
- **ККК** — конец абзаца внутри текста

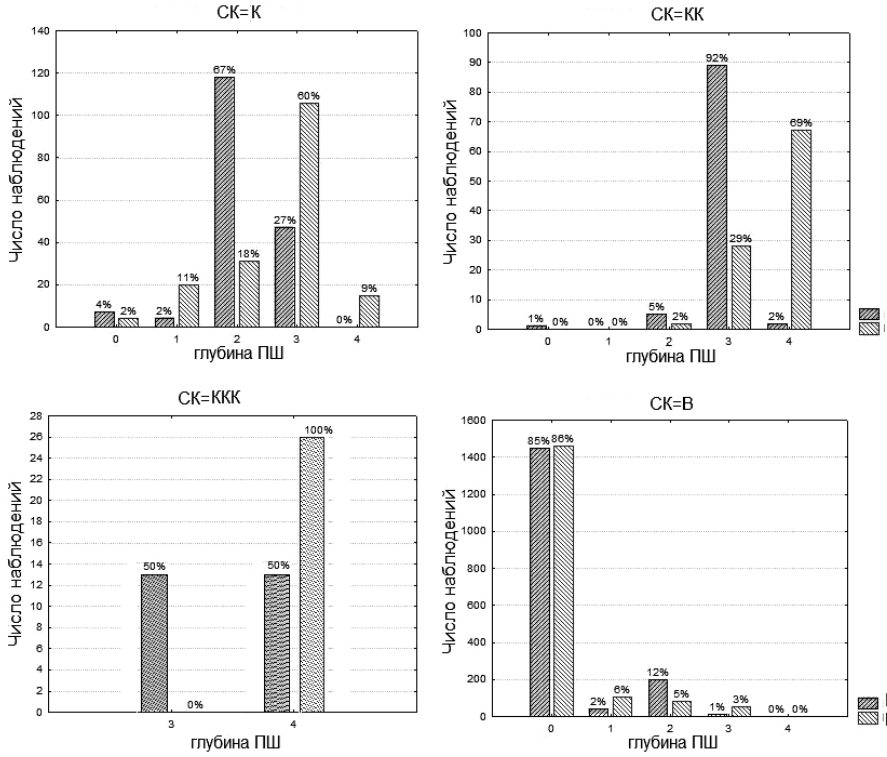


Рис. 4. Частотное распределение силы словоразделов в зависимости от синтаксической границы на словоразделе

Рис. 4 демонстрирует вполне ожидаемую и естественную тенденцию: показатели сегментирующей силы словоразделов возрастают с увеличением глубины синтаксической границы. Действительно, внутри клаузы в обоих типах разметки заметно преобладает **0-й** показатель, хотя встречаются **1** и **2** (около 10 % случаев); в конце самостоятельного предложения внутри абзаца **0**-й показатель встречается редко в обоих типах разметки. Показатели **2** и **3** с некоторым предпочтением первого активно используются в И-разметке, а в П-разметке эти же показатели преобладают, но с предпочтением показателя **3**, при этом встречается и показатель **4**, в отличие от И-разметки. Конец абзаца внутри текста маркируется самыми большими показателями **3** и **4**, причем в П-разметке

используется только максимальный, а в И-разметке оба показателя с равной вероятностью.

2.2.3. Частотное распределение силы словоразделов в зависимости от знака препинания на словоразделе во многом сходно с аналогичным СК-распределением, что естественно, так как пунктуация в русском языке хорошо отражает синтаксическую структуру предложения [Гращенков и др. 2018].

Из гистограмм на **рис. 5** видно, что наличие любого знака препинания на словоразделе значительно снижает частоту **0**-го показателя для обоих типов разметки, но с некоторыми различиями. Неконечные ЗП все же допускают отсутствие просодического маркера (показатель **0**) на словоразделе, в особенности в И-разметке, но наиболее частотным здесь является средний показатель **2**. В П-разметке примерно одинаковые высокие частоты у показателей **2** и **3**, но встречается и показатель **1**, правда, с меньшей частотностью. Конечные ЗП в И-разметке передают заметное частотное лидерство показателю **3** при возможности **2** и **4**. В П-разметке картина в некотором смысле обратная: частотное преимущество у максимального показателя **4**, при возможности также показателя **3**.

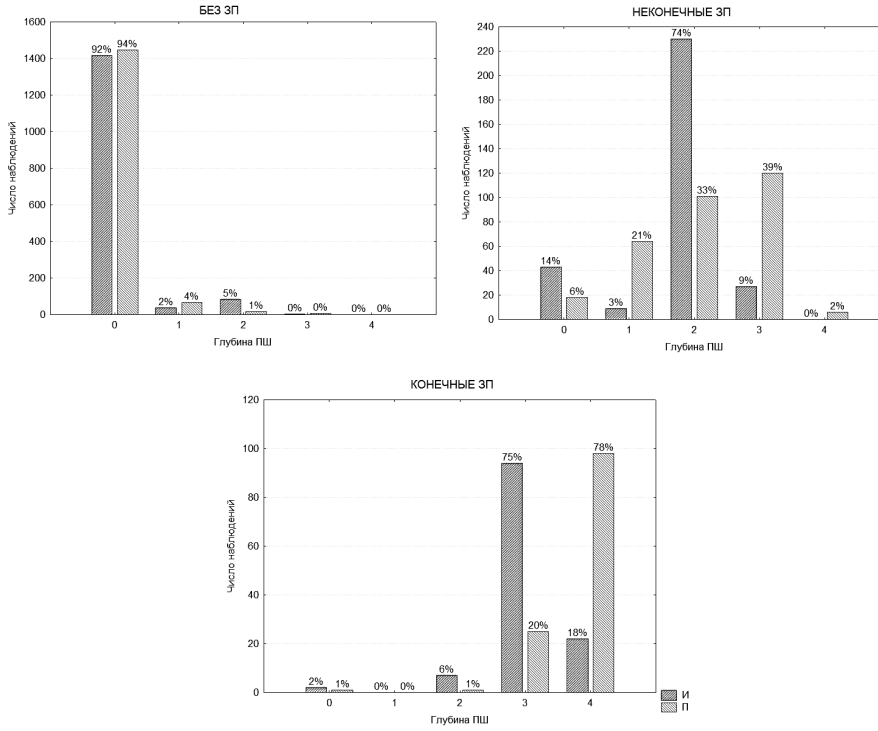


Рис. 5. Влияние ЗП на словоразделе на глубину ПШ. Слева гистограмма для словоразделов без ЗП; в центре — для словоразделов с неконечным ЗП (с явным преобладанием запятой); справа для словоразделов с конечным ЗП (с преобладанием точки)

3. Связь интроспективных и перцептивных показателей словоразделов с наличием паузы на словоразделе и ее длительностью

В предисловии к РИА Аванесов отмечает возможность разграничения четырех категорий пауз, которые характеризуются им как минимальная (факультативная); небольшая; более длительная, чем небольшая; более длительная, чем средняя. Никаких физических коррелятов указанных категорий при этом не дается. 4 категории пауз признаются многими исследователями, что имеет экспериментальное обоснование для разных языков [Кривнова 2015]. В [Кривнова, Смирнова 2018] была предпринята еще одна попытка формальной категоризации длительности пауз. Было показано, что переход от числовой шкалы к укрупненной номинальной дает классы с достаточно устойчивыми центроидами, близкими к наиболее вероятным значениям внутри классов («нет паузы». «минимальная» в среднем ~ 100 мс, «короткая» ~ 400, «средняя» ~ 700, «большая, длинная» ~ 1000, «максимальная» — порядка 2000 мс; паузы из максимального класса часто имеют особую техническую природу. Указанные данные были получены на материале достаточно выразительного чтения повествовательного текста в среднем темпе произнесения и являются округленными.

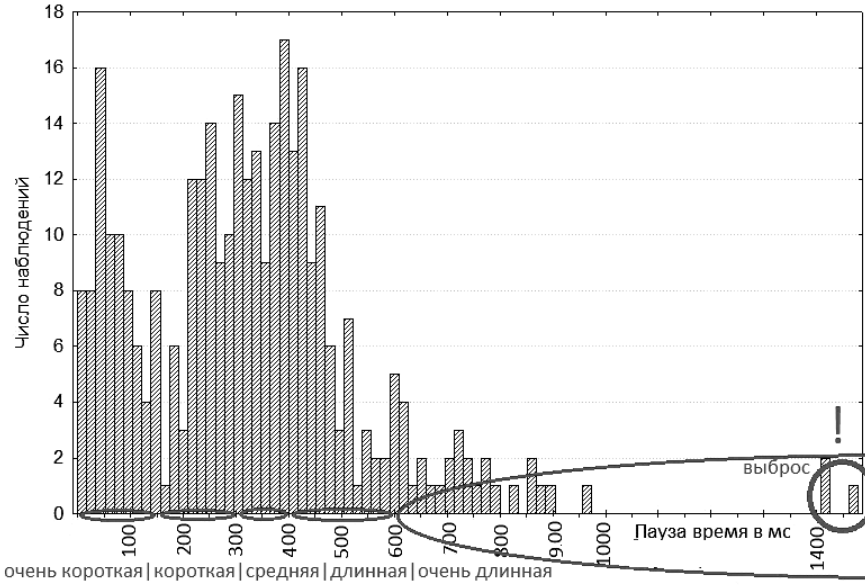


Рис. 6. Гистограмма частот длительности ненулевых пауз на словоразделах, отдельно помечены диапазоны категорированных пауз

В рамках настоящего исследования для анализа связи пауз с показателями глубины ПШ была выбрана категориальная шкала с 6-ю аналогичными диапазонами: «нет паузы», «очень короткая» («минимальная»), «короткая»,

«средняя», «длинная» и «очень длинная» («максимальная») паузы. Надо заметить, что переход к категориальной шкале позволяет в значительной степени не учитывать влияние темпа речи диктора на оценку длительности пауз на словоразделах. Формальная классификация была выполнена методом К-средних с переменным числом классов. В исходной разметке пауз имеется несколько выбросов, вызванных, вероятно, ошибками диктора, однако они не являются паузами неречевой природы и в действительности являются длинными паузами. Классификация производилась без их учета, после чего выбросы были присоединены к классу очень длинных пауз. На **рис. 6** видно, что полученные диапазоны и их «средние значения» довольно точно соответствуют «зубцам» полимодальной гистограммы частот длительности пауз.

Округляя приведенные данные, получим следующие диапазоны: «меньше 150 мс» со средним значением ~65 мс; «150–300» со средним ~240; «300–400» со средним ~350; «400–600» со средним значением 460 и «больше 600» со средним порядка 700 мс. Отличие абсолютных значений от данных в [Кривнова, Смирнова 2018] объясняется высоким темпом речи диктора, озвучивавшего РИА.

Таблица 2. Дескриптивные характеристики диапазонов длительности пауз

ПАУЗА	Дескриптивная статистика для категоризированных пауз (диапазонов)				
	Число наблюдений	Среднее	Медиана	Минимум	Максимум
Очень короткая	69	68,03	**64,00**	**7,00**	**141,00**
Короткая	69	230,54	**240,00**	**145,00**	**293,00**
Средняя	88	352,36	**350,50**	**297,00**	**408,00**
Длинная	71	466,11	**458,00**	**411,00**	**583,00**
Очень длинная	32	718,59	**710,00**	**596,00**	**975,00**
Оч. длинная с выбросом	35	780,74	**721,00**	**596,00**	**1485,00**

Анализ интересующей нас зависимости с помощью критериев типа «хи-квадрат» показывает высокую степень зависимости глубины ПШ и длительности паузы: отсутствию ПШ с практически 100% вероятностью соответствует отсутствие паузы, ПШ глубины **1** и **2** с большой вероятностью непаузальны, хотя на таких словоразделах могут быть и короткие паузы. ПШ большой глубины с большой вероятностью реализуются при помощи длинных и очень длинных пауз, а в середине диапазона возможна сильная вариативность пауз. Тем не менее, можно утверждать, что в этом случае наиболее вероятны средние и длинные паузы. **Таблицы 3** и **4** демонстрируют распределение (в %) ***категорированных*** пауз в зависимости от глубины ПШ, частоты, максимальные для данной категории, выделены жирным шрифтом.

Таблица 3

	Связь глубины ПШ в И-разметке (Аванесов) и категоризированной длительности паузы						
	И-разметка	Без паузы 0	Пауза оч. короткая 1	Пауза короткая 2	Пауза средняя 3	Пауза длинная 4	Пауза оч. длинная 5
%	0	98,77%	0,68%	0,14%	0,21%	0,07%	0,14%
%	1	97,78%	2,22%	0,00%	0,00%	0,00%	0,00%
%	2	56,76%	**15,62%**	**15,32%**	10,21%	2,10%	0,00%
%	3	4,91%	3,68%	9,82%	30,67%	36,81%	**14,11%**
%	4	0,00%	0,00%	0,00%	4,55%	**13,64%**	81,82%
Количество	Все группы	1686	69	69	88	71	43

Вероятность ошибки 1-го рода при отвержении гипотезы независимости <0,00001, коэф. квадратичной связи Крамера 0,535, ранговая корреляция Спирмена 0,735.

Таблица 4

	Связь глубины ПШ в П-разметке (аудиторы) и категоризированной длительности паузы						
	П-разметка	Без паузы 0	Пауза оч. короткая 1	Пауза короткая 2	Пауза средняя 3	Пауза длинная 4	Пауза оч. длинная 5
%	0	99,93%	0,07%	0,00%	0,00%	0,00%	0,00%
%	1	95,45%	4,55%	0,00%	0,00%	0,00%	0,00%
%	2	68,60%	**24,79%**	4,13%	1,65%	0,83%	0,00%
%	3	6,84%	**16,84%**	31,58%	32,11%	11,58%	1,05%
%	4	0,00%	0,00%	3,39%	**21,19%**	40,68%	34,75%
Количество	Все группы	1686	69	69	88	71	43

Здесь вероятность ошибки 1-го рода также меньше 0,00001, а коэффициент квадратичной связи Крамера — 0,585, ранговая корреляция Спирмена — 0,810, т. е. несколько больше полученных для И-разметки, что является следствием того, что аудиторы ориентировались на конкретное озвучивание текста.

4. Заключение

В докладе описаны результаты сравнительного анализа достаточно полярных по своей природе типов разметки просодического членения художественного текста на русском языке[7]. Различия между интроспективной и перцептивной разметками одного и того же текста, зафиксированные в результатах анализа, вполне объяснимы как следствие разных установок эксперта и диктора на выразительность озвучивания текста и разницу в фонетическом опыте всех участников исследования. В то же время анализ выявил и значительное сходство разметок, что свидетельствует о наличии общих когнитивно-фонетических принципов и факторов формирования просодического членения в звучащей речи. К ним могут быть отнесены: интегрирующие механизмы объединения слов в просодические составляющие фразового уровня, зависимость глубины просодических швов на словоразделе от веса и статуса контролирующих факторов просодического членения: частеречного признака слова перед словоразделом, силы синтаксической границы и знака препинания на словоразделе. По данным перцептивной разметки обнаружена также отчетливая связь глубины просодического шва с категориальным признаком паузы на словоразделе, что очень важно для адекватного восприятия и понимания устного текста. Что касается дальнейших исследований, то ввиду небольшого объема экспериментального материала некоторые результаты и выводы, полученные нами, требуют уточнения и увеличения их статистической надежности на более разнообразном текстовом материале с расширением числа участников исследования. Кроме того, в нашей работе значимые факторы, контролирующие силу словоразделов, рассматривались автономно, однако в реальной речи они сложным образом взаимодействуют, и это, скорее всего, находит отражение в просодических разметках и должно учитываться при их сравнительном анализе.

Литература

1. *Аванесов Р. И.* (1972) Русское литературное произношение. Просвещение, М.
2. *Гращенков П. В., Кириллова А. А., Смирнова О. С.* (2018) Влияние синтаксиса на просодию: данные одного эксперимента над русским письменным текстом // Компьютерная лингвистика и интеллектуальные технологии. Материалы ежегодной международной конференции «Диалог». М., РГГУ, 2018, вып. 17, сс. 219–231.
3. *Кривнова О. Ф.* (2015). Глубина просодических швов в звучащем тексте (экспериментальные данные) // Компьютерная лингвистика и интеллектуальные технологии. Материалы ежегодной международной конференции «Диалог». М., РГГУ, 2015, вып. 14, т. 1., сс. 326–338.

[7] Вообще говоря, в триаде «интроспекция-озвучивание-восприятие» много переменных, которые могут быть предметом сравнительного анализа. В настоящей работе мы рассмотрели лишь один из вариантов, который показался нам наиболее интересным и доступным для исследования.

4. *Кривнова О. Ф., Смирнова О. С.* (2018) База дискурсивных признаков словораздела в устной русской речи: структура, состав и опыт применения // Компьютерная лингвистика и интеллектуальные технологии. Материалы ежегодной международной конференции «Диалог». М., РГГУ, 2018, вып. 17, сс. 368–379.
5. *НКРЯ* (2005) Национальный корпус русского языка 2003–2005: Результаты и перспективы. М.
6. *Смирнова* (2017) Статистический анализ результатов перцептивного оценивания глубины просодических швов в русском звучащем тексте // Компьютерная лингвистика и интеллектуальные технологии. Материалы ежегодной международной конференции «Диалог». М., РГГУ, 2017, электронная публикация.

References

1. *Avanesov R. I.* (1972) Russian Literary Pronunciation [Russkoe literaturnoe proiznoshenie], Education, M.
2. *Dutoit T.* (1997) An Introduction to Text-to-Speech Synthesis. Springer, 285 p.
3. *Grashchenkov P. V., Kirillova A. A., Smirnova O. S.* (2018) The influens of syntax on the prosody: the experimental data from a study jf one russian text [Vliyanie sintaksisa na prosodiyu: dannye odnogo ehksperimenta nad russkim pis'mennym tekstom] Computer linguistics and intellectual technologies. Proceedings of the annual international conference "Dialogue" [Komp'juternaja lingvistika I intellektual'nyje tehnnologii. Materialy jezhegodnoj mezhdunarodnoj konferentsii 'Dialog'] M., RGGU, v. 17, pp. 219–231.
4. *Krivnova O. F.* (2015) The depth of prosodic breaks in spoken text (experimental data) [Glubina prosodicheskih shvov v zvuchaschem tekste (eksperimental'nyje dannyje)] Computer linguistics and intellectual technologies. Proceedings of the annual international conference "Dialogue" [Komp'juternaja lingvistika I intellektual'nyje tehnnologii. Materialy jezhegodnoj mezhdunarodnoj konferentsii 'Dialog'] M., RGGU, v. 14, t. 1, pp. 326–338.
5. *Krivnova O. F., Smirnova O. S.* (2018) A database of wordbreaks discursive features in russian oral speech: the structure, composition and application [Baza diskursivnyh priznakov slovorazdela v ustnoj russkoj rechi: struktura, sostav i opyt primenenij] Computer linguistics and intellectual technologies. Proceedings of the annual international conference "Dialogue" [Komp'juternaja lingvistika I intellektual'nyje tehnologii. Materialy jezhegodnoj mezhdunarodnoj konferentsii 'Dialog'] M., RGGU, v. 1, pp. 368–379.
6. *NCRL* (2005) Russian National Corpus 2003–2005: results and prospects. M., 2005. Russkij Natsional'nyj Korpus 2003–2005: rezul'taty I perspektivy] M., 2005.
7. *Smirnova O. S.* (2017) Statistical analysis of the results of prosodic breaks' perceptual evaluation in spoken Russian text [Statisticheskij analiz rezul'tatov pertseptivnogo otsenivanija glubiny prosodicheskih shvov v russkom zvuchaschem tekste]// Computer linguistics and intellectual technologies. Proceedings of the annual international conference "Dialogue" [Komp'juternaja lingvistika i intellektual'nyje tehnikigii. Materialy jezhegodnoj mezhdunarodnoj konferentsii 'Dialog'] M., RGGU, 2017. Electronic publication.

ADAPTATION OF DEEP BIDIRECTIONAL MULTILINGUAL TRANSFORMERS FOR RUSSIAN LANGUAGE

Kuratov Yu., **Arkhipov M.**

Neural Networks and Deep Learning Lab, Moscow Institute of Physics and Technology

Keywords: language models; transfer learning; low-resource languages

АДАПТАЦИЯ ГЛУБОКИХ ДВУНАПРАВЛЕННЫХ МНОГОЯЗЫЧНЫХ МОДЕЛЕЙ НА ОСНОВЕ АРХИТЕКТУРЫ TRANSFORMER ДЛЯ РУССКОГО ЯЗЫКА

Куратов Ю., **Архипов М.**

Лаборатория нейронных систем и глубокого обучения, Московский физико-технический институт (национальный исследовательский университет)

1. Introduction

A large amount of work is devoted to unsupervised pre-training of neural networks on a variety of Natural Language Processing (NLP) tasks. Unsupervised pre-training shows significant improvements in almost every NLP task [3], [4], [8], [10].

At the moment one of the best performing models for unsupervised pre-training is BERT [4]. This model is based on Transformer [16] architecture and trained on a large number of unlabeled texts from Wikipedia to solve Masked Language Modelling task. It shows state-of-the-art results on a wide range of NLP tasks in English. Currently, there are publicly available two monolingual English and Chinese models and a single multilingual model. It is known that monolingual models performance is significantly better than multilingual ones.

In the present work, we consider the possibility of multilingual to monolingual transfer. We use Russian as a target language for transfer. We show that it is possible to train the monolingual model using multilingual initialization. To show this, we evaluated the multilingual model on a number of common NLP tasks from the target

language. The model trained in a monolingual setting achieves substantially better performance compared to the multilingual model.

2. Model Architecture

In the present work, we use BERT [4] model for all our experiments. The model is a Transformer [16] encoder. The basic building blocks of the model is Self-Attention. The model was trained on the Masked Language Modelling and next sentence prediction tasks. We refer readers to check BERT original paper for details about the model [4].

We used 12 layers (Transformer blocks) version of BERT with self-attention hidden size 768, feed-forward hidden size 3,072, and 12 self-attention heads. This setting corresponds to $\text{BERT}_{\text{BASE}}$ model from [4]. Task-specific layers were trained according to the BERT paper.

3. Language transfer

In this work, we consider the transfer of multilingual BERT model to monolingual. Authors of [4] showed that monolingual models show superior performance compared to multilingual one. Furthermore, the BERT model uses the subword segmentation algorithm[14] to cope with large vocabulary problem. Multilingual models use only a small part of the entire vocabulary for a single language. It results in much longer sequences after tokenization compared to the monolingual model. Since the Transformer model has quadratic computational complexity in terms of input sequence length, it is highly undesirable.

We investigated the possibility of using the multilingual model as initialization for the monolingual model. The basic idea is to use knowledge about target language that already captured during multilingual training. It also is known that training model using data from multiple languages can significantly improve the performance of the model [9]. We used the multilingual model from BERT repository[1]. This model was trained on one hundred languages with largest Wikipedias. The target language is Russian. All parameters of the model except word embeddings were initialized from the multilingual model [9].

The new subword vocabulary was obtained using subword-nmt[2]. Training of the subword vocabulary was performed on the Russian part of Wikipedia and news data. The part of Wikipedia data was around 80%. The result of this step is a new monolingual Russian subword vocabulary. This vocabulary contains longer Russian words and subwords compared to multilingual one.

New word embeddings matrix was obtained by assembling monolingual embeddings from multilingual. Namely, embeddings of all tokens from the intersection of multilingual and monolingual vocabulary were left without any changes. The same for special tokens like [UNK] or [CLS]. We replaced all tokens from outside the intersection with tokens from the monolingual vocabulary. These tokens are mostly longer subword units which are combinations of shorter units present in the intersection. New tokens are initialized with the mean value of embeddings from the intersection. For

[1] https://github.com/google-research/bert

[2] https://github.com/rsennrich/subword-nmt

example, there are tokens 'bi' and '##rd' in the intersection of vocabularies, where '##' stands for the continuation of the word. There is also a token 'bird' present in the monolingual vocabulary and absent in the multilingual vocabulary. The embedding of 'bird' is initialized as the mean value of the embeddings 'bi' and '##rd'.

The model with reassembled vocabulary and embeddings matrix was trained on the same data that was used for building of the monolingual vocabulary. The following hyperparameters were used for training:

- batch size: 256
- learning rate: $2 \cdot 10^{-5}$
- optimizer: Adam
- L2 regularization: 10^{-2}

The monolingual Russian model[3] is available as a part of the DeepPavlov library[4].

4. Tasks description

We have chosen three tasks to evaluate our approach: paraphrase identification, sentiment analysis, and question answering. We briefly describe them in this section.

4.1. Paraphrase Identification with ParaPhraser

ParaPhraser [11] is a dataset for paraphrase detection in Russian language. Two sentences are paraphrases if they have the same meaning. This dataset consists of 7,227/1,924 train/test pairs of sentences which are labeled as precise paraphrases, near paraphrases or non-paraphrases. One approach for paraphrase identification is a binary classification: first class is precise and near paraphrases, second class—non-paraphrases.

4.2. Sentiment Analysis with RuSentiment

RuSentiment [13] is a dataset for sentiment analysis of posts from VKontakte (VK), the most popular social network in Russia. Realised in 2018, it became one of the largest sentiment datasets for Russian language with 30,521 posts. Each post is labeled with one of the five classes. The informal language presented in RuSentiment dataset makes it more challenging for our model, trained on Wikipedia and news articles.

4.3. Question answering with SDSJ Task B

As part of Sberbank Data Science Journey[5] 2017 was held a competition with two tasks. Task B was inspired by Stanford Question Answering Dataset (SQuAD) [12]. Organizers collected about 50,000 (train and development set) questions and contexts,

[3] http://files.deeppavlov.ai/deeppavlov_data/bert/rubert_cased_L-12_H-768_A-12_v1.tar.gz

[4] https://github.com/deepmipt/deeppavlov/

[5] https://sdsj.sberbank.ai/

where the answer is always a span from corresponding context. SQuAD dataset encouraged community to develop sophisticated and effective neural architectures such as BiDAF [15], R-NET [17], Mnemonic Reader [5]. All this models are based on attention mechanisms [1], [7] and building joint context-question representation.

5. Results

We evaluated BERT multilingual model and BERT trained with our approach (RuBERT) on three tasks: paraphrase identification, sentiment analysis, and question answering. All reported results were obtained by averaging across 5 runs.

Table 1: ParaPhraser. We compare BERT based models with models in non-standard run setting, when all resources were allowed

model	F-1	Accuracy
Neural networks [11]	79.82	76.65
Classifier + linguistic features [11]	81.10	77.39
Machine Translation + Semantic similarity [6]	78.51	81.41
BERT multilingual	85.48 ± 0.19	81.66 ± 0.38
RuBERT	**87.73** ± 0.26	**84.99** ± 0.35

Table 2: RuSentiment. We used only randomly selected posts (21,268) subset for training

model	F-1	Precision	Recall
Logistic Regression [13]	68.84	69.53	69.46
Linear SVC [13]	68.56	69.46	69.25
Gradient Boosting [13]	68.48	69.63	69.19
NN classifier [13]	71.64	71.99	72.15
BERT multilingual	70.82 ± 0.75	—	—
RuBERT	**72.63** ± 0.55	—	—

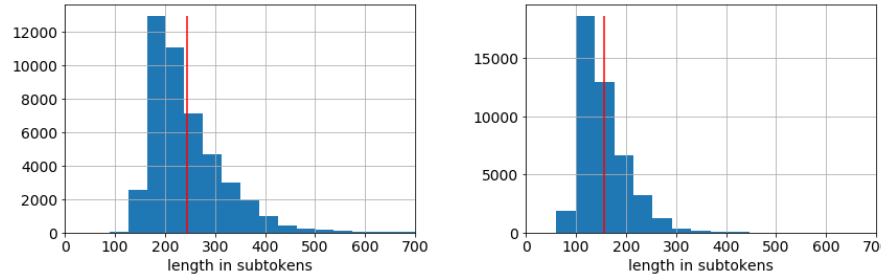
Table 3: Results on question answering with SDSJ Task B. Models performance was evaluated on development set (public leaderboard subset)

model	F-1 (dev)	EM (dev)
R-Net from DeepPavlov [2]	80.04	60.62
BERT multilingual	83.39 ± 0.08	64.35 ± 0.39
RuBERT	**84.60** ± 0.11	**66.30** ± 0.24

SDSJ Task B and ParaPhraser datasets share the same domain with data, which we used for training RuBERT. The RuSentiment dataset is based on posts from a social network and shows to be more challenging for RuBERT. As result, we can see only 1 F-1 point improvement from previous state of the art for RuSentiment in Table 2, comparing to 4-6 F-1 points improvement on SDSJ Task B and ParaPhraser (results in Table 3 and Table 4).

5.1. Vocabulary comparison

BERT multilingual and RuBERT have the same size of vocabulary (about 120k subtokens), but RuBERT vocabulary was built especially for Russian language. Figure 1 shows that RuBERT model allows to reduce mean sequence length in 1.6 times in subtokens, what makes possible to increase batch size or feed longer texts to the model, comparing to BERT multilingual.



(a) BERT multilingual (b) RuBERT

Figure 1: Distribution of lengths in subtokens of contexts with their questions (SDSJ Task B dataset). Red vertical lines represent mean values.

5.2. Training dynamics

In this section we compare training BERT model for Russian language from scratch (random initialization) and initialized with BERT multilingual. Figure 2 shows that BERT multilingual initialization helps to converge faster: about 800 thousand steps is required for random initialized model to get the same loss as at 250 thousand step of multilingual initialization. It takes about two days to train for 250 thousand steps (on Tesla P100 x 8), so it helped us to save six days of computational time. Proposed averaging of new subtokens in vocabulary also has positive effect on the rate of convergence (instead of averaging we could take random initialization for new subtokens).

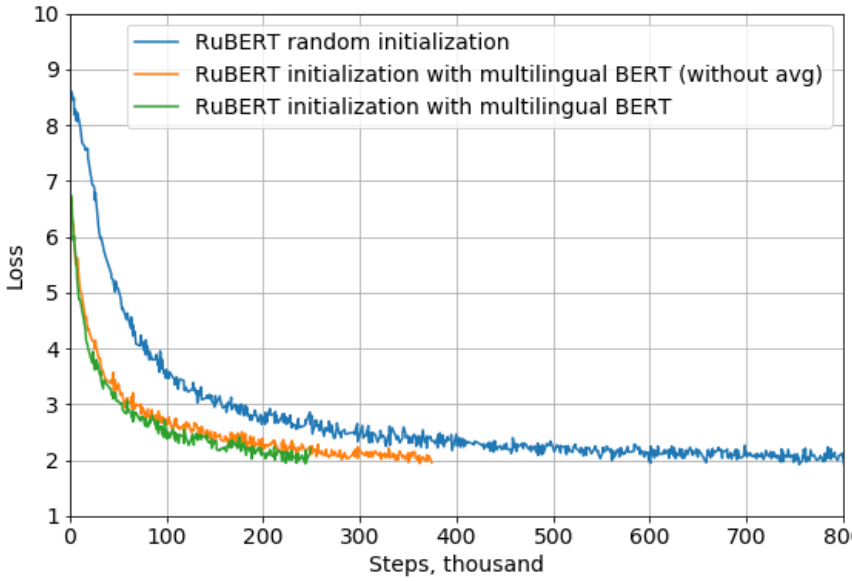


Figure 2: Models training dynamics to get to the same value of loss

6. Conclusion

In this work, we have shown that Transformer network pre-trained on the multilingual Masked Language Modelling task significantly improves performance on a number of Russian NLP tasks compared to existing solutions. Furthermore, language-specific unsupervised training with multilingual initialization results in even better improvements. Pre-trained models for the Russian language are open sourced, as well as code to reproduce our results as part of DeepPavlov library.

7. Acknowledgments

This work was supported by National Technology Initiative and PAO Sberbank project ID 0000000007417F630002.

References

1. *Bahdanau, D. et al.:* Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473. (2014).
2. *Burtsev, M. et al.:* DeepPavlov: Open-source library for dialogue systems. Proceedings of ACL 2018, System Demonstrations. 122–127 (2018).
3. *Dai, A. M., Le, Q. V.:* Semi-supervised sequence learning. In: Advances in neural information processing systems. pp. 3079–3087 (2015).
4. *Devlin, J. et al.:* Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. (2018).
5. *Hu, M. et al.:* Reinforced mnemonic reader for machine reading comprehension. arXiv preprint arXiv:1705.02798. (2017).
6. *Kravchenko, D.:* Paraphrase detection using machine translation and textual similarity algorithms. In: Conference on artificial intelligence and natural language. pp. 277–292 Springer (2017).
7. *Luong, M.-T. et al.:* Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025. (2015).
8. *Mikolov, T. et al.:* Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. (2013).
9. *Mulcaire, P. et al.:* Polyglot semantic role labeling. arXiv preprint arXiv:1805.11598. (2018).
10. *Peters, M. E. et al.:* Deep contextualized word representations. In: Proc. Of naacl. (2018).
11. *Pivovarova, L. et al.:* ParaPhraser: Russian paraphrase corpus and shared task. In: Conference on artificial intelligence and natural language. pp. 211–225 Springer (2017).
12. *Rajpurkar, P. et al.:* Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250. (2016).
13. *Rogers, A. et al.:* RuSentiment: An enriched sentiment analysis dataset for social media in russian. In: Proceedings of the 27th international conference on computational linguistics. pp. 755–763 (2018).
14. *Sennrich, R. et al.:* Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909. (2015).
15. *Seo, M. et al.:* Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603. (2016).
16. *Vaswani, A. et al.:* Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017).
17. *Wang, W. et al.:* Gated self-matching networks for reading comprehension and question answering. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers). pp. 189–198 (2017).

КОНЦЕПТУАЛИЗАЦИЯ НЕ ПОЛНОСТЬЮ КОНТРОЛИРУЕМЫХ СИТУАЦИЙ: ГЛАГОЛЫ И МЕСТОИМЕНИЯ[1]

Кустова Г. И. (galinak03@gmail.com)

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

В работе вводится противопоставление «уровень ситуации» vs. «уровень сценария». В рамках этого противопоставления рассматриваются признаки глаголов неполного контроля *удалось*, *получилось 1*, *вышло 1* (*встретиться*) и *произошло*, *получилось 2*, *вышло 2, случилось (так, что он опоздал)*: управление (инфинитив vs. клауза), взаимодействие с отрицанием и пропозициональными местоимениями. Пропозициональные местоимения *так* и *это* и матричные глаголы, с которыми они сочетаются, предполагают разную концептуализацию ситуации-антецедента: ***Так*** *получилось, что мы проиграли* vs. *Мы хотели победить, и нам* ***это*** *удалось*. *Так* семантически связано с образом действия и в производных значениях предполагает вариативный фактор или аспект.

Ключевые слова: пропозициональное местоимение, контроль, контролируемая ситуация, концептуализация

CONCEPTUALIZATION OF NON-FULLY CONTROLLED SITUATIONS: VERBS AND PRONOUNS

Kustova G. I. (galinak03@gmail.com)

Vinogradov Russian Language Institute of the Russian Academy of Sciences; Moscow, Russia

The paper introduces the opposition “level of the situation” vs. “level of the story”. Within this opposition, features of the verbs denoting non-fully controlled situations are considered (*to succeed* vs. *to happen*): government

[1] Работа выполнена при поддержке РФФИ, проект № 17-29-09154 офи_м «Динамика языковой системы: корпусное исследование синхронной вариативности и диахронических изменений в текстах разных типов». Примеры извлечены из Национального корпуса русского языка, www.ruscorpora.ru.

(infinitive vs. clause), combinability with negation and propositional pronouns. Propositional pronouns *tak* ('so') and *eto* ('it') and the matrix verbs which they are combined with, imply a different conceptualization of the antecedent situation: *My proigrali. Tak poluchilos'* ('We lost. So it turned out') vs. *My hoteli pobedit', i nam eto udalos'* ('We wanted to win, and we succeeded'). *Tak* is semantically related to the mode of action and in other meanings implies a variable factor or aspect.

Key words: propositional pronoun, control, controlled situation, conceptualization

1. Параметры ситуаций и матричные глаголы

Существует определенный набор параметров, которые характеризуют ситуации (реальность-нереальность, многократность, возможность, обычность и под.). Эти параметры (по крайней мере, часть из них) входят в универсальный набор смыслов (ср. [Мельчук 1999], [Плунгян 2011]), которые хотя бы в некоторых языках грамматикализованы. Параметры ситуаций могут выражаться грамматическими показателями (которые присоединяются независимо от лексического значения — ср. наклонение), словообразовательными морфемами (ср. способы действия, которые маркируют интенсивность (*иссохнуть*), повторяемость (*захаживать*), взаимность (*перешептываться*) и т. п.), специализированными глаголами. Эти глаголы являются предикатами второго порядка — они подчиняют пропозициональный (сентенциальный) актант в виде инфинитива или клаузы. Для краткости будем называть их матричными глаголами (МГ). Если «обычных» глаголов тысячи (с учетом многозначности — десятки тысяч лексем), то МГ относительно немного. Они являются промежуточным звеном между обозначениями «обычных» ситуаций — действий, процессов, состояний — и грамматическими или словообразовательными показателями и занимают важное место в системе языка. Будучи в той или иной степени грамматикализованными единицами, МГ имеют достаточно широкую сочетаемость (хотя и не такую широкую, как грамматические показатели) и обозначают основные параметры концептуализации ситуаций, релевантные для данного языка.

Мы сейчас отвлекаемся от матричных глаголов, выражающих отношение человека как субъекта сознания, эмоции, речи к ситуации (реальной или воображаемой) или пропозиции (*X думает / знает / рад / утверждает, что P*). Нас будут интересовать глаголы, которые описывают, так сказать, объективные параметры реализации ситуаций. Очевидными параметрами являются, например, фаза и модальность, — обслуживающие их классы фазовых и модальных глаголов традиционно выделяются в лингвистических описаниях.

Однако бóльшая часть матричных глаголов не имеет общепринятого разбиения на классы и тем более общепринятой терминологии для обозначения этих классов.

2. Концептуализация ситуации как не полностью контролируемой

Задача данной работы — рассмотреть формальные и семантические свойства группы матричных глаголов, которые выражают представление о наличии **случайных**, неконтролируемых **факторов**, которые могут вмешиваться в «**нормальный ход событий**» и так или иначе нарушать его — мешать реализации запланированной ситуации, ср.: *Вчера не получилось встретиться*; приводить к возникновению незапланированных (в том числе нежелательных) ситуаций, ср.: *Так вышло, что мы оба забыли ключи*.

Представление о нормальном ходе событий и его нарушении — один из важных параметров описания внешнего мира. Оно, например, играет ключевую роль в описании противительной и уступительной семантики и соответствующих союзов (ср. *Он целый день бродил по лесу, но не нашел ни одного гриба* — о союзе *но* см., например, [Санников 1989], о семантике уступки — [Апресян 2015]).

Наличие случайных факторов, которые помешали или могли помешать реализации ситуации и — тем самым — нормальному ходу событий выражается безличными глаголами (*не*) *удалось*, (*не*) *получилось*, (*не*) *сложилось*, (*не*) *вышло* [*встретиться еще раз*] (семантические особенности этих глаголов описаны в [Зализняк, Левонтина 2006]). Эти глаголы не имеют общепринятого названия; мы будем называть их (в силу наличия в их семантике идеи случайных факторов) глаголами неполного контроля.

Эти глаголы относятся к классу импликативов, который выделяется по формально-семантическим основаниям — импликативному статусу подчиненной предикации (ср. [Karttunen 1973], [Зализняк 1988], [Разлогова 1988]).

ЗАМЕЧАНИЕ. Класс импликативов довольно пестрый и семантически неоднородный — в него входят, например, глаголы *постеснялся*, *осмелился*, *удосужился*, *соизволил* (+ Инф.), которые не связаны с нарушением контроля. С другой стороны, к глаголам неполного контроля примыкают (тоже импликативные) (*не*) *смог* (СВ! — в отличие от неимпликативного НСВ (*не*) *может*) и (*не*) *сумел*. С другой стороны, есть глаголы, не имеющие (в отличие от глаголов типа *удалось*, *получилось*) пресуппозиции ‘собирался, пытался, планировал’, т. е. описывающие стечение совершенно не контролируемых факторов и обстоятельств — неблагоприятных (*угораздило*) или благоприятных (*посчастливилось*, *повезло*). В силу ограниченности объема статьи мы не можем подробно обсуждать состав класса глаголов неполного контроля, поэтому для описания интересующих нас свойств выбрали наиболее репрезентативные глаголы.

Традиционно понятие контролируемости-неконтролируемости используется, чтобы описать семантическое различие между контролируемыми ситуациями типа *идти в магазин*, *написать письмо, купить билет*, реализация которых зависит от усилий человека, а результат гарантирован (при отсутствии помех!), и неконтролируемыми ситуациями типа *споткнуться, заболеть, упасть, забыть* (к группе неконтролируемых ситуаций по умолчанию примыкают стихийные

события типа *взошло солнце, идет дождь, листья пожелтели, раздался стук* и под.), см. [Кустова 1992]. В этом отношении контролируемость-неконтролируемость можно считать синонимом агентивности-неагентивности.

Однако на контролируемую ситуацию могут влиять случайные факторы, понижая ее статус до не полностью контролируемой.

Но есть более крупный масштаб (более высокий уровень) приложения понятия «контроль» — это уровень не отдельной ситуации, а сценария. Человек, как правило, мыслит не отдельными ситуациями, а, так сказать, нарративами. Ситуация — особенно ситуация с участием человека — обычно встроена в историю, сценарий, цепочку событий (она чем-то вызвана, с чем-то связана и к чему-то приводит). Точно так же, как человек планирует и реализует единичные ситуации, он планирует и реализует сценарии (или «вписывается» в чужие сценарии) — и в этом смысле к сценарию тоже приложимо понятие контроля. В идеале, при отсутствии помех, сценарий состоит из запланированных или ожидаемых ситуаций (не важно, контролирует их сам субъект, ср. *обед*, или другое лицо, ср. *арест*). Но в сценарий, как и в отдельную ситуацию, могут вмешиваться случайные факторы, и тогда в нем появляются «лишние», не запланированные, не ожидавшиеся ситуации или, наоборот, не реализуются запланированные (реализуются не так, как планировалось).

Глаголы, фиксирующие неполный контроль на уровне сценария, — [*так*] *получилось / вышло / сложилось / случилось / произошло, что P* — концептуализуют ситуацию P как такую, которая не предусмотрена сценарием, незаконно «вклинивается» в него, нарушая нормальный / ожидаемый ход событий.

Таким образом, данный параметр концептуализации ситуаций (масштаб, или уровень, рассмотрения) как бы накладывается на другие параметры.

Подчеркнем, что речь идет о различиях в концептуализации. «Физически» ситуация может быть одна и та же. Просто в масштабе ситуации говорящий сообщает, что в реализацию ситуацию могли вмешаться или вмешались случайные факторы, ср. *Вчера получилось / не получилось пообедать у друзей*, а во втором — что сама ситуация является случайной в цепочке событий, в сценарии, ср. *Вчера так получилось, что я обедал у друзей* (*поэтому я не смог пообедать с тобой*), в силу чего этот сценарий часто, хотя и не всегда, оказывается неоптимальным или даже нежелательным: *Мы не смогли улететь одним рейсом. Так получилось*.

Весьма показательно, что оппозиция «масштаб ситуации» — «масштаб сценария» релевантна не только для глаголов неполного контроля. Оно обнаруживается и у других классов матричных глаголов, например фазовых и модальных:

- фазовые: *Он начал готовиться к экзамену* — уровень ситуации vs. *Полоса неудач началась с того, что он опоздал на экзамен* — уровень сценария;
- модальные: *Он может поднять мешок с песком* (‘в состоянии поднять’) — модальная характеристика ситуации vs. *Он может прийти* — модальный статус ситуации (возможность) в сценарии.

ЗАМЕЧАНИЕ. В данной паре в обеих конструкциях *может* употребляется с инфинитивом, однако уровень сценария допускает и придаточное: *Может быть* [*может случиться*], *что он придет*.

Группы глаголов неполного контроля ситуации и неполного контроля сценария частично пересекаются: в обе группы входят — в разных значениях — глаголы *получиться*, *выйти*, *сложиться*. Это не удивительно, поскольку эти глаголы включают в свою семантику идею наличия случайных факторов. Условно назовем эти группы *получилось-удалось* (уровень ситуации) и *получилось-произошло* (уровень сценария). Противопоставление «ситуация vs. сценарий» отражается и в формальных различиях глаголов этих групп.

3. Сходства и различия групп *получилось-удалось* и *получилось-произошло*

3.1. Контролируемость

На уровне ситуации неполный контроль применим только к исходно контролируемым (агентивным) ситуациям: *На этой неделе так и не получилось пообедать вместе*. На уровне сценария «лишняя» ситуация P может быть контролируемой — самим субъектом (*Как-то раз в Ленинград приехал Морис Дрюон с женой. Получилось, что два дня я была их гидом* [Сати Спивакова. Не всё (2002)] (P не планировалось) или другим лицом (*И вот получилось, что меня арестовали точно в тот день, когда я, идя на очередной допрос, окончательно уверилась, что не так страшен черт* [Владимир Шаров. Воскрешение Лазаря (1997–2002)] (P не ожидалось); может быть неконтролируемой: *Нагнули, понимаешь, сосну. Пристегнули зэка к верхушке монтажным ремнём — и отпустили. А зэк в полёте расстегнулся — и с концами. Улетел чуть не за переезд. Однако малость не рассчитал. Надеялся в снег приземлиться у лесобиржи. А получилось, что угодил во двор райвоенкомата*… [Сергей Довлатов. Наши (1983)].

3.2. Синтаксическая конструкция: инфинитив vs. клауза

Глаголы уровня ситуации подчиняют инфинитив, т. е. для них характерна инфинитивная конструкция: ***У Рязанова получилось снять*** *городскую сказку. Хорошую*… [Форум: рецензии на фильм «Служебный роман» (2006–2010)]; клаузу они не подчиняют, ср.: **У нас получилось, что мы взяли билеты в один вагон* (в значении ‘нам удалось’). Глаголы уровня сценария в норме подчиняют клаузу, т. е. участвуют в пропозициональной конструкции: ***Может получиться, что*** *через пять-семь лет муниципалитеты останутся вообще без учителей физики* [«Новгородские ведомости», 2013]. Однако это не правило, а тенденция (в группе *получилось-произошло* инфинитив возможен — с глаголом *случиться*: *Однажды мне* ***случилось участвовать*** *в переговорах с иностранцами — **Так случилось, что** я участвовал в переговорах с иностранцами*).

3.3. Отрицание

На уровне ситуации возможны как утвердительные, так и отрицательные конструкции. Утвердительные конструкции обозначают успех — случайные факторы были, но не помешали: *Наконец удалось / получилось встретиться*. Отрицательные конструкции обозначают неудачу — случайные факторы были и помешали: *Вчера не удалось / не получилось / не вышло встретиться*.

У пропозициональной конструкции нет естественного отрицания в режиме сообщения: *Получилось так, что никого не оказалось на месте* vs. **Не получилось так, что никого не оказалось на месте*. В косвенных модальностях, в том числе в сфере действия вопроса, отрицание возможно, но оно фиктивно и устранимо: *Не получится ли так, что никого не будет на месте?*; *Как бы не получилось / не получилось бы так, что никого не будет на месте*; ср. перифразы без отрицания: *А вдруг получится / а если получится так, что никого не будет на месте?*

Отрицать можно то, что ожидалось, но не произошло, ср.: *Петя пришел? — Не пришел*; *Пришел Петя? — Не Петя*. Высказывания типа *ПЕТЯ пришел!* (с ударением на первом слове), которые целиком составляют новое сообщение (рему), — не имеют естественного отрицания (см. [Падучева 1985], [Янко 2001]). По этой же причине не имеют естественного отрицания конструкции с глаголами группы *получилось-произошло*: ситуация P случайна, и ее содержание не может быть известно заранее, «на ее месте» могло произойти множество других ситуаций. Напротив, в группе *получилось-удалось* ситуация P запланированная, т. е. ее содержание заранее известно.

3.4. Субъект

В группе *получилось-удалось* субъект выражается дативом или группой *у* + Род. при матричном глаголе и кореферентен субъекту инфинитива: *Ему удалось / у него получилось* + Инф. Глаголы группы *получилось-произошло* не имеют синтаксически выразимого субъекта: *Сам режиссёр и автор идеи в интервью говорили об интересе к современному образованию, которое все критикуют. А* ***получилось, что*** *они подняли межнациональные, этические и политические пласты проблем: тему расизма и нелегальной миграции* [Форум: Класс — Франция (2008–2011)]. Субъектом сознания (в смысле [Падучева 1991]) при этом является говорящий, который не обязательно совпадает с субъектом ситуации P, — но даже если совпадает (т. е. если говорящий рассказывает свою собственную историю), он все равно смотрит на ситуацию P со стороны, т. е. выполняет две роли — субъекта и наблюдателя.

Одним из важных показателей различия между масштабом ситуации и масштабом сценария является сочетаемость данных групп глаголов с пропозициональными местоимениями *это* и *так*. Этому различию мы посвятим специальный раздел. Но сначала коротко остановимся на свойствах *это* и *так*.

4. Пропозициональные *так* и *это*

О пропозициональных местоимениях (далее — ПМ), т. е. местоимениях, замещающих пропозиции, написано существенно меньше, чем о местоимениях с предметными референтами (см., в частности, разделы в [Падучева 1985], [Тестелец 2001] и работы [Летучий 2011], [Пекелис 2018], ср. также [Meijer 2018], [Schwabe 2016].

В работе [Летучий 2011] местоимения *это*, *так*, *такое*, *оно* считаются пропозициональными актантами. [Пекелис 2018] усматривает у *так*, по сравнению с *это*, следующую особенность: *так* отсылает к актанту, но имеет свойства сирконстанта, обстоятельства. Подтверждением обстоятельственных свойств *так* считается, например, то, что оно совместимо с *это* в одном предложении: *Так это происходит / Так это бывает, когда…* Надо заметить, однако, что такая конструкция возможна даже не для всех бытийных глаголов, ср.: **Так это получилось, когда к нему пришли с проверкой*. Тем более она невозможна для глаголов других семантических классов, например глаголов речи (**Он так это говорил*; **Так это говорят*) или глаголов мнения (**Он так это считает*; **Так это считается*).

Мы будем рассматривать *так* и *это* (поскольку *такое* и *оно* являются более маргинальными, редкими), причем наибольший интерес для нас представляет *так* — более семантически нагруженная единица, чем *это*.

Это и *так* несимметричны и неравноценны. *Это* почти универсально, сочетается с большинством глаголов, которые вообще допускают ПМ, тогда как для употребления *так* нужны специальные семантические и синтаксические условия: *Почему ты это терпишь?* vs. **Почему ты так терпишь?*; *Этого я и боялся* vs. **Так я и боялся*; *Я не это имел в виду* vs. **Я не так имел в виду*; *Я этому рад* (*Я так рад* означает степень); *Я этого не одобряю* vs. **Я так не одобряю*; *Он этому способствовал* vs. **Он так способствовал* и т. д.

При некоторых глаголах — например, бытийных — возможны и *так,* и *это*, ср.: *В России* ***это бывает, что*** *медведь возьмет арфу, да на арфе и сыграет отлично* [П. В. Анненков. Письма И. С. Тургеневу (1852–1874)] vs. *В футболе* ***так бывает, что*** *даже у самых великих тренеров случаются неудачи* [Известия, 2012.10.16].

С некоторыми глаголами возможно только *так*, например с глаголами мнения: *Я так считаю* (*?это считаю*) */ думаю / полагаю*, — в отличие от глаголов знания (*Они уехали. Я это понял сразу / Я этого не знал — *Я так не знал*).

Так действительно имеет обстоятельственную семантику, как справедливо отмечается в [Пекелис 2018]. Это, вообще говоря, естественно, если учитывать его происхождение от наречия образа действия. Однако вопрос о семантических условиях употребления *так* не ставился. Ниже мы попытаемся предложить объяснение специфики поведения *так*.

5. *Так* и *это* с глаголами неполного контроля

Итак, одним из важных различий глаголов неполного контроля ситуации и неполного контроля сценария является то, что они по-разному сочетаются с ПМ *так* и *это*.

Группа *получилось-удалось* сочетается с *это*: *Мы хотели взять билеты в один вагон, и нам* ***это*** *удалось = нам удалось* ***взять*** *билеты в один вагон*; *Она пыталась вообще не спать: увы, у нее это не получилось* [«Домовой», 2002.08.04] *= у нее не получилось вообще не спать*. Существенно, что *это* находится в постпозиции к Р и в другом предложении.

Одновременное употребление *это* и инфинитива при матричном глаголе в обычных условиях невозможно: **У него* ***это*** *получилось* ***показать*** *первый результат на соревнованиях*. *Это* при матричном глаголе при наличии инфинитива в том же предложении возможно только в специальной конструкции — пояснительной (кроме того, эта конструкция экспрессивная): *Он долго шел к своей цели. И у него* ***это получилось — показать*** *первый результат на соревнованиях*.

Группа *получилось-произошло* сочетается с *так*: *Так получилось, что мы взяли билеты в один вагон* ('не планировали ехать в одном вагоне'), причем к матричному глаголу одновременно присоединяется *так* и пропозициональный актант — зависимая клауза (о постпозитивной конструкции *Мы взяли билеты в один вагон. Так получилось* см. ниже).

Сочетаемость с *это*, вообще говоря, естественна и не требует объяснений. Сочетаемость с *так*, напротив, требует объяснений, но в еще большей степени требует объяснений невозможность для *получилось-произошло* сочетания с *это*: **Это получилось, что мы ехали в одном вагоне*, — при том, что другие бытийные глаголы (*бывает*, *случается*), см. выше, допускают и *это*, и *так*. (Забегая вперед, скажем, что и группа *получилось-произошло* — в других условиях — допускает *это*).

Местоимения *так* и *как* являются очень многозначными и многофункциональными. Исходным для них можно считать значение образа действия. При этом значение образа действия, по определению, встречается только в контролируемых (агентивных) ситуациях, т. е. действиях (об акциональном классе действий см. [Апресян 2006], [Падучева 2004]). Как же произошло, что *так* из показателя образа действия в контролируемых ситуациях превратилось в показатель случайности, неконтролируемости?

В действительности, у *так* есть некоторая общая основа, которая позволяет ему сочетаться как с контролируемыми, так и с неконтролируемыми ситуациями, — это идея выбора, которая в неконтролируемых ситуациях превращается в идею варианта: *так* предполагает наличие вариативного фактора или аспекта.

Образ действия — это вариант реализации ситуации, который агенс выбирает, чтобы достичь результата. У простых ситуаций образ действия, или состав (термин «состав ситуации» мы употребляем в том смысле, как он используется в работе [Филипенко 2003]), более или менее фиксирован. У более абстрактных глаголов конкретный «состав», содержание ситуации может варьироваться, и «выбор» содержания в каком-то смысле аналогичен выбору образа действия — это тоже вариант. Показательно, что *так* сочетается с глаголами мнения (*Я так думаю / считаю*), но не знания (ср. **Я так знаю*), — знание единственно, а мнение — это

всегда вариант ситуации, который может и не реализоваться (или одна из версий, которая может быть ошибочной). Показательно также, что конструкция *Я так и знал* обозначает не знание, а правильное мнение (ср. [Апресян 1995]), которое подтвердилось (‘Я так считал, и оказалось, что правильно’, ср. также *Я так и понял*). Аналогичное значение выражается в конструкции *Я так понял, что они передумали* — это тоже выбор одной из версий (идея выбора реализуется также в контекстах *Я так хочу*; *Я так решил*, но мы не можем на этом останавливаться).

Так всегда предполагает *иначе*, *по-другому*. Разумеется, любая ситуация может произойти так и иначе, по-другому. Но выбирая *так*, говорящий подчеркивает возможность других вариантов, рассматривает ситуацию на фоне этих возможных вариантов — он выбирает именно такую концептуализацию. *Это* не имеет такой семантической добавки, оно «механически» вмещает содержание антецедента.

Однако случайные факторы действуют и на уровне ситуации, и на уровне сценария –именно поэтому эти уровни правомерно сравнивать: они составляют, так сказать, естественную минимальную пару. Почему же глаголы группы *получилось-удалось* не допускают *так* (*Мы победили. * Нам так удалось → это*)?

В группе *получилось-удалось*, несмотря на наличие случайных факторов, в содержании ситуации Р нет элемента случайности — оно известно агенсу заранее; а выбор между вариантами исхода (которых только два — успех и неудача) выражается не местоимением, а отрицанием (см. выше).

В группе *получилось-произошло* возможны и *так*, и (при определенных условиях) *это* — в зависимости от концептуализации. В случае *Так получилось / вышло / случилось / произошло, что Р* сама ситуация Р, как уже говорилось, является случайным фактором, «вторжением» в сценарий. *Так* показывает, что реализовался другой вариант сценария (не тот, который ожидался). ПМ *это* употребляется (по общему правилу) в том случае, если содержание Р уже введено в рассмотрение, известно (а потому безальтернативно), ср.: *Как это получилось* (*случилось*, *вышло*), *что целый район остался без света?* — чтобы задать вопрос, необходимо уже знать содержание Р (*как* в данной конструкции не коррелирует с *так* и не замещает позицию *так*, а имеет значение причины: ‘почему Р? / из-за чего Р? что привело к Р?’); ср. также: *Целый район остался без света. Как это получилось* (*случилось*, *вышло*)?; ответ также предполагает знание Р: *Это получилось из-за аварии на подстанции* (при другой концептуализации — чтобы подчеркнуть, что Р — лишь один из вариантов развития сценария, — говорящий может выбрать *так*: *Так получилось из-за аварии*).

Аналогичный эффект возникает в контексте оценочных наречий (чтобы оценивать ситуацию, нужно знать ее содержание): ***Странно*** *все-таки* ***это получилось, что*** *у меня объявился Милорд* [Юрий Коваль. От Красных ворот (1990)]; — *Да, — сказал мой отец, —* ***это удачно получилось, что*** *ты соль забыл* [Фазиль Искандер. Сандро из Чегема (1989)]. Впрочем, таких примеров в НКРЯ всего два, большинство вхождений — с вопросительными местоимениями.

Постпозитивная конструкция «*Р. Так получилось* (*вышло* и т. д.)» не эквивалентна препозитивному употреблению *так* (*Так получилось, что Р*) и представляет собой своего рода идиому (фразему): «*Совершенно новая конструкция рухнула от одного порыва ветра. Так получилось*». *Так получилось* значит не просто ‘случайно’, но и ‘нежелательно’.

Отрицательная импликатура (‘Р плохо, нежелательно’) возникает из общих соображений: раз Р не планировалось, значит, оно не нужно субъекту и, вполне вероятно, чему-то мешает, что-то нарушает. Наличие отрицательно-оценочной импликатуры видно по тому, что *Так получилось* сочетается с нейтральными событиями, ср.: *Мы ехали в одном вагоне. Так получилось*, и с отрицательными событиями, ср.: *Он не поступил в институт. Так получилось* (‘это плохо’), но странно звучит в контексте положительного события, ср.: *?Он поступил в институт. Так получилось* — такое высказывание получает смысл в ситуации, если субъект не планировал поступить в институт, и это поступление нарушает его планы.

ЗАМЕЧАНИЕ. Конструкция *Это удалось / получилось* тоже имеет оценочную импликатуру (‘хорошо’), ср. существительное *удача*, но собственно оценочные употребления глаголы *получиться*, *выйти* и *удаться* имеют в другой конструкции — подлежащной, где бытийный компонент является пресуппозицией, а оценочный — ассерцией: *Пирог получился* = ‘пирог хороший, вкусный’; *Пирог не получился* = ‘пирог плохой, невкусный’ (пресуппозиция ‘Пирог существует (испечен)’; в отрицательной конструкции возможно, что пирога как такового (как блюда) нет, но была попытка его испечь и есть следы этого, см. [Кустова 2019, в печати]).

Итак, противопоставление уровня ситуации и уровня сценария, проиллюстрированное на материале глаголов неполного контроля, релевантно для разных групп глаголов и имеет различные внешние проявления, касающиеся управления (инфинитив vs. клауза), характера и способов выражения субъекта, взаимодействия с отрицанием и пропозициональными местоимениями. Экспликация этих различий может иметь не только теоретический интерес, но и практическое применение при создании фильтров для автоматического разрешения неоднозначности, ср. [Кустова, Толдова 2008]; [Кустова, Ляшевская, Толдова 2008].

Литература

1. *Apresyan V. Yu.* (2015) Ustupitel’nost’: mekhanizmy obrazovaniya i vzaimodejstviya slozhnyh znachenij v yazyke. M.: YaSK. (In Russ.)
2. *Apresyan Yu. D.* (1995), The problem of factivity: to know and synonyms [Problema faktivnosti: znat’ i ego sinonimy] // «Voprosy yazykoznaniya». 1995. № 4; perepechatano v: Apresyan Yu. D. Selected Works, v. 2 [Izbrannye trudy, t. 2]. Moscow, YaSK, pp. 403–433. (In Russ.)
3. *Apresyan Yu. D.* (2006) Osnovaniya sistemnoj leksikografii // Yazykovaya kartina mira i sistemnaya leksikografiya. Otv. red. Yu. D. Apresyan. M.: YASK, pp. 33–160. (In Russ.)
4. *Bulygina T. V., Shmelev A. D.* (1997) Yazykovaya konceptualizaciya mira (na materiale russkoj grammatiki). — M.: Yazyki russkoj kul’tury. (In Russ.)
5. *Filipenko M. V.* (2003) Semantika narechij i adverbial’nyh vyrazhenij. M.: Azbukovnik. (In Russ.)
6. *Karttunen L.* (1973) La logique des constructions anglaises à complément predicatif // Languages. № 30, pp. 56–80.
7. *Kustova G. I.* (1992) Nekotorye problemy analiza dejstvij v terminah kontrolya // Logicheskij analiz yazyka. Modeli dejstviya. M.: Nauka, pp. 145–150. (In Russ.)

8. *Kustova G. I.* (2019) Glagoly s semantikoj uspekha / neudachi i strategii konceptualizacii agentivnyh situacij // Trudy Instituta russkogo yazyka im. V. V. Vinogradova. Vyp. XX. Gl. red. A. M. Moldovan. M. (v pechati). (In Russ.)
9. *Kustova G. I., Lyashevskaya O. N., Toldova S. Yu.* (2008) Semanticheskie fil'try dlya razresheniya mnogoznachnosti v nacional'nom korpuse russkogo yazyka: glagoly // Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Trudy mezhdunarodnoj konferencii «Dialog-2008». M, pp. 522–529. (In Russ.)
10. *Kustova G. I., Toldova S. Yu.* (2008) Nacional'nyj korpus russkogo yazyka: semanticheskie fil'try dlya razresheniya mnogoznachnosti glagolov // Trudy mezhdunarodnoj konferencii «Korpusnaya lingvistika–2008». SPb., pp. 210–219. (In Russ.)
11. *Letuchij A. B.* (2011) Pronominalizaciya sentencial'nogo argumenta v russkom yazyke // Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoj Mezhdunarodnoj konferencii «Dialog» (2011). Vyp. 10. M.: Izdvo RGGU, pp. 403–413. (In Russ.)
12. *Meijer A. M.* (2018) The Pragmatics and Semantics of Embedded Polar Responses with English so // Proceedings of the 35th West Coast Conference on Formal Linguistics, ed. by Wm. G. Bennett, Lindsay Hracs, and Dennis Ryan Storoshenko, pp. 269–279.
13. *Mel'chuk I. A.* (1999) Opyt teorii lingvisticheskih modelej «Smysl ↔ Tekst». Semantika, Sintaksis. M., «Nauka», 1974. 2-e izd. — M., YaRK. (In Russ.)
14. *Paducheva E. V.* (1985) Vyskazyvanie i ego sootnesennost' s dejstvitel'nost'yu (Referencial'nye aspekty semantiki mestoimenij). M.: Nauka. (In Russ.)
15. *Paducheva E. V.* (1991) Govoryashchij: sub"ekt rechi i sub"ekt soznaniya // Logicheskij analiz yazyka. Kul'turnye koncepty. M.: Nauka. (In Russ.)
16. *Paducheva E. V.* (2004) Dinamicheskie modeli v semantike leksiki. [Dynamic models in lexical semantics]. Moscow, JaSK Publ. 607 p. (In Russ.)
17. *Pekelis O. E.* (2018) Komplement i adverbial odnovremenno: o mestoimenii tak, zameshchayushchem sentencial'nyj aktant // Tipologiya morfosintaksicheskih parametrov. M. (In Russ.)
18. *Plungyan V. A.* (2011) Vvedenie v grammaticheskuyu semantiku: grammaticheskie znacheniya i grammaticheskie sistemy yazykov mira. M. (In Russ.)
19. *Razlogova E. E.* (1988) Eksplicitnye i implicitnye propozicional'nye ustanovki v prichinno-sledstvennyh i uslovnyh konstrukciyah // Logicheskij analiz yazyka. Znanie i mnenie. M., Nauka, pp. 98–107. (In Russ.)
20. *Schwabe, K.* (2016) Sentential proforms and argument conditionals// Frey, W., Meinunger, A., Schwabe, K. (Eds.), Inner-sentential Propositional Proforms: Syntactic Properties and Interpretative Effects. Benjamins, Amsterdam, pp. 211–240.
21. *Testelets Ya. G.* (2001) Vvedenie v obshchij sintaksis. M.: RGGU. (In Russ.)
22. *Yanko T. E. Kommunikativnye strategii russkoj rechi. M.: YaSK, 2001.* (In Russ.)
23. *Zaliznyak Anna A.* (1988) O ponyatii implikativnogo tipa (dlya glagolov s propozicional'nym aktantom) // Logicheskij analiz yazyka. Znanie i mnenie. M., Nauka, pp. 107–121.
24. *Zaliznyak Anna A., Levontina I. B.* (2006) Control and its place in the semantic structure of the predicates of the internal state // Zaliznyak Anna A. Mnogoznachnost' v yazyke i sposoby ee predstavleniya. [Polysemy in language and ways of its presentation]. Moscow, JaSK Publ., pp. 518–524. (In Russ.)

ЛЕКСИЧЕСКИЙ СОСТАВ ТЕКСТОВ УЧЕБНИКОВ РУССКОГО ЯЗЫКА ДЛЯ МЛАДШЕЙ ШКОЛЫ: КОРПУСНОЕ ИССЛЕДОВАНИЕ[1]

Лапошина А. Н. (antonina.laposhina@gmail.com),
Веселовская Т. С. (TSVeselovskaya@pushkin.institute),
Лебедева М. Ю. (m.u.lebedeva@gmail.com),
Купрещенко О. Ф. (ofkupr@gmail.com)

Государственный институт русского языка
им. А. С. Пушкина (Москва, Россия)

В статье представлены первые результаты сравнительного корпусного исследования современных учебников русского языка для младшего школьного возраста. Приводятся статистические данные об объеме и разнообразии лексики, используемой в учебниках, результаты анализа лексики на вхождение в частотные и тематические группы.

Ключевые слова: русский язык, учебник русского языка, корпус учебников, учебные тексты, сложность текста, русский язык в начальной школе, доказательная педагогика, сравнение корпусов

LEXICAL ANALYSIS OF THE RUSSIAN LANGUAGE TEXTBOOKS FOR PRIMARY SCHOOL: CORPUS STUDY

Laposhina A. N. (antonina.laposhina@gmail.com),
Veselovskaya T. S. (TSVeselovskaya@pushkin.institute),
Lebedeva M. U. (m.u.lebedeva@gmail.com),
Kupreshchenko O. F. (ofkupr@gmail.com)

Pushkin State Russian Language Institute (Moscow, Russia)

Annotation: This paper presents the first results of a comparative corpus-based research of the modern Russian language textbooks for primary school children. Volume and diversity statistics of textbooks' vocabulary, the results of the vocabulary's analysis included in frequency and thematic groups are given.

Keywords: Russian language, textbook of the Russian language, corpus of textbooks, educational texts, text complexity, Russian language in elementary school, evidence-based pedagogy, corpora comparison

[1] Работа выполнена при финансовой поддержке РФФИ, проект 17-29-09156

Постановка проблемы

Роль учебника русского языка для начальной школы трудно переоценить: он является не только источником первичной информации о системе языка, но и развивает навыки, необходимые для успешного освоения других предметов школьной программы: умение читать, понимать и анализировать тексты, грамотно писать, выражать свои мысли.

Качество учебных текстов — их доступность, разнообразность, актуальность и занимательность для данного возраста — является основным критерием оценки качества учебника, причем как с позиций теории учебника [Беспалько, с. 97–100], [Михеева с. 176], так и с пользовательских позиций[2]. Попытки объективно оценить качество учебных текстов предпринимаются мировым научным сообществом уже не один десяток лет [Collins-Thompson, 2014], [DuBay, 2007]. Среди исследований последних лет, направленных на тексты российских учебников, стоит отметить исследование И. А. Оборневой, посвященное адаптации формул читабельности для русского языка [Оборнева, 2006], работу на материале учебных текстов по химии [Шпаковский, 2012], исследование сложности текста на материале учебников истории [Иванов et al., 2018]. Однако все вышеупомянутые исследования посвящены анализу читабельности текста учебника как формальному критерию понятности/доступности текста. Таким образом, за рамками проведенных исследований остаются другие параметры, которые могут влиять на доступность и увлекательность учебника — тематическая принадлежность текстов, стилистические особенности, объем агнонимичной, абстрактной лексики.

Помочь в получении объективных данных о текстовом наполнении учебников может частотный анализ коллекции учебников: в настоящее время корпусные методы широко используются в исследовании и создании учебных материалов [Boulton, 2017]; [McEnery et al, 2010]; [Tribble, 2015]. Тем не менее, примеров создания и разметки корпуса учебной литературы даже в мировой практике не так много: стоит упомянуть исследование на материале учебников английского языка [Islam, 2014], корпус учебников японского языка [Sato, 2008], сравнение методических школ Южной и Северной Кореи на материале учебников английского для детей этих стран [Kim, 2017]. Однако внимание исследователей ещё не было направлено на учебные материалы по русскому языку.

Таким образом, данная работа преследует две локальные цели. Во-первых, это постановка проблемы, обнаружение лакуны в объективных данных о содержании учебников русского языка. Во-вторых, иллюстрация варианта решения этой проблемы на примере исследования двух параллелей учебников русского языка для младшей школы методами корпусной лингвистики.

Более широкой, глобальной целью исследования является описание современного состояния методики преподавания русского языка посредством количественного анализа содержания одного из главных компонентов обучающего процесса — учебника.

---

[2] т. н. «родительские» форумы, отзывы об учебниках в интернет-магазинах.

1. Материал исследования

Данное сравнительное исследование лексического состава учебников русского языка было проведено на материале созданного пилотного корпуса учебников русского языка для детей младшего школьного возраста (далее — корпус учебников). Отбор учебников проходил на основании Федерального перечня учебников[3], отзывов сообществ учителей и родителей.

Простейшим элементом корпуса является законченный, визуально отделяемый блок текста. Разметка представляет собой метатекстовую информацию о каждом блоке текста, необходимую для анализа содержания фрагментов и сравнения методических приёмов и содержит следующую информацию:

- тип аппарата учебника (текст, наполнение упражнения, формулировка задания и др.);
- тип текста (поэзия или проза);
- авторство текста (аутентичный, адаптированный, сконструированный);

Пример разметки корпуса представлен в **Таблице 1**.

Таблица 1. Пример метатекстовой разметки корпуса учебников

Текстовый блок	Аппарат учебника	Тип текста	Авторство	Имя автора
Летит над речкой зимородок. Очень красивая птица: брюшко оранжевое, спинка ярко-зелёная, а нос длинный и прямой, как палочка.	Наполнение упражнения. Текст	Проза	Аутентичный	Г. Скребицкий
Определите тему и главную мысль текста. Придумайте к нему заголовок.	Формулировка задания	—	—	—

2. Объём и воспроизводимость лексики в текстах учебников русского языка

Объем размеченного корпуса, на материале которого проводится данное исследование, составляет две полные параллели учебников 1–4 класса, около 300 тысяч токенов, 42 300 предложений, 15 000 текстовых блоков. На **Рис. 1** представлено распределение уникальной лексики в учебниках 1–4 классов двух анализируемых параллелей.

Для удобства визуализации и описания данных здесь и далее мы обозначим параллель учебников под ред. В. П. Канакиной и В. Г. Горецкого [Канакина et al, 2013–2014] как *Канакина* и учебники под ред. М. Л. Каленчук и Н. А. Чураковой [Чуракова et al, 2013–2017] как *Чуракова*.

[3] Приказ Министерства просвещения РФ №345 от 28.12.2018

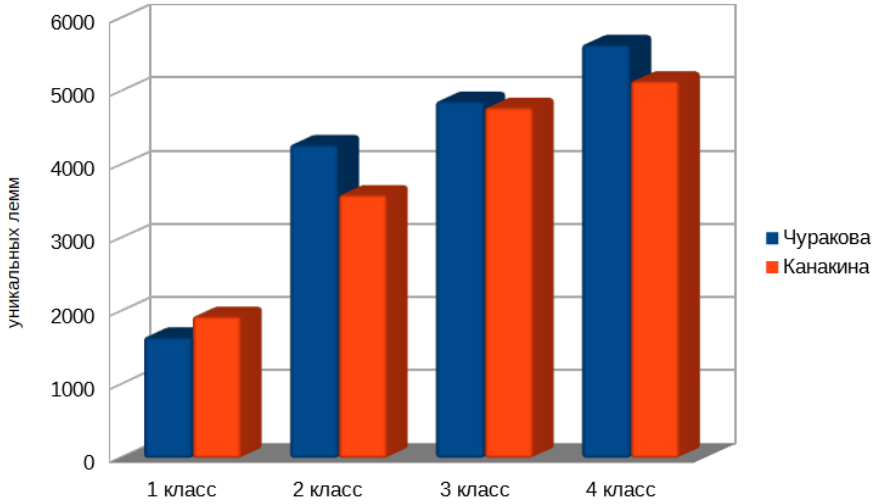


Рис. 1. Кол-во уникальных лемм в учебниках в зависимости от класса

Для параллели Канакиной общее число уникальных лемм составляет 8363, для Чураковой — 9720. При переходе от класса к классу совершенствуется навык чтения учащихся, растет их словарный запас, вместе с тем увеличивается и объем входящей в учебники лексики. На графике (**Рис. 1**) наглядно представлен процесс прироста лексики, при котором максимальный скачок в обоих учебниках приходится на переход из первого класса во второй. Так, например, в параллели Чураковой количество уникальных лемм в первом и втором классах отличается в 2,5 раза. Далее скорость роста объема лексики замедляется.

Логично предположить, что некоторый процент этой входящей лексики будет пересекаться, встречаться сразу в нескольких классах. Эти цифры могут служить показателем согласованности учебников внутри одной параллели между собой: такой общей базой является терминологический (*предложение, звук*) и операционный (*прочитать, списать*) аппарат учебника, отрабатываемые словарные слова, опционально может присутствовать общая сюжетная линия. Так, например, для параллели Чураковой число слов, общих для всех четырех классов даже с учетом общей сюжетной линии и системы персонажей составляет 8% (872 леммы из 9720), цифра же для параллели Канакиной, не объединенной общей сюжетной канвой, несколько выше — 14.8% (1240 лемм из 8360). Это может свидетельствовать как о более высоком уровне преемственности лексики в параллели Канакиной, её воспроизводимости из учебника в учебник, так и о тематическом однообразии текстов параллели, и требует дополнительного исследования.

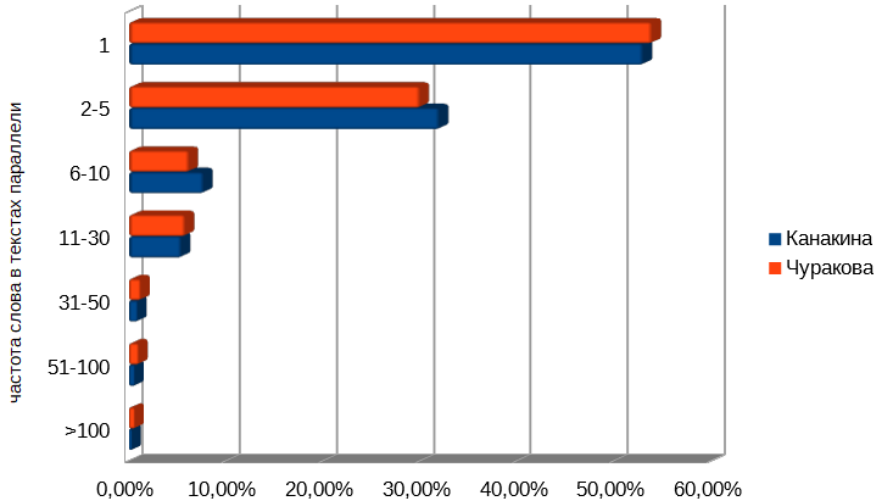


Рис. 2. Распределение слов по абсолютной частоте встречаемости во всех текстах параллели

Рис. 2 демонстрирует распределение всей лексики текстов по абсолютной частоте встречаемости. Общая схема распределения частот для двух параллелей весьма схожа. Так, две группы с наивысшей частотой (более 100) представляют собой предлоги и союзы (т. н. стоп-слова, частотные для всех русских текстов), в параллели Чураковой также к этой группе относятся имена основных персонажей учебников. Показательно, что единственное неслужебное слово, попавшее в эту частотную группу у обеих авторов — *лес*. Эта информация нам пригодится в дальнейшем. В группе с частотой от 51 появляется лексика, оказавшаяся центральной в учебниках русского языка (*земля, человек, идти, русский, берёза, птица*). Самую большую группу (более 50 % всех уникальных лемм) на рис. 2 составляют слова, которые встречаются во всей параллели 1 раз.

Вполне ожидаемо самыми частотными словами в учебниках русского языка оказываются лингвистические термины (*звук, глагол*) и требуемые учебные операции (*спиши, прочитай*). Набор этой лексики весьма ограничен: несмотря на то, что формулировки заданий и справочная информация занимают более 60 % объема учебника, лексика этих блоков составляет 11 % от всех уникальных лемм учебника.

Напротив, текстовые фрагменты, включенные в учебник в качестве иллюстрации лингвистических явлений, отработки орфографии и т. д., составляют около 30 % всего объема учебника, однако являются самой насыщенной и лексически разнообразной категорией: в обеих параллелях 89 % всех уникальных лемм встречается именно в текстах. Если набор терминов хотя бы в какой-то мере регулируется учебными программами, то выбор текстовых фрагментов полностью зависит от авторов учебника, что и представляет большой интерес для исследования. Поэтому в дальнейших разделах речь будет идти о подсчётах, выполненных на материале собственно текстовых фрагментов учебников.

3. Частотный анализ лексики учебников

Учебник русского языка готовит ученика к грамотному использованию родного языка в разных ситуациях и контекстах общения, формирует целостный, социально ориентированный взгляд на мир[4]. В таком случае логично предположить, что учебник должен содержать в себе усредненную модель русского языка, в равной степени отражать различные темы, постепенно наращивая объем этой лексики от простой к более сложной. Однако знакомство с учебными материалами позволяет выдвинуть предварительную гипотезу о тематической несбалансированности текстов в современных учебниках русского языка и господстве темы природы. Проверить эту гипотезу на практике возможно с помощью частотного анализа лексики учебников и сравнения полученных данных с условным «стандартным» русским языком. В качестве такой отправной точки для сравнения мы будем использовать данные Нового частотного словаря русской лексики О. Н. Ляшевской и С. А. Шарова [Ляшевская, Шаров, 2009], далее — частотный словарь.

3.1. Частотная лексика в учебниках русского языка

Обратимся вначале к высокочастотной лексике текстовых фрагментов учебников. **Таблица 1** представляет собой список самых популярных самостоятельных частей речи по частотному словарю, в параллели Канакиной и параллели Чураковой.

Таблица 1. Самые частотные слова самостоятельных частей речи

Частотный словарь	ipm	Параллель Канакиной	ipm	Параллель Чураковой	ipm
быть	12 160	быть	7 540	быть	6 821
год	3 727	слово	6 263	лес	3 427
мочь	2 912	лес	3 873	солнце	2 537
человек	2 723	язык	3 667	сказать	2 504
сказать	2 396	земля	3 172	вода	2 372
ещё	2 323	русский	2 884	день	2 372
уже	2 179	вода	2 801	маленький	2 239
время	2 015	птица	2 554	река	2 207
говорить	1 755	солнце	2 554	жить	2 207
знать	1 713	весна	2 472	идти	2 142
стать	1 621	день	2 389	земля	2 076
дело	1 412	белый	2 389	собака	2 010
жизнь	1 389	снег	2 348	снег	1 845

Первое, что обращает на себя внимание — разница в распределении слов по частям речи. В списках по учебникам значительно меньше глаголов и больше

[4] Приказ Минобрнауки России № 373 от 06.10.2009.

существительных: если в частотном словаре в таблице представлено 3 существительных и 6 глаголов, то в параллели Канакиной соотношение равно 10 и 1, Чураковой — 8 и 4. Это может сигнализировать о преобладании в учебниках описательных текстов над повествовательными, а следовательно, малом количестве действий. Так, примеры 1 и 2 демонстрируют характерные для учебников тексты описания природы без участия глаголов. По этому параметру параллель Чураковой оказывается ближе к корпусу, чем параллель Канакиной.

(1) *Зимний тёплый день. На земле мягкий снег.*
(Канакина et al, 2013–2014, 2 кл., ч. 2, стр. 139)

(2) *А сегодня — что за день! Солнце, птицы! Блеск и счастье! Луг росист…*
(Канакина et al, 2013–2014, 4 кл., ч. 2, стр. 138)

Данные таблицы 1 обнаруживают также явные тематические доминанты текстов учебников. В абсолютные лидеры по частотности в обеих параллелях выходит *лес* (211 ipm по частотному словарю), а также именования стихий и природных явлений: *земля, вода, солнце, снег* (494, 484, 165, 125 ipm по частотному словарю соответственно). На этом этапе становятся видны и личные предпочтения авторов касательно животных: параллель Канакиной обнаруживает интерес к птицам (79 упоминаний вариантов птица/птичка/пичужка + 560 упоминаний конкретных видов птиц, см. подробнее **раздел 3.2**), параллель Чураковой на первый взгляд отдаёт предпочтение собакам, однако при более детальном рассмотрении выявляется небольшое превосходство кошек (95 упоминаний на варианты *собака/пёс/щенок/собачка* и 104 на *кот/кошка/котёнок/котик*).

Таким образом, данные **Таблицы 1**, во-первых, подтверждают гипотезу о значительном уклоне русских учебных текстов в сторону описания природы и погодных явлений, а во-вторых, демонстрируют явный частеречный дисбаланс, который требует отдельного изучения.

Остановимся подробнее на тематической принадлежности наиболее часто встречающихся слов в учебниках. Примерные тематические группы расположены ниже по убыванию количества представителей и их общей частоте, в скобках указаны абсолютные частоты в параллелях авторов Чураковой (число слева) и Канакиной (число справа):

- Природа и животные: *лес* (136/118), *береза* (26/117), *заяц* (34/37),
- *река* (56/45);
- Погода и сезоны: *солнце* (80/86), *снег* (72/79), *зима* (54/44);
- Деревенская лексика: *село* (30/15), *сад* (32/22), *деревня* (13/25), *корова* (22/11), *сено* (4/9), *огород* (20/9), *грядка* (7/6), изба (6/6);
- Школьная лексика: *класс* (17/18), *читать* (16/23), *учительница* (15/4), *карандаш* (10/17);
- Бытовая лексика: *хлеб* (15/23), *шапка* (11/10), *мыло* (12/4), *подушка* (13/5), *магазин* (8/7);
- «Детская» лексика: *мама* (36/26), *мяч* (26/5), *игрушка* (8/3), *кататься* (3/3), *качели* (2/0), *велосипед* (6/2), *коньки* (2/6), *лыжи* (2/4);
- Городская лексика: *город* (28/42), *парк* (3/2), *автобус* (4/0), *музей* (4/1), *квартира* (4/3), *театр* (0/5); *зоопарк* (7/4), *цирк* (3/1);

На первый план выходят два блока лексики, тематически связанные с природой и погодой: они значительно опережают остальные группы по частотности. Также интересно отметить преобладание деревенской лексики над городской. Этот факт может быть связан как с использованием материала сказок, место действия которых обычно связано с деревней, так и с тем, что жизнь в деревне сильнее связана с природой, которая в свою очередь является самой частотной темой. Интересно отметить, что «детская» лексика, связанная с играми, развлечениями и спортом и городская лексика представлены значительно скуднее.

Также стоит отметить частотные по корпусу слова, ни разу не появившиеся в учебниках (например, *документ, информация, система, проблема, центр, цена*). Безусловно, они часто встречаются во «взрослых» публицистических текстах и деловых документах и трудно говорить о методической необходимости их добавления в учебник. Однако эта информация лишний раз подчеркивает стилистическую ориентацию детских учебников на художественные и описательные тексты.

3.2. Много в учебниках, мало в корпусе: осознанный выбор авторов

Редкая по частотному словарю на основе корпуса русского языка, но часто появляющаяся в учебниках лексика особенно интересна для анализа, так как она ярче всего отражает осознанный (или неосознанный) выбор авторов, касающийся тематики текстового наполнения учебника, воспроизводимости лексики, включения слов различных исторических эпох и т. д. **Таблица 2** представляет собой топ-10 слов, часто появляющихся в учебниках, но не вошедших в список 10 тысяч частотных слов русского языка.

Таблица 2. Топ-10 слов из параллели, не входящих в 10 тыс. частотных слов русского языка

Параллель Канакиной		Параллель Чураковой	
лемма	абс. частота	лемма	абс. частота
прилетать	29	**Анишит Йокоповна**	239
метель	25	**волшебница**	104
скворец	24	норка	41
дождик	23	**Асырк**	30
иней	22	соловей	23
клюв	22	ива	20
вьюга	20	**Торк**	18
душистый	19	барашек	17
ласточка	19	дождик	17
соловей	18	радуга	16

Учебники параллели Чураковой связаны между собой центральными персонажами — Машей и Мишей, которые осваивают правила русского языка под руководством волшебницы Анишит Йокоповны и её помощников Асырк

и Торк[5]. Эта сюжетная линия хорошо просматривается в таблице — персонажи заняли 4 из 10 популярных низкочастотных слов параллели. В общей сложности имена персонажей насчитывают 1500 упоминаний за параллель, эту информацию стоит учитывать при расчете объема лексики.

В обеих параллелях снова подтверждается предположение о главенствующих тематиках текстов — природа, погодные явления и животный мир, однако и тут можно заметить ряд любопытных особенностей. Так, параллель Канакиной обнаруживает явное тяготение к описанию зимнего времени года (*метель, иней, вьюга*), осадкам (*дождик, снег, иней*) и птицам (*прилетать, клюв, скворец* и др.), в то время как параллель Чураковой — к лесу (*пенёк, опушка, норка*). Хотелось бы отдельно отметить масштаб интереса к текстам о птицах в учебниках Канакиной: всего в параллели учебников было найдено упоминание 46 видов птиц с общей частотой встречаемости более 560 раз за параллель. Помимо привычных представителей крылатой фауны, таких как *ворона, воробей, сорока, синица*, на страницах учебников обнаруживаются и более незаурядные примеры, такие как *рябчик, сойка, киви, оляпка, деряба*.

Таким образом, предположение о тематической сбалансированности текстов учебника также не находит подтверждения в данных **Таблицы 2**.

3.3. «Одноразовая» лексика

Данные **Рисунка 2** показали, что более 50 % всей лексики анализируемых учебников встречается 1 раз за всю параллель. Предположение, что это повседневная лексика, встречающаяся в текстах, но не требующая повторения и отработки, не подтверждается: 48 % этой лексики не входит в частотный список 10 000. Рассмотрим некоторые случаи употребления такой единичной лексики подробнее:

1. Задания, направленные на изучение или уточнение лексического значения этих слов. Чаще всего это задания на определение «старых» и «новых» слов, где встречаются историзмы (*бурлак*), устаревшая лексика *(ветрило, длань, ямщик, скань).* Интересен и тот факт, что некоторые слова, задуманные авторами как примеры «новых» слов, уже стали историзмами (*дискета, радиотелефон*), но всё ещё присутствуют в современных переизданиях учебника.

2. Иллюстрация специфических языковых явлений. Так, например, слово *гусятница* приводится для отработки суффикса *ниц*, а пример 1, вызвавший резонанс на родительских форумах — парность согласного звука З по твердости-мягкости:

(3) *Зябнет*
Зубр,
Зайчонок
Зябнет,

[5] Выбор имен персонажей становится немного понятнее, если прочитать их наоборот.

Зуй[6] *зазяб,*
Зазяб и
Зяблик.
(Чуракова et al, 2013–2017, 1 кл., стр.45).

3. Выбор авторами фрагмента аутентичного текста или самостоятельное составление текстов с редкой неповторяющейся лексикой для упражнений на анализ текста (найти главную мысль, озаглавить и т. д.)

(4) *Мальчики посадили луковицы гладиолусов и клубни георгинов.* (текст авторов учебника, Канакина et al, 2013–2014, 2 кл., ч. 1, стр. 21).

(5) *Каждый год, принимаясь за огород, тётя Шура кладёт на меже тряпицу и носит в неё всё, что блеснёт под лопатой: бусину, черепок, костяную пуговицу, наконечник стрелы…* (фрагмент аутентичного текста В. Пескова, Чуракова et al, 2013–2017, 4 кл., ч.2, стр. 63).

Описанные примеры показывают, что появление редкой лексики один раз за параллель может быть обусловлено специфическим заданием (группа примеров 1), или быть личным авторским решением. Зачастую наличие устаревшей лексики и историзмов связано с продолжением традиции литературоцентричности в обучении русскому языку и использованием в качестве учебного материала текстов русской классики. Между тем, иллюстрация новых лингвистических явлений (группа примеров 2) или аналитическая работа с текстом (группа примеров 3) на незнакомых школьнику словах и ситуациях, далеких от его жизни, представляется спорным методическим решением. Кроме того, такая лексика с большой вероятностью вызовет трудности при чтении. Поэтому вопрос о целесообразности и объеме включения такой лексики в учебники должен исследоваться отдельно и подтверждаться экспериментально, например, методами айтрекинга или экспериментами на скорость чтения с заданиями на понимание.

4. Выводы

Данное исследование является первым шагом в направлении исследования лексики учебников и в большей степени намечает проблемные точки, чем предлагает готовые ответы. Однако подведем промежуточные итоги исследования лексического состава учебников русского языка для младшей школы на материале двух параллелей учебников.

Текстовые фрагменты являются самым лексически разнообразным блоком учебника, в них содержится 89 % всей уникальной лексики. Более 50 % этой лексики появляется один раз за всю параллель. Явными тематическими доминантами текстов оказываются природа, погодные явления и животные. Этот вывод вряд ли можно назвать революционным, однако он впервые подкреплен реальными цифрами. Деревенская лексика представлена в учебниках шире, чем

[6] Толковый словарь Д. Н. Ушакова: Зуй, зуя, мн. зуи, муж. (обл.). Местное название некоторых птиц из рода куликов. Пример: болотный зуй.

городская и детская. Скудно представлена лексика, свойственная для публицистического стиля. Разница в частотах глаголов и существительных с частотным словарём может сигнализировать о перевесе текстов описательного характера.

В качестве перспективных направлений работы представляется коллокативный анализ учебных текстов, более подробное количественное исследование принадлежности лексики учебников к определенной тематической группе, исторической эпохе, типу текста и функциональному стилю русского языка.

5. Благодарности

Авторский коллектив сердечно благодарит студентов 1 курса магистратуры филологического факультета и С. И. Ельникову, заведующую кафедрой методики преподавания РКИ Государственного института русского языка им. А. С. Пушкина, за помощь в разметке корпуса, а также анонимных рецензентов за ценные комментарии.

Литература

1. *Boulton A.* (2017), Corpora in language teaching and learning.//Language Teaching. 50.4, pp. 483–506.
2. *Collins-Thompson K.* (2014), Computational assessment of text readability: a survey of current and future research. In: François, Thomas and Delphine Bernhard (eds.), Recent Advances in Automatic Readability Assessment and Text Simplification. //Special issue of International Journal of Applied Linguistics, pp. 97–135.
3. *Dax Thomas* (2005), Type-token Ratios in One Teacher's Classroom Talk: An Investigation of Lexical Complexity.
4. *DuBay W. H.* (2007), Smart Language: Readers, Readability, and the Grading of Text. ERIC.
5. *Islam, Md. Zahurul* (2014), Multilingual Text Classification using Information-Theoretic Features PhD Thesis; Goethe University Frankfurt.
6. *Ivanov V. V., Solnyshkina M. I., Solovyev V. D.* (2018), Efficiency of Text Readability Features in Russian Academic Texts//Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018", Moscow, pp. 284–293.
7. *Jeong-ryeol Kim* (2017), A Comparative Look at South and North Korean English Education Based on the Analysis of Textbook Corpus//Corpus Linguistics Research. Vol. 3 (2017.09). pp. 1–22.
8. *McEnery, Tony; Xiao, Richard.* (2010), What corpora can offer in language teaching and learning. / Handbook of Research in Second Language Teaching and Learning. ed. / E. Hinkel. Vol. 2 London & New York: Routledge, pp. 364–380.
9. *Norberg, Cathrine & Nordlund, Marie* (2018), A Corpus-based Study of Lexis in L2 English Textbooks. Journal of Language Teaching and Research. Vol. 9, No. 3, pp. 463–473.

10. *Sato, Satoshi & Matsuyoshi, Suguru & Kondoh, Yohsuke* (2008), Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus.Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, pp. 654–660.
11. *Tribble, Christopher* (2015), "Teaching and language corpora." Multiple affordances of language corpora for data-driven learning, 69, pp. 37–62.
12. *Bespalko V. P.* (1988), Theory of the textbook: Didactic aspect [Teoriya uchebnika: Didakticheskij aspekt]. — M.: Pedagogy [M.: Pedagogika].
13. *Bim I. L.* (1975), To the development of a foreign language textbook theory [K razrabotke teorii uchebnika inostrannogo yazyka], Russian language abroad [Russkij yazyk za rubezhom], Moscow, № 4.
14. *Glinkina L. A.* (2008), Illustrated Dictionary of Forgotten and Difficult Words of the Russian Language [Illyustrirovannyy tolkovyy slovar' zabytykh i trudnykh slov russkogo yazyka]. World of Encyclopedia Avanta + [Mir entsiklopediy Avanta+].
15. *Lerner I. Ya.* (1974), Criteria for the complexity of some elements of the textbook [Kriterii slozhnosti nekotorykh elementov uchebnika]. Problems of the school textbook [Problemy shkol'nogo uchebnika], vol. I, pp. 23–25.
16. *Lyashevskaya O. N., Sharov S. A.* (2009), Frequency Dictionary of the Modern Russian Language (on the materials of the National Corpus of the Russian Language) [Chastotnyy slovar' sovremennogo russkogo yazyka (na materialakh Natsional'nogo korpusa russkogo yazyka)]. M.: Azbukovnik [Azbukovnik].
17. *Matskovsky M. S.* (1976), The Problem of Readability of Printed Material [Problema chitabel'nosti pechatnogo materiala]. The Meaning Perception of a Voice Message in Mass Communication [Smyslovoye vospriyatiye rechevogo soobshcheniya v usloviyakh massovoy kommunikatsii]. M.: Science, pp. 126–141.
18. *Mikk, J. A.* (1981), Optimizing the complexity of an educational text: to help authors and editors [Optimizatsiya slozhnosti uchebnogo teksta: v pomoshch' avtoram i redaktoram]. M.: Education [Prosveshcheniye].
19. *Mikheeva S. A.* (2015), System of formalized criteria for evaluating a school textbook [Sistema formalizovannykh kriteriyev otsenki shkol'nogo uchebnika], Educational Studies [Voprosy obrazovaniya], issue 4, pp. 147–183.
20. *Obornneva I. V.* (2006), Automated assessment of the complexity of educational texts based on statistical parameters [Avtomatizirovannaya otsenka slozhnosti uchebnykh tekstov na osnove statisticheskikh parametrov]. Thesis for Ph.D. [Dissertatsiya na soiskaniye step. K.p. n.], Moscow.
21. *Pedagogical and ergonomic requirements for means of education* [Pedagogiko-ergonomicheskiye trebovaniya k sredstvam obucheniya]. — SPb, M: Chrismas +, IOSO RAO, 2000. — 64 p.
22. *Plungyan V. A.* (2005), Why do we make the National Corpus of the Russian language? [Zachem my delayem Natsional'nyy korpus russkogo yazyka?] Domestic Notes [Otechestvennyye zapiski]. № 2, pp. 296–308.
23. *Order* of the Ministry of Education and Science of Russia dated October 06, 2009 № 373 (as amended on 12/31/2015) "On approval and implementation of the federal state educational standard of primary general education" (Registered

in the Ministry of Justice of Russia December 22, 2009 N 15785) [Prikaz Minobrnauki Rossii ot 06.10.2009 № 373 (red. ot 31.12.2015) «Ob utverzhdenii i vvedenii v deystviye federal'nogo gosudarstvennogo obrazovatel'nogo standarta nachal'nogo obshchego obrazovaniya» (Zaregistrirovano v Minyuste Rossii 22.12.2009 N 15785)].

24. *Order* of the Ministry of Education of the Russian Federation of December 28, 2018 № 345 "On the federal list of textbooks recommended for use in the implementation of state-accredited educational programs of primary general, basic general and secondary general education" [Prikaz Ministerstva prosveshcheniya RF ot 28.12.2018 №345 «O federal'nom perechne uchebnikov, rekomendovannykh k ispol'zovaniyu pri realizatsii imeyushchikh gosudarstvennuyu akkreditatsiyu obrazovatel'nykh programm nachal'nogo obshchego, osnovnogo obshchego, srednego obshchego obrazovaniya»].
25. *Frumin I. D., Dobryakova M. S., Barannikov K. A., Remorenko I. M.* (2018), Universal Competences and New Literacy: What to Teach Today for Success tomorrow. Preliminary findings of the international report on trends in the transformation of school education [Universal'nyye kompetentnosti i novaya gramotnost': chemu uchit' segodnya dlya uspekha zavtra. Predvaritel'nyye vyvody mezhdunarodnogo doklada o tendentsiyakh transformatsii shkol'nogo obrazovaniya]. National Research University Higher School of Economics, Institute of Education. Modern education analytics [Natsional'nyy issledovatel'skiy universitet «Vysshaya shkola ekonomiki», Institut obrazovaniya. Sovremennaya analitika obrazovaniya]. № 2 (19). M.: HSE.
26. *Shpakovsky Yu.* (2012), Estimation of the difficulty of perception and optimization of the complexity of an educational text. On the material of texts in chemistry [Otsenka trudnosti vospriyatiya i optimizatsiya slozhnosti uchebnogo teksta. Na materiale tekstov po khimii]. LAP Lambert Academic Publishing.

Источники анализируемого материала

27. *Канакина В. П.* Русский язык. Коллекция учебников 1–4 класс. Учеб, для общеобразоват. организаций с прил. на электрон, носителе / В. П. Канакина, В. Г. Горецкий. 5-е изд. — М.: Просвещение, 2013–2014 гг.
28. *Чуракова Н. А., Каленчук М. Л.* Русский язык. Коллекция учебников 1–4 класс: учебник / Н. А. Чуракова, М. Л. Каленчук, Т. А. Байкова, О. В. Малаховская — М.: Академкнига/Учебник, 2013–2017 гг.
29. *Канакина В. П.* Русский язык. Рабочие программы. Предметная линия учебников системы «Школа России». 1–4 классы: пособие для учителей общеобразоват. организаций / [В. П. Канакина, В. Г. Горецкий, М. В. Бойкина и др.]. — М.: Просвещение, 2014. — 340 с.

SENTENCE LEVEL REPRESENTATION AND LANGUAGE MODELS IN THE TASK OF COREFERENCE RESOLUTION FOR RUSSIAN

Le T. A. (anhlt@vimaru.edu.vn)[1, 2],
Petrov M. A. (maksimallist@gmail.com)[1],
Kuratov Y. M. (yurakuratov@gmail.com)[1],
Burtsev M. S. (burtcev.ms@mipt.ru)[1]

[1]Neural Networks and Deep Learning Lab—Moscow Institute of Physics and Technology, Moscow, Russia

[2]Faculty of Information Technology—Vietnam Maritime University, Hai Phong, Viet Nam

Coreference Resolution (CR) is one of the most difficult tasks in the field of Natural Language Processing due to the lack of deeply and comprehensively understanding the semantic meaning of the mention in not only the sentence-level context but also the entire document-level context. To the best of our knowledge, the previous proposed models often address the coreference resolution task in two steps: 1) detect all possible mention candidates, 2) score and cluster them into chains. We instead propose a new approach which reforms the coreference resolution task to the task of learning sentence-level coreferential relations. Additionally, by leveraging the power of state-of-the-art language representation models such as BERT, ELMo, it was possible to achieve cutting edge results on Russian datasets.

Key words: coreference resolution, language modeling, sentence-level coreference

1. Introduction

Coreference resolution has a long research history but the quality of solutions is still not really convincing, especially for Russian language. As far as we know, there have been few deep learning-based coreference models that achieved state-of-the-art performances and all of them are studied on English datasets. Kevin et al. [6] proposed a variant of reinforcement learning solution with reward-rescaled max-margin objective to directly optimize a mention-ranking model for coreference evaluation metrics. This model obtained remarkable results on CoNLL2012 dataset [10], 65.73% and 63.88% on English and Chinese test sets, respectively. In the follow up paper [5] the problem was approached in a different way, where the entity-level information was captured with distributed representations of coreference cluster pairs. This model was trained with learning-to-search algorithm. The model performance was not better than the previous one.

Kenton Lee et al. [4], [3] proposed two end-to-end coreference models. The first one is the simple model with two steps: 1) create span representation from context-dependent boundary representations and head-finding attention mechanism, 2) cluster mentions. The second model is an improvement of the first one with inference procedure involving iterations of refining span representations. The model achieved 73% of average F1 on the test set of the English CoNLL-2012 shared task [10].

Starting with Kenton Lee et al. 's work as a baseline, this paper aims to build an end-to-end coreference model for Russian language. Previous works on coreference resolution on Russian language were mostly rule-based or feature-based simple models like random forest [15], [13]. The main contributions of this paper are listed bellow:

- Original model that learns to predict sentence-level coreferential relationships. The model is then directly integrated or generate features for the baseline model.
- Extention of the baseline model with state-of-the-art contextual language models ELMo and BERT trained for Russian language to boost task performance.

To test our proposed models, we participated in the shared task at Dialogue 2019 conference and achieved the best results in both tasks:

- Coreference task: The first place at the round using the gold mention boundaries, the second place at the round using only the raw text.
- Anaphora task: The first place at both rounds with or without using the gold mention boundaries.

2. Models

2.1. Baseline model

In this section we briefly describe the Higher-order Coreference Resolution model, which is used as baseline model (refer to the original papers [3] and [4] for more details).

Coreference resolution task consists of two sub-tasks: mention detection and mention clustering. This model solves both of them in end-to-end manner. Firstly, each text span is encoded by a single vector g_i, which is a concatenation of the first, last and head tokens representations. Secondly, mention score $s_m(i)$ is computed as $s_m(i) = w_{\text{m}} \cdot \text{FFNN}_m(g_i)$, where FFNN is a feed-forward neural network. Then top K (K depends on text length) spans are selected based on mention score. Finally, antecedent score $s_a(i,j)$ is computed for selected top K spans. Antecedent score should be positive if mention j is an antecedent to mention i. Then coreference chains are collected according to antecedent scores.

2.2. Sentence-level Coreferential Relationship-based Model

One of the main difficulties of coreference resolution task in comparison to other NLP tasks is the length of input text. Input of Name Entity Recognition task is one sentence. Question Answering models use several sentences as a context and one

as a question. Meanwhile, input of coreference resolution task is one paragraph, or even one document with several hundred sentences. In order to make predictions correctly, the model need to capture the sentence semantics in the document context. Encoding a sentence in the context of document with several hundred sentences is a long-standing challenge. This challenge leads to the difficulty of applying common deep neural network models. In order to address this problem, we propose Sentence-level Coreferential Relationship-based model (SCRb model) that takes as input a document and outputs a square matrix representing the probabilities of coreferential links of sentences. In the training set this matrix is a binary square matrix (See Fig. 1, 2 for more details). The matrix is then can be used in two ways. In the first one, the probabilities produced by SCRb model are utilized as an input features. In the second way, the SCRb model is directly integrated into the end to end coreference model and both of them are trained jointly.

Figure 1: Visualization of the document chtb_0219.v4_gold_conll in the OntoNotes 5.0 dataset. The mentions with the same highlight color are belong to the same cluster

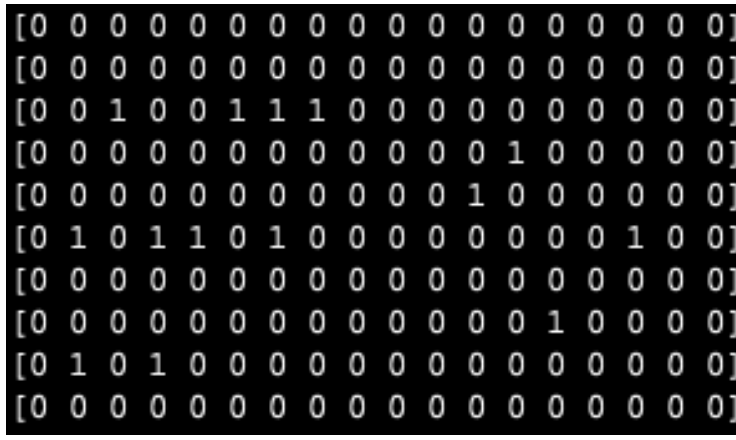


Figure 2: Binary matrix for document chtb_0219.v4_gold_conll representing the sentence-level coreferential relationship that SCRb model learns to predict

Here we describe step by step how SCRb model works:

- The model uses two types of word embedding: 1) free-context word embedding (e_{fc}) and 2) context-based word embedding (e_{cb}). In addition, to represent OOV words better, a convolutional network is utilized to generate character-based word embedding (e_{ch}). All these vectors are then concatenated to create the final word embedding:
$$e_w = [e_{fc}, e_{cb}, e_{ch}] \quad (1),$$
here [,] denotes the concatenation operator.
- The final word embeddings of each sentence are then feed into a Bi-LSTM network to output word vectors representing words in their sentence context:
$$w = \left[o_{\overrightarrow{lstm}}, o_{\overleftarrow{lstm}}\right], \quad (2)$$
here $o_{\overrightarrow{lstm}}$ and $o_{\overleftarrow{lstm}}$ are outputs of forward and backward LSTM networks, respectively.
- A maxpooling layer is used to reduce the word dimension to create the sentence representation:
$$s = max\_pooling(w_i), \quad (3)$$
where $w_i \in s$.
- The second Bi-LSTM network is utilized to capture the final sentence representation in the document context:
$$s_{dc} = \left[s_{\overrightarrow{lstm}}, s_{\overleftarrow{lstm}}\right] \quad (4)$$
- To create the matrix representing sentence relations, we modified Multi-dimensional Self-attention [12]:
  - Let $s_i \in \mathbb{R}^{d_s}$, where d_s denotes the length of sentence vectors outputted by the last Bi-LSTM network, is the vector representing the i^{th} sentence in the document.
  - Let $e_{d_{ij}} \in \mathbb{R}^{d_d}$, where d_d denotes the length of position encoding vectors, is distance embedding between s_i and s_j.
  - Let $W \in \mathbb{R}^{d_s}$, $W_1, W_2 \in \mathbb{R}^{d_s \times d_s}$, $W_d \in \mathbb{R}^{d_s \times d_d}$ are weight matrices, and $b_1 \in \mathbb{R}^{d_s}$, $b \in \mathbb{R}$ are bias terms.
  - The aligment score between s_i and s_j are computed as following formula:
$$f(s_i,s_j) = W^T\sigma(W_1 s_i + W_2 s_j + W_d e_{d_{ij}} + b_1) + b, \quad (5)$$
where σ is the activation function.
  - Final antecedent score is computed as sum of $f(s_{mi}, s_{mj})$ and antecedent score $s_a(i,j)$ of baseline model, s_{m_i}—sentence, which mention i belongs to.

The graphical illustration of SCRb model is shown in **Fig. 3**.

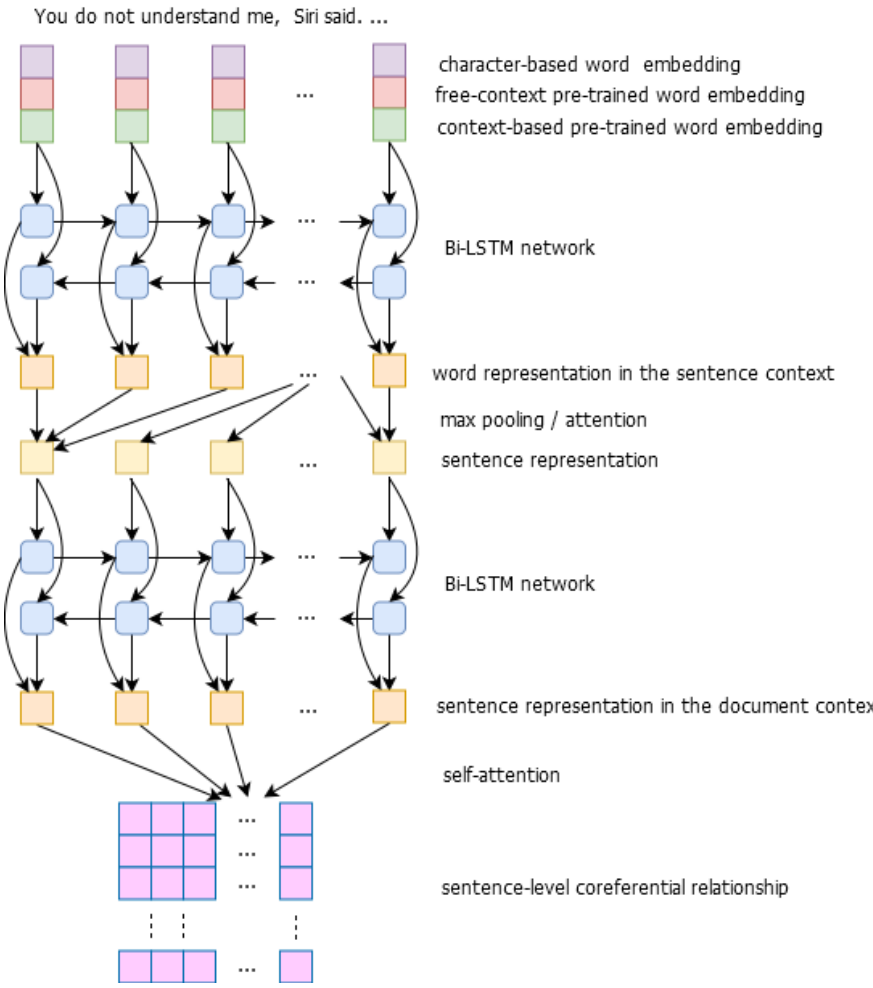


Figure 3: Sentence-level coreferential realtionship-based model

2.3. Model based on Language Modeling

Pretrained language models, such as ELMo [9], GPT [11], BERT [2], showed to be very effective in wide range of tasks from text classification to question answering. ELMo has been already tested on the task of coreference resolution for English language and helped to achieve new state-of-the-art performance [3] on CoNLL-2012 shared task dataset. Pretrained language models are usually used as a provider of contextualized word embeddings instead of usual word embeddings like w2v [8]. Contextualized word embeddings can be computed as weighted sum of outputs from each layer of language model, weights in this sum are trainable parameters, e.g., one scalar variable for word embedding layer and two scalars for BiLSTM layers are trained for

ELMo, three parameters in total. BERT-base model is a 12-layer Transformer network and we did experiments with 1–6–12 and 10–11–12 layers outputs. The Higher-order Coreference Resolution model [3] uses two types of embeddings: word embeddings and contextualized word embeddings. We experimented with contextualized word embeddings from ELMo and BERT models trained for Russian Language (RuBERT).

3. Experiments and Results

3.1. Evaluation Metrics and Datasets

We did our experiments with three datasets, one for English language—CoNLL 2012 Shared Task[1] [10] and two datasets for Russian language: RuCor [15] from Dialogue-21 2014 Shared Task[2] and AnCor from Dialogue-21 2019 Shared Task.[3] As shown in **Table 1**, Russian datasets are 7–10 times smaller than the English one. This makes the CR task for Russian even harder.

Table 1: Coreference resolution datasets. Mentions and chains number computed for train + dev + test sets

Datasets	Language	Mentions	Chains
CoNLL 2012 Shared Task [10]	En	194,480	44,221
RuCor [15]	Ru	16,558	3,638
AnCor	Ru	28,961	5,678

There are three most common metrics for coreference resolution: MUC, B-cube, CEAF [7]. Overall coreference resolution systems performance is usually computed as averaged F-1 measure of these three metrics.

3.2. Experiments details

We used TensorFlow[4] to implement all models in our experiments. We took ELMo and RuBERT[5] models for Russian Language from DeepPavlov library[1]. For experiments with Russian language we used only raw texts without any additional features (like speaker id, morphological tags, etc) or pre-processing steps. All we have to do is to transform mention clusters in the original datasets to binary matrices representing the sentence relationships (as shown in **Fig. 2**).

1 http://conll.cemantix.org/2012/data.html

2 http://www.dialog-21.ru/evaluation/2014/anaphora/

3 http://www.dialog-21.ru/evaluation/

4 https://www.tensorflow.org/

5 http://docs.deeppavlov.ai/en/master/components/bert.html

All experiments were run on GeForce GTX 1080 Ti. The average training time is about one day for Russian datasets and one day and a half for the English one.

3.3. Results

In the first batch of experiments, information about the sentence-level coreferential relationship was supposed to be known before. In other words, we want to evaluate how sentence-level coreferential relationship affects the model performance. To do this, the baseline model is trained on two kinds of datasets: 1) the original OntoNotes 5.0; 2) the OntoNotes 5.0 with sentence-level coreferential relationships. The OntoNotes 5.0 with coreference chains was released as a part of CoNLL 2012 Shared Task. The experiment results pointed out that the information about the sentence relationships is a very useful feature for the coreference resolution task. If this feature is provided with 91% of accuracy the model performance can be boosted by about 2.5%. Under the ideal condition, when training with groundtruth sentence-level coreferential relationship, the model performance can be as large as 78.84% (refer Table 2 for more details).

Table 2: Effect of sentence-level coreferential relationship on the baseline model performance

Dataset	Max. F1 on the dev. set
Original OntoNotes 5.0	73.00
OntoNotes 5.0 + sent.-level coref. relationship with 20% of noise	74.13
OntoNotes 5.0 + sent.-level coref. relationship with 16% of noise	74.73
OntoNotes 5.0 + sent.-level coref. relationship with 9% of noise	75.56
OntoNotes 5.0 + sent.-level coref. relationship with 6% of noise	76.36
OntoNotes 5.0 + sent.-level coref. relationship with 3.5% of noise	77.01
OntoNotes 5.0 + sent.-level coref. relationship with 1.5% of noise	77.92
OntoNotes 5.0 + groundtruth sent.-level coref. relationship	78.84

Results on AnCor and RuCor datasets were obtained by averaging results across 10-folds. RuBERT(1–6–12) with features from 1–6–12 layers showed better performance than RuBERT(10–11–12) in our preliminary experiments. In some experiments on AnCor dataset we also used RuCor dataset as additional training data (+ RuCor in Tables 5 and 6). Sentence-level information showed to be useful on RuCor dataset, it outperforms baseline model with about 1 F-1 point.

We tested our models in two settings:

- Gold mentions—uses gold mention boundaries and builds coreference chains (Tables 3 & 5).
- Full pipeline—includes mentions extraction from texts and building coreference chains (Tables 4 & 6).

Table 3: Results on RuCor dataset, gold mentions

Model	*muc*	*bcube*	*ceafe*	*avg.* F_1
Sysoev [13]	69.28	63.12	55.33	62.58
Toldova [15]	70.25	60.14	—	—
Baseline + ELMo	90.54	79.71	67.81	79.36

Table 4: Results on RuCor dataset, full pipeline

Model	*muc*	*bcube*	*ceafe*	*avg.* F_1
Sysoev [13]	41.90	34.30	29.06	35.10
Baseline + ELMo	67.26	52.29	53.18	57.58
SCRb	66.32	54.09	54.86	58.42

Table 5: Results on AnCor dataset, gold mentions

Model	*muc*	*bcube*	*ceafe*	*avg.* F_1
Baseline + ELMo	90.22	83.41	59.44	77.69
Baseline + RuBERT(1–6–12)	91.04	84.38	63.07	79.50
Baseline + ELMo + RuCor	91.51	84.16	61.33	79.01
Baseline + RuBERT(1–6–12) + RuCor	91.47	84.49	63.81	79.92

Table 6: Results on AnCor dataset, full pipeline

Model	*muc*	*bcube*	*ceafe*	*avg.* F_1
Baseline + ELMo	50.29	48.89	46.99	51.72
SCRb	60.00	48.89	50.39	53.61
Baseline + RuBERT(1–6–12)	60.95	51.08	49.24	53.76
Baseline + ELMo + RuCor	65.01	52.67	50.19	55.96
Baseline + RuBERT(1–6–12) + RuCor	66.74	54.88	51.72	57.78

4. Discussions and Conclusions

In this paper, we presented a new approach to the task of coreference resolution with focus on Russian language. The previous models often address coreference resolution task in two stages: 1) detect all mention candidates, 2) cluster them into chains. We instead build a model to extract the sentence relations in the coreference context. This idea stems from an attempt of achieving sentence-level coreferential relationships to deal with the long term dependency. However, so far, the performance of the SCRb model is not really impressive. By analyzing the weights of the trained model, we found that the combined model tends to ignore the features learned by SCRb model. Hence, we claim that a part of the reason may lie in the way we combine the SCRb model with the baseline model. One more reason is the class imbalance problem that occurs when

transforming mention clusters from original datasets to binary matrices. Although we used a weighted loss function that gives more importance to the minority classes, the problem has not been solved thoroughly. However, this model still has promising potentials to be applied to not only the CR task but also other NLP tasks such as Question Answering as well as Text Summarization. The experiment mentioned in the beginning of the **Section 3** shows that if the quality of the sentence-level coreferential feature is good enough, it can significantly boost the model performance.

In conclusion, we propose a new model that is able to learn the sentence-level coreferential relationships. In addition, we used two cutting edge language representation models (ELMo, BERT) to boost our model for Russian language. Our experiments and results on the shared tasks at Dialogue conference showed that our model achieved state-of-the-art performance on Russian language.

Acknowledgements

This work was supported by National Technology Initiative and PAO Sberbank project ID 0000000007417F630002.

References

1. *Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, et al.* Deeppavlov: Open-source library for dialogue systems. In Proceedings of ACL 2018, System Demonstrations, pages 122–127, 2018.
2. *Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.* Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
3. *Luheng He Kenton Lee and Luke Zettlemoyer.* Higher-order coreference resolution with coarse-to-fine inference. In Proceedings of NAACL-HLT 2018, pages 687–692, 2018.
4. *Mike Lewis Kenton Lee, Luheng He and Luke Zettlemoyer.* End-to-end neural coreference resolution. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 188–197, 2017.
5. *Chirstopher D.* Manning Kenvin Clark. Improving coreference resolution by learning entity-level distributed representations. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 643–653, 2016.
6. *Christopher D. Manning Kevin Clark.* Deep reinforcement learning for mention-ranking coreference models. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2256–2262, 2016.
7. *Xiaoqiang Luo.* On coreference resolution performance metrics. In Proceedings of the conference on human language technology and empirical methods in natural language processing, pages 25–32. Association for Computational Linguistics, 2005.

8. *Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean.* Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119, 2013.
9. *Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer.* Deep contextualized word representations. In Proc. of NAACL, 2018.
10. *Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang.* Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In Joint Conference on EMNLP and CoNLL-Shared Task, pages 1–40. Association for Computational Linguistics, 2012.
11. *Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever.* Improving language understanding with unsupervised learning. Technical report, Technical report, OpenAI, 2018.
12. *Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang.* Disan: Directional self-attention network for rnn/cnn-free language understanding. CoRR, abs/1709.04696, 2017.
13. *A. A. Sysoev, I. A. Andrianov, and Khadzhiiskaia A. Y.* Coreference resolution in russian: State-of-the-art approaches application and evolvement. In Computational Linguistics and Intellectual Technologies. International Conference "Dialogue 2017" Proceedings, pages 317–338, 2017.
14. *S. Toldova, A. Roytberg, A. A. Ladygina, M. D. Vasilyeva, I. L. Azerkovich, M. Kurzukov, G. Sim, D. V. Gorshkov, A. Ivanova, A. Nedoluzhko, and Y. Grishina.* Evaluating anaphora and coreference resolution for russian. In Komp'juternaja lingvistika i intellektual'nye tehnologii. Po materialam ezhegodnoj Mezhdunarodnoj konferencii Dialog, pages 681–695, 2014.
15. *Svetlana Toldova and Ionov Maxim.* Coreference resolution for russian: The impact of semantic features. In Computational Linguistics and Intellectual Technologies. International Conference" Dialogue 2017" Proceedings, pages 339–349, 2017.

ЯЗЫКОВЫЕ МЕХАНИЗМЫ РАСШИРЕНИЯ СОЧЕТАЕМОСТИ: СОЧЕТАЕМОСТЬ ЧАСТИЦЫ *-КА*[1]

Левонтина И. Б. (irina.levontina@mail.ru)
ИРЯ им. В. В. Виноградова РАН, Москва, Россия

RELAXING COOCCURRENCE RESTRICTIONS: THE DISTRIBUTION OF THE RUSSIAN PARTICLE '*-KA*'

Levontina I. B. (irina.levontina@mail.ru)
RLI RAS, Moscow, Russia

The development of corpus linguistics quite often makes it necessary to revisit the items studied and comprehensively described in the "pre-corpus" epoch. As a result we obtain a more voluminous or even radically different picture of their functioning. This is especially true of linguistic units with bizarre compatibility, in a complex way motivated by their semantics, such as the Russian particle *-ka*. It is a study of a large array of linguistic data that makes it possible to notice relatively rare, but regularly arising types of combinations that reveal the semantic potential of this particle. In the present work, we used the Russian National Corpus, as well as Yandex search, which allowed us to assess if this or that type of combination is relevant for nowaday live speech. The study of corpus data not only contributes to our understanding of the properties of linguistic units — in this case, the distribution of a particle, but also makes it possible to observe the linguistic mechanisms involved in relaxing cooccurrence restrictions. Thus, the analysis of the corpus material allowed us to find two fairly common, but very nontrivial types of combinations of *-ka* with non-imperative expressions: *лучше-ка* and *знаешь-ка/знаете-ка*. As we show, their occurrence is due to the effect of completely different linguistic mechanisms.

Key-words: Russian particles, semantic potential, Imperative, relaxing cooccurrence restrictions, corpus studies

[1] Работа выполнена при поддержке РФФИ, Грант № 19-012-00505 «Дискурсивные механизмы русского языка: семантика и механизмы прагматикализации»; рук. Анна А. Зализняк.

1. Вводные замечания

Развитие корпусной лингвистики нередко заставляет возвращаться к единицам, исследованным в «докорпусную» эпоху, чтобы получить более объемную, а иной раз и кардинально иную картину их функционирования. Особенно это касается языковых единиц с причудливой сочетаемостью, сложным образом мотивированной их семантикой, — таких, как частица *-ка*. Именно обследование большого массива языковых данных дает возможность выявить относительно редкие, но регулярно возникающие типы сочетаний, раскрывающих семантический потенциал частицы. В настоящей работе мы пользовались поиском НКРЯ, а в дополнение к корпусным данным — данными Яндекса, позволяющими оценить, насколько тот или иной тип сочетаний актуален для современной живой речи.

Приимперативная частица *-ка* неоднократно становилась объектом лингвистического исследования. В книге В. С. Храковского и А. П. Володина 1986 г. «Семантика и типология императива: Русский императив» содержится обстоятельный обзор точек зрения на частицу *-ка* в лингвистической литературе, начиная с К. С. Аксакова (в частности, у Р. О. Якобсона, В. В. Виноградова и др.). Сами авторы предлагают следующее описание семантики *-ка*: «Употребляя частицу *-ка*, говорящий как бы делает вид, что он игнорирует реальные этикетные отношения, связывающие его со слушающим: он дает понять слушающему, что обращается к нему как к человеку, с которым его связывают внеэтикетные, неформальные и непринужденные отношения. По своей социальной роли говорящий при этом обычно выше слушающего, но возможны случаи, когда говорящий обращается как к равному к слушающему, чья социальная роль выше его собственной» [Храковский, Володин 1986: 183]. Еще ранее на значимость «социальнго» компонента в значении *-ка* указывалось в [Крысин 1983]. В разных описаниях семантика *-ка* передается при помощи признаков непринужденности, фамильярности, а также усиления/смягчения императивности. Особенно интересно, что лингвисты не могут прийти к согласию в вопросе о том, смягчает –*ка* императив или, напротив, усиливает его. Более того, в некоторых случаях прямо указывается, что *-ка* имеет или первое значение, или второе: «Широко используется *-ка* при глаголах повелительного наклонения, придавая волеизъявлению, выраженному ими, различные модально-экспрессивные оттенки: подчеркивает категоричность приказания, или, напротив, смягчает категоричность повеления, или придает ему характер дружеского увещевания, совета, иронии, насмешки т. п.» [Киселев 1976: 122]. В [Левонтина 1991] отмечается, что в значение *-ка* входят две очень важных идеи: что желание возникло **только что** (эта идея высказана также в [Кобозева 1990][2]) и что осуществить его нужно **немедленно**. Эти компоненты во многом

[2] В 1989 г., ровно 30 лет назад, были прочитаны два доклада — И. М. Кобозевой на совещании группы «Логический анализ языке» под руководством Н. Д. Арутюновой о семантике модальных частиц (в связи с проблемой аномальности текста) и мною — в Лексикографическом семинаре Ю. Д. Апресяна, где шла работа над проектом Интегрального словаря русского языка; в моем докладе была представлена словарная статья частицы *-ка* в формате Интегрального словаря. В этих работах И. М. Кобозева и я независимо

отпределяют и сочетаемость частицы (например, сочетания с вводным словом *кстати*, которое указывает на отклонение от темы или на неожиданно возникшую мысль, несочетаемость с частицами, маркирующими повторные просьбы и требования), и ее рематизирующую функцию, и прагматические эффекты, возникающие в результате взаимодействия *-ка* с иллокутиваными характеристиками высказывания.

2. Сочетаемость частицы *-ка*

Энклитика *-ка* имеет довольно ограниченную сочетаемость с точки зрения того, к каким типам единиц она может примыкать. В «докорпусную» эпоху эта сочетаемость была подробно описана. *-Ка* стандартно употребляется как энклитика со следующими формами и словами: с морфологическим императивом (*иди-ка, идите-ка*); с инклюзивным (квази)императивом (*идем-ка, пойдемте-ка*); с формами прошедшего времени с императивным значением (ограниченно, только с формами глагола *пойти*: *Пошел-ка / Пошла-ка ты отсюда*; *Пошел-ка я домой*); с первым лицом будущего времени (*Пойду-ка я / Пойдем-ка мы домой*); с частицами *пусть, ну, дай, давай*; с междометиями *ну, на* и *нате, слышь*; с вводным словом *поди*. Сочетаемость *-ка* сильно лексикализована: вовсе не все единицы с побудительным значением могут присоединять *-ка* [Левонтина 1991: 137–138]. Однако сейчас, когда стали доступны обширные корпусные данные, представления о сочетаемости *-ка* могут быть уточнены[3].

Так, оказалось, что –*ка* сочетается с еще некоторами единицами с побудительным значением — например, частицей *айда*:

(1) — *Ну ты, раскисла от сладости?* ***Айда****-ка собери нам чего-нибудь… Угостим, что ли, блудного сына!* [М. Горький. Фома Гордеев (1899)]

(2) — *Вояка тоже нашелся… Черт их тут разберет… Пискунов-то полна изба…* ***Айда-ка****, айда домой.* [А. Веселый. Россия, кровью умытая (1924–1932)]

(3) — *Ну, времени у меня в обрез.* ***Айда-ка*** *к москвичам. Посмотрим новинки механизации…* [Георгий Радов. Час и рубль // «Огонек». № 15, 1959]

(4) ***Айда ка*** *делай!!* ***айда ка*** *делай!! наше лето.*
https://muzonoff.online/mp3/айда+ка+делай%21%21

сделали отчасти совпадающие наблюдения по поводу частицы *-ка* (в частности, это касается идеи новизны и сочетаемости с разными типами побудительных речевых актов). Результаты были опубликованы в 1990 и 1991 гг. соотвественно.

[3] Мы не ставим перед собой задачу описать весь языковой материал, который можно извлечь из корпусных данных. Так, в этом материале большое количество контекстов, имеющих ярко выраженную региональную и диалектную специфику. Этот аспект мы здесь совсем не затрагиваем. Задача настоящей работы — выявить языковые механизмы расширения сочетаемости.

Широко представлены в материале и сочетания *–ка* с устаревшими частицами *полно* и *полноте*, также с побудительной семантикой:

(5) *— Нет, я спрошу, не нужно ли что. —* ***Полноте-ка!*** *посмотрите, на дворе мгла какая!* [М. Е. Салтыков-Щедрин. Мелочи жизни (1886–1887)]

(6) *Отец приблизился к жене, опустился рядом на скамью, обнял ласково, пытаясь ее «разговорить»: —* ***Полно-ка****, матушка. Голиафа мы с тобой не породили.* [С. Т. Григорьев. Александр Суворов (1939)]

(7) *Он бесстрашно врезался в толпу, всячески стал успокаивать народ: —* ***Полно-ка****, полно, друзья мои. Опомнитесь да подумайте, что это вы затеяли…* [В. Я. Шишков. Емельян Пугачев. (1934–1939)]

(8) ***Полноте-ка****, Клим Климыч, пустое это, с кем греха не бывает.* [Д. А. Фурманов. Драма Луши (1924)]

Можно привести и другие подобные факты. Однако интереснее другое. Исследование большого массива языковых данных не только позволяет дополнить наши представления о сочетаемости языковых единиц — но и дает возможность наблюдать действие языковых механизмов, участвующих в расширении сочетаемости языковой единицы. Так, анализ корпусного материала позволил обнаружить два довольно распространенных, но весьма нетривиальных типа сочетаний *-ка* с **непобудительными** единицами: *лучше-ка* и *знаешь-ка/знаете-ка*.

Рассмотрим их подробнее. Сами по себе *лучше* и *знаешь* типичны для контекстов с частицей *-ка*:

(9) *—* ***Знаете*** *ли что, будемте-****ка лучше*** *говорить потише, а то ведь здесь, поди-ка, и стены уши имеют…* [В. В. Крестовский. Панургово стадо (1869)]

3. *Лучше-ка*

Лучше в роли дискурсивной частицы часто используется в побудительных высказываниях[4] (в том числе высказываниях со значением самопобуждения (*Лучше даже не начинай!; Я лучше пойду*). *Лучше* указывает на то, что ситуация допускает разные варианты действий и говорящий побуждает избрать именно данный вариант:

(10) *Наконец терпение мое лопнуло; я подошел к ним и с убийственным выражением сказал: «Ваши превосходительства, сядемте-****ка лучше*** *в карты!»* [С. Т. Аксаков. История моего знакомства с Гоголем (1856)]

(11) *— Знаете ли что, будемте-****ка лучше*** *говорить потише, а то ведь здесь, поди-ка, и стены уши имеют…* [В. В. Крестовский. Панургово стадо (1869)]

[4] Это не обязательно; ср. *Да он учше удавится, но не заплатит!*

(12) *А потом позвонила свинья: / — Пришлите ко мне соловья. / Мы сегодня вдвоём с соловьем / Чудесную песню споём. / — Нет, нет! Соловей / Не поёт для свиней! / Позови-**ка** ты **лучше** ворону!* [К. Чуковский, Телефон]

По значению *лучше* хорошо сочетается с *-ка* (получается гармоничное сочетание новизны и контраста). Это отмечалось в [Левонтина 1991].

На новом материале было обнаружено, что довольно часто *-ка* примыкает и к самому *лучше*, хотя последнее и не принадлежит к числу единиц с побудительным значением:

(13) *— И охота вам время терять? Поедемте **лучше-ка** пообедаем вместе, а потом, если желаете, отдохнём.* [А. Слаповский. Гибель гитариста (1994–1995)]

(14) *Мне ни разу больше не привелось слышать или видеть Самойлова в разлуке с его иронией, с его элегантным сарказмом… и довольно об этом. **Лучше-ка** припомню еще два более «типичных случая».* [В. Смехов. Театр моей памяти (2001)]

(15) *Только ругаешься. **Лучше-ка** вот я тебе объясню. Ведь мы, Сеня, такие деньги получаем — ты их нигде не заработаешь: ни на заводе, ни в колхозе.* [Г. Владимов. Три минуты молчания (1969)]

(16) *— Вы **лучше-ка** вот что, — сказал Пантелеев, — снимите предохранительные сетки с фар.* [К. Симонов. Так называемая личная жизнь (1956–1965)]

(17) *Все эти вопросы — чепуха! Вы **лучше-ка** отгадайте, как человек полетит на Луну.* [М. Баранова, Е. Велтистов. Тяпа, Борька и ракета (1962)]

Причем это явление абсолютно не новое:

(18) *— Полно, душенька, эрфиксы-то выпускать, — произнес он, — с старыми-то приятелями эдак не встречаются. Вот **лучше-ка** по душе, запросто, без закорючек, обнимемся и поцелуемся.* [И. И. Панаев. Опыт о хлыщах (1854–1857)]

(19) *Ахъ, вообще — отстань отъ меня! **Лучше-ка** сядь на этот чемодан, — добавила она нотой ниже.* [В. В. Набоков. Подвиг (1932)]

(20) *— Не очень берегись, это не Великий Истребитель идет, а **лучше-ка**, чудак, иди-**ка** к кочке* [М. М. Пришвин. Дневники (1929)]

В современной устной речи сочетание *лучше-ка* можно услышать довольно часто. Ср. также следующие примеры из интернет-коммуникации:

(21) *…И так Вы — просто хороши!Сходите, **лучше-ка**, по делу* wplanet.ru/mobile/files/index.php…

(22) ***Лучше-ка**, Таня, попусту ты не плачь (Все «Спорттовары» мира полны мячами)* socratify.net/quotes/alan-ebbot/136451

(23) *@NavalnyMaster **лучше-ка** забаним вас, потому что вы не только дурак, но и хам.* twitter.com/besttoday_ru/status/590473197390299136

Как кажется, механизм здесь вполне ясен: происходит чисто техническое смещение *-ка* от собственно побудительного слова к сопутствующему *лучше*: ср. *Лучше забаним-ка вас = Лучше-ка забаним вас*.

При этом существенно следующее. У *-ка* нет жестких требований к расположению во фразе. Оно может примыкать и к слову, находящемуся в конце фразы; ср.

(24) — *Вадька! Осциллограф* ***включи-ка!*** [А. Солженицын. В круге первом, т. 1, гл. 1–25 (1968) // «Новый Мир», 1990]

(25) *Ты, Кира Петровна, отсюда* ***уйди-ка****, я сама управлюсь*. [И. Грекова. Перелом (1987)]

(26) *Бумажку мне свою сюда* ***дай-ка****, да?* [А. Петров. Воплощение мысли (2003) // «Вслух о...», 2003.08.04]

Однако некоторое тяготение к ваккернагелевской позиции у *-ка* сохраняется, что видно по большинству приведенных выше примеров.

Лучше же может занимать позицию как в начале фразы, так и в конце, как перед предикатом, так и после него. Заметим при этом, что в приведенных примерах *лучше-ка* предшествует побудительному слову. Правда, это тоже лишь тенденция, исключения нечасто, но бывают:

(27) *Выйду* ***лучше-ка*** *я по хозяйским делам, Голос жизни послушать и свой им отдам*. lit-salon.ru/Стихи/...-vselennoi-42356.html

Правда, здесь перед нами стихотворный текст, где возможен и не очень естественный для живой речи порядок слов.

В целом можно предположить, что *лучше*, которое сопровождает побудительные высказывания и ассоциируется с ними, довольно легко может перетягивать на себя *-ка*, особенно в случаях, когда *лучше* предшествует побудительному предикату.

Осмысление корпусного материала позволяет сделать еще одно предположение. Сочетание *лучше-ка* встречается не слишком часто, однако примеры попадаются уже с середины 19 в. Можно думать, что выражение *лучше-ка* существует в языке скорее не в готовом виде, а как возможность смещения *-ка* с «законного» носителя — побудительного слова — на стандартный элемент побудительного контекста. Интересно, что этот сдвиг относится к тем явлениям разговорного синтакисиа, которые почти не осознаются говорящими: люди употребляют сочетание *лучше-ка*, при этом нередко считая, что «так не говорят».

Можно отметить, что не только *лучше* как дискурсивная частица, но и сам по себе компаратив, типичный для побудительных высказываний, в принципе, может перетягивать на себя *-ка*; ср. пример из Николая Рубцова, на который нам указал рецензент «Диалога»: *А ну, поближе-ка / иди к сосне! / Ах, сколько рыжиков! / Ну как во сне...* Правда, такие сочетания не так сводобно используются, как *лучше-ка*. Во всяком случае, НКРЯ не дает других сочетаний *-ка* с компаративом, кроме *лучше-ка*, не считая забавного паразитического результата *подлей-ка*.

Судя по всему, в языке есть ресурс для расшатывания ограничений на сочетаемость *-ка* путем перемещения частицы по фразе. Ср. следующий игровой пример, в котором *-ка* как бы распространяется по всему предложению:

(28) *Anonymous. Дата: 17.04.18 <…> А давайте-**ка лучше-ка** свою-**ка** фотку.* eva.ru/96871967.htm

Но, конечно, в этом случае *свою-ка* представляет собой чисто окказиональное контекстное образование, в отличие от практически стандартного для разговорной речи *лучше-ка*.

4. *Знаешь-ка*

Совсем иная ситуация с сочетаниями *знаешь-ка* и *знаете-ка*.

Обороты с формами *знаешь* и *знаете* очень интересны. Ю. Д. Апресян в статье «Знать» Активного словаря русского языка отмечает, что этот глагол: «в форме 2-Л НАСТ с последующим придаточным дополнительным имеет следующие сдвинутые значения:

а) ‘Говорящий сообщает, что сейчас он скажет нечто, что он считает новым и интересным для своего собеседника’ [иногда с местоимением что]: *Знаешь, лицо как газон. Бывает старый, но ухоженный, а бывает старый и неухоженный* (С. Спивакова); *А знаешь что, зайдем к нему сейчас, он тут за парком* (Ю. Домбровский);

б) ‘Говорящий выражает свое недовольство происходящим или адресатом, которого он считает виновником происходящего’ [часто с местоимением что]: *— Знаешь, Тимоша, — в сердцах проговорил я, — шел бы ты куда подальше* (В. Белоусова); *Знаешь что, давай не шутить на эти темы, ладно?* (А. Слаповский)» [Апресян 2017].

Итак, *знаешь* и *знаете* могут служить для привлечения внимания к тому, что, по мнению говорящего, неочевидно и, возможно, даже неожиданно. Рассмотрим следующий диалог:

(29) *а. — Ты ведь не будешь с этим спорить?*
— Знаешь, буду!

Здесь нейтрально было бы — *Не буду*, и эта реплика соответствовала бы ожиданиям. А ответ — *Буду!* не ожидается, и естественно сопроводить такой ответ каким-то маркером типа *а вот и, как раз, еще как* или *знаешь*.

Любопытно, что возможен и такой диалог:

(29) *б. — Ты ведь не будешь с этим спорить?*
— Знаешь, не буду!

Здесь *а вот и, как раз* невозможны, поскольку они чересчур полемичны и возникало бы противоречие с подтверждением. А реплика — *Знаешь, не буду!* подразумевает, что заранее ответ был не очевиден самому отвечающему, но, услышав вопрос, он подумал и решил согласиться. Любопытно, что такой ответ

звучит не менее полемично, чем ответ без отрицания, но оспаривается здесь то, что спрашивающему ответ казался очевидным (на это указывает частица *ведь*).

Сочетания со формами *знаешь* и *знаете* часто используются как самостоятельные эмоциональные реплики, без всякого продолжения: *Ну, знаешь (ли)!; Знаешь, что!* Говорящий как бы начинает возмущенную тираду, но потом бросает, не находя слов.

Знаешь и *знаете* часто фигурируют в контексте *-ка*, поскольку присущая им функция привлечения внимания хорошо гармонирует с заключенной в *-ка* идеей новизны.

(30) ***Знаете**, что я вас попрошу? Оставьте-**ка** мне это дня на три, ведь он вам их не на один день дал, верно? — Да, конечно, берите, — сказал Корнилов* [Ю. О. Домбровский. Факультет ненужных вещей, часть 3 (1978)].

(31) *А. М, нахмурился, подумал и взялся за телефонную трубку. — Позвоню-**ка** я, **знаете**, Кольцову.* [А. Мильчин. В лаборатории редактора Лидии Чуковской // «Октябрь», 2001]

Но оказалось, что и само *знать* в этой функции хорошо сочетается с *-ка*, хотя в целом формы второго лица *-ка* не присоединяют. Такие сочетания часто встречаются в том числе и в современной живой речи, в частности образуя самодостаточные возмущенные реплики вида *Знаешь-**ка** что!* Ср. пример из современной сетевой литературы:

(32) *— **Знаешь-ка что**, замполит? — чуть привстав, не своим голосом произнес Антон, глядя в лицо своему бывшему заместителю по политической части.* читать-онлайн.com.ua/reedbook...

Часто *знаешь-ка* фигурирует в контексте побудительных высказываний:

(33) *Сын управляющего фабрикой <…> сказал ему утром: — **Знаешь-ка**, я придумал славную штучку, пойдем. И они пошли вниз под обрыв, прямо к реке.* [Л. А. Чарская. Золотая рота (1911)]

(34) *— **Знаешь-ка**, **что**. Добавься ко мне в аську.* proza.ru/2016/10/15/137

(35) *— Чахохбили из кур… Не верю, как говорил Станиславский. Нарубят костей и зальют томатом. **Знаешь-ка**, друг, вот что. Сделай нам по рыбной соляночке и по хорошему куску мяса…* [М. Кураев. Записки беглого кинематографиста // «Новый Мир», 2001]

На основании таких примеров можно было бы думать, что механизм расширения сочетаемости *-ка* здесь тот же, что и в случае с *лучше* — перенос энклитики от «законного» носителя к другому слову: *Знаешь-ка, добавься ко мне в аську = Знаешь, добавься-ка ко мне в аську.*

Однако это не так. Материал показывает, что *знаешь-ка/знаете-ка* вполне свободно используется в совершенно не побудительных контекстах:

(36) *— Ведь эта затея, **знаете**-ка, даже пожалуй, и удивительная. И Иван Яковлевич задумался о том, что вот и ему пришлось вдруг удивиться* [Е. А. Салиас. Аракчеевский подкидыш (1889)]

(37) — ***Знаешь-ка что?*** *— сказал майор Декстер Смит осьминогу. — Ты сегодня получишь истинное наслаждение, если мне удастся кое-кого поймать.* online-knigi.com/Читать книгу/71059

(38) *А* ***знаете-ка что****. Решила вести неделю супер правдивый ЖЖ. Вот прям досконально. Как чую. Каждую мысль без редакторства излагать* gramino.com/instagram/alena____fox/photo/...

(39) *А* ***знаете-ка что****... У меня тут по соседству, в Малом Гнездниковском, кум служит... — Да что вы! В самом деле, какая удача!* velib.com/read_book/akunin_boris...tom...naka_no...chto...

(40) ***Знаете-ка что****, девушка.. Если бы были хоть ЧУТОЧКУ разумны, то вы бы никогда не назвали человека животным с высоты, как вам кажется, своего превосходства над ними только на том основании, что вы верите в Бога*. otvet.mail.ru/question/192926424

Мы предполагаем, что здесь работает совсем другой механизм. Хотя сочетаемость *-ка* достаточно жестко лексически и грамматически ограничена, все же существует возможность ее семантического расширения за счет неканонических способов выражения побудительности. Приведем в качестве примера еще один ранее не отмечавшийся тип побудительных сочетаний с *-ка*, который встречается как раз в современном материале — сочетание *к черту-ка*:

(41) *Поднималась... зачем? А* ***к черту-ка*** *жизнь...не врозь и не рядом... Я билась в открытые двери, оставшись никем, Убитая насмерть фальшивым снарядом.* stihi.ru/2007/09/11/1791

(42) *а* ***к черту-ка*** *всё, нет, ну правда. и что мне неймется? мы хорошие, нет, мы лучшие друзья, это намного лучше, чем что-либо другое, это вернее, это надежней*. liveinternet.ru/users/2222888/page7.html

Выражение *к черту* имеет побудительное значение, поэтому сочетание его с *-ка* нестандартно, но вполне органично.

Ср. также необычный, но очень показательный пример:

(43) *— Слушай, ей плохо.* (Достает таблетки из сумки) ***Воды-ка!*** (Дима уходит) *Выпейте таблетку. Выпейте таблеточку.* [Л. С. Петрушевская. Я болею за Швецию (1977)]

Форма родительного падежа *воды* стандартно используется для выражения просьбы срочно дать попить воды, то есть для выражения побудительного значения. Отсюда и возможность окказионального сочетания с *-ка*. Нечто подобное происходит и в нашем случае. *Знаешь/знаете* используются для привлечения внимания, то есть выполняют своего рода мобилизующую функцию. И это дает основания для расширительного понимания этих единиц как побудительных, что, в свою очередь, открывает возможность для присоединения к ним энклитики *-ка*.

References

1. *Apresjan* (2017). Active dictionary of the Russian language [Aktivnyy slovar' russkogo yazyka]. V. 3 / V. Yu. Apresjan, Yu. D. Apresjan, E. E. Babaeva and others. «Nestor-Story» Moscow, St. Petersburg, 2017. — 768 p.
2. *Khrakovsky V. S., Volodin A. P.* (1986). Semantics and typology of the imperative: Russian Imperative [Semantika i tipologiya imperativa: Russkiy imperativ]. L.: Science.
3. *Kiselev I. A.* (1976). Particles in modern East Slavic languages [Chastitsy v sovremennykh vostochnoslavyanskikh yazykakh]. Minsk: BSU Publishing House.
4. *Kobozeva I. M.* (1990) Pragma-semantic Deviations and Modal Pariicles [Pragmasemanticheskaya anomal'nost' vyskazyvaniya i semantika modal'nykh chastits] // Logicheskiy analiz yazyka. Protivorechivost' i anomal'nost' teksta. Moscow, Nauka, pp. 194–203.
5. *Krysin L. P.* (1983). On the "social" component of lexical meanings [O «sotsial'nom» komponente leksicheskikh znacheniy] Wiener Slawistischer Almanach. Bd. 11, 1983. pp. 169–187.
6. *Levontina I. B.* (1991). Lexical entry of the particle *-ka* [Slovarnaya stat'ya chastitsy -ka] Semiotics and Informatics [Semiotika i informatika], vol. 32, Moscow, 136–140.
7. *Vinogradov, V. V.* (1938). Modern Russian language. Grammatical doctrine of the word [Sovremennyy russkiy yazyk. Grammaticheskoye ucheniye o slove]. Issue 1 M.: State. training ped. Publishing house of the People's Commissariat of Education of the RSFSR.

ДОСТАЛИ ТАК УПОТРЕБЛЯТЬ ИНФИНИТИВ! О НОВОЙ КАУЗАТИВНОЙ КОНСТРУКЦИИ В РУССКОМ ЯЗЫКЕ[1]

Левонтина И. Б. (irina.levontina@mail.ru)
ИРЯ им. В. В. Виноградова РАН, Москва, Россия

Полинская М. С. (mpolinsk@gmail.com)
Национальный исследовательский университет «Высшая школа экономики», Москва, Россия; профессор школы лингвистики факультета гуманитарных наук

DRIVING US CRAZY WITH YOUR INFINITIVES! THE RISE OF A NEW CAUSATIVE CONSTRUCTION IN RUSSIAN

Levontina I. B. (irina.levontina@mail.ru)
RLI RAS, Moscow, Russia

Polinsky M. S. (mpolinsk@gmail.com)
National Research University Higher School of Economics, Moscow, Russia; Professor: Faculty of Humanities / School of Linguistics

Russian has an impressive set of psych-verbs with the general meaning of causing extreme irritation and exhausting one's patience, which we will henceforth refer to as EXASPERATE-verbs: *достать*; *задолбать, заколебать, замучить, бесить*, etc. With these predicates, the experiencer is in the accusative, and the non-salient, inanimate or abstract causer of irritation can be expressed by a noun phrase in the nominative, or by an infinitival clause, e. g., *Меня достало это выражение/разбирать эти выражения*. In addition, these verbs participate in another causative construction, with a salient, agentive causer expressed by a noun phrase in the nominative case, and the manner in which irritation is brought about expressed by the instrumental phrase, with or without a preposition: *Ты меня*

[1] Исследование осуществлено в рамках Программы фундаментальных исследований НИУ ВШЭ в 2019 году (в части работы М. Полинской) и поддержано РФФИ, грант 19- 012-00291А, «Подготовка четвертого выпуска Активного словаря русского языка» (в части работы И. Левонтиной).

достал (*с*) этими *выражениями.* In modern spoken Russian, we also find a new agentive causative construction (NACC): *Ты меня достал ныть!* 'You drive me up the wall by your whining.' The NACC is colloquial and is largely used by younger speakers. Among the verbs that participate in the NACC are vulgar lexical items, which further adds to its colloquial nature. (The use of vulgar expressions to vent frustration is attested cross-linguistically, so Russian is not exceptional in that regard.) We provide a detailed analysis of the syntax of the NACC and argue that it instantiates obligatory adjunct control by the subject. We hypothesize that the rise of the NACC is driven by the analogy with the existing constructions with EXASPERATE-verbs in standard Russian, and we address several other factors that contribute to the development of the new construction.

Keywords: Adjunct control, Causative construction, Modern colloquial Russian, Obligatory subject control, Psych-verb construction, Salient/non-salient causation, Substandard lexicon

1. Постановка вопроса

В русском языке представлена целая группа глаголов, обозначающих каузацию внутреннего состояния, с общим значением «довести до предела терпения»: *достать*; *за*бать* и его заместители *задолбать, заколебать, затрахать, задрать, заманать*; к ним примыкает ряд глаголов, в которых идея «предела терпения» ослаблена: *надоесть; замучить; выбешивать, (вы)бесить*. Как видно из списка, большинство этих глаголов — глаголы совершенного вида. Почти все они употребляются как двухместные, с экспериенцером в винительном падеже, кроме *надоесть* и, возможно, некоторых его более грубых аналогов, при которых экспериенцер в дательном падеже. Источник раздражения (назовем его «каузатор») выражен либо ИГ в именительном падеже, (1), либо группой с предикатом в инфинитивной форме (2). В позиции каузатора нередко выступает неодушевленный референт: конкретный предмет или абстрактное событие. Нейтральный порядок слов в этой конструкции предполагает постановку экспериенцера в начало предложения, и уж определенно до появления каузатора; отклонения от этого порядка возможны, но маркированы с точки зрения актуального членения предложения.

(1) *Меня* *заколебали эти жалобы*.
ЭКСПЕРИЕНЦЕР КАУЗАТОР

(2) *Меня* *заколебало отвечать на жалобы*.
ЭКСПЕРИЕНЦЕР КАУЗАТОР

Глаголы этой группы также могут выступать в другой каузативной конструкции, где форма глагола не меняется. В этой конструкции одушевленный каузатор, доводящий кого-то до предела терпения, представлен в именительном падеже; экспериенцер (тот, чье терпение оказывается исчерпанным) в винительном (иногда дательном), и наконец, способ, которым экспериенцер выведен из терпения, в творительном. Здесь порядок слов уже иной, каузатор обычно предшествует экспериенцеру:

(3) *Студенты заколебали* меня *своими жалобами*.
КАУЗАТОР ЭКСПЕРИЕНЦЕР СПОСОБ

Конструкцию, представленную в (3), можно прямо соотнести с конструкцией в (1). В дополнение к этому, в устной форме современного русского языка все чаще появляется параллель конструкции, показанной в примере (2) выше. Новая каузативная конструкция (далее просто НКК) впервые была зафиксирована в работе [Левонтина 2017]. Каузатор здесь по-прежнему в именительном падеже, экспериенцер в винительном/дательном, а причина раздражения выражена инфинитивным оборотом (далее ИО; здесь и далее ИО показан в квадратных скобках):

(4) *Студенты заколебали* *меня* [*жаловаться на задания*].
КАУЗАТОР ЭКСПЕРИЕНЦЕР СПОСОБ

НКК принадлежит к явлениям устного синтаксиса и если и получает письменную фиксацию, то в основном в интернет-коммуникации, особенно в социальных сетях, где вообще происходит сближение письменной речи с устной.

Поскольку рассматриваемая конструкция принадлежит экспрессивному разговорному синтаксису, во многих случаях она фигурирует во фразах, где позиция экспериенцера, а нередко и каузатора, не заполнена, и они только подразумеваются:[2]

(5) pro *задолбал* pro *уже* [*ныть*]!
КАУЗАТОР ЭКСПЕРИЕНЦЕР СПОСОБ

Однако конструкция допускает и реализацию с полным заполнением:

(6) [agd-ardin, nick] *во вторых они меня задолбали* [*будить по ночам*]
[Твоя подпись в статусе (ICQ переписка) (2007.11.24)].

В обсуждении нашего синтаксического анализа мы будем использовать и предложения с полной реализацией конструкции.

(7) Для начала приведем несколько примеров НКК из нашей довольно богатой коллекции (некоторое количество примеров НКК есть и в НКРЯ;[3] орфография и пунктуация оригинала здесь и далее сохранены):

---

[2] При невыраженных аргументах может возникнуть впечатление, что конструкция (2) и (4) едва различимы, ср. (i) *Задолбало уже* [*всё раскапывать и закапывать*]! и (ii) *Задолбали уже* [*всё раскапывать и закапывать*]! Однако даже здесь видны существенные различия. Во-первых, согласование глагола в примере (i) всегда по среднему роду (или 3 лицу ед. числа), тогда как в примере (ii) согласование с неопределенным подлежащим в 3 лице мн. ч. (мы вернемся к этой теме в **разделе 3**). Во-вторых, в (i) подразумеваемый исполнитель действий «раскапывать и закапывать» совпадает с экспериенцером, а в НКК в (ii)—с невыраженным каузатором раздражения; мы вернемся к этой черте НКК в **разделе 2**.

[3] Отметим при этом, что наше исследование не позиционируется как корпусное. Корпусное исследование НКК — дело будущего.

(8) *Олег Кашин реально заебал* [*не давать ссылок на источники*]
(Фейсбук Антона Носика 5 июня 2017)
https://www.facebook.com/nossik/posts/10156279443587942

(9) *любите блять русский язык! заебали его* [*каверкать*]*!*
killdos07072009

(10) *Знаешь а ты достала я хочу тебе сказать*
Ты достала [*врать ничего не понимать*]
Нас больше нет иди теперь своей дорогой
(Песня) webkind.ru/text/30873167_941573807p971361604_text…

(11) *Девочки, как же Вы заколебали* [*жаловаться на своих мужей*]! *Ну если они у вас такие плохие, как Вы описываете на фига Вы с ними живете?*
askmamas.ru/answers.php?qid=1452609

(12) *Мой кот задолбал* [*орать каждый день*]*, что делать?* otvet.mail.ru›question/631891drive2.ru›l/3432801/

(13) *блин задрал* [*спрашивать*] *иди к тренеру который в тренажёрном зале и попроси составить программу и питание.* rusbody.com/forum/bodybuilding5/tema10491/5.html

(14) *— А наколка зачем? — Да затрахали* [*спрашивать*]…
(Коллекция анекдотов: сексопатологи (1970–2000))

(15) *Сними с меня галочку, когда пишешь, плиз, заманал* [*ныть*]. *Че за мужики пошли.* https://twitter.com/norimyxxxo/status/936734258181394435

(16) *Плюс ещё изменили форму мушки — убрали зуб, т. к. замучил* [*цепляться за ткань кармана, в котором я частенько таскаю револьвер*].
forum.guns.ru/forummessage/86/376026.html 23 окт. 2008 г.

(17) *На меня мой все утро орал, тип, надоела* [*жрать по ночам*].
bzik.info/post/37008/

(18) *Дашуля, ты лучшая. Но бесишь* [*орать на меня и посылать*]
@Лера | ask.fm/Dasha_princess 1

В следующем разделе мы разберем синтаксис НКК, а в разделе 3 рассмотрим возможные причины распространения этой конструкции.

2. Синтаксис

Мы предлагаем рассматривать НКК как конструкцию с обязательным синтаксическим контролем, в которой подлежащее главного предложения кореферентно подразумеваемому подлежащему ИО. Сам ИО находится в позиции адъюнкта (сирконстанта) финитного глагола. В подразделах этого раздела мы рассмотрим основные синтаксические признаки НКК.

2.1. Предикат главного предложения каузативен

Ряд каузативных глаголы в русском языке имеет особую морфологию, например, *расти—растить, сидеть — сажать* и т. д. С точки зрения морфологии может показаться странным, что глаголы, рассматриваемые здесь, не маркированы как каузативные. Однако морфологическое оформление каузатива в глаголе—не единственный признак каузативной конструкции. В русском языке нулевые (немаркированные) каузативы обнаруживаются и за пределами глаголов внутреннего состояния [Летучий 2006]. Ср.:

(19) a. *Его ушли на пенсию.*
b. *покакать/пописать младенца/собаку/кошку*
c. *Кто девушку ужинает, тот ее и танцует*.[4]

Нулевое оформление каузатива в глаголе характерно для глаголов внутреннего состояния, что было отмечено как для русского языка [Markman 2004]; [Lavine 2010], так и для конструкций с глаголами внутреннего состояния в типологическом освещении [Belletti & Rizzi 1988].

Еще одно подтверждение каузативного характера НКК состоит в том, что от НКК можно образовать декаузатив, с сохранением базового значения «дойти до предела терпения»:

(20) *Я заколебалась/затрахалась* [*чинить это ржавое корыто*]

Соответственно, в значении глаголов НКК, можно выделить два абстрактных семантических компонента: (a) психологическое состояние потери терпения и (b) каузация этого состояния. Формы совершенного вида предпочтительны для описания состояния (a), достигнутого неким действием (b), а описание состояния раздражения и составляет важный семантический компонент рассматриваемых конструкций. Таким образом, семантическая композиционность глаголов типа *заколебать* позволяет объяснить предпочтительность совершенного вида в рассматриваемых конструкциях. Однако поскольку речь идет о семантике, а не о синтаксисе, употребление совершенного вида все же предпочтение, а не строгое правило.

2.2. Деление НКК на клаузы

Целый ряд наблюдений свидетельствует о том, что НКК состоит из двух клауз и что ИО представляет собой отдельную клаузу, что является одним из признаков конструкций с синтаксическим контролем [Polinsky 2013].

ИО в НКК может быть замещен ситуативно-анафорическим местоимением *это,* (20); может перемещаться в предложении как единое целое, (21), а его части не могут подвергаться скрэмблингу (scrambling) в главном предложении, (22).

(21) *Студенты меня этим заколебали*

(22) [*Жрать по ночам*] *уже реально достала…*

[4] Большой игровой потенциал подобных преобразований отмечается в [Эпштейн 2007].

(23) *Присылать$_i$ студенты заколебали *меня* [t_i *свои жалобы*].

Доказательством того, что ИО в НКК является полной клаузой, является также поведение родительного падежа при отрицании. Родительный падеж при отрицании возможен в пределах одной клаузы, (23a), а также с инфинитивным дополнением глаголов субъектного контроля, (23b,c).

(24) a. *Мои дети не ожидают похвал за хорошее поведение*.
b. *И не обещай мне* [*писать длинных писем*]!
c. *Он не старается* [*употреблять умных слов*].

Однако родительный при отрицании совершенно невозможен в НКК:

(25) a. **И не доставай меня* [*писать длинных писем*]!
b. **Он тебя еще не заколебал* [*употреблять умных слов*]?

Эти наблюдения указывают на то, что НКК делится на главную (матричную) клаузу с каузативным глаголом и инфинитивную клаузу, которая, несмотря на отсутствие лексических единиц на поверхности, должна быть интерпретирована как полная клауза. Таким образом, каузативный глагол в НКК имеет следующую структуру: он сочетается с двумя ИГ и допускает клаузу, выраженную ИО.

(26) *достать* [$X_{\text{Каузатор.NOM}}$, $Y_{\text{Экспериенцер.ACC/DAT}}$, (CP)]

2.3. НКК — конструкция с синтаксическим контролем

ИО в НКК содержит подразумеваемый субъект, который не может быть выражен несмотря на то, что в принципе дативные подлежащие при инфинитиве возможны (см. также [Тестелец 2001: гл. V]).

(27) *Для меня главное* [*тебе не заболеть*].

(28) **Она меня задолбала* [*ее детям постоянно шуметь*].

Это указывает на то, что подразумеваемое подлежащее в ИО в НКК должно быть интерпретировано как кореферентное одному из аргументов главного предложения. Такая кореференция возможна только с подлежащим. Даже когда контекст наводит на мысль о кореференции с дополнением, синтаксис не допускает этой кореференции:

(29) *#Мать его достала* [*снова чувствовать себя пятиклассником/неженатым*].

Пассивизация предложения приводит к потере исходного смысла, что тоже является характеристикой конструкций с синтаксическим контролем:[5]

---

[5] Здесь существенны ограничения на пассивизацию именно в присутствии ИО. Конструкции без ИО пассивизуются без труда, ср. некоторые примеры из нашей коллекции:
(i) *Если жизнью ты задолбан и проблемы одолели…* http://www.sektam.net/forum/topic/2247-o-radastee-stihami-esli-zhiznyu-ty-zadolban/.
(ii) 10 апр. 2012 г. — @2012Esipov *Я познакомился с ним прошлой весной,уже тогда он был задолбан вопросами про Месси…*
(iii) *Линукс задолбан фанатами винды…* http://forum.ixbt.com/topic.cgi?id=15:50920.

(30) a. *Мать его замучила/задолбала* [*спрашивать про работу*]
b. **Он замучен/задолбан матерью* [*спрашивать про работу*]

Синтаксический контроль, в отличие от подъема (raising), не сохраняет целостности идиоматических выражений, и это ограничение соблюдается в НКК. Подлежащные идиоматические выражения *мухи дохнут* или *жаба душит* теряют свое некомпозиционное значение, появляясь в НКК:

(31) a. *#Мухи достали дохнуть у нее на занятиях*.
b. *#Жаба заколебала душить*.

Невыраженное подлежащее при контроле, как правило, должно быть представлено одушевленным именем, выражающим агентивного (сентиентного) участника ситуации (о некоторых отклонениях от этого общего правила, см. [Langacker 1995]). Большинство примеров с НКК действительно содержат в качестве подлежащего указания на людей, чаще всего — участников речевого акта (которые в этой связи, как мы уже упоминали, могут быть опущены), но также и не участвующих в диалоге соседей, чиновников, партии и т. д. Однако, учитывая экспрессивный характер НКК, в ней возможны и неагентивные и даже неодушевленные каузаторы, причем они появляются в НКК чаще, чем в менее экспрессивной конструкции с ИГ способа в творительном падеже, ср.

(32) a. *Телефон достал отключаться*.
b. *%Телефон достал отключениями*.

В таких случаях каузатор всегда прагматически окрашен, он «одушевляется» и «очеловечивается» говорящим. Примеры из живой речи действительно содержат в качестве каузатора в НКК названия животных и насекомых, а также предметов, воспринимаемых как обладатели собственной воли (автомобилей, двигателей, компьютеров, телефонов, растений) и подчас более обезличенных предметов (ковров, чемоданов и пр.). Однако примеры последнего типа вызывают колебания в суждениях даже у тех носителей, которые свободно владеют НКК. Это указывает на интерпретацию каузатора как наделенного собственной волей.

(33) a. *Мухи достали* [*жужжать*].
b. *Телефон достал* [*жрать трафик*].
c. *%Жасмин уже задолбал* [*лезть на дорожку*].
d. *%Чемодан задолбал* [*опрокидываться*].

Итак, ИО в НКК является клаузой облигаторного синтаксического контроля (obligatory control), чье невыраженное подлежащее интерпретируется как кореферентное подлежащему-каузатору в главном предложении. Теперь нам осталось определить, какая структурная позиция приписывается ИО в главном предложении.

2.4. ИО является адъюнктом глагольной группы

По отношению к каузативному глаголу ИО выступает как адъюнкт со значением способа действия, ср. (25) выше. Он легко может быть опущен, что

характерно для адъюнктов в целом. Он также соотносится с адъюнктом в творительном падеже в более традиционной конструкции, ср.

(34) a. *Он меня уже достал* [(*со*) своими *приставаниями*].
b. *Он меня уже достал* [*приставать*].

Более того, ИО и адъюнкт в творительном падеже могу быть сочинены, что указывает на сходство их категориальных статусов. Следующий неподтвержденный пример зафиксирован в живой речи, правда, многие носители НКК его отвергают:

(35) *%Мать достала* [*орать и уборкой*]…

Кроме того, имеются и более тонкие синтаксические свидетельства того, что ИО в НКК является сентенциальным адъюнктом. Это видно из сравнения между этим ИО с одной стороны и сентенциальным дополнением, с другой. Эти два типа подчиненных предложений различаются в отношении образования вопросов (Wh-movement).

В разговорном русском языке допускается образование вопросов к составляющим сентенциального дополнения, при котором вопросительное слово перемещается из сентенциального дополнения в главное, тем самым пересекая границу клаузы. Иначе говоря, сентенциальное дополнение «проницаемо» для передвижения. Например:[6]

(36) a. *%Что$_i$ ты боишься, [что она у тебя возьмет t_i без спроса*]?
b. *%Кого$_i$ он требует,* [*чтобы я позвала t_i в гости*]?

С другой стороны, подобное передвижение через границу клаузы совершенно недопустимо в НКК; ее ИО «непроницаем»:

(37) a. **Что$_i$/Чей телефон$_i$ он задолбал* [*брать t_i без спроса*]?
b. **Кого$_i$ он достал [звать t_i в гости*]?

Напомним читателю о недопустимости скрэмблинга составляющих ИО, вынесенных в главное предложение НКК; об этом свойстве НКК мы уже писали выше (см. **раздел 2.2**). Это проявление того же ограничения: адъюнкты, в отличие от аргументов, представляют собой синтаксические острова, и «покидать остров» грамматика не позволяет [Тестелец 2001: гл. XII]; [Митренина и др. 2012].

Сентенциальные адъюнкты могут размещаться в разных позициях в структуре предложения, примыкая к разным группам, от глагола до флективной вершины. В случае НКК адъюнкт расположен достаточно низко в структуре предложения, так как он указывает на способ, к которому прибегает каузатор, доводя экспериенцера до потери терпения. Это соображение, основанное на семантике (см. **раздел 2.1**), подтверждается и синтаксическими данными. При эллипсисе глагольной группы, ИО в НКК тоже стирается, что доказывает, что этот оборот находится внутри глагольной группы. Так, следующее

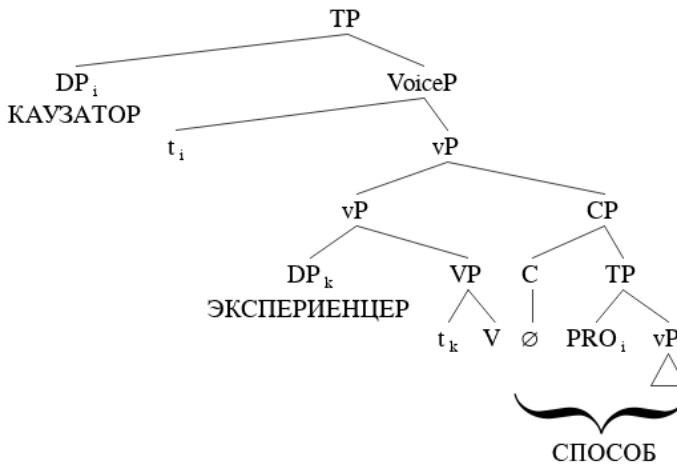
[6] Подобные вопросительные предложения распространены, но носители, более тяготеющие к стандартной норме, часто их отвергают, отсюда знак % в наших примерах. Существенно, что вопросы типа (35) легко и без сомнений допускаются теми, кто владеет НКК.

предложение должно быть интерпретировано как то, что и мать, и свекровь досаждают говорящему вопросами о работе; понимание, что свекровь досаждает вообще, иными способами, здесь неуместно.

(38) *Мать заколебала спрашивать про работу, и свекровь тоже ~~заколебала~~ ~~*(спрашивать про работу)~~.*

Напомним, что каузативный предикат в НКК состоит из двух частей и выражает два сцепленных события: событие каузации и состояние, вызванное каузацией (в нашем случае — потеря терпения). ИО модифицирует событие каузации: спрашивая про работу, мать вызывает в говорящем раздражение и т. п. Это означает, что ИО должен присоединяться к каузативной части предложения.

Если представить НКК иерархически, то мы получим структуру, представленную ниже (несущественные элементы структуры опущены). В этой структуре мы обозначаем функциональную вершину, ответственную за каузативное значение, как *v*; внешний аргумент вводится вершиной Voice и задает собой каузативную часть ситуации.



Далее возникает вопрос, чем заполнена вершина C в предлагаемой структуре. Мы полагаем, что это некая нулевая лексическая единица, по значению сходная с русским *при помощи* или английским *by*: нулевой субординатор. Тогда как прочие компоненты структуры хорошо обоснованы конкретными фактами, допущение относительно нулевого субординатора носит более произвольный характер. Однако нулевые субординаторы нередко постулируются в самых различных структурах, как в русском языке [Тестелец 2001]; [Szucsich 2001, 2002]; [Bailyn 2012, и др.], так и в типологических исследованиях [Cristofaro 2003], так что принятое здесь решение не необычно.

3. Развитие и распространение НКК в разговорной речи

Разговорная НКК преимущественно используется в речи подростков и молодежи. Набор управляющих инфинитивом в НКК глаголов таков, что вся конструкция имеет не просто разговорную, а сниженную окраску. Следует отметить, что вообще выражение разного рода фрустраций большим количеством вульгарных и сленговых единиц характерно для разных языков (Allan 2019), так что русский в этом отношении не исключение.

Есть некоторые основания полагать, что инфинитивное управление у глагола *за*бать* существовало довольно давно, возможно, даже около полувека. По понятным причинам получить письменные подтверждения этого предположения нам пока не удалось, есть лишь устные свидетельства, которые невозможно проверить. Мы опросили около 20 носителей 1950–1960 гг. рождения, и они подтверждают, что знакомы с грубым выражением в (39a), которое в их идиолекте чередуется с (39b).

(39) a. *За*ал п*деть!*
b. *Хорош п*деть!*

Возможно, что единичные образования типа (39a) были зерном НКК. В последние 25 лет конструкция активизировалась и сначала распространилась на эвфемизмы матерного глагола, а затем и на слова *замучить, надоесть, выбешивать* и др., хотя доказать это мы пока не можем.

(40) Тем не менее, большинство носителей стандартного литературного языка воспринимают эту конструкцию как очень странную и «нерусскую». Любопытно, что типичной реакцией на нее является утверждение, что это, «конечно, англицизм», — при том что на самом деле в английском языке нет конструкции, которая могла бы послужить непосредственным прототипом для *достал ныть*; ее ближайшим аналогом является конструкция с герундием, вводимым предлогом *by*:

(41) *You drive me up the wall* [*by your whining/*by (to) whine*].

Мы полагаем, что НКК возникает на основе конструкции психологического состояния вида (2) и по структурной аналогии с ней. По своей структуре конструкция вида (2) неоднозначна. ИО, выражающий каузатор, может быть сентенциальным подлежащим, вынесенным вправо (т. н. экстрапозиция подлежащего), как показано в (41), либо адъюнктом, выражающим способ, которым экспериенцер выведен из терпения, (42). В последнем случае подлежащее выражено нулевым местоимением со значением «стихия» [Мельчук 1974].

(42) *Меня заколебало* ec$_i$ [$_{CP}$ *отвечать на жалобы*]$_i$
ЭКСПЕРИЕНЦЕР КАУЗАТОР

(43) *Меня заколебало* pro [$_{CP}$ *отвечать на жалобы*]
ЭКСПЕРИЕНЦЕР КАУЗАТОР СПОСОБ

Вторая структурная интерпретация одновременно приближает конструкцию вида (2) и к НКК (в обеих конструкциях присутствует ИО), и к каузативной

конструкции с агентивным каузатором вида (3), так как в обеих конструкциях выражен способ, которым достигается раздражение. Иначе говоря, продуктивность новой модели поддерживается возможным смешением конструкций:

(44) *(Мне) надоело скандалить + (Ты мне) надоел (со) скандалами = Ты (мне) надоел скандалить*.

Помимо смешения конструкций, кодирующих ситуации, где экспериенцер доведен до предела терпения, надо отметить и другие факторы, которые могли повлиять на распространение НКК в речи. Таких факторов по крайней мере два.

Во-первых, семантика глаголов, обозначающих каузацию достижения предела терпения, такова, что провоцирует их использование в косвенном речевом акте требования: *Ты меня достал* естественно понимается как требование прекратить соответствующее действие. Поэтому на управление вида *Достал ныть!* может влиять использование инфинитива во фразах типа *Прекрати / хватит / хорош ныть* (ср. (39b))

Во-вторых, инфинитив вообще активизировался в последнее время; ср. разнообразные и весьма активные инфинитивные конструкции в современной разговорной речи:

(45) a. *Ты рехнулся туда ходить?!*
b. *Я уже охренел там сидеть.*
c. *Ты с ума сошел с ним спорить.*
d. *У меня нога болит по лестницам ходить.*
e. *Я уже в ледышку превратилась тебя ждать.*
f. *Так есть ты решила не есть?*
g. *Нужно ещё оставить время на поесть и позаниматься*, и др.

Надо заметить, что активизация «словарных» форм — именительного падежа и инфинитива вписывается в общую тенденцию к аналитизму, которую давно отмечают лингвисты в русской грамматике; см. [Panov 1971]. Получается, что формы косвенных падежей из одних позиций вытесняются номинативом (*Йогурт малина клюква; Спасибо, что пользуетесь Аэроэкспресс*), а из других — инфинитивом (*Задолбали орать*). Особенно это естественно для разговорной речи, ведь инфинитив вообще более разговорная форма, чем творительный отглагольного слова (*нытьем, опозданиями* и т. п.).

***

Мы благодарны Владимиру Ивановичу Беликову, Ирине Бурукиной, Полине Касьяновой, Екатерине Лютиковой, Оре Матушанской, Полине Плешак и Татьяне Филипповой за плодотворное обсуждение этой статьи. Мы также признательны Софье Левиной, Варваре Левонтиной и Александру Полинскому за суждения по поводу отдельных примеров, приведенных в этой работе.

References

1. *Allan, Keith, ed.* (2019), The Oxford Handbook of Taboo Words and Language. Oxford: OUP.
2. *Bailyn, John F.* (2012), The syntax of Russian. Cambridge: Cambridge University Press.
3. *Belletti, Adriana, and Luigi Rizzi.* (1988), Psych-verbs and θ-Theory. Natural Language and Linguistic Theory 6, pp. 291–352.
4. *Cristofaro, Sonia.* (2013), Subordination. Oxford: Oxford University Press.
5. *Epstein, Mikhail* (2007), On the creative potential of the Russian language. Transitivity grammar and a transitive society [O tvorčeskom potenciale russkogo jazyka. Grammatika pepexodnosti i tranzitivnoje obščestvo]. Znamja 2007, 3, available at: http://magazines.russ.ru/ZNAMIA/2007/3/ep18.html.
6. *Fowler, George, and Michael Yadroff* (1993) The argument status of accusative measure nominals in Russian. Journal of Slavic Linguistics 1, pp. 251–279.
7. *Gribanova, Vera* (2013), Verb-stranding verb phrase ellipsis and the structure of the Russian verbal complex. // Natural Language and Linguistic Theory 31, pp. 91–136.
8. *Ionin, Tania, and Tatiana Luchkina* (2018), Focus on Russian scope: An experimental investigation of the relationship between quantifier scope, prosody, and information structure. Linguistic Inquiry 48, pp. 741–779.
9. *Langacker, Ronald* (1995), Raising and transparency. // Language 71, pp. 1–62.
10. *Lavine, James* (2010), Case and events in transitive impersonals. // Journal of Slavic Linguistics 18, pp. 101–130.
11. *Letuchiy, Alexander* (2006), Lability in Pussian: an exception or a rule? [Labilnost' v russkom jazyke: slučajnost' ili zakonomernost'?], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2006" [Komp'juternaja Lingvistika i Intellektual'nye Texnologii: Trudy Meždunarodnoj Konferencii "Dialog 2006"], Bekasovo, available at: http://www.dialog-21.ru/digests/dialog2006/materials/html/Letuchy.htm.
12. *Levontina, Irina* (2017), New developments in "marginal" syntax: Mama dostala orat' i uborkoj [Novoje v marginal'nom sintaksise: Mama dostala orat' i uborkoj] // Magrinalia-2017: The limits of culture and text. Proceedings. Institute for Russian Language. Moscow, p. 59.
13. *Markman, Vita* (2004), Causatives without causers and Burzio's Generalization. Proceedings of the North East Linguistics Society 34. Amherst, MA: GLSA Publications, pp. 425–440.
14. *Mel'čuk, Igor* (1974), On silent elements in syntax. [O sintaksičeskom nule] // Typology of causative constructions [Tipologija kauzativnyx konstruktsij]. Moscow, Nauka, pp. 343–362.
15. *Mitrenina, Olga, Elena Romanova, Natalia Slusar* (2012), Introduction to generative grammar [Vvedenije v generativnuju grammatiku]. Moscow, URSS.
16. *Panov, Mikhail* (1971), On analytical adjectives. [Ob analitičeskix prilagatel'nyx] // Phonetics. Phonology. Grammar. [Fonetika. Fonologija. Grammatika] Moscow, Nauka, pp. 240–253.

17. *Polinsky, Maria*. (2013). Raising and control. In Marcel Den Dikken (Ed.), The Cambridge Handbook of Generative Syntax. Cambridge: Cambridge University Press, pp. 577–606.
18. *Szucsich, Luka* (2001), Adjunct positions of nominal adverbials in Russian. In Gerhild Zybatow et al. (eds.) Current Issues in Formal Slavic LinguisticsFrankfurt am Main: Peter Lang, pp. 106–116.
19. *Szucsich, Luka* (2002), Case licensing and nominal adverbials in Slavic. In Jindrich Toman (ed.) Formal Approaches to Slavic Linguistics 10: The Second Ann Arbor Meeting. Ann Arbor, MI: Michigan Slavic Publications, pp. 249–270.
20. *Testelets, Yakov* (2001), Introduction to general syntax [Vvedenije v obščij sintaksis]. Moscow, RSUH.

AUTOMATIC VOCABULARY POSITIONING IN A THESAURUS

Likhonosov A. (andrew.likhonosov@abbyy.com),
Indenbom E. (Eugene_I@abbyy.com),
Yudina M. (maria_yu@abbyy.com)
ABBYY, Moscow, Russia

Thesauri are one of the most widely used resources in natural language processing. At the same time, many of them are built manually, which takes a lot of time and, due to human errors, can affect their quality and completeness. We propose a procedure for automatic positioning of vocabulary in the ABBYY Compreno thesaurus using large monolingual corpora, a regular bilingual dictionary and a subset of already positioned words.

Key words: vocabulary positioning, thesaurus, word embeddings, supervised bilingual dictionary induction

АВТОМАТИЧЕСКОЕ ПОЗИЦИОНИРОВАНИЕ ЛЕКСИКИ В СЛОВАРЕ ТЕЗАУРУСНОГО ТИПА

1. Introduction

Thesauri are one of the most useful resources in natural language processing. However, most of them are crafted by hand, which brings up problems of incompleteness, human errors and time costs. A perfectly complete thesaurus is inherently impossible, as language changes with time, new words appear to describe new objects and phenomena, while others disappear. Moreover, human brains are not designed for enumerating objects, in this case, new words. So, a machine might be of great help. The machine can crunch corpora with infinite patience and precision and produce a perfect list of unknown words and even position them into a thesaurus. A human linguist would only have to supervise the process.

On the other hand, available language resources (and thesauri in particular) are unequally distributed between languages. English resources are by far more rich, complete and diverse than resources for any other language. So, the problem of knowledge transfer to languages other than English is well recognized in linguistic community, and several attempts to automate the process has been made, mainly for

WordNet (Fellbaum, 1998). For example, (Farreres et al., 1998) described the process for Catalan and Spanish. (Patel et al., 2018) did it for Hindi, they used the idea of linear transformation of word embeddings between languages, as originally proposed in (Mikolov et al., 2013b). And (Niemi et al., 2012) did the English-Finnish transfer, where they built the Finnish version of WordNet mostly manually and then used bilingual resources to extend and to improve it.

In building ABBYY Compreno Semantic Hierarchy (a kind of thesaurus, which is described below) we also face the problem of knowledge transfer from languages that we already have in our system to languages that are new to it. Doing it manually takes a lot of time and resources.

In this paper we describe a method for automatic positioning of a language vocabulary in a thesaurus using knowledge transfer technique from positioning of English and Russian languages in the same thesaurus. We report results for the ABBYY Compreno Semantic Hierarchy and the German language.

1.1. ABBYY Compreno Semantic Hierarchy

The ABBYY Compreno Semantic Hierarchy is the key element of the ABBYY Compreno linguistic model. It can be thought of as a tree of universal notions called "semantic classes". In a sense, these semantic classes are like Plato's "ideas", as opposed to real world objects, "shadows", but applied to natural language. For example, there is a semantic class "BULLDOG", which, as a pure idea, resides in the world of ideas, in universal language. You can think of a BULLDOG as a dog with all the necessary breed characteristics. But for real languages we have actual words, for example, "bulldog" in English and "Bulldogge" in German. So, for every language these semantic classes are filled with actual words. Since natural languages have synonymy, sometimes there are several words in a semantic class.

Apart from regular words and classes there are also "collocations" and "idioms" under semantic classes. Both collocations and idioms are stable multiword expressions, the difference between them lies in compositionality principle of semantics. Collocations can represent stable concepts, specifying in some way the meaning of the core, adding more information to it and thus forming a new concept: "буря:STORM" -> "песчаная буря". It is not rare that in other language the same notion is expressed by whole word: "снежная буря"—"blizzard" under semantic class 'SNOWSTORM'. On the contrary, the meaning of the idiom is never composed from the meanings of its parts ("белая ворона", "вбить в голову"). In the Semantic Hierarchy collocations are positioned under the semantic class of the root (or main) word. Idioms are positioned under semantic classes, according to the meaning of the whole expression. For more detailed information on how the Semantic Hierarchy is designed see [Manicheva et al., 2012], [Petrova et al., 2018], [Goncharova et al., 2015]. In this article we treat all collocations as their roots and idioms as distinct language units.

Another important property of the Semantic Hierarchy is its hierarchical structure. All classes are organized according to the hyper-hyponym relationship. To continue with our BULLDOG example, the semantic class can be found under the following path (arrow "->" designates the hyper-hyponym relation): PHYSICAL_OBJECT->BEING->ANIMAL->CHORDDATA->WARM_BLOODED->PREDATORS-

>CANIDAE->CANINAE->DOG->BULLDOG. A lot of semantic properties are described at the level of semantic classes. It allows us to describe a concept once and then all the descendants will inherit its semantic (and syntactic, when applied to a particular language) properties. This structure simplifies the process of semantic description, making the positioning of new words the most challenging part. As there are almost 200k semantic classes, this task, done manually, is quite resource-intensive, and there is need for optimizations.

1.2. Word Embeddings

Distributional vector space semantic models, or word embeddings, prove to be useful in many natural language processing tasks and are de facto standard for modern deep learning researches in NLP domain [Collobert et al., 2011]. According to the distributional hypothesis [Harris, 1954], words with similar distribution tend to have similar meanings, thus, can be considered synonyms. Mikolov [Mikolov et al., 2013a] suggested a method that scales well on billion-word size corpora and allows to capture distributional properties of a huge vocabulary. Basing on Harris hypothesis, we assume that word embeddings encode word relationship information and can be used to position words into a thesaurus.

As we aim to map words (and word embeddings) of a 'new' language into an existing thesaurus with bindings to 'known' languages, the task is essentially cross-lingual knowledge transfer. A good overview of cross-lingual word embedding models is given in [Ruder et al., 2017]. Our own method is described below.

2. Proposed Method

In this paper we describe the following approach to automatic positioning of new words in the Semantic Hierarchy. We assume there is at least one (almost) fully described language in ABBYY Compreno (as of today, we consider the description of the first languages, i.e. Russian and English, to be almost full). From now on we will refer to it as 'source' language, and, similarly, to the language of interest—as 'target' language.

First, we train semantic class embeddings for the source language. We use our Compreno Syntactic and Semantic parser [Anisimovich, et al., 2012] to extract semantic classes from a big corpus. We train embeddings on these semantic classes as tokens using the SkipGram algorithm [Mikolov et al., 2013a]. During training all proper names were generalized to their hypernyms (e.g. 'Smith' => 'PERSON_BY_LASTNAME', 'Intel' => 'ORGANIZATION_BY_NAME'). 82396 embedding were trained on the corpora of $3{,}5 \cdot 10^9$ words.

Second, we train lemma embeddings for the target language. Technically speaking, objects to be positioned in the Semantic Hierarchy are lexemes, but lemmas are rather good approximation for our purposes. Lexeme is a word with all its morphological forms with implied lexical grammatical characteristics, i.e. part of speech. Lemma is the text of the dictionary form of the lexeme. So, different lemmas must correspond to different lexemes, but different lexemes may share the same lemma

(with, for example, different parts of speech). This can cause some problems with homonyms, for example, 'address' could be either verb or noun. We cannot distinguish between such homonyms while training word embeddings, nor between their corresponding vectors. Since there is only one vector for "address", it will represent some average meaning, which can be something completely different from what we expect. We decided not to deal with this problem in this paper and leave it for future research. The corpora used for lemma embedding training contained $5{,}7 \cdot 10^9$ words.

Finally, we train a binary classifier for pairs of Semantic Classes and Lemmas. The classifier, given a pair (SemanticClass, Lemma), will output a number between 0 and 1. This way, for each lemma from a target language, and a set of N hypotheses of semantic classes, we feed this classifier with pairs [(SemanticClass1, Lemma), (SemanticClass2, Lemma),…,(SemanticClassN, Lemma)], and use its output to sort the hypotheses and, finally, find the best semantic class candidates.

3. Experiments

3.1. Data preparation

As we stated earlier, our method requires at least one fully described language (which will be used as a source language), but in this paper we ended up using English and Russian as source languages. We trained semantic class embeddings on a big English corpus and we used a German-Russian bilingual dictionary. Since the word list of this dictionary was already manually positioned in the Compreno system, we use manual positioning data as reference markup for our evaluation. It would be interesting to measure how the two components—semantic class embeddings and a bilingual dictionary—affect the final quality. For example, it would be interesting to run the same experiment with semantic class embeddings trained on a Russian corpus and the same (or another) German-Russian dictionary, it is the task for the future research.

As previously mentioned, while training semantic class embeddings, we treat collocations and idioms in a special way: we extract semantic class which corresponds to the collocation (or idiom), not semantic classes of the collocation's parts. Thus, for example, the whole idiom "to beat around the bush" will be extracted as the class "TO_BEAT_AROUND_THE_BUSH" which is located under the class "TO_EVADE", and "BUSH" (as a plant) will not be extracted.

For the classification problem we generate a positive and a negative sample of pairs (Semantic class, Lemma). We use the ABBYY Lingvo Universal (De-Ru)[1] dictionary, which contains Russian translations for German words. An article in the Lingvo dictionary (Universal (De-Ru)) looks like this:

[1] Universal (De-Ru) (for ABBYY Lingvo x6). The comprehensive German-Russian dictionary contains over 80,000 entries. © ABBYY, 2013, website: www.lingvo.ru/european/dictionary/

Dendrit
m, <-en, -en>
1) *геол*, , *мет* дендрит
2) *анат* дендрит, древовидный отросток нервной клетки мозга

Figure 1

For “Dendrit” there are two classes in Semantic Hierarchy called “DENDRITE” and “DENDRITE_AS_CRYSTAL”. These pairs (Dendrit, DENDRITE), (Dendrit, DENDRITE_AS_CRYSTAL) will be added to our positive sample.

For the negative sample we created a simple hypotheses generator.

3.2. Simple hypotheses generator

We propose the following procedure for generating hypotheses for German words. The ideal hypotheses generator should be simple but must produce all true and not so many negative classes, so that was what we were aiming for.

For the Russian part, we run the Compreno Syntactic and Semantic Parser and extract all possible classes for the Russian translation. For example, the first meaning from the dictionary article, “дендрит”, is parsed like “DENDRITE” or “DENDRITE_AS_CRYSTAL”, so these classes are added to the set of hypotheses. The second meaning, “дендрит, древовидный отросток нервной клетки мозга” is parsed like this:

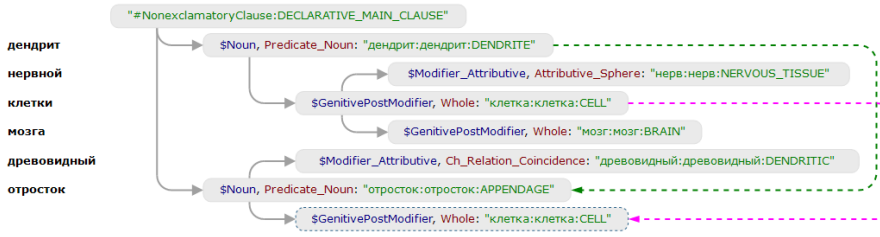


Figure 2

We don’t pay any attention to the structure, but simply extract all classes for all words. This gives us another addition to the hypotheses set: {DENDRITE, NERVOUS_TISSUE, CELL, BRAIN, DENDRITIC, APPENDAGE, CELL}.

We found that adding neighbor classes (parents, or hypernyms, and children, or hyponyms) to the hypotheses set improves chances to generate all the true classes. But it expands the hypotheses set too much, which affects the final quality of the classifier. So, we decided to do it another way: we don’t add neighbor classes to the hypotheses set, but we still consider these classes as true. Thus, we are allowing our hypotheses generator to be a little bit imprecise and be one-level wrong in the Semantic Hierarchy.

Then we parsed redirects—constructions like “см Apsis”—“see Apsis”. We merge such redirects to the word where these redirects point to. This way, we added all hypotheses, as well as true classes, for “Apside” to the destination word “Apsis”.

As in general there is no gold standard for morphological description, dictionary entries are not necessarily primary forms in our morphological system. Therefore

if dictionary entry does not represent primary form in our morphological system we assign semantic classes extracted from dictionary article to all possible primary forms of the dictionary entry. For example, there are a lot of past participles (Partizip II) of German verbs in this Lingvo dictionary which may or may not have another translation. For example, “gezählt”, being the past participle of “zählen”, has also another meaning “встречающийся в самом большом [маленьком] количестве”. So we take all the classes from this translation and add them to the lemma “zählen”. And “gezählt” is removed completely from the list of lemmas.

Finally, for every word that begins with a capital letter (and there are a lot of them in German, since all nouns in German are capitalized) we add classes that contain named objects, like “PERSON_BY_NAME”, “TOWN_BY_NAME” and so on. We needed them because the case when our hypotheses generator was not able to generate true classes for named objects was quite frequent.

This way, we have achieved a recall of 0.81, that is, our suggest generator was able to cover 81% of true classes across all the German words from the Lingvo dictionary.

3.3. Neural network architecture

For the classifier we used a simple feed-forward multilayer perceptron. It takes as input a concatenation of two embeddings: one for a target (German) word and the other for a semantic class. At the output layer, we have a single neuron with the sigmoid activation function. We interpret the value of this sigmoid as probability for a given pair (lemma, class) to be a good positioning suggestion. We use the log loss as a loss function in our neural network.

The best architecture was chosen by a randomized hyperparameter search procedure called Tree of Parzen Estimators (TPE). We used the implementation provided by a python package called Hyperopt [Bergstra, 2013]. The varying parameters included layers’ neuron count, batch size, learning rate, activation function, optimizer algorithm, dropout rate and batch normalization momentum. The final architecture has 6 layers of 2048, 1024, 512, 256, 128 and 64 neurons, the learning rate of 0.01, the batch size of 128 examples, dropout rate of 0.3, leaky relu activation function with alpha parameter 0.1 and the Momentum optimizer.

3.4. Training

The whole dataset of positioning hypotheses (pairs of lemma and class, positive and negative) was split into three datasets: training, validation and test. We sort the German lemmas list by frequency and then perform the split based on lemmas’ ranks.

We take first 1,000 lemmas for training. Then we split next 14,000 lemmas (from 1,000th to 15,000th) into parts: one for training, with 7,000 lemmas in it and the other part is added to the rest of lemmas. The rest of lemmas is split in half: one part for validation and the other for test dataset.

So we have 8k lemmas for training, and these lemmas are among the most frequent ones. In real world we usually start with description of some “core” words of a language, so by performing these manipulations we tried to simulate these conditions.

3.5. Results

We measure our classifier's performance on train, validation and test datasets. As we said above, our classifier evaluates pairs of lemmas and classes. Table 1 is an example output of our classifier for several German dictionary entries and a number of Semantic Classes. Suggests generator proposed from 10 to 400 Semantic Classes for each lemma. We sorted suggested Semantic Classes by score and cut the list at score 0.1 or last Semantic Class from markup, whichever comes last. Semantic Classes matching markup (the manually built Semantic Hierarchy contains the lemma in this Semantic Class) are marked 'Yes' in 'Markup' column and printed bold. Scores are given in probability scale.

Table 1

Lemma	Semantic Class	Representatives/Description	Score	Markup
Kranich	**CRANE**	'crane' (as a bird)	0.51	Yes
	CONSTELLATION_BY_NAME	'Grus'	0.18	Yes
	PERSON_BY_LASTNAME	surnames, i.e. 'Smith'	0.14	No
	PERSON_BY_FIRSTNAME	first names, i.e. 'Michelle'	0.05	No
	COMPANY_BY_NAME	company name, i.e. 'Intel'	0.02	No
Eruption	**TO_ERUPT**	'to erupt' (about a volcano)	0.71	Yes
	RASH	'rash' (spots on the skin)	0.57	Yes
	ACUTE_STAGE	'burst', 'explosion' (as top point of some process)	0.49	No
	LAVA	'lava' (as substance)	0.45	No
	TO_CUT_AS_TO_APPEAR	'eruption' (as process of appearance of the teeth)	0.42	Yes
	OUTLIER	'outlier' (as a sudden peak in a graph)	0.24	No
	FALLOUT	'fallout', 'emission' (as waste)	0.17	No
	TO_FLAKE	'to flake', 'to exfoliate'	0.14	No
	TO_THROW	'to throw', 'to heave', 'to toss'	0.13	No
	TO_POUR_SMTH_FRIABLE	'to dust', 'to strew'	0.06	No

Lemma	Semantic Class	Representatives/Description	Score	Markup
verschlucken	**TO_SWALLOW**	'to swallow', 'to gulp'	0.35	Yes
	TO_ABSORB	'to absorb', 'to ingest' (something inedible)	0.33	Yes
	TO_DEVOUR	'to guttle', 'to devour' (as eat greedily)	0.31	No
	TO_DIE_AWAY	'to muffle', 'to drown' (to diminish about sound and light)	0.30	No
	TO_TAKE	'to take' (in general meaning)	0.28	No
	TO_HIDE	'to hide', 'to conceal'	0.24	No
	ANGER	'anger' (as emotion)	0.20	No
	TO_MAKE_INVISIBLE	'to envelop', 'to haze'	0.18	No
	TO_TAKE_AWAY_BY_FORCE	'to bereave', 'to deprive'	0.15	No
	TO_SUPPRESS_FEELINGS	'to suppress', 'to swallow down'	0.10	Yes
	TO_BE_FULL_ABSORBED	'to absorb' (about work, activity)	0.10	No
	TO_CRITICIZE	'to criticize'	0.09	No
	TO_CONSUME	'to consume', 'to absorb' (about resources, i.e. fuel)	0.09	Yes
obsolet	**UP_TO_DATENESS**	'modern', 'outdated'	0.32	Yes
	TO_USE	'to use' (in general meaning)	0.27	No
	SUPERFLUOUS	'superfluous', 'excessive'	0.26	Yes
	TURN_OUT_AS_BE	'to turn out' (as to prove to be)	0.14	No
	TO_GET_RID_FROM_ DIFFICULTY	'to extricate' (to get someone out of a difficult or unpleasant situation)	0.08	No
Mysterium	**MYSTERY**	'mystery', 'secret'	0.57	Yes
	MYSTERY_AS_RELIGIOUS_ CEREMONY	'mystery', 'rite'	0.56	Yes
	SACRAMENT	'sacrament' (as rite)	0.03	No
	PERSON_BY_FIRSTNAME	first names, i.e. 'Michelle'	0.02	No
	PERSON_BY_LASTNAME	surnames, i.e. 'Smith'	0.02	No
Statistik	**STATISTICAL_DATA**	'statistics' (as data)	0.82	Yes
	STATISTICS	'statistics' (as science)	0.60	Yes
	DATA	'data'	0.50	No
	STATISTICIAN	'statistician'	0.31	No
	NATURAL_SCIENCE	sciences like 'physics'	0.28	No
	SCIENCE	'science' (in general meaning)	0.13	No
	QUALITY_PROPERTY	'characteristic', 'quality'	0.11	No

Lemma	Semantic Class	Representatives/Description	Score	Markup
Bestellung	**TO_ORDER_GOODS**	'to order' (as to buy something)	0.50	Yes
	TO_DELIVER	'to convey', 'to deliver'	0.28	No
	ORDER_AS_RESULT	'order' (as a result of making an order)	0.25	Yes
	WARRANT	'order', 'warrant' (as permission or command to do something)	0.23	No
	TO_INFORM	'to report', 'to inform'	0.18	No
	MESSAGE_COMMUNICATION	'message' (as quantum of information)	0.18	No
	TO_PROCESS_INFORMATION	'to process', 'to handle' (things like reports and claims)	0.15	No
	REPORT	'report' (as official description of something)	0.12	No
	TRANSPORT_ COMMUNICATIONS	'service' (as transport communication)	0.11	No
	TO_GIVE	'to give' (in general meaning)	0.11	No
	ORDER	'order' (as a thing that is ordered or bought)	0.07	Yes

Table 2

Dataset	acc	precision	recall	f1	top1	top3	top5
Train	1.00	0.85	0.38	0.53	0.85	0.89	0.92
Dev	1.00	0.50	0.19	0.28	0.59	0.79	0.87
Test	1.00	0.47	0.22	0.30	0.61	0.80	0.88
test_with_true_class	1.00	0.51	0.20	0.29	0.59	0.78	0.86

In the Table 2 we summarized the results for our classifier on different parts of the dataset. Train, dev and test set were described earlier. The "test_with_true_class" is a filtered test dataset, containing only those lemmas for which our suggest generator was able to generate at least one true class. The accuracy, precision, recall and F1 are simple classifier metrics for classifying pairs (Lemma, Semantic class).

In the Table 3 we provide more information on the distribution of target lemmas in the test dataset and on our classifier performance on these parts of the test dataset. "Core lexis" column indicates that this part of dataset contains only core lexis, i.e. set of lemmas with frequency ranks less than 15,000. "Polysemous" indicates that words have more than one true class.

Table 3

core lexis	polysemous	lemmas	top1	top3	top5
yes	yes	1,102	0.71	0.60	0.70
yes	no	2,435	0.65	0.85	0.92
no	yes	1,475	0.57	0.57	0.70
no	no	22,244	0.56	0.78	0.86

TopN are metrics for ranking hypotheses and positioning. First, while calculating these metrics, we excluded lemmas for which either there is no true hypothesis, or the number of negative hypotheses is less than N. Thus, we leave only those German lemmas, where our method can fail. For example, if we included German lemmas with only 3 negative hypotheses and 2 positive ones, the top5 metric would always be 1, no matter how good or bad the final ranking is. TopN is the average of true hypotheses shares in the top N ranked hypotheses over all German lemmas:

$$top(N) = \frac{1}{lemmasCount} \sum_{lemma} trueHypothesesShareInTop(N, lemma), \text{ where}$$

$$trueHypothesesShareInTop(N, lemma) = \frac{numberOfTrueHypothesesInTop(N, lemma)}{\min(totalPositiveHypotheses, N)},$$

where $numberOfTrueHypothesesShareInTop(N, lemma)$ is the number of true hypotheses found in top N hypotheses in a ranked list of hypotheses.

So, for example, top1 of 0.61 on the test means that on the test dataset our classifier was able to rank hypotheses for 61% of lemmas so that a true class was on the top of the list. On average, we were able to guess 80% of classes across all lemmas in top3. And 88% of classes across all lemmas in top5. The positioning (topN) metrics on "test_with_true_class" dataset are two percent worse, which means that lemmas where our hypotheses generator produced a true class were positioned worse than when only negative hypotheses were generated and the true class was obtained from the Semantic Hierarchy. This means that we can position a significant amount of lexis totally automatically. The rest must be verified by a linguist, and for the 88% of cases the right decision lies within first 5 results and can easily be reached.

4. Conclusion

In this article we tried to apply embeddings to the problem of positioning new words in an existing thesaurus. Our method requires semantic-syntactic parsing for a large monolingual corpus of texts in the source language (English in our case), morphological parsing for a large monolingual corpus in the target language (we used German), a number of already positioned words and a regular bilingual dictionary (German-Russian) for hypotheses generation. We expect our method to position automatically about 61% of lemmas of a new language. On average, lemmas will be assigned to 88% of correct classes when using top-5 results.

We see two main directions for further research. Firstly, we plan to compare performances of our classifier on different source language data (different corpora for semantic class embeddings and different bilingual dictionaries). Secondly, we plan to run similar experiments using words alignment statistics from parallel corpora instead of manually positioned core lexis words, avoiding the necessity to position initial word set manually.

All in all, our results show that the process of positioning (and thus adding) new lexis to the thesaurus-like dictionary can be automated to a significant extent. This is crucial when we deal with a language that is completely new to the linguists involved. Otherwise it speeds up and simplifies considerably the work on the lexical description of a language that is new to the system.

References

1. *Anisimovich, K., Druzhkin K. Ju., Minlos F. R., Petrova M. A., Selegey V. P., & Zuev K. A.* (2012). Syntactic and semantic parser based on ABBYY Compreno linguistic technologies. Computational linguistics and intellectual technologies: Proceedings of the International Conference "Dialog 2012", Vol. 2, pp. 91-103.
2. *Bergstra, J. Y.* (2013). Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. 30th International Conference on Machine Learning (ICML 2013).
3. *Collobert, R., Weston, J., Bottou, L., Karlen, M., & Kuksa, P.* (2011). Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research 12(12).
4. *Farreres, X., Rigau, G., & Rodríguez, H.* (1998). Using WordNet for Building WordNets. WordNet@ACL/COLING.
5. *Fellbaum, C.* (1998). A semantic network of English verbs, WordNet: An electronic lexical database., (pp. pp. 153–178.).
6. *Goncharova, M., Kozlova, E., Payukov, A., Garashchuk, R., & Selegey, V.* (2015). Model-Based WSA as Means of New Language Integration into a Multilingual Lexical-Semantic Database with Interlingua. Papers from the Annual International Conference "Dialogue", (pp. vol. 1, pp. 169-182).
7. *Harris, Z. S.* (1954). Distributional structure. Word, 10(23), 146-162.
8. *Manicheva, E., Petrova, M., Kozlova, E., & Popova, T.* (2012). The Compreno Semantic Model as an Integral Framework for a Multilingual Lexical Database. Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon, 215-230.
9. *Mikolov, T., Chen, K., Corrado, G., & Dea, J.* (2013a). Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at ICLR.
10. *Mikolov, T., Le, Q. V., & Sutskever, I.* (2013b). Exploiting similarities among languages for machine translation.
11. *Niemi, J., Linden, K., & Hyvarinen, M.* (2012). Using a Bilingual Resource to Add Synonyms to a Wordnet: FinnWordNet and Wikipedia as an Example.
12. *Patel, K., Kanojia, D., & Bhattacharyya, P.* (2018). Semi-automatic WordNet Linking using Word Embeddings.
13. *Petrova, M., Druzhkina, A., Garashchuk, R., & Yudina, M.* (2018). Semi-automatic Integration of a new Language into a multilingual NLP model: the case of Japanese. Proceedings of the International Conference “Dialogue 2018”. Moscow.
14. *Ruder, S., Vulić, I., & Søgaard, A.* (2017). A Survey of Cross-lingual Word Embedding Models. The Journal of Artificial Intelligence Research.

АНАЛИЗ ПРОСОДИЧЕСКИХ ПРИЗНАКОВ ЭМОЦИОНАЛЬНОЙ ИНТОНАЦИИ С ИСПОЛЬЗОВАНИЕМ СИСТЕМЫ «INTONTRAINER» (на примере русскоязычных фраз)

Лобанов Б. М. (Lobanov@newman.bas-net.by),
Житко В. А. (zhitko.vladimir@gmail.com)
Объединённый институт проблем информатики НАН Беларуси, Минск, Беларусь

ANALYSIS OF PROSODIC FEATURES OF THE EMOTIONAL INTONATION USING “INTONTRAINER” SYSTEM (on the example of Russian phrases)

Lobanov B. M. (Lobanov@newman.bas-net.by),
Zhitko V. A. (zhitko.vladimir@gmail.com)
United Institute of Informatics Problems NAS Belarus, Minsk, Belarus

The main results of the update of the IntonTrainer system for the purposes of analyzing and studying the prosodic signs of emotional intonation are described. A distinctive functional feature of the updated system is the creation of an expanded set of prosodic signs of emotional intonation. The paper presents preliminary assessments of their effectiveness using the created experimental database of emotional phrases of Russian speech.

Keywords: Speech intonation, basic emotions, emotional intonation, melodic portrait, intonation analysis, software model

1. Введение

Хорошо известно, что человеческая речь передает не только смысловую, но и эмоциональную информацию. В теории эмоций многочисленные эмоциональные состояния часто отображаются в двухмерном пространстве: «приятное — не приятное», «активное — пассивное» [1], Существует множество различных дискретных наборов эмоций. Однако, большинство исследований ограничиваются анализом просодических характеристик следующих 6-ти эмоциональных состояний.

Нейтральность (спокойствие, сдержанность …) — уравновешенное состояние ума, никаких беспокойств, сомнений, волнений, забот.

Радость (восторг, упоение…) — положительное эмоциональное состояние, связанное с возможностью достаточно полного удовлетворения фактической потребности.

*Грусть (*тоска, уныние…) — негативное эмоциональное состояние, связанное с полученной информацией о невозможности удовлетворения важных жизненных потребностей

Гнев (возмущение, ярость…) — эмоциональное состояние, отрицательное по признаку, возникающее в форме аффекта и вызванное внезапным появлением серьезного препятствия.

Страх (испуг, тревога…) — негативное эмоциональное состояние, которое возникает, когда субъект получает информацию о реальной или воображаемой опасности.

Удивление (изумление, ошеломление…) — эмоциональная реакция на неожиданные обстоятельства.

Имеется также ряд эмоций, относимых довольно часто к основным, такие как*: Страдание, Отвращение, Презрение, Стыд,* а кроме того, многочисленные оттенки перечисленных выше эмоций.

До настоящего времени не имеется достаточных знаний о деталях акустических моделей, которые описывают определенные эмоции человеческого голоса. Типичные акустические характеристики, которые, как считается, вовлечены в этот процесс, включают следующее [2]–[4]:

- уровень, диапазон и форма контура основной частоты (F0);
- уровень вокальной энергии голоса;
- темп речи.

В последнее время исследованы новые важные речевые характеристики, которые содержат информацию о эмоциях, такие как частоты формант, коэффициенты линейного прогнозирования и коэффициенты мел-частотного кепстра [5]–[7]. Во многих работах особое место отводится анализу поведения мелодической кривой. В одной из последних работ, посвящённых анализу просодических характеристик эмоций [8], предложено использовать следующее описание контура высоты тона:

- Количество максимумов в контуре основного тона в вокализованном сегменте;
- Среднее значение и дисперсия пиковых значений;
- Средний наклон;
- Средний градиент между двумя точками выборки на кривой основного тона;
- Дисперсия градиентов основного тона.

Вопросам анализа и сопоставления контуров высоты тона различных интонационных конструкций посвящена также работа [9]. С начала 2018 года на вебсайте (см. https://intontrainer.by) выложена демо-версия программной системы «IntonTrainer». Она ориентирована на использование её в качестве компьютерного средства обучения интонации устной речи. В состав программного комплекса входят подсистемы, включающие наборы эталонных фраз, которые представляют основные интонационные модели русской, английской (британский и американский варианты), немецкой и китайской речи. В процессе обучения «IntonTrainer» осуществляет сравнение и оценку интонационного сходства произнесённой и эталонной фраз. Оценка интонационного сходства производится на основе представления тонального контура в виде универсального (унифицированного) мелодического портрета (УМП) [9]. Представляет интерес соответствующая доработка и модернизация системы «IntonTrainer» для целей анализа и исследования просодических признаков эмоциональной интонации.

2. Основные направления и результаты модернизации системы «IntonTrainer»

В задачу модернизации существующей системы входит создание такого программного средства, которое обеспечивает анализ и визуализацию эффективного набора просодических признаков эмоциональной интонации, а также даёт возможность предварительной оценки их эффективности с использованием доступных БД эмоциональной речи. Для того чтобы модернизированная система позволяла достаточно эффективно анализировать и визуализировать эмоциональные признаки интонации, в систему добавлен ряд новых дополнительных функций.

На **рис. 1** показан вид начального окна после загрузки системы.

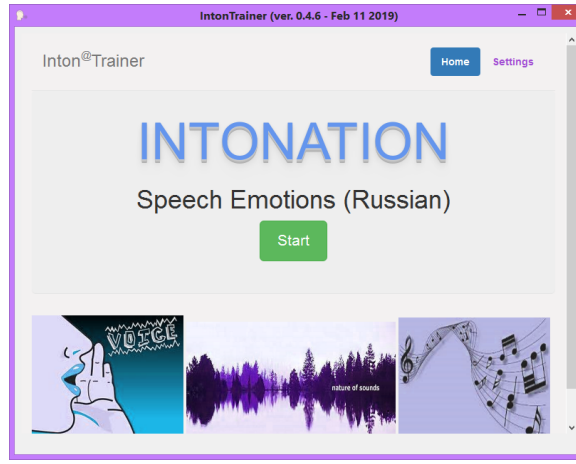


Рис. 1. Начальное окно

После нажатия кнопки «**Start**» открывается главное окно Программы, содержащее структурированный перечень эталонных фраз с указанием наименования и номера БД, имени диктора, названия эмоции и текста фразы, в которой она отражена (**рис. 2**).

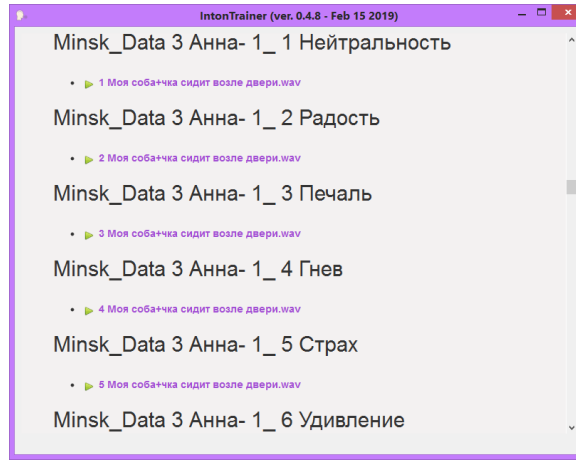


Рис. 2. Главное окно

Путём выбора с помощью курсора требуемой директории, например:

«Minsk_Data 3 Anna-1_1 Нейтральность 1 *Моя соба+чка сидит возле двери*»

открывается окно, в котором отображается в графическом виде результаты интонационного анализа этой фразы (**рис. 3**)

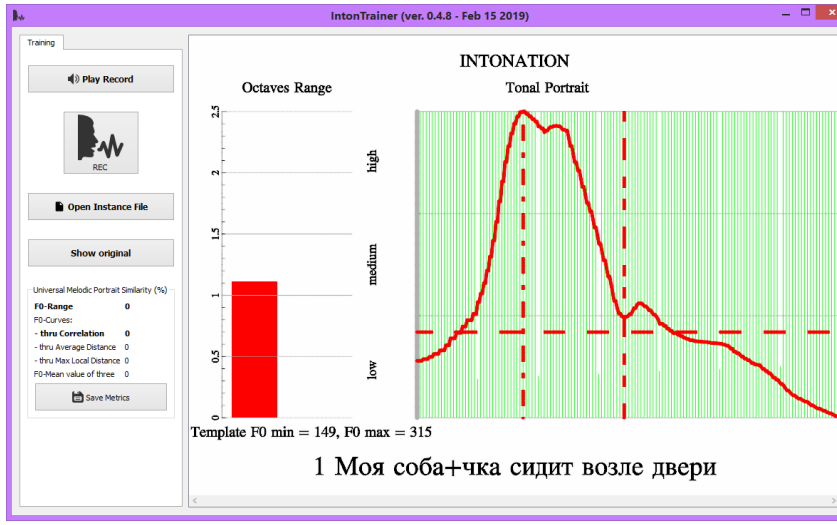


Рис. 3. Окно отображения кривой НМП: диктор «Анна-1» (*Нейтральность*)

Непрерывная кривая отображает траекторию изменения ЧОТ на голосовых участках фразы — нормированный мелодический портрет (НМП). Построение кривой НМП в отличие от универсального мелодического портрета УМП не потребует «ручной» разметки фразы на участки пред-ядра, ядра и заядра. При выборе в разделе **Setting** режима **Auto Marking** сегментация сигнала на голосовые регионы осуществляется автоматически на основе информации о наличии периодичности в сигнале (голоса) при одновременном присутствии достаточно высокой амплитуды сигнала — **A0(t)**.

Горизонтальная штриховая линия на кривой НМП показывает среднее значение нормированной ЧОТ. Две вертикальные линии характеризуют положение центра кривой и её ширину (размытость) на нормированной временной оси. Высота столбика (слева от НМП) показывает диапазон изменения ЧОТ в октавах.

В левой части на рис. 3 показаны кнопки управления, с помощью которых доступно осуществление следующих функций:

- «**Play Record**» — прослушивание перечня эталонных фраз.
- «**Rec**» — оперативная запись фраз пользователя через микрофон,
- «**Open Instance File**» — вызов тестовых фраз из папки «**TEST**»,
- «**Show original**» — просмотр исходных сигналов,
- «**Save Metrics**» — сохранение данных об измеренных просодических признаков.

При нажатии кнопки «**Save Metrics**» появляется дополнительный значок и открывается страница в EXCEL, на которой записывается полный набор из 10-ти просодических признаков эталонной фразы (см. табл. 1). Полученные данные сохраняются в той же папке, где хранится исследуемая эталонная фраза.

Таблица 1. Результаты вычисления просодических признаков: Анна-1 (*Нейтральность*)

№	Названия просодических признаков	Names of Prosodic Features	Results
1	Диапазон изменения ЧОТ [F0max/Fomin]	Pitch Range F0	**2,11**
2	Регистр ЧОТ [(F0max +F0min)/2]	Register F0 [Hz]	**232,00**
3	Среднее значение кривой НМП	Mean Value of the curve NMP	**27,97**
4	Положение центра кривой НМП	Center of the curve NMP	**36,99**
5	Эффективная ширина кривой НМП	Width of the curve NMP	**23,89**
6	Среднее значение кривой d/dt (НМП)	Mean Value of the Derivative curve NMP	**32,68**
7	Положение центра кривой d/dt (НМП)	Centh rangeer of the Derivative curve NMP	**46,26**
8	Ширина кривой d/dt(НМП)	Width of the Derivative curve NMP	**55,24**
9	Средний уровень звонких звуков	Voiced Sounds Level	**0,15**
10	Суммарная длительность звуков	Voiced Sounds Duration	**239,00**

В таблице 1 кроме данных о параметрах исходной кривой НМП, представлены также данные о величине её производной по времени — d/dt (НМП). Сравнительный вид этих кривых представлен на **рис. 4**. Дополнительный анализ параметров производной от НМП оказывается полезным для учёта динамических характеристик движения ЧОТ, характерных для некоторых видов эмоций.

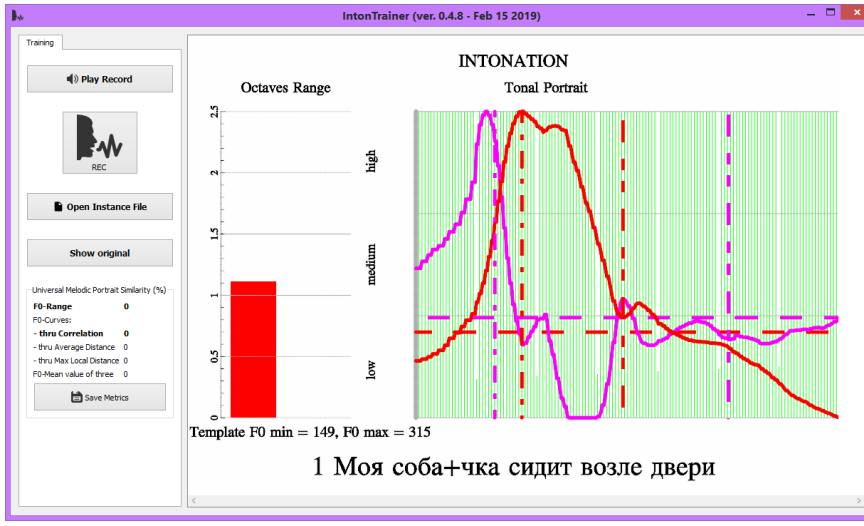


Рис. 4. Пример отображения кривых НМП и d/dt (НМП): Диктор «Анна-1» (*Нейтральность*)

При нажатии кнопки «**Open Instance File**» осуществляется вызов тестовых фраз из папки «**TEST**», в которой для сравнительного отображения различных эмоций могут быть помещены те же аудио файлы, что и в папке **PATTERNS**. На рис. 5 представлен пример сравнения фразы «*Моя собачка сидит возле двери*» с эмоцией «Нейтральность» (светлая линия) с той же фразой с эмоцией «Гнев» (тёмная линия). Как видно из рисунка, отличие этих эмоций наблюдается в форме кривых НМП и в их параметрах, таких как средние значения, центры кривых и их ширины.

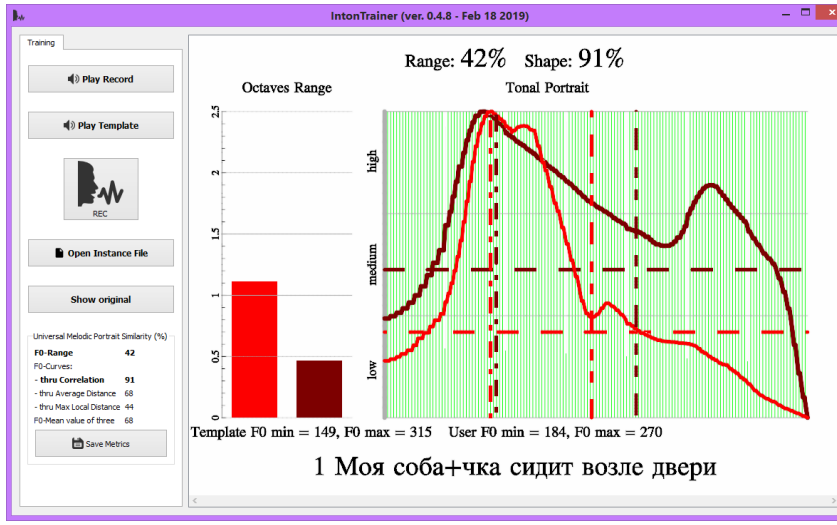


Рис. 5. Пример сравнения кривых НМП:
Диктор «Анна-1» (*Гнев/Нейтральность*)

Детальную информацию о результатах сравнения просодических признаков этих фраз можно получить при нажатии кнопки «**Save Metrics**» (см. таблицы 2, 3, 4).

В таблице 2 приведен пример результатов вычисления численных значений просодических признаков фразы, произнесённой с выражением эмоции «**Гнев**».

Таблица 2. Результаты вычисления просодических признаков: Анна (*Гнев*)

№	Названия просодических признаков	Names of Prosodic Features	Results
1	Диапазон изменения ЧОТ [F0max/Fomin]	Pitch Range F0	**1,45**
2	Регистр ЧОТ [(F0max +F0min)/2]	Register F0 [Hz]	**228,00**
3	Среднее значение кривой НМП	Mean Value of the curve NMP	**54,47**

№	Названия просодических признаков	Names of Prosodic Features	Results
4	Положение центра кривой НМП	Center of the curve NMP	**46,00**
5	Эффективная ширина кривой НМП	Width of the curve NMP	**37,40**
6	Среднее значение кривой d/dt (НМП)	Mean Value of the Derivative curve NMP	**49,28**
7	Положение центра кривой d/dt (НМП)	Center of the Derivative curve NMP	**44,31**
8	Ширина кривой d/dt (НМП)	Width of the Derivative curve NMP	**50,73**
9	Средний уровень звонких звуков	Voiced Sounds Level	**0,19**
10	Суммарная длительность звуков	Voiced Sounds Duration	**226,00**

В **таблице 3** приведен пример результатов вычисления по данным, приведенным в **таблицах 1** и **2**, относительных значений для просодических признаков пары фраз с эмоциями «**Гнев/Нейтральность**», выраженных в децибелах. Использование относительных величин позволяет осуществлять сравнение пары фраз с различными эмоциями, используя просодические признаки различной природы с оценкой в различных единицах измерения.

Таблица 3. Численные значения относительных признаков: Анна (*Гнев/Нейтральность*)

№	Названия просодических признаков	Names of Prosodic Features	Relations [dB]
1	Диапазон изменения ЧОТ [F0max/Fomin]	Pitch Range F0	**−0,67**
2	Регистр ЧОТ [(F0max +F0min)/2]	Register F0 [Hz]	**0,92**
3	Среднее значение кривой НМП	Mean Value of the curve NMP	**−1,59**
4	Положение центра кривой НМП	Center of the curve NMP	**−0,09**
5	Эффективная ширина кривой НМП	Width of the curve NMP	**2,38**
6	Среднее значение производной кривой НМП	Mean Value of the Derivative curve NMP	**0,65**
7	Положение центра производной кривой НМП	Center of the Derivative curve NMP	**1,41**
8	Эффективная ширина производной кривой НМП	Width of the Derivative curve NMP	**1,55**
9	Средний уровень звонких звуков	Mean Value of Voiced Sounds Level	**−0,07**
10	Общая длительность звуков фразы	Total duration of phrase sounds	**−0,26**

В **таблице 4** приведен также пример результатов вычисления численных мер сходства и расстояний между 2-мя предъявленными реализациями эмоциональных фраз (в данном случае пара: (*Нейтральность — Гнев*). Описание способов вычисления и соотвтствующие им формулы приведены в [9].

Таблица 4. Результаты вычисления мер сходства и расстояний: Анна (*Гнев/Нейтральность*)

№	Способ сравнения сходства	Type of the proximity	Prox-imity	Dis-tance
1	Коэффициент взаимной корреляции	Cross correlation coefficient	**91**	**9**
2	Интегральное сравнение кривых НМП	Integral comparison of NMP curves	**68**	**32**
3	Локальное сравнение кривых НМП	Local comparison of NMP curves	**44**	**56**
4	Среднее значение 3-х способов сравнения	Average of the three above proximities	**68**	**32**
5	Сравнение диапазонов изменения ЧОТ	Comparison of pitch ranges	**42**	**58**

3. Экспериментальная оценка разработанного набора просодических признаков

Для экспериментальной оценки эффективности разработанного набора просодических признаков эмоциональной интонации на примере русскоязычных фраз совершенно необходима соответствующая БД, подобная, например, англоязычной БД эмоциональных фраз [10], которая доступна для бесплатного скачивания. Существуют по крайней одна работа [11], в которой указывается на существование русскоязычной БД эмоциональных фраз, однако условия доступа к ней нам не известны. По этой причине для целей экспериментальной оценки разработанного набора просодических признаков мы решили создать собственную экспериментальную БД небольшого объёма по следующей методике.

Созданы специальные текстовые сценарии, провоцирующие диктора на выражение одной из 6-ти видов эмоций (*Нейтральность, Радость, Грусть, Гнев, Страх, Удивление)* при произнесении фразы «*Моя **собачка** сидит возле двери*» с фразовым акцентом на втором слове. Выбранная фраза является вольным переводом с английского фразы, используемой для тестирования в англоязычной БД эмоциональных фраз [10].

Таблица 5. Тексты сценариев, провоцирующих различные эмоции

№	Эмоция	1-й образец фразы	2-й образец фразы
1	Нейтраль	*Наконец я на даче.* *Кажется, я вижу её.* Моя **собачка** сидит возле двери.	*Да, действительно, это она:* Моя **собачка** сидит возле двери.
2	Радость	*Ура! Какое счастье!* *Она вернулась!* Моя **собачка** сидит возле двери!	*Маша, посмотри!* *Это же наш Шарик!* Моя **собачка** сидит возле двери.

№	Эмоция	1-й образец фразы	2-й образец фразы
3	Печаль	*Какая жалость...* *Её не взяли с собой…* Моя **собачка** сидит возле двери…	*Что же теперь делать...* *Бедная, бедная…* Моя **собачка** сидит возле двери…
4	Гнев	*Эй! Кто её выпустил!?* *Посмотри!* Моя **собачка** сидит возле двери!!!	*Ты что, не слышишь?!* Моя **собачка** сидит возле двери!!!
5	Страх	*Боже мой! Волки! Я боюсь!* *Она же щенок!* Моя **собачка** сидит возле двери!	*Волки уже близко!* *Что же делать?!!* Моя **собачка** сидит возле двери!
6	Удивлен	*Да, что ты говоришь? Это она?* Моя **собачка** сидит возле двери!?	*Неужели, правда? Неужели, она?* Моя **собачка** сидит возле двери!?

В качестве дикторов были привлечены 5 мужчин и 5 женщин из числа студентов театрального института и преподавателей русского языка для иностранцев. Перед ними была поставлена задача прочесть представленные в **таблице 5** тексты с максимально полной имитацией эмоциональных состояний, подсказываемых контекстом. В результате получены аудио записи фразы «*Моя **собачка** сидит возле двери*» по 2 варианта для каждой из 6 эмоций. Затем все записи в случайном порядке были представлены 3-м аудиторам, в задачу которых входило распознавание одной из 6 эмоций в предъявляемых образцах.

По результатам аудирования были отобраны наилучшие результаты: с женским голосом — «Анна» и с мужским — «Борис», которые использовались при проведении экспресс-исследования эффективности анализа просодических признаков эмоциональной интонации модернизированной системой «**IntonTrainer**». Однако, даже только для двух дикторов, полный анализ полученных данных и их графическое представление в данной работе представляется не реальным. Как видно из **таблиц 1–4**, разработанная система анализа просодических признаков генерирует для каждой пары эмоций 35 количественных показателей. Анализ же всевозможных пар становится возможным только с привлечением специальных программ обработки больших данных, например, с использованием нейросетевых алгоритмов.

Тем ни менее, для достаточно наглядной оценки качества предложенного набора просодических признаков эмоций мы посчитали возможным ограничиться здесь нижеследующим набором. На **рисунках 6–10** представлены графические диаграммы нормированных значений различных пар признаков, рассчитанных и усреднённых по двум реализациям: «Борис-1» и «Борис-2».

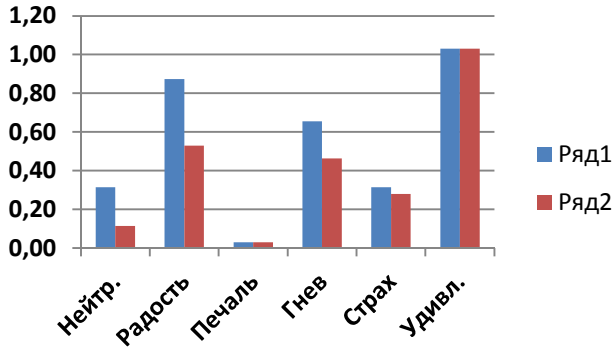


Рис. 6. Диапазон изменения ЧОТ (ряд 1), Регистр ЧОТ (ряд 2)

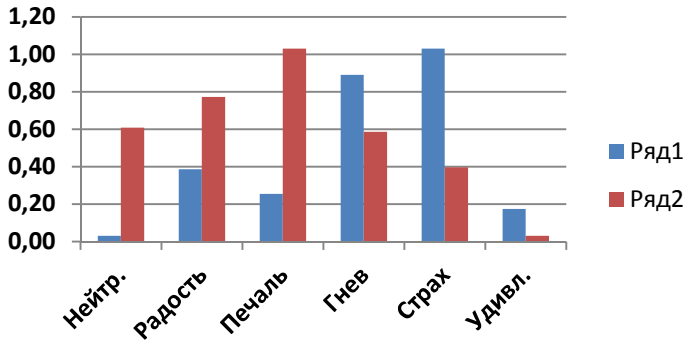


Рис. 7. Среднее значение кривой НМП (ряд 1) и её производной (ряд 2)

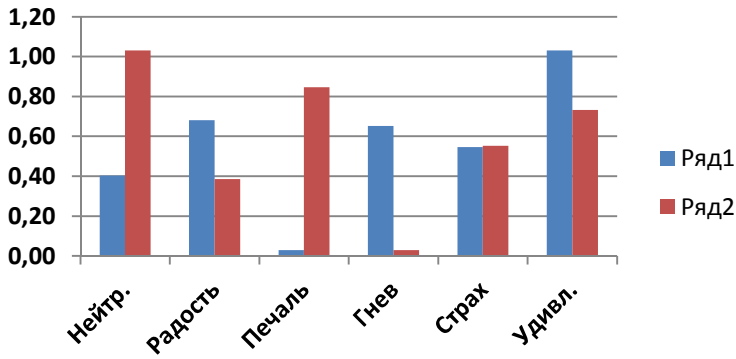


Рис. 8. Эффективная ширина кривой НМП (ряд 1)
и её производной (ряд 2)

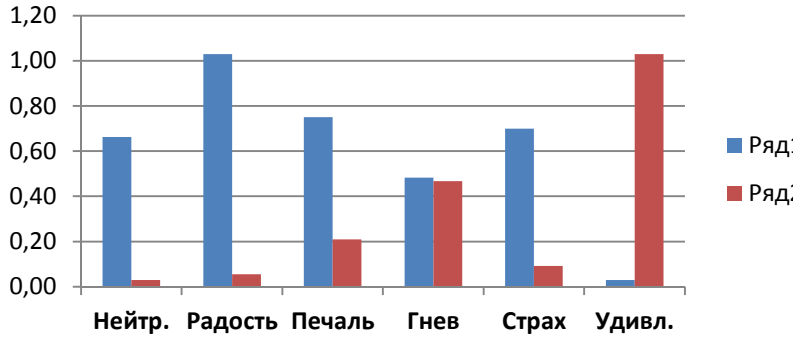


Рис. 9. Положение центра кривой НМП (ряд 1) и её производной (ряд 2)

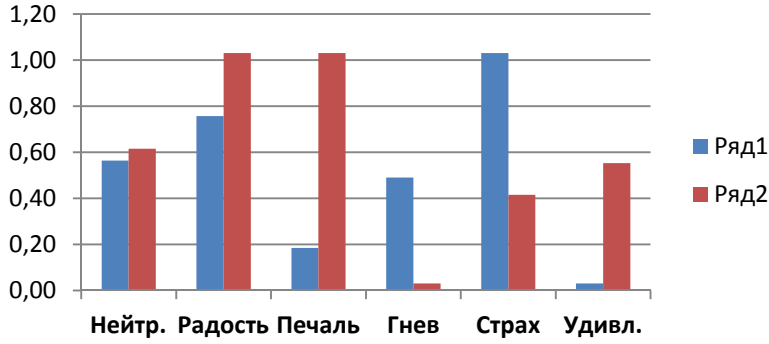


Рис. 10. Средний уровень звонких звуков МП (ряд 1) и их длительность (ряд 2)

Как видно из **рисунков 6–10**, полученные пары значений 10-ти просодических признаков сравнительно слабо коррелированы и характеризуются значительными различиями для каждого из 6-ти видов эмоций. Отметим некоторые очевидные результаты сопоставления просодических признаков эмоций. Из **рис. 6** видно, что эмоции «Радость», «Гнев» и «Удивление» характеризуются расширенным диапазоном изменения и высоким регистром ЧОТ. В то же время наименьшие значения этих признаков характерны для эмоций: «Печаль», «Нейтральность» и «Страх», что вполне соответствует нашим интуитивным представлениям. Менее очевидные результаты отражены на **рис. 7–10**. Их статистическая достоверность может быть проверена лишь после проведения более масштабных экспериментов.

4. Заключение

В настоящей работе описаны основные результаты модернизации системы «IntonTrainer» для целей анализа и исследования просодических признаков эмоциональной интонации. Отличительной функциональной особенностью обновлённой версии системы является реализация возможности расчёта численных значений расширенного набора просодических признаков, а также их сохранения в формате EXCEL таблиц. Модернизированная система установлена на сайте https://intontrainer.by под именем «***Russian Emotions IntonTrainer***» и доступна для бесплатного скачивания. На сайте помещена также подробная инструкция для пользователя.

В задачу модернизации системы не входило создание действующей модели распознавания речевых эмоций. Конечная цель доработки ограничивалась созданием такого программного средства, которое бы обеспечивало анализ и визуализацию расширенного набора просодических признаков эмоциональной интонации, и которое могло бы быть использовано как новое инструментальное средство для фонетических исследований речи. Не исключены, мы полагаем, и некоторые прикладные аспекты применения системы, например, в задачах обучения требуемой эмоциональной интонации актёров, а также людей различных профессий, стремящихся к повышению своего, так называемого, «эмоционального интеллекта (EQ). Этот новый термин появился сравнительно недавно и уже обсуждаются достоинства его применения в дополнение к IQ при оценке не только человека, но и системы искусственного интеллекта.

Литература

1. *Scherer, K. R., Schorr, A., Johnstone, T.* (2001) Appraisal Processes in Emotion: Theory, Methods, Research.Oxford University Press, New York and Oxford.
2. *Banse, R., Sherer, K. R.* (1996) Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology 70 (3), 614–636.
3. *Banse, R., Sherer, K. R.* (1996) Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology 70 (3), pp. 614–636.
4. *Abelin, A., Allwood, J.* (2000) Cross-linguistic interpretation of emotional prosody. In: Proceedings of the ISCA Workshop on Speech and Emotion.
5. *D. Ververidis, C. Kotropoulos, and I. Pitas* (2004) “Automatic emotional speech classification”, in Proc. 2004 IEEE Int. Conf. Acoustics, Speech and Signal Processing, vol. 1, pp. 593–596, Montreal, May 2004.
6. *Xiao, Z., E. Dellandrea, Dou W.,Chen L.* (2005) “Features extraction and selection for emotional speech classification”. 2005 IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 411–416, Sept 2005.
7. *T.-L. Pao, Y.-T. Chen, J.-H. Yeh, P.-J. Li* (2006) “Mandarin emotional speech recognition based on SVM and NN”, Proceedings of the 18th International Conference on Pattern Recognition (ICPR’06), vol. 1, pp. 1096–1100, September 2006.

8. *C. Rinaldi, R. Tedesco, M. Matteucci, and A. Trivilini* (2014) Extracting Emotions and Communication Styles from Prosody. Springer-Verlag Berlin Heidelberg, pp. 21–42.
9. *Lobanov, B. A.* (2018) Prototype of the Software System for Study, Training and Analysis of Speech Intonation / B. Lobanov, V. Zhitko, V. Zahariev // Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings / — Springer, 2018. — P. 337–346.
10. *Livingstone SR, Russo FA* (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.
11. *Veronika Makarova and Valery A. Petrushin* (2002) RUSLANA: a database of Russian emotional utterances / 7th International Conference on Spoken Language Processing / Denver, Colorado, USA, September 16&20, 2002.

A REUSABLE TAGSET FOR THE MORPHOLOGICALLY RICH LANGUAGE IN CHANGE: A CASE OF MIDDLE RUSSIAN[1]

Lyashevskaya O. N. (olesar@yandex.ru)

National Research University Higher School of Economics; Vinogradov Institute of the Russian Language RAS, Moscow, Russia

The paper discusses the standardization efforts to create a morphological standard for the Middle Russian corpus, which is part of the historical collection of the Russian National Corpus (RNC). To meet the needs of different categories of corpus researchers as well as NLP developers, we consider two styles of the morphological annotation (RNC schema and Universal Dependencies schema). A number of specifications of the feature list proposed to facilitate data reusability, linking and conversion.

Key words: full morphology tagging, pos-tagging, lemmatization, tagset, historical corpora, Russian National Corpus, Universal Dependencies, Old Russian, Middle Russian

МНОГОЦЕЛЕВОЙ МОРФОЛОГИЧЕСКИЙ СТАНДАРТ РАЗМЕТКИ ДЛЯ ЯЗЫКА С МЕНЯЮЩЕЙСЯ ГРАММАТИЧЕСКОЙ СТРУКТУРОЙ: СЛУЧАЙ СТАРОРУССКОГО КОРПУСА

Ляшевская О. Н. (olesar@yandex.ru)

Национальный исследовательский универститет «Высшая школа экономики»; Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Статья посвящена созданию морфологического стандарта для разметки Старорусского корпуса, который входит в состав исторических корпусов Национального корпуса русского языка (НКРЯ). Для того,

[1] The article was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) and supported within the framework of a subsidy by the Russian Academic Excellence Project '5-100'.

чтобы сделать разметку удобной для лингвистов, работающих с историческими и современными корпусами, а также для разработчиков систем автоматической обработки исторических текстов, мы предусматриваем две параллельные схемы морфологической разметки, в нотации НКРЯ и Универсальных зависимостей (Universal Dependencies). Предлагается ряд спецификаций тагсета для облегчения совмещения разметок разных корпусов, связывания и конвертирования данных.

Ключевые слова: лексико-грамматическая разметка, частеречная разметка, лемматизация, тагсет, исторические корпуса, Национальный корпус русского языка, древнерусский язык, старорусская письменность

1. Introduction

Middle Russian Corpus (MidRus) is part of the Russian National Corpus (http://ruscorpora.ru) included in the collection of historical corpora [Sichinava 2014]. The MidRus contains over 4,700 texts of different genres written mostly between 1,300 and 1,700 (over 7 million words). Up to now, only a simple search for word forms and their parts has been available in the corpus interface. The paper represents the first attempt to develop the full morphology annotation standard for the MidRus.

Tagging the parts of speech, inflectional grammatical categories, and lemmas in historical corpora is a challenging task, since from one period to another, the grammatical structure changes: some grammatical forms drop out of use whereas new categories and grammatical patterns appear, the structure of the intra- and interparadigmatic homonymy varies. Furthermore, grammar and lexicon varies across schools and manuscripts, the texts often have noticeable dialect and stylistic features as well as varying and unstable spelling. While developing the full morphology annotation of the MidRus, we take into account the academic interests of the different categories of users including:

- researchers in the Middle Russian period of the language;
- researchers of the older periods of Russian accustomed to the annotation schemas of the Old Russian RNC corpus (OldRus) and the Old Novgorodian/East Slavic birchbark letters RNC corpus (OldNovg);
- researchers of the modern language who are interested in the micro-diachrony studies and are used to the tagset of the RNC Main corpus (ModernRus);
- NLP researchers who would be likely to use the Middle Russian data in their computational experiments, including comparative ones based on various paleoslavic data collections.

What makes things more challenging is that the annotation standards for the corpora of the earlier period and the modern period of Russian are well established but differ with regard to the lists of tags, the boundaries of lexical classes to which they apply, attested combinations of tags representing particular grammatical forms, and lemmatization rules. Therefore, we need to adopt existing schemas while evaluating contradicting data and clarifying the boundaries of the phenomena.

The last, but not the least issue that deserves attention is data reusability and customization. In recent years, new cross-language standards have gained popularity in NLP as they allow one to accumulate data of different origin and reuse and deploy the language technologies developed in the community.

To meet these new trends, the morphological annotation standard of the MidRus adopts two tagsets in parallel:

- RNC-MidRus: RNC Middle Russian tagset close to those of the Main RNC corpus, Old Russian, and Old Novgorod corpora;
- UD-MidRus: Universal Dependencies (UD) tagset close to those of the UD-Church Slavic and UD-Russian data collections.

As for the tagset customization, we distinguish among the core annotation schema (RNC and UD), an extended schema (RNC-ext and UD-ext), and a simplified schema encompassing only a selection of tags shared by the UD-MidRus and other UD corpora (UD-s).

The paper is structured as follows. **Section 2** outlines the state of the art in the field of historical Russian corpora and available NLP technologies. **Section 3** focuses on the part-of-speech tagging, **Section 4** covers the core grammatical tags, and **Section 5** is devoted to the analytical forms. The optional tags, extended and simplified annotation schemas are discussed in **Sections 6**, **7**, and **8**, respectively. Unless otherwise stated, the paper will refer to the core annotation schema, and the UD tags will be explicitly marked UD, if needed.

2. Historical Russian corpora and tagging methods

In this section, we overview the known historical corpora for Russian and methods for their tagging. Apart from the MidRus, there are three diachronic corpora in the RNC: OldRus, OldNovg, and Church Slavic (ChurchSlav) corpus [Moldovan 2015]. The Old Russian corpus [Mishina, Pichkhadze 2015] is provided with manual lemmatization and morphological annotation. The tool Morphy [Arkhangelsky et al. 2014] suggests annotations known from the texts which were tagged before. The original (Russian) tags are then translated into the (latin) tags used by the RNC search engine. The tagsets of the OldNovg [Sichinava 2018] and ChurchSlav [Dobrushina et al. 2015] are similar to those used in OldRus but differ in details. The annotation of the OldNovg is done semi-manually whereas the ChurchSlav is tagged automatically. An additional annotation of ambiguous word boundaries, fragmented tokens and comments on possible interpretations is available in OldNovg and, to a lesser extent, in OldRus. Moreover, the analyses in the OldNovg are most theoretically motivated, since they are based on the foudational work by [Zaliznyak 2004].

The annotation of the Northern Russian hagiographic corpus SCAT [Alexeeva, Azarova 2013] is done manually and follows an in-house extension of the TEI schema [Alexeev 2011]. The annotation features labeling the declension types.

The web page of the Regensburg Russian Diachronic Corpus mentions a “best bet” method based on the output of three taggers: Regensburg Old Church Slavonic tagger, Regensburg Old Russian guesser, and the modern Russian model of TreeTagger.

[Meyer 2011] adds that the main source is the annotation projection from modern translations.

The corpus Manuscript [Baranov et al. 2007] is partially tagged using a sophisticated rule-based pipeline which is powered by the Old Russian grammatical dictionary, modern grammatical dictionary, and a dictionary of pseudo-units. The tool carries out lemmatization and provides normalized orthographic representations.

The TOROT treebank [Eckhoff, Berdičevskis 2015] is an Old Russian add-on to the PROIEL Old Church Slavonic (OCS) treebank, which uses the same annotation environment and tagset. The texts are tagged manually, lemmas and annotations being provided with the aid of statistical preprocessing [Berdičevskis et al. 2016]. Currently, the data are released offline in MULTEXT-East XML format, and the PROIEL OSC data are also converted into the UD-CONLL format (the Old Russian TOROT data are planned to be released in UD in 2019).

To sum up, the morphological tagsets for many corpora described above are hardly available (see also detailed reviews in [Mitrenina 2014], [Eckhoff forthc.]). The most popular tagset is RNC (which exists in a few slightly different versions); MULTEXT-East and UD schemas are most accessible for NLP purposes due to the open license of the TOROT data.

Among the tagging methods, labeling by precedents, dictionary- and rule-based systems, and the projection of the modern Russian annotations are widely used. However, remarkably, other methods pave the way for the statistical learning: [Berdičevskis et al. 2016] compares the output of the HMM-based probabilistic tagger TnT and a hybrid system that makes use of the grammatical dictionary. [Scherrer et al. 2018] run computational experiments using conditional random fields method (CRF, tagger MarMoT) and deep neural network learning (char-embedding BLSTM). It is worth noting that since the amount of machine readable data is very modest and the historical data do not have a homogeneous structure with respect to their tagsets, this could potentially foster the interest of NLP developers to the material. Thus, the harmonization of data annotation is obviously crucial for improving the quality of tagging.

3. Parts of speech

The lists of part-of-speech (pos) tags and core grammatical features is available at: https://github.com/olesar/UD_MidRussian/blob/master/MidRussianUD.md. The document also reports the mapping between the RNC and UD tags. To evaluate the mismatches in the corpus annotation practice, we compared all attested combinations of pos-tags and features as well as their association with lemmas (lexical coverage) in OldRus, OldNovg, TOROT, UD-Church Slavic, and ModernRus.

In general, the RNC pos-list can be mapped to the UD UPOS list almost straightforwardly. The pos-tags for adjectives (A), ordinal numerals (ANUM), and the most part of predicative words (PRAEDIC, see below) are mapped to ADJ in UD; the pos-tags for adverbs and parenthetic words (ADV, ADVPRO, PARENTH) are mapped to ADV in UD. The noun tags (S in RNC) are mapped to NOUN (common nouns) and PROPN (proper nouns), and the verb tags (V in RNC) are splitted between VERB and AUX (auxiliaries)

in UD. The RNC tag CONJ is splitted between CCONJ (coordinate conjunction) and SCONJ (subordinate conjunction). The non-words (NONLEX) are splitted into X (foreign words, unknown words) and SYM (symbols). Besides that, the punctuation marks are explicitly tagged PUNCT in UD.

In the remainder of the section, we consider the mismatches in the annotation schemas with respect to the lexical coverage of pos categories in RNC and UD.

3.1. Pronominal words

И, *е*, *я* are tagged `spro` (UD: `pron`), the same way as in OldRus. Similarly, *иже*, *еже*, *яже* are tagged SPRO in RNC and PRON in UD. (In OldRus, they are tagged eigher APRO or SPRO, but we follow the principle to label a lemma uniformly as much as possible).

The relative pronouns *который*, *кыиждо*, *кыиже* are tagged APRO in RNC and PRON in UD. The reason is that they have the morphological properties of an adjective and the syntactic properties of a noun (nominal head), and this solution has already been implemented in the modern Russian UD [Droganova et al. 2018].

The possessives *его*, *ея*, *ихъ*, etc. are tagged as the Genitive forms of *онъ*, *оно*, *она*, *онъ*, *они*: SPRO, `gen` (UD: PRON, `Case=Gen`). (In OldRus, they are tagged as the Genitive forms of *и*; in ModernRus, they are tagged as indeclinable adjectival pronominals *его*, *ее*, *их*).

The list of APRO (UD: DET) includes:

- interrogative, relative, negative adjectival pronouns, quantifiers: *каковый*, *никакий*, *вьсь*;
- deictic (demonstrative) words: *сей*, *овъ*, *таковый*, etc.;
- possessive adjectival pronouns: *мой*, *свой,* etc.

The numeral *одинъ* is tagged ANUM in RNC and NUM in UD. In tagging it ANUM, we follow the practice of ModernRus (*один* has an adjective-like paradigm and is used as an attribute: for example, in the Nominative, it does not govern the Genitive case of the noun phrase compared to other numerals, see [Zaliznyak 2003]). However, in the UD treebanks the pos-tag NUM is applied consistently to the lexical equivalents of *один*. In OldRus, *одинъ* is labeled NUM as well.

3.2. Predicative words

Since there is no general mapping for the RNC PRAEDIC class to UPOS tags in UD, we use the conventions similar to those of the modern Russian UD standard:

- *-о*, *-е/-ь* forms (cf. (*ночью*) *тяпло*, *пригоже*, *явно*) that have corresponding adjectival forms are tagged as the short neutral forms of adjectives (UD: ADJ, `Gender=Neut, Number=Sing, Variant=Short`);
- the modal words—*можно*, *льзх*, *надобно*, *уне*—and the negative existentials *нхтъ*, *нх* are tagged VERB in UD;
- nouns such as *пора* used predicatively (cf. *пора идти*) are tagged as s in RNC and NOUN in UD;
- interjections, onomatopoeic words used predicatively are tagged as INTJ.

3.3. Auxiliaries

AUX in UD is used to tag:

- the auxiliary use of *быти*, *имѣти*, *хотѣти* in the analytical verb forms; this also includes the conditional markers *бы*, *бъ*—originally, the forms of *быти*, too, which got grammaticalized as indeclinable particles by the end of the Middle Russian period;
- the copula use of *быти* in nominal clauses;
- the reflexive markers (clitics) *си*, *ся*.

Only the existential and locative uses of the verb *быти* are tagged VERB in UD.

In the RNC schema, *бы* and *бъ* are subject of a double tagging strategy: they are labeled as verbs (lemma *быти*) and particles.

3.4. Named entities

The patronymics, last (family) names, nicknames and family nicknames and the like are tagged s (UD: PROPN): *Васильевичю*, *Колюбакинымъ*. This also applies to naming formulae with non-agreeing and agreeing possessive forms such as *Ивану Ильину сыну Челищева*, *Семену Васильеву сыну Власьеву*. The only exception are forms with full adjectival endings such as *Борисовую* in *княгиню Борисовую* and *Ондрѣевскую* in *Ефросинию, княж Ондрѣевскую жену Ивановича* which are considered adjectives (cf. the same practice in OldRus: *бабы* (*своеи*) *Романовои*). Note that in TOROT, the patronymics are sometimes considered adjectives.

4. Core grammatical tags

This section highlignts only key grammatical categories that distinguish the annotation schema of MidRus from those of OldRus or ModernRus.

4.1. Animacy

Animacy (`anim`; UD: `Animacy=Anim`) is tagged in the Accusative construction in which the form of Accusative is equal to the Genitive form, cf. *брата нашег*[*о*] *молодшег*[*о*]. In OldRus, such forms aree tagged `accgen`. The opposite case, when the Accusative case form is equal to the Nominative form, is not marked in MidRus.

4.2. L-form (indeclinable perfective participle)

L-participles (cf. *взялъ*) are tagged `perf` (UD: `VerbForm=PartRes, Tense=Past`), to distinguish them from other participles (cf. *взявъ*: `past partcp`; UD: `VerbForm=Part, Tense=Past`). The tense tag in UD will allow one to map the MidRus l-forms to the ModernRus past forms. L-forms are used both on their own and within the analytical forms, see below.

4.3. Gerundive (indeclinable adverbial participle)

Following [Zaliznyak 2004], forms such as *уповая*, *слышев* are considered indeclinable gerundives: `ger` (UD: `VerbForm=Conv`).

5. Analytical forms

The analytical forms are annotated as two (or more) tokens cross-linked at the morphological (in OldRus) and syntactic (in UD) level. All tokens are tagged `analyt` (UD: `Analyt=Yes`) and the grammatical features of the analytical form as a whole are labeled on the content word, cf. the annotation of the clause (1) *а будет не дошла* 'And if it won't reach (you)' in RNC (**Fig. 1**) and UD (**Fig. 2**).

(1)
```
а       <ana lex="а" gr="CONJ"></ana>
будет   <ana lex="быти" gr="V,3p,act,analyt,fut,indic,intran,sg" gr_ext="IN:FUT2+3312"></ana>
не      <ana lex="не" gr="PART"></ana>
дошла   <ana lex="доити" gr="V,act,analyt,f,fut2,intran,perf,pf,sg" gr_ext="IN:FUT2+3310"></ana>
```

Figure 1: A sample annotation in RNC-MidRus

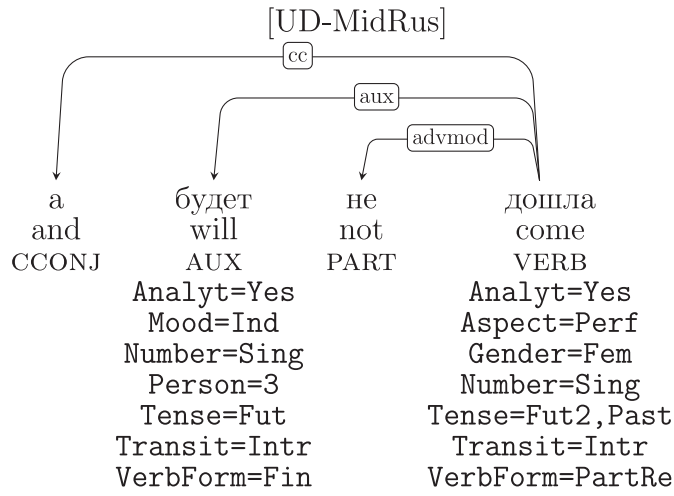


Figure 2: A sample annotation in UD-MidRus

In example (1), number, person, and future tense are labeled on the auxiliary *будет*: `sg`, `3p`, `fut` (UD: `Number=Sing`, `Person=3`, `Tense=Fut`), and gender, number, l-form are labeled on the content verb *дошла*: `f`, `sg`, `perf` (UD: `Gender=Fem`, `Number=Sing`, `Tense=Past`, `VerbForm=PartRes`): these are the intrinsic grammatical values of the tokens. The content word is also labeled by the tense of the whole analytical form `fut2` (UD: `Tense=Fut2`). Furthermore, *будет* is tagged AUX (part of speech) and `aux` (dependency relation) in UD.

The list of analytical forms includes:

- analytical future (new form, attested starting from the 1600s): infinitive + the future form of *быти* (*буду*, *будешь*), cf. *буду просить*: in new future forms, the content verb is tagged `fut` (UD: `Tense=Fut`);
- future 1: infinitive + the auxiliary nonpast forms of *хотѣти* and *имѣти*, cf. имет *обидѣти*: in: in the future 1 forms, the content verb is tagged `fut1` (UD: `Tense=Fut1`);
- future 2: l-form + the future form of *быти* (*буду*, *будешь*), cf. *боудеш*[ь] *послал*, *боудоу задѣла*: in the future 2 forms, the content verb is tagged `fut2` (UD: `Tense=Fut2`). Note that in OldRus, the analytical forms with почати, *начати*, *учати*, *стати*, *яти* are also labeled as the future 1 or future 2 forms, but we do not consider them as such in MidRus;
- analytical perfect: l-form and the 1st and 2nd person auxiliary in the present tense (*есмь*, *еси*, etc.), cf. *взял еси*;
- pluperfect (plusquamperfect): l-form + the perfect form of *быти*, cf. *дал еси был*, the content verb is tagged `pqperf` (UD: `Tense=Pqp`);
- subjunctive (conditional): l-form + *бы*, *бъ*, other aorist forms of *быти* (or conjunctions that incorporate *-бы*: *чтоб*(ы), *абы*, etc.), cf. *я бы сталъ*, *чтоб онъ пожаловалъ*: in conditional forms, the content verb is tagged `cond` (UD: `Mood=Cnd`);
- subjunctive (conditional) 2: l-form + *бы еси*, *бы есте* (2nd person forms of *быти*), cf. *держали бы есте* (*веру христианскую*): in conditional 2, the content verb *бы* is tagged `cond2` (UD: `Mood=Cnd2`).

The optative construction (*да* + non-past), the periphrastic comparative constructions of adjectives and adverbs are not considered analytical forms, nevertheless, they can be labeled with specific optional tags.

6. Optional tags

6.1. Features not available in automatic annotation

The following categories used of OldRus and OldNovg can be identified only in a particular context, often with the assistance of encyclopedic knowledge. In MidRus, they are used optionally in manual annotation:

- `as _ S` (UD: `AdjType=Subst`)—substantivized use;
- `as _ persn` (UD: `AdjType=Persn`, `NounType=Persn`)—used as a personal name. In particular, old nicknames such as *Мономахъ* are not counted as the last names and tagged `as _ persn`;
- `as _ topn` (UD: `NounType=Topon`)—used as a toponym;
- `as _ ethn` (UD: `NounType=Ethn`)—used as an ethnonym;
- `as _ ADV` (UD: `NounType=Adv`, `AdjType=Adv`)—used as an adverb, cf. (*придоша Ветроу*) *вечеръ*, (*по*) *готовоу*;
- `as _ PART` (UD: `VerbType=Part`)—used as a particle, cf. *хотя*;

- as _ PARENTH (UD: AdvType=Parenth; pos-tag PARENTH in RNC-ext, see below)—parenthetical use;
- as _ PRAEDIC (UD: AdjType=Praedic; pos-tag PRAEDIC in RNC-ext)—predicative use;
- as _ deb (UD: VerbType=Debit)—used as a debitive, cf. *да не погубиши мьзды своеа*.

The following tags are used optionally and only in the RNC-style annotation:

- husbn—distinguishes the name given by husband's name from patronymics, cf. OldNovg (*оу*) *тоудоровъи*;
- in _ persn—used within a personal name, cf. *анастасу корсунянину*;
- in _ ethn—used within an ethnonym, cf. *Черніи Клобуци*;
- in _ topn—used within a geographic name, cf. (в) *Константинъ градъ*;
- in _ ADV—used in an adverbial phrase, cf. *тако же*;
- in _ NUM—used within a complex numeral, cf. *двъма на десяте*;
- in _ CONJ—used within a multitoken conjunction, cf. *егда како*;
- in _ PR—used within a multitoken preposition, cf. *в мъсто*.

In UD, there are ways to encode most of such cases with the dependency relation tags (e. g. flat:name and fixed).

6.2. Spelling and non-standard variants

The feature abbr (UD: Abbr=Yes) is used to tag abbreviated words including those marked by titlo.

- The feature ciph is used in RNC schema to label cardinal and ordinal numerals expressed by (Euro-Arabic) digits and Cyrillic letters. In UD-MidRus, the corresponding tag NumForm=Digit is used to label cardinal and ordinal numerals expressed by (Euro-Arabic) digits (*за 5 верстъ, 5-ти дней, лета 7030-го июля в 9 день*);
- NumForm=Cyril—used to tag numerals expressed by Cyrillic letters (*КЕ ал, по Д чысло*);
- NumForm=Word—used to tag numerals expressed by words (*одинъ, первый, лѣта семь тысячь девятаго*).
- The feature distort (UD: Typo=Yes) is used to label distorted words and words guessed by the editors of the historical manuscrupts. Specific cases include (RNC-style only):
- damaged—guessed words (if the text segment is damaged);
- crossed _ out—crossed out, cf. OldRus: (*и ко полотьску*)
- redundant—redundant word (*не не сподобилъ же еси*). Note that in UD, the feature Echo can be used to label various kinds of repetitions.

The feature anom (not tagged in UD) is used to tag grammatically anomalous forms. However, what is considered 'grammatically anomalous' in the historical data is controversial and theory-specific. Therefore, this tag should be used with caution.

Finally, `oov` (cf. `bastard` in ModernRus, not tagged in UD) is a specific kind of tags which is used to label words not seen in the training data or the grammatical dictionary of the tagger.

7. Extended annotation schema

We introduce the notion of cross-features (or x-features) that can be added into the schema to make the annotations in different corpora comparable. For example, in micro-diachronic studies, the data of the modern language are compared against the historical data. Even if a certain grammatical category is under development and it is not evident if it is present or absent in the data, x-features allow one to look for the potentially interesting patterns. In the current Middle Russian standard RNC-ext, the x-features include:

- `anim$` and `inan$` (UD: `Animacy[lex]=Anim, Animacy[lex]=Inan`): classifying features that correspond to anim and inan in the ModernRus annotation. This category is not to be mixed with `anim` (UD: `Animacy=Anim`) that is applicable only to the Accusative constructions (see above). There are cases in which the lexically animated nouns (`anim$`) are not tagged as `anim`;
- the transitivity tags `tran` and `intr` (UD: Transit=Tran, `Transit=Intr`). The transitivity is tagged often inconsistently in modern corpora, and the situation is even worse in historical corpora. However, this is an interesting category under development that allows a user to study various morphosyntactic phenomena.

Another example is the use of cross-features to make the data conversion between different formats more straightforward. So, in the intermediate schema UD-ext, an extended list of parts of speech is used which includes ANUM, PRAEDIC, PARENTH. Further, a number of cross-features under the category NounType are introduced in UD-ext to reflect RNC tags such as `persn, patrn, famn, zoon, ethn, topon` (e.g. NounType=Ethn).

8. Simplified annotation schema

An alternative option to make data compatible is reducing the lists of tags. This is particularly useful in NLP evaluation tasks since the dominance of features carefully designed for human research but rarely attested in corpora can cause the drop in tagging performance. In order to make the tagsets of historical corpora available in UD (UD Church Slavic (UD-PROIEL), UD-TOROT and UD-MidRus) compatible, the following features can be excluded from annotation:

- `Aspect` (verb aspect)
- `Reflex` (reflexivity labeled on verbs and pronound)
- `Animacy` (Acc=Gen)
- `PronType` (pronominal type)
- `Variant` (long/short forms)
- `Strength` (a rough equivalen for Variant in UD-PROIEL/TOROT)

Except for Variant/Strength and Animacy, these features are lexical (classifying) and do not add to the identification of which paradigm cell the form fills. Obviously, extended and optional features are out of the simplified list as well.

In addition, the tense forms of aorist (`Tense=Aor`) and imperfect (`Tense=Imp`) should be relabeled as `Tense=Past` according to the universal UD guidelines (and thus mirroring the annotation in UD-PROIEL/TOROT).

9. Conclusion

We have presented the annotation standard for the Middle Russian corpus, detailing guidelines to the tagging of part-of-speech and morphological features in RNC and UD schemas and introducing a mapping between the RNC and UD tags. We distinguish between core, extended and simplified tagsets and show that different categories of users can benefit from them.

The annotation schemas were evaluated and corrected while doing the manual annotation of the MidRus gold standard [Lyashevskaya 2018], on the one hand, and carrying out computational experiments in automatic tagging and training data amplification [Scherrer et al. 2019], on the other hand. The test sample was annotated manually in both standards, RNC and UD, in parallel. After data conversion from RNC to UD-s, the inter-annotator agreement was calculated over a total of 400 tokens. The ratio of equivalent annotations was considerably high (95%).

A pilot version of the gold standard MidRus data is released with open license in Universal Dependencies, v2.4.

Acknowledgements

We are grateful to Irina Juryeva, Roman Ilushin, Maria Skachedubova, Elizaveta Bunina, and Dmitri Sitchinava who contributed to the annotation of the Middle Russian gold standard data and revision of the annotation guidelines. We would also like to thank Anna Pichhadze, Alexandr Moldovan, Vladimir Plungian, Roman Krivko, Yves Scherrer, Achim Rabus, Hanne Eckhoff for fruitful discussion and advice.

References

1. *Arkhangelsky T. A., Mishina E. A., Pichkhadze A. A.* (2003), A tool for the electronic grammatical annotation of Old Russian and Church Slavonic texts and its use in web resources [Sistema elektronnoj grammaticheskoj razmetki drevnerusskikh i tserkovnoslavjanskikh tekstov i jejo ispol'zovanie v veb-resursakh], Baranov V. A., Zheljazkova V., Lavretiev A. M. (eds.), Textual heritage and information technologies. El'Manuscript–2014 [Pismenoto nasledstvo i informatsionnite tekhnologii. El'Manuscript–2014]. Proceedings of the 5th International research conference, Sofia, Izhevsk, 2014.

2. *Alexeev V. A.* (2011), Expansion and implementation of the format for describing the grammatical and graphic data of the SKAT corpus [Rasshirenie i realizatsija formata opisanija grammaticheskikh i graficheskikh dannykh korpusa SKAT]. Master's thesis, St.-Petersburg, St.-Petersburg state university.
3. *Alekseeva E. L., Azarova I. V.* (2013), Peculiarities of the morpho-syntactic annotation for the Old Russian hagiographic texts [Osobennosti morfo-sintaksicheskoy razmetki drevnerusskikh agiograficheskikh tekstov], Proceedings of the International conference "Corpus linguistics-2013", St.-Petersburg, pp. 157–164.
4. *Baranov V. A., Mironov A. N., Lapin A. N. et al.* (2007), Automatic morphological analyzer of Old Russian language: linguistic and technological solutions [Avtomaticheskij morfologicheskij analizator drevnerusskogo jazyka: lingvistichekie i tekhnologicheskie reshenija] 10th jubilee international conference EVA 2007, Moscow.
5. *Berdičevskis A., Eckhoff H. M., Gavrilova T.* (2016), The beginning of a beautiful friendship: rule-based and statistical analysis of Middle Russian, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2016", Moscow, pp. 99–111.
6. *Dobrushina E. R., Kravetsky A. G., Poljakov A. E.* (2015), A corpus and a frequency grammatical corpus-based dictionary of Church Slavonic in the collection of the Russian National Corpus [Korpus i chastotnyj grammaticheskij korpusnyj slovar' tserkovnoslavjanskogo jazyka v sostave Nacional'nogo korpusa russkogo jazyka], Research papers of Vinogradov Institute od the Russian Language [Trudy Instituta russkogo jazyka im. V. V. Vinogradova], Vol. 6 (6).
7. *Droganova K., Lyashevskaya O., Zeman D.* (2018), Data Conversion and Consistency of Monolingual Corpora: Russian UD Treebanks Proceedings of the 17th IntGernational Workshop on Treebanks and Linguistic Theories (TLT 2018), Oslo, pp. 52–65.
8. *Eckhoff H. M.* (forthc.), Historical corpora and the re-evaluation of Slavonic language history.
9. *Eckhoff H. M., Berdičevskis A.* (2015), Linguistics vs. digital editions: The Tromsø Old Russian and OCS Treebank, Scripta & e-Scripta, Vol. 14–15, pp. 9–25.
10. *Lyashevskaya O.* (2018), A test dataset for the automatic morphological analysis of the Middle Russian texts [Testovaja kollektsija dlja zadach avtomaticheskogo morfologicheskogo analiza tekstov starorusskoj pis'mennosti], The academic heritage of V. A. Bogoroditsky and the modern vector of research of the Kazan linguistic school [Nauchnoje nasledije V. A. Bogoroditskogo i sovremennyj vektor issledovanij Kazanskoj lingvisticheskoj shkoly], Works and materials of int. conf., Kazan: Kazan University, pp. 131–135.
11. *Meyer R.* (2011), New wine in old wineskins? Tagging Old Russian via annotation projection from modern translations, Russian linguistics, Vol. 35 (2), pp. 267–281.
12. *Mishina E. A., Pichkhadze A. A.* (2015), Old Russian subcorpus of the Russian National Corpus [Drevnerusskij podkorpus Nacional'nogo korpusa russkogo jazyka], Research papers of Vinogradov Institute od the Russian Language [Trudy Instituta russkogo jazyka im. V. V. Vinogradova], Vol. 6 (6).
13. *Mitrenina O.* (2014), The corpora of Old and Middle Russian texts as an advanced tool for exploring an extinguished language, Scrinium: Journal of Patrology, Critical Hagiography, and Ecclesiastical History, Vol. 10 (1), pp. 455–461.

14. *Moldovan A. M.* (2015), Old Russian manuscripts in the Russian National Corpus [Pamjatniki drevnerusskoj pis'mennosti v Natsional'nom korpuse russkogo jazyka], Research papers of Vinogradov Institute od the Russian Language [Trudy Instituta russkogo jazyka im. V. V. Vinogradova], Vol. 6 (6).
15. *Nivre J., De Marneffe M. C., Ginter F., Goldberg Y., Hajic J., Manning C. D., McDonald R. T., Petrov S., Pyysalo S., Silveira N., Tsarfaty, R.* (2016), Universal Dependencies v1: A Multilingual Treebank Collection, Proceedings of LREC 2016.
16. *Nivre J., Abrams M., Agić Ž. et al.* (2018), Universal Dependencies 2.3, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-2895.
17. *Polyakov A. E.* (2012), A stemmer for the pre-reform Russian orthography [Lemmatizator dlja doreformennoj russkoj orfografii], Baranov V. A., Varfolomeyev A. G. (eds.), Proceedings of the international conference Information Technologies and Textual Heritage El'Manuscript-12 [Informatsionnye tekhnologii i pis'mennoe nasledie: materialy IV mezhdunarodnoj nauchnoj konferentsii], Petrozavodsk, Izhevsk, pp. 211–215.
18. *Sichinava D. V.* (2014), Historical corpora of the Russian National Corpus as a tool for diachronic grammatical studies [Istoricheskie korpusa Natsional'nogo korpusa russkogo jazyka kak instrument diakhronicheskikh issledovanij grammatiki], Baranov V. A., Zheljazkova V., Lavretiev A. M. (eds.), Textual heritage and information technologies. El'Manuscript–2014 [Pismenoto nasledstvo i informatsionnite tekhnologii. El'Manuscript–2014]. Proceedings of the 5th International research conference. Sofia, Izhevsk, 2014.
19. *Scherrer Y., Rabus A.* (2019), Variation in pre-modern Slavic corpus data and accuracy of neural tagging, Proceedings of the conference "Historical Corpora and Variation", Cagliari, 2019.
20. *Sichinava D. V.* (2014), Historical corpora of the Russian National Corpus as a tool for diachronic grammatical studies [Istoricheskie korpusa Natsional'nogo korpusa russkogo jazyka kak instrument diakhronicheskikh issledovanij grammatiki], Baranov V. A., Zheljazkova V., Lavretiev A. M. (eds.), Textual heritage and information technologies. El'Manuscript–2014 [Pismenoto nasledstvo i informatsionnite tekhnologii. El'Manuscript–2014]. Proceedings of the 5th International research conference. Sofia, Izhevsk, 2014.
21. *Sichinava D. V.* (2018), The corpus/database of Old East Slavic birchbark letters, El'Manuscript 2018 Book of Abstracts, Vienna, Krems.
22. *Zaliznyak, A. A.* (2003), A Grammatical Dictionary of Russian [Grammaticheskij slovar' russkogo jazyka], Moscow.
23. *Zaliznyak, A. A.* (2004), Old Novgorod Dialect, Moscow, Languages of Slavonic Culture.

ПОСЛЕЛОЖНЫЕ КОНСТРУКЦИИ ТАТАРСКОГО ЯЗЫКА: МЕТОДИКИ ОЦЕНКИ ВНУТРИЯЗЫКОВОГО ВАРЬИРОВАНИЯ[1]

Лютикова Е. А.

МГУ имени М. В. Ломоносова, МПГУ,
Гос. ИРЯ им. А. С. Пушкина

Герасимова А. А.

МГУ имени М. В. Ломоносова, Гос. ИРЯ им. А. С. Пушкина

В статье рассматриваются методы исследования внутриязыковой вариативности на материале падежного варьирования в послеложных конструкциях в татарском языке. Татарские послеложные конструкции обнаруживают дифференцированное падежное маркирование зависимого: выбор падежной формы определяется морфолого-синтаксическим классом зависимой именной синтагмы. При этом для послелогов, синхронно связанных с существительными, выбор падежного оформления зависимого осложняется варьированием притяжательной и непритяжательной форм. В данном исследовании с помощью корпусных и экспериментальных методик устанавливается распределение падежных вариантов, а также определяется статус варьирования с синхронной и диахронической точек зрения. Экспериментальные данные свидетельствуют о том, что процесс грамматикализации послеложных слов еще не завершен. Так, послеложные конструкции проявляют ряд свойств именных изафетных конструкций: это касается как выбора падежа зависимого, так и наличия/отсутствия согласовательных показателей на послелоге. Полученные результаты позволяют охарактеризовать вариативность в синхронном срезе языка, а также провести сравнение методик исследования внутриязыкового варьирования. В статье показывается, что использование различных исследовательских методик, как корпусных, так и экспериментальных, не только дает возможность определить направление развития языка в конкретных проблемных областях, но и может служить эмпирической базой для создания описательных грамматик нового типа.

Ключевые слова: послелоги, послеложная конструкция, падежное варьирование, корпусная лингвистика, лингвистический эксперимент, татарский язык

[1] Исследование выполнено в рамках проекта РНФ № 18-18-00462 «Коммуникативно-синтаксический интерфейс: типология и грамматика», реализуемого в Гос. ИРЯ им. А. С. Пушкина. Авторы выражают искреннюю благодарность Д. А. Зариповой и А. М. Галиевой за неоценимую помощь в подготовке материалов для экспериментального исследования.

POSTPOSITIONAL CONSTRUCTIONS IN TATAR: METHODOLOGIES FOR MEASURING INTRALINGUAL VARIATION[2]

Lyutikova E. A.
MSU, MPSU, Pushkin State Russian Language Institute

Gerasimova A. A.
MSU, MPSU, Pushkin State Russian Language Institute

The paper addresses the issue of intralingual variation in Tatar postpositional phrases. The nominal in Tatar postpositional phrases demonstrates differential case marking: the choice between genitive and unmarked case form is determined by the morphosyntactic class of the nominal. With postpositions derived from nouns with locative or abstract semantics variation in case assignment is accompanied by presence/absence of the ezafe marker on the postposition. In this paper we use corpus-based and experimental methods to investigate the distribution of grammatical variants and estimate the current status of the variation. We argue that the existing grammatical descriptions do not capture the current state of affairs.

We show that pronouns and nouns do not form a homogeneous class with respect to case marking in the postpositional phrase. The genitive case marking is common for 1st/2nd person personal pronouns and 3rd person singular personal pronoun. All other pronouns and nouns are primarily used in an unmarked form, an observation supported by both corpus and experimental data.

We argue that the grammaticalization of denominal postpositions is not complete. In both corpus and experimental studies, we observe a wide range of features that unite postpositional phrases with nominal embedding ezafe constructions. First, genitive case marking for the complement is acceptable for non-personal pronouns and nouns. Second, the absence of the ezafe marker is acceptable only with 1st / 2nd person personal pronouns and partially with 1st / 2nd person reflexive pronouns. Third, the case marking of the nominal and the choice of the ezafe marker for the postposition are interrelated. When the complement is genitive, speakers prefer the agreeing form of the postposition. When the complement is unmarked, the postposition shows no agreement with the possessor. This contrast reflects the opposition between ezafe-3 and ezafe-2 constructions, respectively.

Interestingly, the denominal postpositions demonstrate different degrees of grammaticalization. For instance, the postposition *turında* 'about' is mostly used with a possessive affix that shows no agreement. We suppose that the form with the non-agreeing ezafe affix is reanalyzed by the speakers as uninflected.

Another crucial observation concerns the reflexive pronoun *üz*. In both experiments 1st / 2nd person reflexive pronouns show syntactic behavior

[2] The study has been supported by RSF, project #18-18-00462 "Communicative-syntactic interface: typology and grammar" at the Pushkin State Russian Language Institute. We thank Diana Zaripova and Alfia Galieva who greatly helped in preparing the materials for the experimental study.

similar to the one of personal pronouns, while 3rd person singular reflexive pronoun patterns with interrogative pronouns.

As the result of the study, we compare different methodologies for investigation of the intralingual variation. We suggest that the combination of different sources of data, both corpus-based and experimental, provides the fuller description for cases of intralingual variation than a single method. The experimental methods that we used differ in sensitivity to various aspects of language phenomena: the elicited production is better in distinguishing deviation from the grammatical pattern; the acceptability judgements show to what extent a grammatical innovation is used. Remarkably, the comparison of the different sources of data allows us to determine the direction of language change and estimate the current status of the variation.

Key words: postposition, postpositional phrase, case variation, ezafe, corpus linguistics, experimental linguistics, Tatar

1. Введение

Как при решении задач, связанных с автоматической обработкой естественных языков, так и при построении языковой теории исследователь неизбежно сталкивается с проблемой языкового разнообразия. Для решения этой проблемы разные по своим задачам области лингвистики используют концептуально сходные методы. Так, в компьютерной лингвистике при переходе от одного языка к другому производят настройку параметров языковой модели в соответствии с новым набором данных, а в теоретических подходах для описания межъязыкового варьирования успешно используется метод параметризации грамматических систем конкретных языков [Лютикова, Циммерлинг, Коношенко 2016].

Метод параметризации также применим к описанию внутриязыкового разнообразия. Подобный подход основывается на предположении о том, что язык представляет собой совокупность грамматических систем. Различная обусловленность вариантов грамматик позволяет выделить несколько типов внутриязыковой вариативности [Герасимова 2016]. Во-первых, вариативность можно рассматривать в пределах языковой общности. Распределение грамматик в этом случае определяется внешними факторами: территориальными (диалекты), социальными (социолекты), функционально-стилистическими (стили, жанры), историческими (исторический срез языка). Во-вторых, существует распределение грамматик по носителям языка: каждый индивид обладает идиолектом — результатом индивидуальной параметризации. Не исключается сосуществование грамматик и их тесное взаимодействие: так, в пределах одного языка существуют синтаксически омонимичные конструкции, выбор между которыми определяется частными языковыми лицензорами.

Цель статьи — исследовать внутриязыковую вариативность в татарских послеложных конструкциях. Для решения этой задачи используются три метода: корпусный анализ, эксперимент на порождение и эксперимент на оценку приемлемости, — что позволяет не просто установить распределение вариантов, но и определить статус явления с синхронной точки зрения и в перспективе языкового развития.

2. Послеложные конструкции в татарском языке

Послеложные конструкции в татарском языке образует гетерогенная категория лексических единиц, в разной степени грамматикализованных в качестве функциональных. [Татарская грамматика 1993: 309]; [1995: 47] выделяет собственно послелоги и послеложные слова, которые «...отличаются от послелогов тем, что имеют живые словообразовательные отношения и лексико-семантические связи с знаменательными частями речи...». Соответственно, послеложные слова имеют знаменательный эквивалент, обладающий лексическим значением: «В предложении *Язу өстендә озак утырды* ‘Долго сидел над письмом’ *өстендә* — послеложное слово; в предложении *Су өстендә көймә йөзә* ‘Лодка плывет на поверхности воды’ то же самое *өстендә* — имя существительное» [Татарская грамматика 1993: 11].

Послелоги и послеложные слова управляют именной группой в определенной падежной форме. Так, например, послелог *таба(н)* ‘к’ и послеложное слово *күрә* ‘ввиду, из-за, по’ управляют дативом, а послелоги *башка* ‘кроме’ и *соң* ‘после’ — аблативом.

Большой класс послелогов и послеложных слов употребляются с именной группой в немаркированной форме, совпадающей с номинативом (татар. *баш килеш*) — падежом подлежащего. В этот класс входят как собственно послелоги, например, *белән* ‘с’, *өчен* ‘для’, так и послеложные слова, соотносимые с существительными с локативным или абстрактным значением, например, *өстендә* ‘над’ (ср. *өс* ‘верх’), *янына* ‘рядом, к’ (ср. *ян* ‘бок’), *урынында* ‘вместо’ (ср. *урын* ‘место’), *ярдәмендә* ‘благодаря’ (ср. *ярдәм* ‘помощь’). Зачастую послеложные слова образуют группы, объединенные общей основой и различающиеся падежными аффиксами, напр. *янына* ‘к’ (датив), *янында* ‘около’ (локатив), *яныннан* ‘мимо’ (аблатив); *артына* ‘за, вслед’ (датив), *артында* ‘за, позади’ (локатив), *артыннан* ‘из-за’ (аблатив).

Татарская грамматика [1993: 253] указывает, что с послелогами и послеложными словами этого класса местоимения-существительные используются в форме генитива. Таким образом, в послеложной конструкции возникает падежное варьирование: именные группы на основе существительного (или субстантивированного атрибута, отглагольной номинализации и т. п.) демонстрируют немаркированную форму, в то время как местоимения-существительные — форму генитива: «… *аның белән (абый белән) әшләү* ‘делать с ним (с братом)’, *аның өчен (дустым өчен) тырышу* ‘стараться ради него (друга)’». К местоимениям-существительным Татарская грамматика относит личные местоимения 1–2 лица *мин* ‘я’, *син* ‘ты’, *без* ‘мы’, *сез* ‘вы’, местоимения 3 лица *ул* ‘он’ и *алар* ‘они’, возвратное местоимение *үз* ‘сам’, вопросительные местоимения *кем* ‘кто’, *нәрсә* ‘что’, а также образованные на основе вопросительных местоимений серии неопределенных и универсальных местоимений. Дифференцированное маркирование дополнения в послеложной конструкции, таким образом, лицензируется формальным фактором — его морфосинтаксической категорией.

Отличие личных местоимений от прочих субстантивов проявляется также в том, что послеложные слова демонстрируют лично-числовое (притяжательное) согласование с генитивным местоименным дополнением. В примерах (1a)–(1b)

показаны согласованные по лицу и числу с местоименным дополнением формы послелога *алдында* ‘перед’; в (1c) — тот же послелог в дефолтной форме 3 лица.

(1) a. *Юк, сез-нең* ***минем*** *алд-**ым**-да бер гаеб-егез дә юк.*
нет вы-GEN я.GEN перед-1SG-LOC один вина-2PL емрн нет
‘Нет, вы передо мной ни в чем не виноваты.’ [TNC]

b. *Әфәнде-ләр, хәзер мин* ***сез-нең*** *алд-**ыгыз**-да*
господин-PL сейчас я вы-GEN перед-2PL-LOC
бик зур эш ача-чак-мын.
очень большой дело открывать-FUT-1SG
‘Господа, сейчас я открою вам (= перед вами) одну очень важную вещь.’ [TNC]

c. ***Кыз-лар*** *алд-**ын**-да ясалма йөр-гән-не*
девушка-PL перед-3-LOC искусственно ходить-PART-ACC
ярат-м-ый-м.
любить-NEG-PRS-1SG
‘Не люблю выпендриваться (= лживо ходить) перед девушками.’ [TNC]

Причиной вариативной структуры конструкций с послеложными словами является, безусловно, их диахронический источник. Послеложная конструкция представляет собой результат грамматикализации именной синтагмы, возглавляемой локативным или абстрактным существительным в одной из падежных форм. Зависимое в послеложной конструкции, соответственно, является приименным зависимым и образует с вершиной посессивную конструкцию.

Посессивная конструкция в татарском языке также демонстрирует дифференцированное маркирование аргумента-посессора. Традиционная грамматика выделяет два типа посессивных конструкций — изафетную конструкцию 2 и изафетную конструкцию 3[3]. В изафетной конструкции 2 зависимое выступает в немаркированной форме, вершина несет на себе изафетный (посессивный) показатель (2a). В изафетной конструкции 3 зависимое имеет форму генитива, вершина несет на себе изафетный показатель (2b).

(2) a. *укычы дәфтәр-е*
ученик тетрадь-3
‘ученическая тетрадь, тетрадь ученика’

b. *укычы-ның дәфтәр-е*
ученик-GEN тетрадь-3
‘тетрадь ученика’

[Татарская грамматика 1993: 32–37] отмечает, что местоимения-существительные образуют только изафетную конструкцию 3, но не 2, ср. (3a)–(3b); если посессором выступают местоимения 1–2 лица, в разговорной речи возможно опущение изафетного показателя в изафетной конструкции 3 (3c).

[3] В изафетной конструкции 1, образованной соположением двух существительных (*таш йорт* ‘каменный дом’), зависимое не может ветвиться, присоединять показатели числа и принадлежности, то есть является вершиной. Ее свойства нерелевантны для обсуждения послеложной конструкции.

Структуры, в которых изафетный показатель на вершине возникает в отсутствие выраженного генитивного посессора, естественно рассматривать как изафетную конструкцию 3, в которой в позиции посессора находится нулевое местоимение *pro* соответствующего лица и числа, контролирующее согласование изафетного показателя (3d).

(3) a. *без-нең* *мәктәб-ебез*
мы-GEN школа-1PL
'наша школа'
b. * *без* *мәктәб-ебез*
мы школа-1PL
c. *без-нең* *мәктәп*
мы-GEN школа
'наша школа'
d. [*pro*] *мәктәб-ебез*
pro.1PL.GEN школа-1PL
'наша школа'

Еще один класс посессоров, которые возможны только в изафетной конструкции 3, но не 2, — это именные группы, сами представляющие собой изафетную конструкцию 3. Это, во-первых, собственно посессивные конструкции, такие как в (3), а также субстантивированные кванторные и атрибутивные конструкции, такие как в (4):

(4) a. *(а-лар-ның)* *күб-есе*
он-PL-GEN много-3
'многие из них'
b. *(без-нең)* *иң* *акыл-лы-быз*
мы-GEN самый ум-ATR-1PL
'самый умный из нас'
c. *кыз-лар-ның* *кайсы-сы*
девушка-PL-GEN который-3
'которая из девушек'

Остальные типы именных групп могут быть зависимыми в обеих изафетных конструкциях. Выбор между генитивом и немаркированной формой определяется как определенностью зависимой именной группы, так и семантическим отношением между вершиной и зависимым. Так, например, имена собственные обычно выступают в изафетной конструкции 3 (*Марат**(*-ның*) *ата-сы* 'отец Марата'), однако при обозначении наименования используется изафетная конструкция 2 (*Марат урам-ы* 'улица Марата').

В [Pereltsvaig, Lyutikova 2014]; [Лютикова, Перельцвайг 2015]; [Lyutikova, Pereltsvaig 2015]; [Lyutikova 2017] показано, что различия между двумя конструкциями не сводятся к падежному оформлению, но затрагивают различные характеристики посессоров — их структурную позицию в именной группе, их собственный категориальный статус, возможные интерпретации и способность к выражению различных тематических отношений с именной вершиной. Указанные кластеры свойств возникают не случайно, но выводятся из категориального

статуса именной группы-посессора. DP-посессоры насыщают аргументные позиции, выражают тематические отношения, получают конкретно-референтную или обобщенно-кванторную интерпретацию, получают падеж, контролируют посессивное согласование показателя изафета и располагаются в крайней левой позиции в именной группе (Spec, DP). Посессоры малой структуры вводятся особой функциональной вершиной Poss, выражают широкий спектр отношений между двумя именными группами, уточняемых на основе энциклопедических знаний, имеют предикатную интерпретацию, не нуждаются в падеже, неспособны контролировать посессивное согласование и располагаются в своей базовой позиции, правее атрибутивных модификаторов (Spec, PossP). Таким образом, дифференцированное маркирование посессора в татарском языке существенно отличается от дифференцированного маркирования дополнения послелога: в послеложной конструкции генитивом оформляются только местоимения-существительные, в то время как в посессивной конструкции — любые именные составляющие, имеющие статус DP, в том числе — изафетная конструкция 3, имена собственные, другие определенные именные группы. Ср. примеры в (5)–(6), демонстрирующие послеложные группы, и (7) с именными группами.

(5) a. *минем / *мин* *өчен*
я.GEN / я для
‘для меня’

b. *ата-м / *ата-м-ның* *өчен*
отец-1SG / отец-1SG-GEN для
‘для моего отца’

c. *Марат / *Марат-ның* *өчен*
Марат / Марат-GEN для
‘для Марата’

(6) a. *минем / *мин* *урын-ым-да*
я.GEN / я вместо-1sg-loc
‘вместо меня’

b. *ата-м / *ата-м-ның* *урын-ын-да*
отец-1SG / отец-1SG-GEN вместо-3-loc
‘вместо моего отца’

c. *Марат / *Марат-ның* *урын-ын-да*
Марат / Марат-GEN вместо-3-loc
‘вместо Марата’

(7) a. *минем / *мин* *мәктәб-ем-дә*
я.GEN / я школа-1SG-LOC
‘в моей школе’

b. **ата-м / ата-м-ның* *мәктәб-ен-дә*
отец-1SG / отец-1SG-GEN школа-3-LOC
‘в школе моего отца’

c. **Марат / Марат-ның* *мәктәб-ен-дә*
Марат / Марат-GEN школа-3-LOC
‘в школе Марата’

Таким образом, послелоги и послеложные слова, присоединяющие именные группы в немаркированной форме или генитиве, могут стать источником особенно интересного материала для изучения внутриязыковой вариативности. Во-первых, грамматические свойства конструкции включают в себя выбор между генитивным или немаркированным оформлением зависимой именной группы. Даже если этот выбор однозначно предопределен морфологическими характеристиками именной группы, варьирование может быть функционально специфицированным и передавать особое семантическое противопоставление (так, например, выбор формы аккузатива, опирающийся на признак одушевленности, в русском языке создает возможность тонкой семантической дифференциации между конструкциями, ср. *увидел трех цыплят* vs. *съел три цыпленка*). Во-вторых, сосуществование в языке послеложных слов и омонимичных им существительных, деривационная связь между которыми, по свидетельству авторов Татарской грамматики, ощущается носителями татарского языка, позволяет предположить, что способы оформления именной синтагмы будут оказывать влияние и на оформление послеложной группы.

Для проверки этих предположений были проведены корпусное и экспериментальное исследование управления послелогов и послеложных слов интересующего нас класса. Далее в статье мы опишем полученные нами результаты.

3. Корпусное исследование

Материалы для исследования были получены нами из Татарского национального корпуса «Туган тел» («Родной язык»)[4], включающего в свой состав размеченные тексты на литературном татарском языке и насчитывающего более 180 млн. словоупотреблений. Корпус включает тексты различных жанров: художественную литературу, тексты СМИ, учебную литературу, научные публикации и др. В корпусе реализован морфологический поиск, есть возможности поиска по лексеме и словоформе, а также поиска нескольких лексем или словоформ, удовлетворяющих определенным линейным или структурным характеристикам (например, на расстоянии 2 слов, в пределах одного предложения, после запятой и т. п.).

Для исследования были отобраны две группы лексических единиц. Во-первых, это послелоги, не имеющие позиции для лично-числового согласования и синхронно нечленимые на именную основу, изафетный показатель и падежный аффикс: *аркылы* ‘через, поперек, посредством’, *белән* ‘с’, *булып* ‘подобно, в качестве’, *катыш* ‘вперемежку с’, *кебек* ‘как’, *өчен* ‘для’, *сәбәпле* ‘из-за, по причине’, *сымак* ‘как’, *төсле* ‘подобно’, *шикелле* ‘как’. Во-вторых, это послеложные слова, формально совпадающие с падежной формой притяжательного склонения синхронно функционирующих существительных: *алдында* ‘перед’ (*ал* ‘перед’), *аркасында* ‘из-за’ (*арка* ‘спина’), *артында* ‘за’ (*арт* ‘зад’), *астында* ‘под’ (*ас* ‘низ’), *нигезендә* ‘благодаря, за’ (*нигез* ‘основа’), *өстендә* ‘над’ (*өс* ‘верх’), *тарафыннан* ‘посредством’ (*тараф* ‘сторона’), *тирәсендә* ‘рядом’

[4] http://tugantel.tatar/; в статье примеры из этого корпуса имеют помету [TNC].

(*тирә* ‘окрестность’), *турында* ‘о’ (*туры* ‘прямо, напротив’), *урынында* ‘вместо’ (*урын* ‘место’), *хакында* ‘о’ (*хак* ‘право, правда’), *янында* ‘сбоку, около, у’ (*ян* ‘бок’), *ярдәмендә* ‘благодаря’ (*ярдәм* ‘помощь’).

На корпусном материале изучались возможные с данными послелогами и послеложными словами конструкции при разных морфолого-синтаксических классах зависимых. В качестве зависимых выступали:

- личные местоимения 1–2 лица *мин* ‘я’, *син* ‘ты’, *без* ‘мы’, *сез* ‘вы’;
- местоимения 3 лица *ул* ‘он’ и *алар* ‘они’;
- возвратное местоимение *үз* ‘сам’ с изафетным показателем *үзем* ‘я сам’, *үзең* ‘ты сам’, *үзе* ‘он сам’;
- вопросительные местоимения-существительные *кем* ‘кто’ и *кемнәр* ‘кто.PL’;
- субстантивированные атрибутивные местоимения *кайсыбыз* ‘который из нас’ и *кайсыгыз* ‘который из вас’;
- существительное *кыз* ‘девушка’.

В корпусе осуществлялся поиск следующих конструкций. Для послелогов строились запросы двух типов: субстантив в генитиве плюс послелог и субстантив в немаркированной форме плюс послелог (напр., *безнең белән* и *без белән* ‘с нами’). Для послеложных слов, помимо падежного варьирования, учитывались также варьирование притяжательной и непритяжательной форм послелога (напр., *аркасында* и *аркада* ‘из-за’), а также согласование послелога в притяжательной форме в случае, если субстантив имеет граммему 1–2 лица, как интерпретируемую (у личных местоимений), так и согласовательную (у возвратного местоимения и субстантивированных атрибутивных местоимений). Таким образом, для послеложных слов проверялись конструкции 6 типов: *үземнең аркада*, *үзем аркада*, *үземнең аркамда*, *үзем аркамда*, *үземнең аркасында* и *үзем аркасында* ‘из-за меня самого’.

Из грамматического описания, представленного в предыдущем разделе, следует, что дистрибуция конструкций должна выглядеть следующим образом. С послелогами мы ожидаем противопоставления местоимений-существительных, получающих генитив, собственно существительным, выступающим в немаркированной форме. Статус субстантивированных атрибутивных местоимений не вполне очевиден: они не относятся к разряду местоимений-существительных, упомянутых в грамматике, однако являются, с одной стороны, местоимениями, а с другой стороны, производными субстантивами. С учетом этого в **Таблице 1** их дистрибуция определена условно, как совпадающая с дистрибуцией существительных.

Таблица 1. Ожидаемая дистрибуция падежных форм с послелогами

Пример	*безнең белән* (GEN)	*без белән* (NOM)
личные местоимения 1–2 лица	+	–
личные местоимения 3 лица	+	–
возвратное местоимение	+	–
вопросительные местоимения-существительные	+	–
субстантивированные атрибутивные местоимения	(–)	(+)
существительные	–	+

Для послеложных слов мы ожидаем такой же падежной дистрибуции. Противопоставление притяжательной и непритяжательной форм послелога, по-видимому, должно быть ортогонально дистрибуции падежных форм. Лично-числовое согласование притяжательной формы предположительно возможно только в случае, когда признаки лица и числа контролера согласования сами не являются согласовательными. Такая логика действует, например, в предикативном (8a) и посессивном (8b) согласовании.

(8) a. *Без-нең укытучы-быз кил-де(*-к).*
мы-GEN учитель-1PL приходить-PST(-1PL)
‘Наш учитель пришел.’

b. *без-нең укытучы-быз-ның дәфтәр-е / *-ебез.*
мы-GEN учитель-1PL-GEN тетрадь-3 / (-1PL)
‘тетрадь нашего учителя’

Ожидаемая дистрибуция конструкций с послеложными словами представлена в **Таблице 2**; ячейки, соответствующие невозможным комбинациям, закрашены.

Таблица 2. Ожидаемая дистрибуция конструкций с послеложными словами

Пример	Непритяжательная форма		Притяжательная согласуемая форма		Притяжательная несогласуемая форма	
	GEN	NOM	GEN	NOM	GEN	NOM
	үземнең аркада	*үзем аркада*	*үземнең аркамда*	*үзем аркамда*	*үземнең аркасында*	*үзем аркасында*
личные местоимения 1–2 лица	+	–	+	–	–	–
личные местоимения 3 лица	+	–			+	–

Пример	Непритяжательная форма		Притяжательная согласуемая форма		Притяжательная несогласуемая форма	
	GEN	NOM	GEN	NOM	GEN	NOM
	үземнең аркада	*үзем аркада*	*үземнең аркамда*	*үзем аркамда*	*үземнең аркасында*	*үзем аркасында*
возвратное местоимение	+	–	–	–	+	–
вопросительные местоимения-существительные	+	–			+	–
субстантивированные атрибутивные местоимения	(–)	(+)	(–)	(–)	(–)	(+)
существительные	–	+			–	+

Результаты, полученные в корпусном исследовании, отличаются от ожидаемых в нескольких отношениях.

Во-первых, падежное оформление зависимого как послелогов, так и с послеложных слов противопоставляет личные местоимения 1–2 лица и местоимение 3 лица ед. ч. *ул*, с одной стороны, и все прочие субстантивы, включая местоимение 3 лица мн. ч. *алар*, вопросительные и возвратные местоимения, с другой стороны. В Таблице 3 показаны результаты поиска конструкций с послелогом *өчен* ‘для’; прочие послелоги дают аналогичную картину.

Таблица 3. Результаты поиска для послелога *өчен* ‘для’

	GEN	NOM
мин ‘я’	2530	0
син ‘ты’	835	0
без ‘мы’	1189	0
сез ‘вы’	427	0
кем ‘кто’	0	148
кемнәр ‘кто.pl’	0	22
кайсыбыз ‘который из нас’	0	1
кайсыгыз ‘который из вас’	0	0

	GEN	NOM
үзем ‘я сам’	0	337
үзең ‘ты сам’	0	127
үзе ‘он сам’	0	730
ул ‘он’	2740	0
алар ‘они’	0	905
кыз ‘девушка’	0	66
атасы ‘его/ее отец’	0	13
балалар ‘дети’	0	1152
дәүләт ‘государство’	0	27
эш ‘работа’	0	279

Во-вторых, противопоставление притяжательной и непритяжательной форм послелога оказывается значимым: непритяжательная форма возможна только в сочетании с личными местоимениями 1–2 лица (см. Таблицу 4 для

послеложного слова *тарафыннан* ‘посредством’)[5]. Очевидно, что эта дистрибуция повторяет закономерности, выявленные для именной синтагмы: допустимость опущения изафетного показателя в изафетной конструкции 3 с посессором 1–2 лица. При этом, однако, послеложная конструкция не может быть отождествлена с именной синтагмой: так, в примере (9a) послеложная конструкция *безнең тарафтан* ‘нами’ используется для выражения агентивного дополнения при пассиве; аналогичный пример (9b) демонстрирует послеложную конструкцию *минем аркада* ‘из-за меня’.

(9) a. *Ул* *без-нең* *тарафтан* *әзерлә-н-еп*,
он мы-GEN посредством готовить-PASS-CONV
«Сөембикә» журналында хәзерге чор укучыларына да ирешкән иде.
‘Она (статья) была нами подготовлена и представлена (=сделана доступной) современному читателю в журнале «Сююмбике».’ [TNC]

b. *Минем аркада өч гөнаһ-сыз кеше* *теге дөнья-га* *кит-те.*
я.GEN из-за 3 вина-ATR человек тот мир-DAT уходить-PST
‘Из-за меня 3 безвинных человека отправились на тот свет.’ [TNC]

Таблица 4. Результаты поиска для послеложного слова *тарафыннан* ‘посредством’

	Непритяжательная форма		Притяжательная согласуемая форма		Притяжательная несогласуемая форма	
	GEN	NOM	GEN	NOM	GEN	NOM
мин ‘я’	68	0	3 (39)[6]	0	0	0
син ‘ты’	9	0	9 (12)	0	0	0
без ‘мы’	95	0	0 (4)	0	0	0
сез ‘вы’	18	0	1 (1)	0	0	0
кем ‘кто’	0	0			0	60
кемнәр ‘кто.PL’	0	0			0	19
кайсыбыз ‘который из нас’	0	0	0	0	0	0
кайсыгыз ‘который из вас’	0	0	0	0	0	0
үзем ‘я сам’	0	0	0	0	0	0
үзең ‘ты сам’	0	0	0	0	0	0
үзе ‘он сам’	0	0			0	57

[5] В целях экономии места мы показываем данные в табличной форме только для некоторых послелогов и послеложных слов. Внутри каждого из двух классов — послелогов и послеложных слов, исключая *турында* ‘о’ — различия между лексическими единицами в отношении дистрибуции конструкций статистически незначимы (*ANOVA*, $p > 0{,}05$). Об особенностях конструкций с послеложным словом *турында* ‘о’ см. ниже.

[6] Число в скобках обозначает количество примеров, содержащих послелог в согласуемой форме с опущенным местоимением 1–2 лица и отсутствием выраженного зависимого. Мы полагаем, что в таком случае зависимым в послеложной конструкции выступает *pro*, аналогично посессивной конструкции (3d).

	Непритяжательная форма		Притяжательная согласуемая форма		Притяжательная несогласуемая форма	
	GEN	NOM	GEN	NOM	GEN	NOM
ул ‘он’	0	0			140	0
алар ‘они’	0	0			0	68
кыз ‘девушка’	0	0			0	8
атасы ‘его/ее отец’	0	0			0	6
балалар ‘дети’	0	0			0	10
дәүләт ‘государство’	0	0			0	89
эш ‘работа’	0	0			0	0

В-третьих, выяснилось, что возвратное местоимение *үз*, принимая согласовательные показатели 1–2 лица, в конструкциях с послеложными словами может демонстрировать свойства личных местоимений 1–2 лица. Во-первых, оно может выступать не только в форме номинатива, но и в форме генитива. Такие конструкции зафиксированы, в частности, для послеложных слов *алдында* ‘перед’, *аркасында* ‘из-за’, *артында* ‘за’, *астында* ‘под’, *өстендә* ‘над’, *урынында* ‘вместо’, *янында* ‘сбоку, около, у’. Во-вторых, если *үзем* ‘я сам’ и *үзең* ‘ты сам’ выступают в форме генитива, то они могут контролировать согласование притяжательной формы послеложного слова (10a)–(10b) или выступать с непритяжательной формой послеложного слова (11a)–(11b). Оба эти свойства отмечаются только у личных местоимений 1–2 лица.

(10) a. *Мин барыбер а-ны үз-ем-нең ян-ым-да*
я все.равно он-ACC сам-1SG-GEN рядом-1SG-LOC
ит-еп сиз-ә-м.
делать-CONV чувствовать-PRS-1SG
‘Я все равно чувствую его рядом с собой.’ [TNC]

b. *Үз-ең-нең ян-ың-да бит, тырыш, тыйнак,*
сам-2SG-GEN рядом-2SG-LOC вот старательный скромный
күз-ең-ә генә кара-п лә̆ббә̆йкә
глаз-3SG-DAT EMPH смотреть-CONV «слушаюсь!»
ди-еп тора.
говорить-CONV AUX
‘Вот он рядом с тобой, старательный, скромный, заглядывает в глаза, говорит «Слушаюсь!».’ [TNC]

(11) a. *…үз-ем-нең хак-та — эш хак-ын-да артык нәрсә юк.*
сам-1SG-GEN о-LOC работа о-3-LOC лишний что нет
‘О себе — о работе <я не сказал> ничего лишнего.’ [TNC]

b. *Хуҗалык эш-ләр-е тулысынча үз-ем-нең өс-тә.*
хозяйство работа-PL-3 полностью сам-1SG-GEN над-LOC
‘Хозяйственные заботы полностью на мне.’ [TNC]

В-четвертых, послеложное слово *турында* ‘о’ проявляет особые свойства (см. **Таблицу 5**). Обращает на себя внимание появление личных местоимений

1–2 лица в конструкции с несогласуемой притяжательной формой, что невозможно для остальных исследованных послеложных слов. Такого рода употребления (ср. пример (12)) встречаются чаще, чем конструкции с согласуемой притяжательной формой.

Таблица 5. Результаты поиска для послеложного слова *турында* ‘о’

	Непритяжательная форма		Притяжательная согласуемая форма		Притяжательная несогласуемая форма	
	GEN	NOM	GEN	NOM	GEN	NOM
мин ‘я’	254	0	9 (14)	0	48	0
син ‘ты’	157	0	10 (18)	0	71	0
без ‘мы’	75	0	0 (4)	0	9	0
сез ‘вы’	102	0	2 (2)	0	9	0
кем ‘кто’	0	0			0	93
кемнәр ‘кто.PL’	0	0			0	16
кайсыбыз ‘который из нас’	0	0	0	0	0	0
кайсыгыз ‘который из вас’	0	0	0	0	0	0
үзем ‘я сам’	0	0	0	0	0	130
үзең ‘ты сам’	0	0	0	0	0	106
үзе ‘он сам’	0	0			0	468
ул ‘он’	4	0			1061	0
алар ‘они’	0	1			0	408
кыз ‘девушка’	0	0			0	78
атасы ‘его/ее отец’	0	0			0	16
балалар ‘дети’	0	0			0	65
дәүләт ‘государство’	0	0			0	3
эш ‘работа’	0	0			0	147

(12) *әйт-егез* *әле,* ***минем*** *тур-**ын**-да* *академия-гә*
говорить-2PL ка я.GEN о-3-LOC академия-DAT
нәрсә *яз-ды-гыз?*
что писать-PST-2PL
‘Скажите-ка, вы что обо мне в академию написали?’ [TNC]

Очевидно, что появление подобных конструкций говорит о дальнейшей грамматикализации послеложного слова *турында* ‘о’, при которой его внутренняя форма становится непрозрачной для носителей языка, а позиция для лично-числового согласования утрачивается. Тем не менее, противопоставление притяжательной и непритяжательной форм послеложного слова сохраняется, так что непритяжательная форма остается доступной только для личных местоимений 1–2 лица.

Таким образом, корпусное исследование позволило нам сделать целый ряд нетривиальных обобщений об употреблении послеложных конструкций разной

структуры в текстах на литературном татарском языке, не нашедших отражения в грамматиках. Вместе с тем очевидны и ограничения корпусного исследования. Во-первых, не удалось выяснить, как ведут себя в контексте послеложных слов субстантивированные атрибутивные местоимения: они встретились только в контексте послелогов *белән* ‘с’ и *өчен* ‘для’ в немаркированной форме, в количестве, явно недостаточном для обобщений (4 употребления и 1 употребление, соответственно). Во-вторых, в рамках корпусного исследования возникает проблема демаркации послеложных слов и имён, особенно в семантически неоднозначных контекстах. Так, в примере (13) только привлечение более широкого контекста позволяет понять, что *минем аркамда* является здесь послеложной конструкцией ‘из-за меня’, а не именной синтагмой ‘у меня на спине’.

(13) *Баш-ы-н* *капша-ды* — *җылы* *кан.*
голова-3-ACC щупать-PST горячий кровь
«Минем *арка-м-да* *харап* *бул-ды», —* *дип*
я.GEN из-за-1SG-LOC загубленный стать-PST COMP
уйла-ды *Мисбах,* *кот-ы* *чыг-ып.*
думать-PST Мисбах душа-3 выходить-CONV
‘Пощупал ему голову — горячая кровь. «Из-за меня погиб», — подумал Мисбах, ужаснувшись.’ [TNC]

Кроме того, неясно, в какой степени не встретившиеся в корпусе или встретившиеся в незначительном количестве конструкции воспринимаются носителями татарского языка как неприемлемые. Преодолеть указанные ограничения позволяет экспериментальное исследование, направленное на изучение использования и оценки послеложных конструкций разной структуры носителями татарского языка.

4. Эксперимент на порождение

Для того, чтобы изучить стратегии падежного маркирования для разных морфолого-синтаксических классов зависимых в послеложной конструкции, а также связанное с этим варьирование притяжательной и непритяжательной форм послелога, мы провели эксперимент с многофакторным дизайном. Первый фактор «Тип вершины в послеложной конструкции» имеет 2 уровня: послелоги без позиции для лично-числового согласования, послеложные слова (в эксперименте мы использовали те же группы послелогов и послеложных слов, для которых собиралась информация в корпусе). Второй фактор «Тип зависимого в послеложной конструкции» имеет 6 уровней: личные местоимения 1–2 лица, возвратное местоимение 1–2 лица, возвратное местоимение 3 лица, вопросительные местоимения-существительные, субстантивированные атрибутивные местоимения, имена существительные собственные. Комбинация двух факторов дает 12 условий, на каждое из которых было подобрано два экспериментальных предложения.

Респонденты видели предложения, в которых на месте послеложных конструкций был пропуск, зависимое и послелог находились в скобках в словарной

форме (14). Задание заключалось в том, чтобы заполнить пропуски, раскрыв скобки наиболее естественным образом.

(14) Кичә мин (син, турында) _____ _____ мәкалә укыдым.
‘Вчера я прочитал статью о тебе’.

В эксперименте участвовало 119 респондентов, из них релевантные ответы дали 109 человек (средний возраст 23; SD = 7; мин. возраст 17, макс. возраст 61; 85 женщин и 23 мужчины). Прежде чем приступить к эксперименту, каждый респондент отвечал на вопросы социолингвистической анкеты. Результаты анкетирования позволили нам контролировать уровень знания татарского языка: установлено, что 90 из 109 респондентов родились и проживают в Республике Татарстан и ежедневно общаются по-татарски в семье и в месте учебы или работы.

При обработке результатов мы провели логлинейный анализ, который показал значимое взаимодействие экспериментальных факторов ($p = 0{,}001$). Во-первых, значимым оказывается противопоставление послелогов и послеложных слов. В частности, при послелогах используется преимущественно немаркированная форма возвратных местоимений, а при послеложных словах обе падежные формы становятся доступны, причем при возвратных местоимениях 1–2 лица значимо преобладает генитив (критерий χ^2, $p << 0{,}001$). Тип вершины в послеложной конструкции оказывает значимое влияние на выбор падежной формы вопросительных местоимений-существительных и субстантивированных атрибутивных местоимений (критерий χ^2, $p < 0{,}001$): при послеложных словах становится допустимым использование генитива.

Таблица 6. Результаты эксперимента на порождение для послелогов и послеложных слов

	Послелоги		Послеложные слова	
	GEN	NOM	GEN	NOM
мин ‘я’, *син* ‘ты’	189	11	142	1
имена сущ. собственные	0	202	3	212
үзем ‘я сам’, *үзең* ‘ты сам’	8	208	116	81
үзе ‘он сам’	14	193	95	104
кем ‘кто’, *нәрсә* ‘что’	0	143	16	183
кайсыбыз ‘который из нас’, *кайсыгыз* ‘который из вас’	3	156	27	176

Во-вторых, выяснилось, что личные местоимений 1–2 лица преимущественно используются с притяжательной несогласуемой формой послеложного слова. Интересно, что в корпусе такая конструкция встретилась только для послелога *турында* (**Таблица 5**), а для всех остальных послеложных слов встречалась непритяжательная форма, которая в эксперименте использовалась в совсем незначительных количествах.

В-третьих, обнаружились интересные свойства возвратных местоимений. Так, при послелогах наблюдается такое же распределение падежных форм, как

у существительных и вопросительных местоимений. При послеложных словах ситуация меняется: в случае использования формы генитива возвратные местоимения 1–2 лица контролируют согласование послеложного слова, а также могут сочетаться с непритяжательной формой послеложного слова, т. е. демонстрируют свойства личных местоимений 1–2 лица. В случае немаркированных форм *үзем* ‘я сам’, *үзең* ‘ты сам’ преимущественно выбирается форма послеложного слова без согласования (критерий χ^2, $p << 0{,}001$). Таким образом, в случае возвратных местоимений 1–2 лица наиболее ярко видно противопоставление послелогов и послеложных слов: с послелогами дистрибуция падежных форм возвратных местоимений 1–2 лица совпадает с дистрибуцией падежных форм существительных, а с послеложными словами у возвратных местоимений 1–2 лица начинают проявляться свойства личных местоимений 1–2 лица.

При субстантивированных атрибутивных местоимениях респонденты, напротив, не допускают согласования послеложного слова, хотя, как и в случае возвратных местоимений, используют генитив с притяжательной несогласуемой формой. В целом экспериментальные данные подтверждают, что субстантивированные атрибутивные местоимения демонстрируют те же свойства, что и вопросительные местоимения-существительные.

Отметим отдельно, что респонденты практически не использовали непритяжательные формы послелогов, хотя немногочисленные случаи употребления были именно при местоимениях 1–2 лица, как предсказывало корпусное исследование.

Таблица 7. Соотношение форм послеложных слов, использованных в эксперименте на порождение

	Непритяжательная форма		Притяжательная согласуемая форма		Притяжательная несогласуемая форма	
	GEN	NOM	GEN	NOM	GEN	NOM
мин ‘я’, *син* ‘ты’	5	0	38	0	99	1
имена сущ. собственные	0	0			3	212
үзем ‘я сам’, *үзең* ‘ты сам’	9	0	85	16	22	65
үзе ‘он сам’	0	0			95	104
кем ‘кто’, *нәрсә* ‘что’	0	0			16	183
кайсыбыз ‘который из нас’, *кайсыгыз* ‘который из вас’	0	0	0	0	27	176

5. Эксперимент на оценку приемлемости

Во втором эксперименте мы просили респондентов оценить предложения с послеложными конструкциями по шкале Ликерта от 1 до 5. В эксперименте на оценку использовался тот же факторный дизайн, что и в эксперименте на порождение, но количество стимульного материала возросло до 42 предложений: добавился фактор падежа (2 уровня: генитив и немаркированная форма), а также на каждое послеложное слово теперь приходилось до трех форм (непритяжательная, притяжательная согласуемая для субстантивов с граммемой 1–2 лица и непритяжательная несогласуемая). В эксперименте на оценку приемлемости участвовал 31 респондент-участник предыдущего эксперимента, а также 7 новых участников (средний возраст 24; SD = 8; мин. возраст 17, макс. возраст 62; 30 женщин и 8 мужчин).

Многофакторный дисперсионный анализ показал, что все выделенные факторы значимо влияют на оценки приемлемости ($p << 0{,}001$)[7]. Для всех морфолого-синтаксических классов зависимых в послеложной конструкции с послелогами, кроме возвратных местоимений 1–2 лица, результаты двух экспериментов совпадают. Для *үзем* ‘я сам’, *үзең* ‘ты сам’ при послелогах используется немаркированная форма, но в оценках значимого различия между двумя падежными формами нет.

Для послеложных слов результаты эксперимента также частично расходятся с данными порождения. Оказывается, что для личных местоимений 1–2 лица как наиболее приемлемые оцениваются предложения с послеложным словом в непритяжательной форме, в то время как при порождении выбиралась притяжательная форма послеложного слова без согласования. Напомним, что предпочтительность непритяжательной формы послеложного слова с личными местоимениями 1–2 лица прослеживается и в корпусе. Примечательно, что возвратные местоимения 1–2 лица снова группируются с личными местоимениями: для них генитив более приемлем, чем немаркированная форма. Однако возможность контроля согласования для возвратных местоимений 1–2 лица в оценках не обнаруживается, хотя наблюдается при порождении. Таким образом, противопоставление двух типов конструкций, генитива возвратного местоимения 1–2 лица *үземнең / үзеңнең* при форме послелога с согласовательным показателем и немаркированной формы *үзем / үзең* при форме послелога без согласовательного показателя, которое наблюдалось при порождении, в оценках никак не проявляется.

Для возвратного местоимения 3 лица, напротив, оценки выявляют противопоставление генитива и немаркированной формы как при послелогах, так и при послеложных словах: более высокие оценки получает номинатив, в то время как при порождении обе формы были одинаково допустимы. Таким образом, в эксперименте на оценку приемлемости респонденты воспринимают возвратные местоимения 3 лица как существительные.

---

[7] Для сравнения оценок использовался *t*-критерий Стьюдента; для всех значимых различий, которые упомянуты в статье, p-value << 0,001. Оценки были нормализованы (z-score transformed).

Наконец, результаты эксперимента выявляют интересные свойства у субстантивированных атрибутивных местоимений: хотя при порождении в конструкции с послеложным словом в несогласуемой притяжательной форме предпочтение отдается немаркированной форме местоимения, более высокие оценки получает форма генитива. Другими словами, распределение оценок для данного типа зависимого существенно отличается от распределения для вопросительных местоимений-существительных.

Распределение оценок для разных конфигураций послеложных конструкций показано на **Рис. 1**; звездочкой отмечены статистически значимые различия при попарном сравнении оценок падежных форм зависимого.

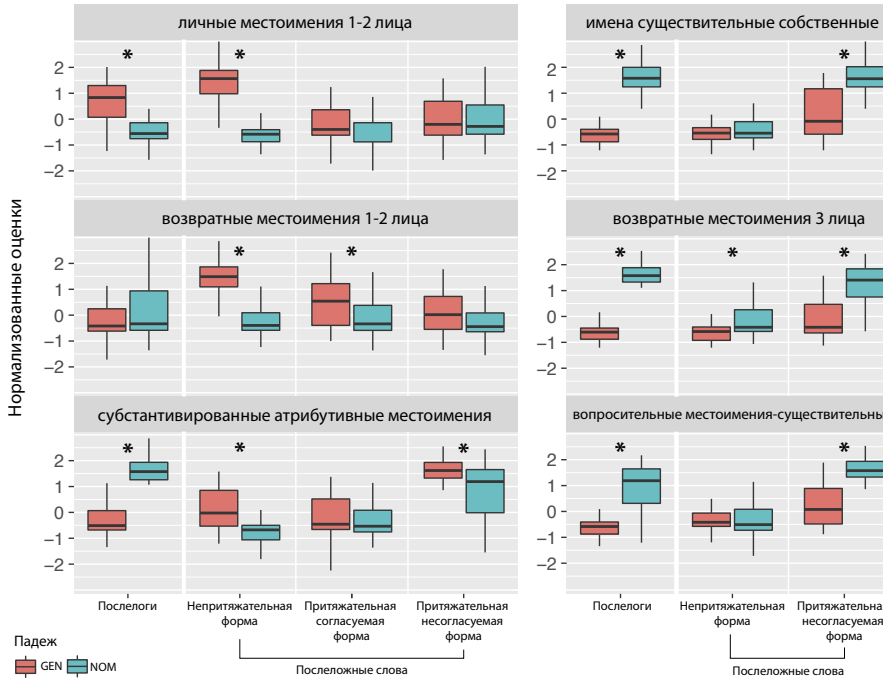


Рис. 1. Результаты эксперимента на оценку приемлемости для различных типов зависимых в послеложной конструкции

6. Выводы

Проведенное исследование показало существенные отличия между представленными в грамматическом описании и наблюдаемыми в языковом поведении носителей татарского языка характеристиками послеложной конструкции. Отличия касаются как выбора падежной формы зависимого, так и возможностей согласования послеложного слова.

Во-первых, местоимения-существительные не образуют единого класса в отношении падежного оформления в послеложной конструкции. Форма генитива характерна для личных местоимений 1–2 лица и местоимения 3 лица ед. числа; прочие местоимения-существительные демонстрируют явную тенденцию к употреблению в немаркированной форме, что подтверждается как корпусным, так и экспериментальными исследованиями. Эта закономерность прослеживается как для послелогов, так и для послеложных слов, хотя и в разной степени.

Во-вторых, можно заключить, что процесс грамматикализации послеложных слов еще не завершен. Несмотря на то, что в корпусном исследовании не учитывались омонимичные послеложным словам существительные в своем лексическом значении (контроль производился вручную), а в экспериментальных исследованиях функциональное значение конструкции было достаточно жестко навязано контекстом, наблюдается целый ряд свойств, объединяющих послеложные конструкции с именными изафетными конструкциями. Это: (1) возможность выбора генитива зависимого для не-личных местоимений и существительных; (2) предпочтение непритяжательной формы послелога только с личными местоимениями 1–2 лица и, частично, с возвратными местоимениями 1–2 лица; (3) корреляция между выбором генитива зависимого и согласуемой притяжательной формой, с одной стороны, и выбором немаркированной формы зависимого и несогласуемой притяжательной формой, с другой стороны, отражающая противопоставление изафетных конструкций 3 и 2, соответственно.

В-третьих, сами послеложные слова, по-видимому, грамматикализованы в разной степени. Особенно ярко в этом отношении выглядит контраст между послеложным словом *турында* ‘о’ и практически синонимичным ему послеложным словом *хакында* ‘о’: у послелога *турында* ‘о’ в корпусе фиксируются несогласованные притяжательные формы в конструкциях с личными местоимениями 1–2 лица. По-видимому, это означает, что в синхроннных грамматиках носителей татарского языка несогласованная притяжательная форма фиксируется как неизменяемая. Можно предположить, что и другие послеложные слова в дальнейшем будут утрачивать позицию для лично-числового согласования. На это указывают, в частности, данные эксперимента на порождение, в котором не только *турында* ‘о’, но и другие послеложные слова были употреблены с личными местоимениями без согласования притяжательной формы послеложного слова. Более того, в эксперименте на оценку приемлемости не было обнаружено значимых различий между согласованной и несогласованной формой послеложного слова в таком контексте.

Особый интерес представляет поведение возвратного местоимения *үз* ‘сам’ в притяжательной форме 1–2 и 3 лица. С точки зрения внутреннего синтаксиса данных форм *үзем* ‘я сам’ и *үзең* ‘ты сам’ представляют собой изафетную конструкцию 3 с генитивным посессором 1/2 лица, а *үзе* ‘он сам’ — изафетную конструкцию 2 или 3 с посессором 3 лица. Соответственно, эти формы — именные группы 3 лица с субстантивной вершиной, а граммема 1–2 лица у них является согласовательной. Полученные в корпусном исследовании данные

не противоречат такой трактовке этих местоимений. Однако и в эксперименте на порождение, и в эксперименте на оценку приемлемости наблюдается тенденция унификации возвратных местоимений 1–2 лица с личными местоимениями: они получают генитив, допустимы с непритяжательной формой послеложного слова, а с притяжательной формой контролируют согласование послеложного слова по лицу. Возвратное местоимение 3 лица *үзе* ‘он сам’ не отличается по своему поведению от вопросительных местоимений-существительных: оно получает номинатив и не может употребляться с непритяжательной формой послеложного слова.

Субстантивированные атрибутивные местоимения с изафетным показателем 1–2 лица *кайсыбыз* ‘который из нас’, *кайсыгыз* ‘который из вас’ демонстрируют существенное расхождение результатов экспериментов на порождение и оценку приемлемости. Если при порождении они единообразно интерпретировались как именные группы 3 лица (номинатив, несогласуемая притяжательная форма послеложного слова), то оценки приемлемости показывают противопоставление послелогов и послеложных слов. С послелогами *кайсыбыз* ‘который из нас’ и *кайсыгыз* ‘который из вас’ предпочтительны в форме номинатива; с послеложными словами же оценки генитива значимо выше оценок номинатива, а употребление их с непритяжательной формой послеложного слова расценивается как допустимое. Возможность согласования послеложного слова по лицу изафетного показателя не выявляется ни при порождении, ни в оценках приемлемости, что противопоставляет субстантивированные атрибутивные местоимения возвратному местоимению.

Предпринятое обсуждение позволяет также сделать выводы о возможных методиках оценки внутриязыкового варьирования. Представляется, что корпусное исследование и различные экспериментальные методики не исключают, а дополняют друг друга. Так, например, достоверные данные о характеристиках послеложных конструкций с субстантивированными атрибутивными местоимениями и возвратными местоимениями удалось получить только экспериментальным путем. С другой стороны, корпусное исследование позволило выявить классы проблемных контекстов и отобрать лексический материал для экспериментальных исследований.

В целом можно заключить, что квантитативные методы исследования способны существенно дополнить прескриптивный подход к языку. Сравнивая результаты использования этих методов, мы можем определять направление развития языка в данной проблемной области, видеть тенденции, характеризующие сегодняшнее состояние языка, и экстраполировать их, предсказывая следующие состояния. Кроме того, использование экспериментальных методов и методов корпусного анализа позволяет значительно усовершенствовать существующие описательные грамматики. В отечественной лингвистике уже есть положительный опыт использования корпусных данных в качестве эмпирической базы в проекте корпусной грамматики русского языка (rusgram.ru). Мы предполагаем, что методы экспериментального синтаксиса также способны расширить грамматические описания и приблизить их к наблюдаемому языковому поведению носителей. Причем, как показывает настоящее исследование,

применение экспериментальных методик может быть особо эффективным для языков, обладающих более скромными текстовыми ресурсами по сравнению с официальными. Сами экспериментальные методики также различаются по «чувствительности»: методика порождения более тонко улавливает возможные отклонения от грамматического образца, а методика оценки приемлемости позволяет нивелировать элемент случайности и оценить распространенность инновации в языковом сообществе. Отдельно стоит отметить, что экспериментальные методики, особенно в сочетании с корпусными, могут быть использованы для собственно лингвистической оценки состояния языка, его статуса как находящегося вне опасности или потенциально угрожаемого. Все эти соображения позволяют рассматривать корпусные и экспериментальные методики как мощный инструмент исследования внутриязыкового варьирования и его параметризации.

Литература

1. *Герасимова А. А.* (2016) Параметрический подход к внутриязыковому варьированию: проблемы и методы // Рема. Rhema. № 3. С. 63–74.
2. *Лютикова Е. А., Перельцвайг А. М.* (2015) Структура именной группы в безартиклевых языках: универсальность и вариативность // Вопросы языкознания. № 3. С. 52–69.
3. *Лютикова Е. А., Циммерлинг А. В., Коношенко М. Б.* (2016) Языковое разнообразие в зеркале параметрической грамматики // Е. А. Лютикова, А. В. Циммерлинг, М. Б. Коношенко (ред.). Типология морфосинтаксических параметров, Вып. 3. Материалы международной конференции «Типология морфосинтаксических параметров 2016». М.: МПГУ. С. 5–15.
4. *Татарская грамматика* (1993) Т. 2: Морфология / Под ред. М. Ф. Закиева. Казань: изд-во Казанского гос. ун-та.
5. *Татарская грамматика* (1995) Т. 3: Синтаксис / Под ред. М. Ф. Закиева. Казань: изд-во Казанского гос. ун-та.
6. *Lyutikova E.* (2017) Agreement, case and licensing: Evidence from Tatar. Урало-алтайские исследования. Т. 25. № 2. С. 25–45.
7. *Lyutikova E., Pereltsvaig A.* (2015) The Tatar DP. Canadian Journal of Linguistics. Vol. 60. No. 3. P. 289–325.
8. *Pereltsvaig A., Lyutikova E.* (2014) Possessives within and beyond NP: Two ezafe–constructions in Tatar. Advances in the syntax of DPs: Structure, agreement, and case, ed. by A. Bondaruk, G. Dalmi and A. Grosu. Amsterdam: Benjamins. P. 193–219.

References

1. *Gerasimova A. A.* (2016). A parametric approach to intralingual variation: problems and methods [Parametricheskii podkhod k vnutriyazykovomu var'irovaniyu: problemy i metody]. Rhema, № 3, pp. 63–74.
2. *Lyutikova E.* (2017). Agreement, case and licensing: Evidence from Tatar. Ural-Altaic Studies [Uralo-altaiskie issledovaniya], Vol. 25, № 2, pp. 25–45.
3. *Lyutikova E., Pereltsvaig A.* (2015). The Tatar DP. Canadian Journal of Linguistics, Vol. 60, № 3, pp. 289–325.
4. *Lyutikova E. A., Pereltsvaig A. M.* (2015). Noun phrase structure in articleless languages: universality and variation [Struktura imennoi gruppy v bezartiklevykh yazykakh: universal'nost' i variativnost']. Voprosy jazykoznanija, № 3, pp. 52–69.
5. *Lyutikova E. A., Zimmerling A. V., Konoshenko M. B.* (2016). Linguistic Diversity from the Perspective of Parametric Grammar [Yazykovoe raznoobrazie v zerkale parametricheskoi grammatiki]. Typology of morphosyntactic parameters, Vol. 3. Proceedings of the international conference "Typology of morphosyntactic parameters 2016" [Tipologiya morfosintaksicheskikh parametrov, Vypusk. 3. Materialy mezhdunarodnoi konferentsii «Tipologiya morfosintaksicheskikh parametrov 2016»], MPGU, Moscow, pp. 5–15.
6. *Pereltsvaig A., Lyutikova E.* (2014). Possessives within and beyond NP: Two ezafe–constructions in Tatar. Advances in the syntax of DPs: Structure, agreement, and case. Benjamins, Amsterdam, pp. 193–219.
7. *Zakiev M. F. ed.,* (1993). Tatar grammar. Vol. 2: Morphology [Tatarskaya grammatika. T. 2: Morfologiya], Izdatel'stvo KGU, Kazan.
8. *Zakiev M. F. ed.,* (1995). Tatar grammar. Vol. 3: Syntax [Tatarskaya grammatika. T. 3: Sintaksis], Izdatel'stvo KGU, Kazan.

ПРОИЗВОДНЫЕ ЗНАЧЕНИЯ РУССКОГО НЕОПРЕДЕЛЕННОГО НАРЕЧИЯ *КАК-ТО*: ОПЫТ КОРПУСНОГО АНАЛИЗА[1]

Микаэлян И. Л. (irina-mikaelian@yandex.ru)
Университет штата Пенсильвания, США

Зализняк Анна А. (anna.zalizniak@gmail.com)
Институт языкознания РАН, Москва;
Институт проблем информатики ФИЦ ИУ РАН, Москва

Доклад посвящен производным значениям русского неопределенного наречия *как-то*, не отмеченным или недостаточно описанным в существующих словарях. Помимо исходного значения неопределенного образа или способа действия (*грабитель как-то проник в дом*), слово *как-то* имеет два производных значения: значение неопределенного момента времени (*он как-то рассказал мне эту историю*) и значение «общей неопределенности», внутри которого различаются два варианта: а) *как-то* маркирует недоопределенность признака (значение «аппроксимации»): ‘в каком-то отношении /в каком-то смысле/в какой-то степени/ можно сказать’ (*как-то тревожно*; *как-то странно посмотрел*; *как-то смутился*; *как-то по-братски обнял*); б) *как-то* акцентирует неконтролируемость описываемой ситуации (= ‘так вышло, что’): *как-то упустил из виду*; *как-то не собрался*. Были выявлены контекстные условия реализации перечисленных значений слова *как-то*. Проведенный корпусный анализ в частности показал, что значение неопределенного образа действия для слова *как-то* является наименее частотным и преимущественно реализуется одновременно со значением «общей неопределенности». Исследование проводилось на материале НКРЯ, в том числе, параллельного английского подкорпуса. Высокая доля случаев, когда слово *как-то* остается без перевода и, наоборот, когда это слово появляется в переводе на русский язык при отсутствии в тексте оригинала какого-либо определенного «стимула», свидетельствуют о высокой степени лингвоспецифичности данной языковой единицы в целом и, в частности, дискурсивной стратегии маркирования неопределенности.

Ключевые слова: неопределенные наречия, русский язык, неопределенность, аппроксимация, неконтролируемость, дискурсивная стратегия, параллельные корпуса, перевод

[1] Работа выполнена при частичной поддержке РФФИ, грант № 19-012-00505.

DERIVATIVE MEANINGS OF THE RUSSIAN INDEFINITE ADVERB *KAK-TO*: A CORPUS-BASED STUDY

Mikaelian I. L. (irina-mikaelian@yandex.ru)
Pennsylvania State University, USA

Zalizniak Anna A. (anna.zalizniak@gmail.com)
Institute of Linguistics of the RAS; Institute of Informatics Problems of the FRC CSC RAS, Moscow, Russia

The paper analyzes derivative meanings of the Russian indefinite adverb *kak-to*, which are insufficiently described in the existing grammars and dictionaries. Besides its primary meaning of indefinite manner, cf. *grabitel’ kak-to pronik v dom* ‘the buglar somehow got into the house’, *kak-to* has two derivative meanings. 1) It can refer to an indefinite moment in time, cf. *on kak-to mne rasskazal etu istoriju* ‘he told me this story once’; 2) it can function as a discursive marker of ‘general indefiniteness,’ which has two varieties: a) *kak-to* can point to an underspecified aspect of a situation—‘in some respect/in some mesure/kind of’ (*ona kak-to stranno posmotrela na menja, on kak-to smutilsja, on kak-to po-brastki obnjal menja* ‘she gave me an odd glance, he felt somewhat confused, he hugged me in a kind of brotherly way’); b) it can accentuate the idea of uncontrollability of a situation (‘it happened so’): *ja kak-to upustil iz vidu* ‘I somehow overlooked’. Using data form the RNC, we have identified contexts correlating with each of the meanings of *kak-to.* We have also demonstrated that its use as a discursive marker is much more frequent than its occurrences as an adverb of manner proper. We used data from Russian-English and English-Russian parallel subcorpora to demonstrate that in many instances, translators from Russian leave the discursive *kak-to* without a translation, and, vice-versa, translators into Russian frequently insert *kak-to* without a specific stimulus for it in the original English text. We conclude that usage of *kak-to* is regulated by a highly language specific discursive strategy in Russian.

Key words: indefinite pronouns, Russian language, indefiniteness, approximation, non-controllability, discourse strategy, parallel corpora, translation

1. Основные значения слова *как-то*

Как-то — наиболее частотное из всех неопределенных наречий русского языка. Мы выделяем у слова *как-то* следующие три основных значения.

1) Исходное значение неопределенного образа (или способа) действия (далее — НОД);

производные значения:

2) значение неопределенного момента времени (темпоральное значение);

3) значение «общей неопределенности» (далее — ОН), в вариантах (которые могут реализоваться одновременно); *как-то P* =
   а) ‘в каком-то отношении /в каком-то смысле/в какой-то степени/ можно сказать’
   б) неконтролируемости/неопределенной причины (= ‘так вышло, что’).

В значении 3а) наречие *как-то* относится к числу «показателей приблизительности», или «аппроксиматоров» (ср. [Сахно 1983], [Sakhno 2010], [Меркантини 2016]; ср. также категорию «признаковой неопределенности» в [Арутюнова 1998: 814–823]). Используемая нами формула ‘можно сказать’ — это метаязыковой ярлык, означающий ‘говорящему кажется, что воспринимаемые им внешние или внутренние признаки свидетельствуют о наличии ситуации, которая может быть обозначена как P, но он не уверен в точности этой номинации’.

Значение неопределенного способа действия и момента времени регулярно отмечаются словарями; значение общей неопределенности отсутствует в МАС; в словаре Ожегова оно обозначено как «в некоторой степени, несколько»: *Говорит как-то непонятно. Здесь как-то неуютно.*

Реализацию указанных значений *как-то* можно проиллюстрировать следующими примерами с глаголом *решить*[2]:

(1) Видно, вопрос портянок в России из той категории неразрешимых, что и дороги. Армии всех стран ***как-то решили*** этот вопрос, но для нас это просто недостижимая цель. [коллективный. Носи носки // «Огонек», 2013] (значение НОД = ‘неизвестным или неважно каким способом’)

(2) Мы ***как-то решили*** собраться группой после сессии у Аньки. [Маша Трауб. Не вся la vie (2008)] (темпоральное значение = ‘как-то раз’)

(3) Мы долго не пускали её к котам, а потом ***как-то решили*** пожалеть. [Эльвира Савкина. Если впрягаюсь, то основательно (2002) // «Дело» (Самара), 2002.05.03] (значение ОН = ‘так вышло, что решили’)

В примере (1) глагол *решить* выступает в значении ‘найти решение’, которое совместимо с обстоятельством образа действия (*каким образом решили? = какое решение нашли?*), а в примерах (2) и (3) глагол *решить* такого обстоятельства не допускает: он обозначает ментальный акт принятия решения, содержание которого уже определено инфинитивным дополнением. В (2) *как-то* указывает на некоторый неопределенный момент времени в прошлом, когда это решение было принято, а в примере (3) *как-то* скорее понимается в значении ‘так вышло, что’, т. е. указывает на неопределенность или несущественность причины сложившейся ситуации.

Проведенный корпусный анализ показал, что количественно примеры реализации значений слова *как-то* распределены очень неравномерно. А именно, основное значение неопределенного образа действия редко встречается в чистом виде; чаще оно совмещается со значением ‘в каком-то отношении / в какой-то степени’, ср. *как-то поможет, как-то повлияет* (см. ниже).

---

[2] Здесь и далее примеры со ссылкой в квадратных скобках взяты из Национального корпуса русского языка (www.ruscorpora.ru).

В значении неопределенного образа действия *как-то* может иметь сферой действия только глагол, в значении общей неопределенности — глагол или наречие (в том числе, адвербиальное выражение любого типа); в темпоральном значении *как-то* является детерминантом предложения в целом.

Нами были проанализированы первые 100 примеров выдачи по запросу «*как-то*» из основного корпуса НКРЯ. Количественное распределение употреблений слова *как-то* в разных значениях оказалось следующее (цифра в квадратных скобках обозначает количество — и одновременно процент от выборки — примеров данного типа):

1) значение неопределенного образа (способа) действия; *как-то* относится к глаголу [6]:

(4) Чтобы ***как-то*** *убить время*, я решил пойти к доске объявлений и почитать, что там написано. [Запись LiveJournal (2004)] (НОД = ‘каким-то образом’);

2) темпоральное значение; *как-то* относится к глаголу или к предложению в целом [14]

(5) *Тормозит меня* ***как-то*** на посту на выезде из города гаишник. [АвтоБайка. База, это семнадцатый… (2003) // «Марийская правда» (Йошкар-Ола), 2003.01.14] (темпоральное =’как-то раз’)

3) значение общей неопределенности; *как-то* относится к глаголу [24]:

(6) Я ***как-то*** с этим и *не сталкивалась*. [Наши дети: Подростки (2004)] (ОН = ‘так вышло, что’);

значение ОН; *как-то* относится к наречию [46]:

(7) Тащить слонёнка было тяжело и ***как-то*** *бессмысленно*, и ураган Бык бросил его на островке посреди океана. [Александр Дорофеев. Эле-Фантик // «Мурзилка», 2003] (ОН = ‘в каком-то смысле’)

значения НОД и ОН совмещены; *как-то* относится к глаголу [10]:

(8) Я не думаю, что […] те возможные изменения в мексиканском законодательстве, которые сейчас обсуждаются, могли бы ***как-то*** *повлиять* на наше сотрудничество. [В. В. Путин. Заявление для прессы по окончании российско-мексиканских переговоров // «Дипломатический вестник», 2004] (НОД+ОН= ‘каким-то образом’ + ‘в какой-то степени’)

Таким образом, значение общей неопределенности «в чистом виде» представлено в 70 % предложений со словом *как-то*; если учитывать случаи его совмещения со значением неопределенного образа действия — в 80 %.

1.1. Значение неопределенного образа действия vs. общей неопределенности

Значение неопределенного образа действия (в том числе, когда оно реализуется одновременно со значением общей неопределенности) возникает у *как-то* только тогда, когда предикат допускает обстоятельство образа действия, т. е. когда имеется некоторое множество выбора вариантов осуществления действия (*как-то упорядочить, как-то решить проблему, как-то помочь*).

Диагностическим тестом на значение НОД, позволяющим отличить его от значения ОН, может служить возможность замены *как-то* на *как-нибудь*, которое также указывает на выбор одного из вариантов осуществления действия[3]. Этот тест работает только в контекстах «снятой утвердительности» (по [Падучева 1985], т. е. в будущем времени, вопросе, предположении, побудительном высказывании, модальном контексте и т. п.), поскольку вне этих контекстов употребление местоимений на *-нибудь* невозможно. В том случае, когда *как-то* выражает значение образа действия «в чистом виде», оно допускает вполне эквивалентную замену, ср. …*как-нибудь отличиться* для примера (9), …*как-нибудь убить время* для (4) — в отличие от (10), где *как-то* имеет значение ‘в какой-то степени’, и замена на *как-нибудь* возможна лишь с утратой смысла ‘в какой-то степени’ (= ‘хотя бы немного’).

(9) Проштрафившейся сотруднице таможни нужно было ***как-то*** *отличиться*. [Лариса Кислинская. Мерседесы в ловушке-2 (2003) // «Совершенно секретно», 2003.09.01]

(10) Чтобы ***как-то*** *сократить* число «влажных уборок», можно использовать специальный блеск. [Татьяна Булгакова. Цветочная «косметичка» (2003) // «Сад своими руками», 2003.01.15]

Другим диагностическим тестом является возможность вопроса с *Как…?*: он допустим в случае значения образа действия и не допустим для значения общей неопределенности, ср. *Грабитель как-то проник в дом — Как проник?*, но *Он как-то смутился — *Как смутился?*[4].

В значении ОН как-то акцентирует или усугубляет неопределенность, уже содержащуюся в приписываемом признаке. Это значение, как уже было сказано, может реализоваться в двух вариантах.

а) В значении ‘в каком-то отношении /в каком-то смысле/в какой-то степени/можно сказать’ *как-то* выступает в качестве показателя приблизительности, или аппроксиматора. Для этого значения наиболее характерны сочетания с наречиями образа действия. *Как-то* выполняет в этом случае дискурсивную функцию, внося дополнительный элемент субъективной модальности, отражающий точку зрения говорящего или субъекта предложения. При этом *как-то* тяготеет к наречиям, которые сами указывают на высокую степень

[3] См. анализ слова *как-нибудь* в [Зализняк, Денисова, Микаэлян 2018].

[4] Ср. [Арутюнова 1998: 819] о недопустимости вопросов с *как*? по отношению к сочетаниям типа *как-то яростно, как-то как молчаливо*.

неопределенности признака (*странно, особенно, по-другому, таинственно, неопределенно*), а также к отрицательным (по содержанию) признакам (*нелепо, глупо, неуютно, холодно*), в особенности, к наречиям, характеризующим внутреннее состояние субъекта (*неуверенно, нехотя, робко, виновато*).

Сочетание *как-то* с качественными наречиями, обозначающими нейтральный или неоднозначный, а также положительный признак, также возможна (ср. (11), (12)), но менее характерна.

(11) Она ***как-то*** *радостно и привычно*, словно нас ничто не разлучало, потянулась ко мне. [Ю. М. Нагибин. Дневник (1984)]

(12) И тут мне сразу стало легче, как будто гора с плеч свалилась! Стало даже ***как-то*** *весело и приятно*. [Валерий Медведев. Баранкин, будь человеком! (1957)]

Особо отметим частотность сочетания слова *как-то* с наречиями сравнительной степени: *как-то* в этом случае акцентирует смысловой компонент ‘в какой-то степени’, ср.:

(13) Мой старший нашёл работу в фирме по оптовым поставкам техники Hi-End, а младший чахнет над уроками — мы в школе *как-то легче* жили. [Письмо мужчины к женщине (2003)]

(14) С тех пор, как она сменила своё обожание Павла Алексеевича на полное его неприятие, он даже *как-то серьёзнее* стал к ней относиться. [Людмила Улицкая. Казус Кукоцкого [Путешествие в седьмую сторону света]

Значение общей неопределенности может реализоваться также в контексте глаголов различной семантики, не допускающих обстоятельства образа действия. Наиболее частотны здесь предикаты внутреннего состояния, в контексте которых на первый план выступает идея приблизительности номинации, которую мы обозначили метаярлыком ‘можно сказать’: говорящий, на основании внешних (в случае 3-го л.) или внутренних (в случае 1-го л.) признаков подбирает подходящую номинацию, но у него все же остается некоторая неуверенность в ее точности.

(15) Секретарша ***как-то смутилась***, а потом шёпотом сообщила мне по большому секрету: [И. К. Архипова. Музыка жизни (1996)]

(16) Но здесь, в школьном классе, глядя на ребятишек, на кипенно-белые банты в косичках крохотной Маринки Башелуковой, он ***как-то оттаивал***, теплело на сердце. [Борис Екимов. Фетисыч // «Новый Мир», 1996]

(17) Но потом жену вспомнил — хотя и сквозь сон-укол, и это меня ***как-то взвинтило***, так что приободрился даже. [Александр Иличевский. Бутылка (2005) // «Зарубежные записки», 2008]

Для *как-то* характерно употребление в контексте сочиненной группы из двух близких по значению предикатов, усиливающей эффект неопределенности номинации, ср. пример (18), а также (12) и (41).

(18) Была она по-прежнему мила, но ***как-то придавлена и надломлена***.
[Зоя Масленикова. Близкие Бориса Пастернака (1968–2000)]

Показательно, что выдача на поисковый запрос «*как-то* +глагол действия» содержит преимущественно глаголы действия в переносном значении, в котором они характеризуют внутреннее состояние; ср. примеры (15)–(18), а также:

(19) Нинино ***сердце*** сразу стало тяжелым и ***как-то ухнуло*** *вниз*, словно оторвалось. [Анна Сапегина. Еще раз о Бунине // «Сибирские огни», 2012]

(20) Да я читала, но всё равно ***как-то резануло*** Ваше «у историков всё что плохое — от запада». [Женщина + мужчина: Брак (форум) (2004)]

(21) Да, Владик печален теперь, видно, что он задет, ***как-то ранен***,
он загрустил, он рассеян за столом и тоскливо поглядывает вправо,
глаза какие-то растерянные, а густые брови сомкнуты на переносице.
[Са и Со (2002) // «Домовой», 2002.08.04]

Заметим, что про *ранен* в буквальном смысле (допускающем обстоятельство образа действия) вряд ли можно сказать *как-то*.

b) Другой вариант значения общей неопределенности реализуется в контексте предикатов, обозначающих неконтролируемое действие или событие, чаще всего включающих отрицание. Это значение, совмещающее идею неконтролируемости и неопределенной причины, которое мы выше обозначили перифразой ‘так вышло’. Как отмечает [Арутюнова 1998: 821], «(с)ообщения о спонтанном развитии событий, в которые человек вовлечен помимо своей воли, или вопреки ей, регулярно включают НМ [неопределенные местоимения]: *как-то не получилось, как-то само собой вышло, как-то не заладилось, как-то вдруг вырвалось, как-то все произошло само собой, как-то не довелось* и т. п. Например: *Он не видел ее ни разу: как-то не случилось* (Чехов); *Веришь ли, старик, не могу, как-то не получается* (Довлатов). […] Неведомая причина порождает непредвиденное течение событий и может блокировать планы человека. Поэтому сообщения о неуправляемых событиях часто содержат отрицание.» Ср. следующие характерные употребления слова *как-то* в контексте выражений, описывающих неконтролируемые положения вещей, наступившие по непонятной причине, и при этом содержащих отрицание:

(22) Зачем рассказал свой секрет? Теперь они все завтра сядут там, где ты сказал, а нам ловить будет негде. — Ты думаешь? А мне ***как-то*** *не пришло в голову*. [Владимир Солоухин. Григоровы острова (1963)]

(23) […] велосипед же собираются починить для другого мальчика,
который теперь живет здесь вместо когдатошнего Пети Лыпова, но все
как-то *руки не доходят*. [Алексей Слаповский. День денег (1998)]

(24) Друзей у меня ***как-то*** *нету*. Кто был из мужиков, с кем вместе на заводе работал, поумирали или потерялись куда-то.
[Роман Сенчин. Квартирантка с двумя детьми (2010)]

(25) Знаете, а в меня ***как-то*** *не влюблялись*. [Красота, здоровье, отдых: Медицина и здоровье (форум) (2005)]

Употребление *как-то* может служить риторическим приемом отсылки к *якобы* неопределенной причине с целью создания иронического эффекта. В этом употреблении оно выступает как «смягченный» вариант риторического *почему-то*[5], *ср.:*

(26) Англичане вот ***как-то*** *не стесняются* транскрибировать точно, не боясь, что простой народ их не поймет. (А. Н. Барулин. Лингвистическая пыль. Транскрипция русской литературной речи. Краткий предварительный отзыв на Большой орфоэпический словарь русского языка. https://bit.ly/2I52fk2)

Особо отметим пример (27), содержащий сразу несколько предикатов, позволяющих представить собственное действие как не полностью контролируемое (*собирался*, *не собрался*, *успеется*, описанных в [Зализняк, Левонтина 1996]; ср. также *получилось* и *удалось* в примерах (29), (30). Слово *как-то* в контексте таких предикатов акцентирует идею неконтролируемости.

(27) Об этом он собирался как-нибудь переговорить с Ефремом, написать Верусе, посоветоваться с Ильёй Финогенычем, вообще крепко и серьезно подумать, но ***как-то*** *не собрался* и все утешал себя: «Успеется!» [А. И. Эртель. Гарденины, их дворня, приверженцы и враги (1889)]

Без отрицания в контексте предикатов, содержащих сему неконтролируемости, *как-то* совмещает в себе значение неопределенного образа действия и общей неопределенности, ср.:

(28) А у меня сразу стоит ком в горле от того, что мальчик не видел яблоки и впервые увидел их в посылке, и ***как-то*** *догадался* об этом. [коллективный. Форум: Обсуждение фильма «Уроки французского» (1978) (2007–2011)]

(29) — Я их пригласил. Нечаянно. Само ***как-то*** *получилось*. [М. С. Аромштам. Мохнатый ребенок (2010)]

(30) Ему ***как-то*** *удалось* пройти в 1921 году свою первую партийную чистку, не раскрывая всей правды ни об отце и брате — генералах, ни о дворянке-жене, дочери жандармского полковника… [А. Г. Колмогоров. Мне доставшееся: Семейные хроники Надежды Лухмановой (2012)]

[5] Объем настоящей статьи не позволяет предложить более подробный анализ данного феномена. Ср. аналогичное риторическое значение местоимения *что-то*, описанное в [Зализняк, Падучева 2019].

1.2. Темпоральное значение *как-то*

Темпоральное прочтение для *как-то* обычно возникает в контексте предикатов, не сочетающихся с обстоятельством образа действия. Особенно типичен контекст глагола *сказать* и других глаголов речи.

(31) — Вы ***как-то сказали***, что у вас нет любимых фильмов,
а есть те, которые запомнились по атмосфере на съёмках.
[Екатерина Иванова. Оперативник Гармаш дослужился
до «Любовника» (2002) // «Финансовая Россия», 2002.09.19]

(32) Борис Ильич ***как-то рассказал*** мне фронтовой эпизод из своей жизни.
[И. Э. Кио. Иллюзии без иллюзий (1995–1999)]

(33) Один испанский писатель ***как-то пошутил***: «Хирург оперирует в маске
и перчатках, чтобы в случае неудачи сохранить инкогнито».
[Светлана Чечилова. Вещь в себе (1999) // «Здоровье», 1999.03.15]

При наличии при глаголе обстоятельства образа действия темпоральное значение тоже возможно; в этом случае может возникать омонимия со значением неопределенного образа действия, которая обычно разрешается за счет контекста; так, из общего смысла предложения следует, что в (34) реализовано темпоральное значение, а в (35) — значение образа действия.

(34) Мой дед был охотником и ***как-то*** *вырезал* мне из берёзы «летающую
бабочку», которая сразу же стала моей любимой игрушкой.
[Эльвира Савкина. Если впрягаюсь, то основательно (2002) //
«Дело» (Самара), 2002.05.03]

(35) ***Как-то*** *отбившись* от развязной солдатни, он заполз в заросли
терновника, где его мучительно стошнило, и уснул.
[Василь Быков. Главный кригсман (2002)]

Однако неоднозначность может сохраняться; так, (36) может означать ‘как-то раз так случилось, что…’ и ‘каким-то образом случилось’.

(36) ***Как-то*** *случилось*, что эти паспарту закончились, и отец предупредил
Фрадкиса, что если завтра не будет паспарту, то на работу пусть
не приходит и считает себя уволенным.
[И. Э. Кио. Иллюзии без иллюзий (1995–1999)]

2. Дискурсивная функция *как-то*: данные английского параллельного корпуса

В качестве дополнительного инструмента анализа семантики русского *как-то* нами был использован «монофокусный» метод контрастивного анализа, опирающийся на идею, что перевод на другой язык может служить своего рода толкованием исследуемой единицы русского языка, а сопоставление перевода на русский язык с текстом оригинала позволяет выявить те фрагменты

и/или признаки текста оригинала, которые послужили «стимулом» появления этой единицы в русском переводе[6]. Особый интерес при этом представляют случаи, когда русское *как-то* остается без перевода и, в особенности, когда оно возникает в переводе на русский без какого-либо непосредственного «стимула» в оригинале.

2.1. Направление перевода «русский → английский»

В качестве объекта анализа был взят текст романа Булгакова «Мастер и Маргарита»; рассматривались два его перевода на английский язык: Mikhail Bulgakov.Master and Margarita, Transl. by Michael Glenny, 1967 (далее — GL) и Mikhail Bulgakov. Master and Margarita. Transl. by Richard Pevear, Larissa Volokhonsky, 1979 (далее — P&V)[7].

В романе имеется 33 вхождения слова *как-то*[8]. *Из них 30 раз как-то* употреблено в значении общей неопределенности в разных ее вариантах и 3 раза в темпоральном значении. В значении собственно неопределенного образа действия слово *как-то* не встретилось ни разу. Нас будут далее интересовать только употребления *как-то* в значении общей неопределенности.

Выбранные переводы выполнены в двух различных стратегиях: в GL переводчик очевидно стремился передать общий смыл высказывания, выразив его наиболее естественным для себя способом. В P&V переводчики стремились максимально точно передать значение русской фразы и по возможности сохранить все ее элементы. Соответственно, данные относительно использованных моделей перевода кардинальным образом расходятся. А именно, в GL в 21 одном случае из 30 нет никакого эквивалента, а в P&V, наоборот, в 24 случаях из 30 переводчики используют наиболее очевидный словарный эквивалент *как-то — somehow*.

Опущение *как-то* в переводе может приводить к частичной утрате смысла. Так, в примере (37) *как-то* указывает, что происходящее воспринимается как ненормальное самим Варенухой. В переводе P&V *somehow* выполняет ту же функцию, тогда как в GL сохраняется только «внешнее» повествование.

(37) Тут в кабинетике [Варенухи] ***как-то*** *быстро стало темнеть*.
Here *it* ***somehow*** *began to grow dark very quickly* in his little office (P&V)
At that moment his office *began to darken*. (GL)

В примере (38) *как-то* подчеркивает, что глагол *смягчился* (как и эпитет *бесовский*) — выражают восприятие происходящего Иваном, и ту же функцию выполняет *somehow* в переводе P&V.

---

[6] В [Бунтман и др. 2014] этот принцип был назван «унидирекциональным».

[7] Перевод P&V имеется в НКРЯ; перевод GL доступен онлайн.

[8] В тексте оригинала, размещенном на сайте НКРЯ, есть еще одно предложение с *как-то*, которое мы не учитываем, поскольку оно опущено в обоих переводах.

(38) Иван опять прилег и сам подивился тому, как изменились его мысли. ***Как-то*** *смягчился* в памяти проклятый бесовский кот, не пугала более отрезанная голова
Ivan lay down again and marvelled himself at how changed his thinking was. The accursed, demonic cat ***somehow*** *softened* in his memory, the severed head did not frighten him any more (P&V)
Ivan lay down again. He was amazed to notice how his mental condition had changed. The memory of the diabolical cat *had grown indistinct*, he was no longer frightened by the thought of the decapitated head. (GL)

В примере (39) *как-то* подчеркивает неконтролируемость действия председателя, и P&V в той же функции используют слово *somehow*, в то время как GLвообще устраняет эту идею из текста (оставляя без перевода не только *как-то*, но и *не удержавшись*).

(39) Пересчитав деньги, председатель получил от Коровьева паспорт иностранца для временной прописки […] и, ***как-то*** *не удержавшись*, стыдливо попросил контрамарочку
After counting the money, the chairman received from Koroviev the foreigner's passport for temporary registration […] and, ***somehow*** *unable to help himself*, sheepishly asked for a free pass (P&V)
Having counted the money the chairman took the stranger's passport to be stamped with his temporary residence permit […] and asked shyly for a free ticket to the show. (GL)

Как мы видим, опущение слова *как-то* при переводе приводит к потере части смысла исходного предложения и возможно даже к его искажению — утрате отсылки к субъекту номинации. Таким образом, проведенный анализ позволяет более определенно выявить функцию русского *как-то* как средства акцентирования неопределенности (в том числе, странности или неконтролируемости) описываемой ситуации, причем эта установка может принадлежать не только говорящему (повествователю), но также и персонажу.

2.2. Направление перевода «английский → русский»

В отличие от переводов с русского языка, где, как мы убедились на примере анализа двух переводов романа «Мастер и Маргарита» на английский язык, употребление той или иной модели перевода принципиальным образом зависит от переводческой стратегии, в переводах *на* русский язык интересующее нас слово появляется в результате решения задачи передачи смысла оригинального предложения в целом, и тем самым может служить более надежным источником информации о семантике интересующей нас русской языковой единицы. Для анализа нами была использована сплошная выборка 100 первых примеров из выдачи по запросу «как-то» из английского параллельного подкорпуса НКРЯ в направлении перевода «английский → русский».

Проведенный анализ показал, что в качестве «стимула» появления в русском переводе слова *как-то* (в значении НОД и/или ОН) могут выступать слова

somehow, *something*, *almost*, *just*, *really*, *in any way*, неопределенный артикль *a* и аппроксиматор *sort of*. Однако чаще всего, а именно, в 44 примерах из 100 в оригинальном английском тексте нет никакого фрагмента, который соответствовал бы русскому *как-то* (что свидетельствует о высокой степени лингвоспецифичности данной единицы)[9]. Ср.:

(40) [...] the children gradually realized that it was a slightly overweight mailman, pointing down the street and *looking at the children fearfully*. [Lemony Snicket. The Erzatz Elevator (2001)]
Дети не сразу поняли, что перед ними почтальон с несколько избыточным весом. Он ***как-то*** *боязливо смотрел на детей*. [Лемони Сникет. Липовый лифт (А. Ставиская, 2005)] (значение ОН: приблизительность номинации внутреннего состояния)

(41) The camerlegno *loosened* like a tall ship that had just run sheets first into a dead calm. [Dan Brown. Angels and Demons (2000)]
Камерарий весь ***как-то*** *обмяк и обвис*. Так обвисают паруса корабля, неожиданно попавшего в мертвый штиль. [Дэн Браун. Ангелы и демоны (Г. Косов, 2004)] значение ОН: приблизительность номинации признака)

(42) “We need to *refine the parameters further*”, Gettum said, stopping the search. [Dan Brown. The Da Vinci Code (2003)]
— Нам следует ***как-то*** *сузить круг поиска*, — заметила Геттем. [Дэн Браун. Код Да Винчи (Н. Рейн, 2004)] (НОД+ОН: ‘каким-то образом’ + ‘в какой-то степени’)

(43) Count Olaf said in his raspy voice, and the Baudelaire orphans were too stunned *to defend themselves*. [Lemony Snicket. The Erzatz Elevator (2001)]
Бодлеры были настолько ошеломлены, что не могли вымолвить ни слова, чтобы *хоть* ***как-то*** *себя защитить*. [Лемони Сникет. Липовый лифт (А. Ставиская, 2005)] (НОД+ОН: ‘каким-то образом’ + ‘в какой-то степени’)

Однако при отсутствии в предложении-оригинале непосредственного стимула в нем могут присутствовать определенные элементы, которые обусловили появление *как-то* в русском переводе; назовем их «контекстными стимулами» или «стимулирующими контекстами». Это: глагол *to manage*, обозначающий действие с не полностью контролируемым результатом; слова *strangely, vague*, *dubious* и т. п., в значение которых входит компонент неопределенности признака; глаголы *to seem*, *to sound* и *to look*, маркирующие тот факт, что описываемый признак идентифицирован на основании интерпретации перцептивных данных, которая может быть неточной, а также обозначения неконтролируемых внутренних состояний с отрицанием (*hadn’t thought*, *had not considered,* и т. п.). Ср.:

[9] О понятии степени лингвоспецифичности языковой единицы и количественных методах ее оценки см. Зализняк 2015, Инькова 2017.

(44) “The initials,” Langdon whispered, eyeing her ***strangely***. [Dan Brown. The Da Vinci Code (2003)]
— Инициалы, — прошептал Лэнгдон, ***как-то*** *странно* глядя на нее. [Дэн Браун. Код Да Винчи (Н. Рейн, 2004)]

(45) No bars or anything.” She laughed, but it ***sounded*** *hollow*. [Lauren Weisberger. The Devil Wears Prada (2003)]
Никаких тяжелых железных решеток или чего-то в этом роде. — Она засмеялась, но ***как-то*** *вымученно*. [Лорен Вайсбергер. Дьявол носит Прада (М. Маяков, Т. Шабаева, 2006)]

(46) I was trying to say the bare minimum, since *it* ***seemed*** *incredibly strange* to be talking on the phone in front of Miranda. [Lauren Weisberger. The Devil Wears Prada (2003)]
— Я старалась произносить как можно меньше слов, говорить по телефону в присутствии Миранды было ***как-то*** *дико*. [Лорен Вайсбергер. Дьявол носит Прада (М. Маяков, Т. Шабаева, 2006)]

(47) “*I* ***hadn’t thought*** *about him*”, Violet said. “He always follows instructions”. [Lemony Snicket. The Erzatz Elevator (2001)]
— *Я* ***как-то*** *не подумала о нем*, — смутилась Вайолет. — Он всегда строго придерживается инструкции [Лемони Сникет. Липовый лифт (А. Ставиская, 2005)]

В примере (48) в исходном английском предложении нет непосредственного стимула не только для *как-то*, но и для другого лингвоспецифичного русского слова, появившегося в переводе — *ухитриться*. При помощи сочетания *как-то ухитрился* переводчик эксплицировал идею неполной контролируемости, присутствующую в английской фразе имплицитно — в форме ‘сделал, хотя это было трудно’ (ср. *struggling against his handicap — reached*).

(48) Struggling against his handicap Kohler reached down and carefully twisted Vetra’s frozen head. [Dan Brown. Angels and Demons (2000)]
Колер, кряхтя и задыхаясь, все же ***как-то ухитрился***, оставаясь в кресле, склониться и осторожно повернуть прижатую к ковру голову Ветра. [Дэн Браун. Ангелы и демоны (Г. Косов, 2004)]

Как проницательно отметила в свое время [Н. Д. Арутюнова 1998: 823], «[о]билие НМ [=неопределенных местоимений], относящихся к признаковым значениям составляет важную характеристику русского дискурса — разговорной речи и письменного текста, которую можно определить как свойство открытости. НМ — это знаки невыраженных или невыразимых смысловых компонентов: невскрытых причин событий, неясных мотивов, поступков, неопределенных и неопределимых вариантов признаков, следы действия неведомых сил».

Проведенный нами анализ русского неопределенного местоимения *как-то* позволяет придать этому утверждению более веские основания. Тот факт, что в переводе почти в половине случаев слово *как-то* возникает за счет того, что переводчик эксплицирует имплицитно содержащийся во фразе

смысл, и даже иногда привносит его, свидетельствует о том, что употребление русского *как-то* определяется, в значительной степени, характерной для русского языка дискурсивной стратегией маркирования и акцентирования неопределенности, в том числе и в особенности — неконтролируемости.

Авторы благодарят анонимных рецензентов за высказанные замечания, которые были по возможности учены в окончательном тексте статьи.

Литература

1. *Арутюнова Н. Д.* (1998), Язык и мир человека. М., 1998.
2. *Бунтман Н. В., Зализняк Анна А., Зацман И. М., Кружков М. Г., Лощилова Е. Ю., Сичинава Д. В.* (2014), Информационные технологии корпусных исследований: принципы построения кросс-лингвистических баз данных // Информатика и ее применения. Т. 8, вып. 2, 2014. С. 98–110.
3. *Зализняк Анна А.* (2015), Лингвоспецифичные единицы русского языка в свете контрастивного корпусного анализа // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции Диалог’2015. М., 2015. С. 651–662.
4. *Зализняк Анна А. , Денисова Г. В., Микаэлян И. Л.* (2018), Русское как-нибудь по данным параллельных корпусов // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции Диалог’2018. С. 803–817.
5. *Зализняк Анна А., Левонина И. Б.* (1996), Отражение «национального характера» в лексике русского языка (размышления по поводу книги: Anna Wierzbicka. Semantics, Culture, and Cognition. Universal Human Concepts in Culture-Specific Configurations. N.Y., Oxford, Oxford Univ. Press, 1992) // Russian Linguistics, vol. 20, 1996, pp. 237–264.
6. *Зализняк Анна А., Падучева Е. В.* (2019), Русское что-то как дискурсивное слово // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции Диалог’2019.
7. *Инькова О. Ю.* (2017), Принципы определения степени лингвоспецифичности коннекторов. // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции Диалог’2017. М., 2017.
8. *Меркантини С.* (2016), Семантическая категория аппроксимации и средства ее выражения в современном итальянском языке. Дисс. …канд. филол. наук. М., 2016
9. *Падучева Е. В.* (1985), Высказывание и его соотнесенность с действительностью. М.: Наука, 1985.
10. *Сахно С. Л.* (1983), Приблизительное именование в естественном языке. // Вопросы языкознания, №6, 1983. С. 29–36.
11. *Sakhno S.* (2010), Les avatars du sens et de la function dans le phénomène de la grammaticalization: Description systématique du lexème russe vrode “dans le genre de” comparé à d’autres lexèmes russes grammaticalisés à fonctionnement proche. Nanterre, Université Paris Ouest, 2010. <hal-00765376>

AN ATTENTION-BASED APPROACH TO AUTOMATIC GAPPING RESOLUTION FOR RUSSIAN

Movsesyan A. A. (derise@iitp.ru)

Institute for Information Transmission Problems (Kharkevich Institute), Russian Academy of Sciences, Moscow, Russia

Gapping is a type of ellipsis in which a finite verb is elided in a coordinate structure. Reconstruction of the elided material is essential for different NLP tasks. However, from a practical point of view, the problem did not receive considerable attention for Russian language because of lack of training data and rarity of the phenomenon itself. This paper is one of the first works of deep learning-based automatic gapping resolution in Russian as a part of AGRR-2019 competition. We used a recurrent neural network-based approach to determine presence/absence of gapping in a sentence and for the full annotation we applied a Universal Transformer neural network that combines self-attention mechanism with recurrence in depth. Also using pretrained fastText word embeddings, we achieved 85% standard F-measure on test set for binary classification task and 62% symbol-wise F-measure for full annotation task. We assume that fixed word embedding like fastText does not contain enough syntactic information to properly match remnants in sentences with gapping. Also we show that our model generalize better if punctuation marks were ignored during training and evaluation.

Key words: gapping, Universal Transformer, fastText, deep learning, NLP

АВТОМАТИЧЕСКОЕ РАЗРЕШЕНИЕ ЯВЛЕНИЯ ГЭППИНГА ДЛЯ РУССКОГО ЯЗЫКА С ИСПОЛЬЗОВАНИЕМ МЕХАНИЗМА ВНИМАНИЯ

Мовсесян А. А. (derise@iitp.ru)

Институт проблем передачи информации РАН им. А. А. Харкевича, Москва, Россия

Гэппингом называют тип эллипсиса, когда при сочинении опускается финитный глагол. Восстановление опущенного предиката является важным для различных задач обработки естественного языка. Однако,

эта задача, с практической точки зрения, не привлекала существенного внимания исследователей ввиду редкости самого явления и отсутствия соответствующих корпусов текстов для русского языка. В данной работе осуществляется одна из первых попыток разрешения явления гэппинга для русского языка с использованием методов глубокого обучения в рамках соревнования AGRR-2019. Для определения наличия гэппинга в предложении мы применили подход на основе рекуррентной нейронной сети, а для полной аннотации использовали нейросетевую архитектуру Universal Transformer, основанную на механизме внимания с рекуррентными связями в глубину. Используя также предобученные векторные представления слов fastText, мы получили результат 85% (стандартная F-мера) для задачи бинарной классификации и 62% (посимвольная F-мера) — для задачи полной аннотации. Мы предполагаем, что фиксированные векторные представления слов как fastText не содержат достаточно синтаксической информации для корректного сопоставления «остатков» с их коррелятами в предложениях с гэппингом. Мы также видим основания считать, что наша модель имеет более высокую способность к генерализации, если не учитывать пунктуацию при обучении и проверке модели.

Ключевые слова: гэппинг, Universal Transformer, fastText, глубокое обучение, NLP

1. Introduction

According to [Ross, 1970], gapping is a type of ellipsis in which a repeated main verb is elided in one or more conjuncts of a coordinate structure, such as in the example (1).

(1) *Moj otec znal ego otca, moj ded — ego deda.*
My father knew his father, my grandfather — his grandfather.

Despite the fact that this phenomenon was widely discussed from a theoretical point of view, there is still no consensus on some cases. For example, gapping can occur in comparative constructions and “short” answers. Moreover, the differentiation between gapping and other types of ellipsis (such as VP-ellipsis and stripping) is not trivial. We refer to [Johnson, 2014] for more details.

However, from a practical point of view, it pose challenges as well. First of all, it is not obvious how a conjunct with the elided main verb should be presented in a sentence’s dependency representation since all dependency representations consider a verb to be the head of a clause. Different approaches were proposed to address this issue, including adding empty nodes [Boguslavsky et al., 2002] and incorporating new or adapting existing dependency relations [Schuster et al., 2018].

The second problem is connected with syntactic parsers. Continuing with the example (1), one option is to reconstruct the verb *znal (knew)* in the sense of its wordform and linear position in the sentence. Then standard parsers should perform well, but such reconstruction is a challenging task itself. Another option is to develop a parser that correctly deals with a clause with a gap and then reconstructs the gap. It is possible not to reconstruct the gap in the latter approach, but indicating the elided

material is essential for downstream tasks such as semantic role labeling [Matthew Lamm and Liang, 2018] and semantic parsing [Ge and Mooney, 2009]. The rarity of the phenomenon of ellipsis in natural languages [Droganova and Zeman, 2017] and lack of training data makes the latter approach even more difficult for statistical parsers.

It is worth to mention that there were a couple of attempts to address this phenomenon in Russian language from practical point of view. In a recent paper [Droganova et al., 2018] the authors trained two existing statistical parsers on a corpus pre-enriched with sentences with gapping. They obtained some improvements of the parsing accuracy of gapping in Russian compared to the baseline where the corpus was not enriched with the gapped sentences, but the improvement was not significant. Another attempt is presented in paper [Bogdanov, 2012] and is basically an extension to an existing rule-based parser. Unfortunately, the paper lacks any evaluation and the approach is strongly dependent on the parser.

In this paper, we propose an automatic gapping resolution system for Russian based on recently proposed Universal Transformer neural network architecture. Our model was evaluated during AGRR-2019 competition. The paper is structured as follows. In **Section 2** we give an overview of data and task description. Moreover, since there is no generally acceptable theoretical definition of gapping, we formulate a working definition based on the data provided. In **Section 3** we describe our approach in details. The results are presented in **Section 4** and the conclusion is provided in **Section 5**.

2. Data and task description and gapping definition

As was mentioned in **section 1**, the evaluation of the proposed model was performed during the AGRR-2019 competition. The organizers provided a corpus of several thousands of sentences from texts of different genres. The corpus statistics is shown in **Table 1**.

Table 1: AGRR-2019 corpus statistics

	Training set	Development set	Test set
Sentences with gapping	5,542	1,382	636
Sentences without gapping	10,864	2,760	1,409

Because some cases of gapping are controversial from theoretical point of view, automatic gapping resolution cannot be held in its entirety. So it is necessary to provide a working definition of gapping that, according to the data provided, can be formulated as follows.

Definition 1 (Working definition of gapping in Russian) *Gapping is a type of ellipsis in which a repeated finite verb, possibly along with contiguous portions of its verb phrase, is elided in one or more clauses conjoined to the right of a clause containing the same verb, with a remnant material at least to the right of the gap.*

Here, the remnant material is the contiguous overt material in a gapped clause. Since the elided material is contiguous, there are no more than two remnants in a gapped clause. Not only a main verb can be omitted in a gapped clasue of a sentence in Russian such as in the example (2), but there are not such examples in the data provided. But at the same time stripping and left node raising are considered as gapping.

(2) *On ee v stol položit, a my voz'mem da v škap pereložim...*
He it in table put, and we to cupboard moved...
(He put it in his table, and we moved it to the cupboard...)
[M. E. Saltykov-Ŝedrin. Gospoda Golovlevy (1875–1880)]

Each sentence in the corpus is annotated as follows:

1. There is a label indicating whether a sentence contains a gap or not.
2. If there is a gapping construction in a sentence, character offsets for annotation borders for each gapping element are provided. Namely, these elements are: the elided predicate (V) with its remnants (R_1, R_2) for every gapped clause; the head of the correspondent predicate (cV) with the correlates of the remnants (cR_1, cR_2) for the initial conjunct.

So, since there is a gapping construction in the sentence from the example (1), the annotation will look as follows:

(3) $[_{cR_1}$*Moj otec*$]$ $[_{cV}$*znal*$]$ $[_{cR_2}$*ego otca*$]$, $[_{R_1}$*moj ded*$]$ — $[_{V}]$ $[_{R_2}$*ego deda*$]$.

Overall, we can classify (see also **Table 2**) gapping constructions presented in the corpus by:

1. type of gap:
   (a) single predicate;
   (b) predicate with portions of its verb phrase (contiguous material);
2. number of gapped clauses:
   (a) one clause;
   (b) more than one clause;
3. number of remnants:
   (a) one remnant;
   (b) two remnants.

Table 2: Extended AGRR-2019 corpus statistics. Only sentences with gapping are included. Number of sentences with different types of gap were estimated by distance between the head of the verb phrase in the initial conjunct and one of the correlates of the remnants.

	Training set	Development set	Test set
Type of gap			
Single predicate	4,581	1,141	583
Predicate-arguments	961	241	97
Number of gapped clauses			
One clause	5,173	1,292	632
More than one clause	369	90	48
Number of remnants			
One remnant	77	27	17
Two remnants	5,465	1,355	663

Three tasks were presented by the organizers.

1. Binary classification. For a given sentence decide if there is a gapping construction in it.
2. Gap resolution. Predict the position of the elided predicate and the correspondent predicate in the antecedent clause.
3. Full annotation. In each clause with the gap predict the linear position of the elided predicate and annotate its remnants. In the antecedent clause find the constituents that correspond the remnants and the predicate that corresponds the gap.

3. Model description

Recurrent and convolutional neural networks has shown promising results in natural language processing tasks in recent years [Yin et al., 2017], [Young et al., 2018]. Despite the fact that every hidden state update in RNN takes previous states into account, however, combining attention mechanism [Bahdanau et al., 2014] with RNNs has become a standard for solving different tasks, especially for encoder-decoder based machine translation systems [Wu et al., 2016]. It led to developing network architectures based solely on attention without any recurrence or convolution. One such model, the Transformer [Vaswani et al., 2017], has established new state-of-the-art results on machine translation tasks. However, one limitation of the network is that it does not generalize well to input lengths not encountered during training. To make the network computationally universal and, in particular, to overcome the mentioned issue, the Universal Transformer with recurrence over depth (unlike RNNs in which recurrence is over time) was recently proposed [Dehghani et al., 2018]. The latter model was also shown to be able to capture dependency structure of a sentence, outperforming the vanilla Transformer significantly. Since the gapping phenomenon

is considered to be purely syntactic, we decided to use a part of the Universal Transformer network in our model.

Speaking about the task, the main observation is that the position for the elided predicate is known after we found the offsets for its remnants: it immediately precedes the second remnant (or the first if it is the only one) in each gapped clause. It implies that all three proposed tasks could be treated as one and it is now straightforward to formulate the joint task as a sequence labeling problem. Namely, the labels are $\{R_1, R_2, cV, cR_1, cR_2, nG\}$, where nG is a label for a word not connected with gapping phenomenon. Now, the example (3) transforms into

(4)

Moj	*otec*	*znal*	*ego*	*otca*	,	*moj*	*ded*	—	*ego*	*deda*	.
cR_1	cR_1	cV	cR_2	cR_2	nG	R_1	R_1	nG	R_2	R_2	nG

Below we explain the models we evaluated to solve the task in more details. The code is publicly available on GitHub[1].

3.1. Data representation

Since the input data is raw text (splitted into sentences), some data preprocessing must be made. First of all, we tokenized the sentences, using NLTK library [Loper and Bird, 2002] with external tokenization model for Russian[2]. But the problem is that the model can generalize worse paying too much attention to punctuation. Even a simple binary classifier that predicts whether a sentence contains a gapping construction in it based solely on presence of a dash achieves precision and recall of about 70% on the training set. That is why the second option we tried is to just ignore all punctuation and treat every character sequence surrounded by non-alphanumeric characters as a word.

Secondly, we used extended fastText word embedding [Bojanowski et al., 2017] pretrained on Wikipedia and Common Crawl [Grave et al., 2018]. It is based on skip-gram model [Mikolov et al., 2013] but each word is represented as a bag of character n-grams along with the word itself. Incorporating the subword information has two important advantages connected with the task:

1. it captures morphological information, improving performance on syntactic tasks, especially for morphologically rich languages such as Russian;
2. the model can produce word vectors for out-of-vocabulary (OOV) words treating a word as a set of n-grams.

[1] https://github.com/Derise/agrr

[2] https://github.com/Mottl/ru_punkt

3.2. Model architecture

We tried two different models. Both models assign a label for each word in a sentence, but one of them is divided into two submodules: one is a binary classifier (solves task 1) and another one is a multi-class classifier trained only on gapped sentences. Multi-class classifiers in both models share the same architecture.

3.2.1. Binary classifier

We used a 2-layer bidirectional gated recurrent neural network (biGRU) [Cho et al., 2014] with dropout [Gal and Ghahramani, 2016]. The hidden state on the last time step of the second layer is an input to a fully-connected layer with sigmoid activation. The cost function is cross entropy.

We used Adam optimizer [Kingma and Ba, 2014] with learning rate $\alpha = 0.000625$. We did not change learning rate during training: instead we adopted the approach presented in [Smith et al., 2017]. Namely, if the validation loss after an epoch is not minimal compared to all previous losses, the batch size is doubled. The upper bound for the batch size is limited to the memory size.

3.2.2. Multi-class classifier

The mentioned above Universal Transformer architecture consists of encoder and decoder with the same basic structures. But since we formulated our task as a sequence labeling problem, no decoder is needed. So we applied a softmax layer directly after the output of the encoder. No changes were made to the encoder architecture compared to the original version; therefore we skip the detailed description of the encoder and refer to the original paper [Dehghani et al., 2018].

Whether the classifier is trained on all sentences from the training corpus or only gapped ones, the configurations are the same. The hidden size is the size of word embedding (300), the number of self-attention heads is 6 and the model depth is 12. Adam optimizer is applied with the same learning rate as for binary classification model. The Universal Transformer takes more memory to train compared to biGRU with the same number of parameters (because recurrence over time steps in RNN is not parallelizable) and we could not increase the batch size. Instead, we used step decay scheme: every 5 epochs the learning rate is decreased by a factor of 2 if loss did not improve.

4. Results and error analysis

The results are shown in **Table 3**. Two different models were evaluated:

- biGRU+UT: biGRU as a binary classifier is evaluated and then the Universal Transformer encoder as a multi-class classifier is applied to tag each word in sentences which were predicted as gapped ones.
- UT(joint): the Universal Transformer encoder were trained on all sentences. If there are no words with tags R_1, cR_1 or cV in a given sentence, it is considered as the one without gapping in it.

Table 3: Standard (for binary classification) and symbol-wise (for other tasks) F-measure of different models trained and evaluated on AGRR-2019 dataset

Model	Train			Dev			Test		
	Binary	Gap	Full	Binary	Gap	Full	Binary	Gap	Full
NLTK tokenizer									
biGRU+UT	0.97	0.82	0.78	**0.92**	**0.73**	**0.69**	**0.85**	**0.64**	**0.60**
UT(joint)	—	—	—	0.76	0.52	0.49	0.65	0.41	0.39
UT_{LARGE}(joint)	—	—	—	0.72	0.51	0.48	0.70	0.49	0.45
Simple tokenizer, no punctuation									
biGRU+UT	0.89	0.51	0.54	0.82	0.44	0.46	0.82	0.45	0.47
UT(joint)	0.70	0.41	0.41	0.66	0.34	0.34	0.65	0.35	0.34

Additionally, we tried bigger model of UT(joint) by adding fully connected layer before UT input, increasing the hidden size from 300 to 512 (increasing total parameters from 1.5M to 3M). Moreover, as was mentioned in **section 3.1**, two tokenization techniques were used: one is using NLTK library and another is a simple approach where all punctuation marks are ignored and each sentence is splitted into words by non-alphanumeric characters.

For binary classification task the metric is standard f-measure. For other tasks the metric is symbol-wise f-measure, here is the example from the organizers: "if the gold standard offset for certain gapping element is 10:15 and the prediction is 8:14, we have 4 true positive chars, 1 false negative char and 2 false positive chars and the resulting f-measure equals 0.727".

Drawing attention to the biGRU+UT model, we take into consideration only those sentences which were correctly predicted by the binary classifier. Almost all *V* tags are correct (in terms of recall) and it turns out that the most challenging task for the model was remnants matching. The most frequent errors are:

1. Not determining an elision of dependents of a verb (5): here and throughout, the example (5a) is correct and the example (5b) is the output of the model.

(5) a. [$_{cR_1}$*Sverh"estestvennoe vmešatel'stvo v dela prirody*] [$_{cV}$*kazalos'*] *emu* [$_{cR_2}$*istočnikom užasa*] *, a* [$_{R_1}$*bessmertie*] — [$_{V}$] [$_{R_2}$*fatal'nym dlja nadeždy izbavit'sja ot boli*].

b. [$_{cR_1}$*Sverh"estestvennoe vmešatel'stvo v dela prirody*] [$_{cV}$*kazalos'*] [$_{cR_2}$*emu istočnikom užasa*] *, a* [$_{R_1}$*bessmertie*] — [$_{V}$] [$_{R_2}$*fatal'nym dlja nadeždy izbavit'sja ot boli*].

(The supernatural intervention in the affairs of nature seemed a source of horror to him, and the immortality—fatal for hope to get rid of pain.)

2. Incorrect remnants matching (6).

(6) a. *Hotja oni* [$_{cV}$*nazyvali*] [$_{cR_1}$*tebja*] [$_{cR_2}$*drugom*] *, a* [$_{R_1}$*ee*] [$_{V}$] [$_{R_2}$*podrugoj*]*!*

b. [$_{cR_1}$*Hotja oni*] [$_{cV}$*nazyvali*] [$_{cR_2}$*tebja drugom*] *, a* [$_{R_1}$*ee*] [$_{V}$] [$_{R_2}$*podrugoj*]*!*

(Although they called you a friend, and her a friend!)

3. Determining multiple gaps when there is only one (7).

(7) a. *V obŝem,* [$_{cR_1}$*politiki i generaly*] [$_{cV}$*pozabotilis'*] [$_{cR_2}$*o svoih mestah, o svoej kar'ere, o svoih kreslah*] , *a* [$_{R_1}$*kto-to eŝe*] [$_{V}$] [$_{R_2}$*i o svoih karmanah*].

b. [$_{cR_1}$*V obŝem,* [$_{cR_1}$*politiki i generaly*] [$_{cV}$*pozabotilis'*] [$_{cR_2}$*o svoih mestah*], [$_{cR_1}$*o svoej kar'ere*], [$_{V}$] [$_{cR_2}$*o svoih kreslah*], *a* [$_{R_1}$*kto-to*] [$_{V}$] [$_{R_2}$*eŝe i o svoih karmanah*] .

(Well, the politicians and the generals have taken care about one's places, one's career, one's armchairs, and somebody also about one's pockets.)

To compare the results with other participants of the competition we used biGRU+UT model with NLTK tokenizer with minor hyperparameters tuning for the UT part. The comparative results (obtained for the test set) are shown in **Table 4**.

For this final model we also reviewed the results for the sentences with gapping in the test set according to the classification mentioned in **Section 2**. It is shown in **Table 5**.

The most conspicuous result is for full annotation of sentences with one remnant. The reason is that the model almost always finds two remnants (8).

(8) a. [$_{cV}$*Dobavljaem*] [$_{cR_1}$*muku, krahmal i razryhlitel'*] , *a v konce* — [$_{V}$] [$_{cR_1}$*smetanu*].

b. [$_{cV}$*Dobavljaem*] [$_{cR_1}$*muku, krahmal*] [$_{cR_2}$*i razryhlitel'*] , *a* [$_{cR_1}$*v konce*] — [$_{V}$] [$_{cR_2}$*smetanu*].

(Add flour, starch and baking powder, and sour cream at the end.)

Another interesting observation is that full annotation performance is weaker if the elided material includes portions of the VP along with the main verb itself. That is because these portions tend to become a part of remnants' correlates (5), (9).

(9) a. [$_{cR_1}$*Ono*] [$_{cV}$*dolžno*] *zahvatit'* [$_{cR_2}$*vas*] , *a* [$_{cR_1}$*ne vy*] [$_{V}$] [$_{cR_2}$*ego*].

b. [$_{cR_1}$*Ono*] [$_{cV}$*dolžno*] [$_{cR_2}$*zahvatit' vas*] , *a* [$_{cR_1}$*ne vy*] [$_{V}$] [$_{cR_2}$*ego*].

(It must capture you, and not you him.)

Table 4: The comparative AGRR-2019 results of competitors' models. The best three results in each task are in bold.

	Binary			Gap resolution	Full
Team	**Precision**	**Recall**	**F-measure**	**F-measure**	**F-measure**
fit_predict	0.97	0.95	**0.96**	**0.90**	**0.89**
EXO	0.90	0.96	**0.93**	**0.81**	**0.79**
Koziev Ilya	0.78	0.90	0.83	**0.68**	**0.65**
Derise	0.80	0.91	**0.85**	0.66	0.62
Meanotek	0.89	0.78	0.83	0.64	0.51
МГУ-DeepPavlov	0.93	0.64	0.76	0.60	0.59
Vlad	0.78	0.92	0.84	0.57	—
MorphoBabushka	0.76	0.62	0.68	0.47	0.44
nsu-ai	0.49	0.126	0.20	0.04	0.04

Table 5: Extended results of the final model for the test set. Since the results are shown only for the sentences with gapping, precision and standard F-measure for binary classification are not shown.

	Binary	Gap resolution	Full
	Recall	F-measure	F-measure
Type of gap			
Single predicate	0.91	0.82	0.79
Predicate-arguments	0.89	0.80	0.69
Number of gapped clauses			
One clause	0.90	0.81	0.76
More than one clause	0.98	0.84	0.76
Number of remnants			
One remnant	0.76	0.69	0.24
Two remnants	0.91	0.82	0.78

With regard to binary classification the main observation is that the more information related to the phenomenon of gapping is presented in a sentence, the higher predictions are made by the classifier.

5. Conclusion

In this paper we proposed the approach to automatic gapping resolution in Russian. Looking at the results it is clear that:

1. RNN-based approach achieves reasonable performance on binary classification task.
2. Increasing the number of parameters in the Universal Transformer did not improve the results. One explanation is that fixed word embedding like fastText does not contain enough syntactic information. Probably, contextual representations such as ELMo [Peters et al., 2018] or BERT [Devlin et al., 2018] would perform better.
3. Remnants matching is the most challenging task for the model.
4. Our model is not capable to deal with sentences with gapping with one remnant, possibly in part owing to the small number of such examples presented in the corpus.
5. Our model generalize better if punctuation marks are removed. Unfortunately, it did not improve the overall performance since the fastText model was pretrained on texts that contain punctuation marks.

Acknowledgements

This work is supported by the Russian Science Foundation under grant 16-18-10422.

References

1. *Bahdanau, D. et al.:* Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473. (2014).
2. *Bogdanov, A.:* Description of gapping in a system of automatic translation. In: Computational linguistics and intellectual technologies: Papers from the annual conference “dialogue”. (2012).
3. *Boguslavsky, I. et al.:* Development of a dependency treebank for russian and its possible applications in nlp. In: Proceedings of the third international conference on language resources and evaluation (lrec-2002). pp. 852–856, Las Palmas (2002).
4. *Bojanowski, P. et al.:* Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics. 5, 135–146 (2017).
5. *Cho, K. et al.:* Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078. (2014).
6. *Dehghani, M. et al.:* Universal transformers. arXiv preprint arXiv:1807.03819. (2018).
7. *Devlin, J. et al.:* BERT: pre-training of deep bidirectional transformers for language understanding. CoRR. abs/1810.04805, (2018).
8. *Droganova, K. et al.:* Mind the gap: Data enrichment in dependency parsing of elliptical constructions. In: Proceedings of the second workshop on universal dependencies (udw 2018). pp. 47–54 (2018).
9. *Droganova, K., Zeman, D.:* Elliptic constructions: Spotting patterns in ud treebanks. In: Proceedings of the nodalida 2017 workshop on universal dependencies (udw 2017). pp. 48–57 (2017).
10. *Gal, Y., Ghahramani, Z.:* A theoretically grounded application of dropout in recurrent neural networks. In: Advances in neural information processing systems. pp. 1019–1027 (2016).
11. *Ge, R., Mooney, R. J.:* Learning a compositional semantic parser using an existing syntactic parser. In: 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: Volume 2-volume 2. pp. 611–619 Association for Computational Linguistics (2009).
12. *Grave, E. et al.:* Learning word vectors for 157 languages. In: Proceedings of the international conference on language resources and evaluation (lrec 2018). (2018).
13. *Johnson, K.:* Gapping, (2014).
14. *Kingma, D. P., Ba, J.:* Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. (2014).
15. *Loper, E., Bird, S.:* NLTK: The natural language toolkit. arXiv preprint cs/0205028. (2002).
16. *Matthew Lamm, D. J., Arun Chaganty, Liang, P.:* QSRL: A semantic role-labeling schema for quantitative facts. In: El-Haj, M. et al. (eds.) Proceedings of the eleventh international conference on language resources and evaluation (lrec 2018). European Language Resources Association (ELRA), Paris, France.
17. *Mikolov, T. et al.:* Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013).

18. *Peters, M. E. et al.:* Deep contextualized word representations. arXiv preprint arXiv:1802.05365. (2018).
19. *Ross, J. R.:* Gapping and the order of constituents. Progress in linguistics: A collection of papers. 43, 249–259 (1970).
20. *Schuster, S. et al.:* Sentences with gapping: Parsing and reconstructing elided predicates. arXiv preprint arXiv:1804.06922. (2018).
21. *Smith, S. L. et al.:* Don't decay the learning rate, increase the batch size. arXiv preprint arXiv:1711.00489. (2017).
22. *Vaswani, A. et al.:* Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017).
23. *Wu, Y. et al.:* Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144. (2016).
24. *Yin, W. et al.:* Comparative study of cnn and rnn for natural language processing. arXiv preprint arXiv:1702.01923. (2017).
25. *Young, T. et al.:* Recent trends in deep learning based natural language processing. IEEE Computational Intelligence Magazine. 13, 3, 55–75 (2018).

СЛОВО *ЭТО* В ЧАСТНОМ ВОПРОСЕ: О ПРИЗНАКАХ, ОТЛИЧАЮЩИХ ЧАСТИЦУ ОТ МЕСТОИМЕНИЯ[1]

Пекелис О. Е. (opekelis@gmail.com)
Российский государственный гуманитарный университет, Москва, Россия

THE WORD *ÈTO* IN A *WH*-QUESTION: ON THE DIFFERENCES BETWEEN A PRONOUN AND A PARTICLE

Pekelis O. E. (opekelis@gmail.com)
Russian State University for the Humanities, Moscow, Russia

The paper examines the grammatical and semantic features of the word *èto* when it precedes or follows a *wh*-word (cf. *Gde èto ty byl?*). In this context, *èto* is usually considered to be a particle, with the only—and not clear-cut—exception being a question with the *wh*-words *kto* and *čto*. However, the data presented below suggest that as many as four different types of *èto* used in an interrogative context have to be distinguished. It is demonstrated that these types differ in their meaning, their syntactic distribution, and their position within the "pronoun-particle" continuum.

Key words: particle, pronoun, *wh*-question, corpus, information structure

1. Введение

Слово *это,* употребленное в препозиции (1) или постпозиции (2) к вопросительному слову, обычно характеризуется как частица [Падучева 1982: 77][2].

(1) *Это кто же тебя обучил таким полетам?* [НКРЯ]

(2) *Почему это мальчики всегда ссорятся?* [НКРЯ]

[1] Работа выполнена при поддержке гранта РГНФ № 17-04-00517(а).

[2] Примеры с указанием источника здесь и далее, если не сказано иное, заимствованы из Национального корпуса русского языка [НКРЯ].

Исключение, согласно [Ibid.: 83], составляют некоторые конструкции с вопросительными словами *кто* и *что*, как в (3), однако их отличие от конструкций с *кто* и *что*, в которых *это* признается частицей (ср. (1)), остается не вполне понятным.

(3) *Машенька, это кто пришел?* [НКРЯ]

Казалось бы, частица должна отличаться от местоимения тем, что только последнее получает статус члена предложения. Однако отсутствие этого статуса, в действительности, не обязательно является признаком частицы. Так, в одном из утвердительных контекстов употребление *это* признается местоименным несмотря на то, что *это* не соответствует члену предложения [Ibid.: 82].

В настоящей работе предпринята попытка выявить семантические и синтаксические признаки, по которым различаются местоимение и частица *это* в составе частного вопроса. Мы надеемся показать, что такие признаки требуют различать четыре разновидности *это*, занимающих разные позиции на шкале «частица-местоимение».

Статья имеет следующую структуру. В **разделе 2** выявляется регулярное семантическое различие между препозитивным и постпозитивным *это*, не являющимся членом предложения (далее *это*$_{\text{нечп}}$). В **разделе 3** демонстрируются, а в **разделе 4** — анализируются синтаксические свойства *это*$_{\text{нечп}}$ в контексте вопросительного слова *кто* (поскольку именно такое *это*, по наблюдению Е. В. Падучевой, может обладать особым статусом). В **разделе 5** выделенные разновидности *это* ранжируются на шкале «частица-местоимение».

2. Препозитивное и постпозитивное *это*$_{\text{нечп}}$: семантические различия

В настоящем разделе предлагается семантическая интерпретация препозитивного и постпозитивного *это*$_{\text{нечп}}$ (**раздел 2.1**), которая затем уточняется на материале косвенного вопроса (**раздел 2.2**).

2.1. Предварительная семантическая интерпретация

Препозитивное и постпозитивное *это*$_{\text{нечп}}$, как кажется, представляют собой разные частицы с разным семантическим вкладом. (Исключение составляют некоторые употребления *это*$_{\text{нечп}}$ при *кто* и, возможно, *что*, о которых речь пойдет в **разделах 3** и **4**). Назначение **препозитивного *это*$_{\text{нечп}}$** предположительно состоит в том, чтобы сигнализировать адресату, что задаваемый вопрос касается положения дел, поступившего в фонд знаний говорящего в предтексте. Семантический вклад **постпозитивного *это*$_{\text{нечп}}$** состоит в выражении недоумения, порожденного положением дел, которое в предтексте поступило в фонд знания говорящего, а также в маркировании вопроса как проявления этого недоумения.

В соответствии с этой интерпретацией, тонкое различие между предложениями (4а), (4б), (4в) состоит в следующем. В (4а) препозитивное *это*$_{\text{нечп}}$ эксплицитно указывает на то, что основанием для вопроса служит информация, активированная в предтексте (‘Шестак нас держит’). В (4б), в отличие от (4а), выражено значение недоумения, порожденного той же информацией. Наконец, в (4в), где *это* нет, не выражено ни то, ни другое значение.

(4) а. *<— Как везде, — вставил Валерий, чтобы быть в беседе. — Ждут, козлы, пока местные богодулы сами дёрнут отсюдова или сдохнут, а Шестак нас держит. Всю деревню спасает.> — Это как он вас держит? — Кирилл повернулся к Сане.* [НКРЯ]
б. *<А Шестак нас держит. Всю деревню спасает.> — Как это он вас держит? — Кирилл повернулся к Сане.*
в. *<А Шестак нас держит. Всю деревню спасает.> — Как он вас держит? — Кирилл повернулся к Сане.*

Приведем еще две иллюстрации.

В примере (5а), с препозитивным *это*, целью вопроса является получение информации о местонахождении. В примере (5б), с постпозитивным *это*, выражено недоумение в связи с местонахождением, породившее вопрос. (Заметим, что в (5а), (5б) *это* отсылает не к предтексту, а к тому, что находится перед глазами говорящего, т. е. употреблено скорее дейктически, чем анафорически.) То же различие — только запрос информации в (а) и недоумение в (б) — видится между вопросами в (6).

(5) а. *— Что, брат Триша, — проговорил Прохор Васильевич, осматриваясь, — это где мы? — У меня на фатере.* [НКРЯ]
б. *Слушай, где это мы? Как нас сюда занесло?* [НКРЯ]

(6) а. *Это кому крышу кроют?*
б. *И по железу кровельщики гремят, споро-споро… кому это крышу кроют?* [НКРЯ]

Важно отметить, что не любой вопрос может быть маркирован препозитивным *это*: истинность положения дел, к которому отсылает такое *это*, должна быть эксплицитно заявлена в предтексте (или в ситуации речи, ср. (5а))[3]. Так, препозитивное *это* в составе вопроса в (7а) (пример рецензента «Диалога») неуместно, потому что в предтексте не содержится утверждения ‘мне понравилась Казань’ (при том что вопрос без *это* допустим). Однако такое *это* приемлемо в составе (7б), поскольку вопрос опирается на утверждение, принадлежащее предтексту.

(7) а. *— Я вчера вернулся из Казани. — (*Это) Как тебе там понравилось?*
б. *— Мне очень понравилась Казань. Я вчера оттуда вернулся. — Это чем она тебе понравилась?*

[3] Мы признательны рецензенту «Диалога», обратившему наше внимание на необходимость этого уточнения.

Постпозитивное *это*$_{\text{нечп}}$ не требует столь же эксплицитного указания на истинность соответствующего положения дел, какого требует препозитивное *это*$_{\text{нечп}}$. Этим объясняется, по-видимому, допустимость постпозитивного *это* в (8а) и (9а) и неуместность препозитивного *это* в (8б) и (9б). Истинность положений дел ‘ты мне тыкаешь’ и ‘я забыл, что Митей звали второго сына’ следует из предтекста (в (8а) — следует не только из того, что адресат вопроса обращается к автору вопроса на «ты», но и из того, что он разговаривает фамильярно), однако не выражается в явном виде. Ср. (8в) с препозитивным *это*$_{\text{нечп}}$, где истинность положения дел, вызвавшего вопрос (‘он — неудачник’), эксплицитно утверждается[4].

(8) а. *Какого *** ты притащила ее сюда? <…> И тут я ему говорю: «А почему это ты всё время мне тыкаешь?»* [НКРЯ]
б. <…> *?И тут я ему говорю: «Это почему ты всё время мне тыкаешь?»*
в. *И вместе с тем он — неудачник! — А это почему ты так решил?* [НКРЯ]

(9) а. *— Значит, Митей звали второго сына! Как это я мог забыть?!* [НКРЯ]
б. <…> *?Это как я мог забыть?*

Вместе с тем, если вопрос касается положения дел, сведения об истинности которого даже в качестве импликации не поступили в фонд знаний говорящего, постпозитивное *это* не используется так же, как и препозитивное. Так, постпозитивное *это* уместно в составе вопроса в (7б), но не в (7а).

Таким образом, хотя и препозитивное, и постпозитивное *это*$_{\text{нечп}}$ предполагают связь вопроса с предтекстом, препозитивное *это*$_{\text{нечп}}$ требует более наглядной манифестации этой связи — посредством эксплицитного указания на истинность положения дел, стимулировавшего вопрос[5].

---

[4] Граница между имплицитным и эксплицитным выражением того или иного положения дел не является четкой, и ее определение представляет собой отдельный вопрос. Так, препозитивное *это*$_{\text{нечп}}$ уместно в составе вопроса в (i), но не в (ii). По нашему предположению, это связано с тем, что указание на истинность положения дел ‘у королевы есть дела’ содержится в предтексте в (i) в более явном виде, чем указание на истинность положения дел ‘у курицы есть мозги’ содержится в предтексте в (ii).

(i) *Однажды утром наша королева пошла по своим делам. — А какие <*$^{\text{ок}}$*это какие> у королевы дела?* [НКРЯ]

(ii) *— Да, куриными мозгами ты пошла в мать, — протянула бабушка. — А какие <*$^{??}$*это какие> у курицы мозги?* [НКРЯ]

[5] Исключение составляет некомпозициональное сочетание с вопросительным словом *какой*, выражающее риторический вопрос: в (i) стимулировавшее вопрос положение дел само представляет собой вопрос и, значит, не утверждается как истинное.

(i) *— Виктор, вы водку пьете? <…> — Это какой же мужик водку не пьет? — весело спросил супругу полковник и подмигнул Терентьеву.* [НКРЯ]

При композициональном употреблении *какой* в вопросе с препозитивным *это* условие истинности соблюдается. Так, в (ii) вопрос с *какой* не является риторическим, а стимулировавшее его положение дел (‘у тебя есть свидетели’) заявлено в предтексте как истинное:

(ii) *У меня, ваше благородие, свидетели есть. Они могут подтвердить, что я правду говорю. — Это какие же у тебя свидетели, откуда? — спрашивает следователь.* [НКРЯ]

2.2. *Это*$_{\text{нечп}}$ в косвенном вопросе

Употребление препозитивного *это*$_{\text{нечп}}$ затруднено в косвенном вопросе. Ср. сомнительный пример (10б), полученный из (10а) переносом *это*$_{\text{нечп}}$ в препозицию:

(10) а. *Вскоре его вызвал президент Ельцин и строго спросил, почему это преступность в стране возросла на 20%.* [НКРЯ]
б. ??*Вскоре его вызвал президент Ельцин и строго спросил, это почему преступность в стране возросла на 20%.*

В НКРЯ отсутствуют примеры с матричными глаголами *знать, понимать, понять, спрашивать, спросить, недоумевать* (свободно присоединяющими придаточное косвенного вопроса) и препозитивным *это*$_{\text{нечп}}$ при вопросительных словах *как, почему, зачем, когда*. Постпозитивное *это*$_{\text{нечп}}$ в том же контексте представлено по меньшей мере 65 вхождениями.

Закономерным образом, запрет на употребление препозитивного *это*$_{\text{нечп}}$ в косвенном вопросе ослабляется, если косвенный вопрос сближается с прямым. Так, в (11) близость к прямому вопросу обеспечивается совпадением субъекта вопроса с говорящим и выражается пунктуационно — вопросительным знаком.

(11) *Виват, брат, — прервал тут Гоголь, — но я не совсем в толк взял, это кто же пляшет?* [НКРЯ]

Приведенные данные позволяют уточнить семантическое различие между препозитивным и постпозитивным *это*$_{\text{нечп}}$, сформулированное в **разделе 2.1**. Препозитивное *это*$_{\text{нечп}}$ сигнализирует о том, что вопрос касается положения дел, поступившего в фонд знаний говорящего в предтексте. Однако переход от прямого вопроса к косвенному обозначает переход от диалогического режима интерпретации к гипотаксическому (в терминологии [Падучева 2009]). При этом предметом вопроса может оказаться то, что поступило в фонд знаний субъекта главной клаузы (*президент Ельцин* в (10а)), но не то, что узнал говорящий. Тот факт, что в косвенном вопросе препозитивное *это*$_{\text{нечп}}$ затруднено, позволяет заключить, что препозитивное *это*$_{\text{нечп}}$ всегда маркирует вопрос о происходящем между говорящим и адресатом, являясь диалогическим показателем.

Напротив, недоумение и порожденный им вопрос, которые, по нашему предположению, маркирует постпозитивное *это*$_{\text{нечп}}$, интерпретируемы и в условиях гипотаксической проекции: они могут касаться не только того, что произошло между говорящим и адресатом, но и того, что произошло между участниками ситуации, обозначенной главной клаузой.

Отметим, что препозитивное *это* затруднено в косвенном вопросе прежде всего в том случае, если оно не является членом предложения. Препозитивное *это* в позиции члена предложения (далее *это*$_{\text{чп}}$) в косвенном вопросе допустимо. В НКРЯ все вхождения препозитивного *это* в контексте бесспорного

косвенного вопроса (т. е. такого, который не сближен с прямым) характеризуются тем, что *это* соответствует в них члену предложения[6]. Ср.:

(12) *Знаешь, это кто? Борода, десятник из Злых Щелей!* [НКРЯ]

(13) *Знаете, это как называется? Дво-е-же-нец!* [НКРЯ]

(14) *Я замер. Я знал, это чье.* [НКРЯ]

Правда, и в этом случае препозиция *это* допускается ограниченно, поскольку является результатом инверсии, которая, как известно, избегает гипотаксического контекста [Green 1976 и др.][7]. Однако ограничение на употребление в косвенном вопросе препозитивной *это*-частицы представляется более весомым, ср. (12)–(14) с (10б)[8].

3. *Это*$_{\text{нечп}}$ в контексте *кто*: синтаксические свойства

Конструкция с препозитивным *это*$_{\text{нечп}}$ и вопросительным местоимением *кто* чувствительна к степени активации невопросительных актантов (т. е. актантов, не выраженных вопросительным словом). Именно так представляется уместным интерпретировать следующий факт: в случае, если невопросительные актанты выражены местоимениями, конструкция с препозитивным *это*$_{\text{нечп}}$ может быть более приемлема, чем если они выражены полными именными группами. На уместность конструкции с постпозитивным *это*$_{\text{нечп}}$ этот фактор не влияет. Ср. примеры (15)–(17), иллюстрирующие эту закономерность.

(15) а. *Это кто вам такое рассказал?*
б. ?*Это кто вам рассказал правду?*
в. *Кто это вам рассказал правду?*

(16) а. *Это кого он ему представил?*
б. ?*Это кого Вася представил Пете?*
в. *Кого это Вася представил Пете?*

---

[6] Хотя к членам предложения относят и актанты (дополнения), и сирконстанты (обстоятельства), фактически словоформа *это* (форма именительного или винительного падежа единственного числа местоимения *это*) в вопросительной конструкции может соответствовать, по-видимому, только актанту.

[7] Мы благодарны рецензенту «Диалога» за это наблюдение.

[8] Отдельного рассмотрения заслуживает вопрос о том, употребляется ли в косвенном вопросе такое препозитивное *это*$_{\text{нечп}}$, которое, как мы надеемся показать в **разделе 3**, является местоимением несмотря на то, что не соответствует члену предложения. Речь идет об употреблении *это* при вопросительном слове *кто* и одноактантном глаголе. Сравнение (11) с (i) как будто показывает, что в контексте косвенного вопроса, не сближенного с прямым, такое *это* в препозиции существенно затруднено:

??*Я знал, это кто пляшет.*

Вместе с тем, приведенный материал не позволяет делать уверенных выводов о сравнительной приемлемости в косвенном вопросе препозитивного местоименного *это*$_{\text{нечп}}$ и *это*$_{\text{чп}}$ (ср. (12)–(14)).

(17) а. *Это кому ты такое подарил?*
б. ?*Это кому Вася подарил книгу?*
в. *Кому это Вася подарил книгу?*

Вне этого обобщения тривиальным образом остаются конструкции с одним актантом, поскольку невопросительных актантов в этом случае не оказывается. Так, в конструкции с *кто* в позиции подлежащего при одноактантном глаголе (18) позиция *это* не влияет на приемлемость[9]:

(18) а. — *Это кто звонил?* [НКРЯ]
б. — *Кто это звонил?* [НКРЯ]

Однако одноактантные конструкции отличаются от конструкций с двумя и более актантами и по другому формальному признаку, носящему, как кажется, уже не случайный характер: только в одноактантных конструкциях *это*$_{\text{нечп}}$ сочетается с частицей *еще*. Ср. примеры с постпозитивным *это* в (19) и препозитивным *это* в (20):

(19) а. *Кто это еще командует?* [НКРЯ]
б. *Кто это еще приехал?* [НКРЯ]

(20) а. *А это еще кто во дворе появился?* [НКРЯ]
б. *А это еще кто тут сидит, свернувшись в клубок?* [НКРЯ]

Частица *еще* в таких примерах семантически взаимодействует со словом *это*. На это указывает тот факт, что элиминация *это* ведет к изменению значения *еще*: из экспрессивно окрашенного оно становится нейтральным («аддитивным», в терминологии [Богуславский 1996: 259]). Ср. (19а), (19б) с вариантами без *это*:

(21) а. *Кто еще командует?*
б. *Кто еще приехал?*

В (21), по сравнению с (19)–(20), меняется и сфера действия *еще*. В (21б) *еще* привносит презумпцию ‘наряду с тем, о ком спрашивает говорящий, приехал кто-то другой’, т. е. аддитивная семантика *еще* распространяет свою сферу действия на подлежащее. В (19)–(20), между тем, в сферу действия *еще* входит целая ситуация. Так, (19б) выражает идею о том, что наряду с ситуацией ‘кто-то приехал’ имеет место другая ситуация — как правило, неприятная для говорящего (но не выражает идеи о том, что приехал кто-то другой). Причиной этого семантического различия видится то, что в (21) *это* модифицирует

[9] Одноактантным можно отнести также конструкции с *кто* в позиции дополнения при безличном глаголе, ср.:

(i) *Это кого убило-то?* [НКРЯ]

(ii) *Кого это принесло?!* [НКРЯ]

По-видимому, *это* в составе таких конструкций имеет те же свойства и тот же статус, что и *это* при *кто*-подлежащем и одноактантном предикате. Рамки статьи не позволяют проиллюстрировать это предположение.

подлежащее, тогда как в (19) и (20) посредством *еще* модифицируется само слово *это* (см. подробнее **раздел 4**).

Между тем, в конструкциях с двух- и трехактантными предикатами частица *еще* при слове *это* кажется мало уместной. Так, приемлемость *еще* при слове *это* вызывает сомнения в примерах (22) и (23), различающихся линейной позицией *это* (варианты (а), (б) и (в) различаются синтаксической позицией вопросительного слова — подлежащее, прямое и косвенное дополнение, соответственно).

(22) а. *Это (?еще) кто вам говорил про кошек?* [НКРЯ]
б. *Это (?еще) кого ты к нам такого хорошего привел?* [НКРЯ]
в. *Это (?еще) кому он знак подает?* [НКРЯ]

(23) а. *И кто это (?еще) его царапнул по щеке?* [НКРЯ]
б. *Кого это (?еще) она зовёт?* [НКРЯ]
в. *Кому это (?еще) он мигает?* [НКРЯ]

Данные НКРЯ согласуются с этой интуицией: примеры типа (22) и (23) — с *еще* при *это*$_{\text{нечп}}$ в составе вопросительной конструкции с *кто* и более, чем одним актантом — в корпусе отсутствуют[10].

Следует отметить, что, помимо одноактантной конструкции, *это* сочетается с *еще* в контексте типа (24):

(24) а. — *Это еще кто такая?* [НКРЯ]
б. *Взбесились, да и только! Кто это еще?!* [НКРЯ]

В этом случае, однако, *это* является членом предложения, выступая в качестве компонента биноминативного предложения, и, значит, имеет местоименный статус [Падучева 2016]. Таким образом, с точки зрения сочетаемости с частицей *еще* препозитивное и постпозитивное *это*$_{\text{нечп}}$ при одноактантном предикате ведут себя как *это*-местоимение. Мы вернемся к этому выводу в **разделе 4**.

4. *Это*$_{\text{нечп}}$ в контексте *кто*: объяснение синтаксических свойств

Будем исходить из того, что только местоимение, но не частица *это* может присоединять частицу *еще* (похожим образом, в [Падучева 1982: 77] ударность *это* рассматривается как признак местоименности). В таком случае в контексте *кто* с местоимением сближается **препозитивное и постпозитивное *это*$_{\text{нечп}}$**

[10] Граница между одноактантными и двухактантными конструкциями с точки зрения приемлемости *еще* не является четкой, ср. пример рецензента «Диалога»: *Это кто еще нами тут командует?* Этот факт видится следствием того, что и особый статус *это* в одноактантных конструкциях не ограничен строго этими конструкциями, но может в специальных условиях возникать у *это* в двухактантных конструкциях. См. подробнее **раздел 4** и в особенности **сноску 15**.

при одноактантном предикате (поскольку именно такое *это*, как показано в **разделе 3**, присоединяет *еще*).

Если это верно, чем может объясняться такое распределение разных употреблений *это*$_{\text{нечп}}$?

Наша основная гипотеза состоит в том, что *это*$_{\text{нечп}}$ получает статус местоимения в тех и только тех вопросительных контекстах, которым можно сопоставить утвердительные контексты с *это*-местоимением. Покажем, что так оно и есть: среди вопросов с *кто* только вопросы с одноактантным предикатом могут считаться вопросительными «аналогами» утвердительной конструкции с *это*-местоимением.

На роль утвердительного аналога вопроса с *это*$_{\text{нечп}}$ и *кто* a priori могут претендовать всего два употребления *это*: «выделительное» *это* (25) и *это* в роли «квази-подлежащего» (26) (в терминологии [Падучева 1982], [Падучева 2016])[11]. Термин «квази-подлежащее» призван отразить тот факт, что такое *это* выступает аналогом подлежащего при «квази-сказуемом», выраженном двусоставным предложением (ср. *венецианцы погуляли* в (26)).

(25) *Раньше я думал, что это я тебя учу, а оказалось наоборот.* [НКРЯ]

(26) *Обрати внимание, у многих местных — голубые глаза. Это венецианцы погуляли.* [НКРЯ]

Эти употребления различаются, во-первых, грамматическим статусом *это*: выделительное *это* признается частицей, а квази-подлежащее — местоимением [Ibid.]. Во-вторых, два употребления различаются коммуникативной структурой. Выделительное *это* присоединяется к контрастной реме, ср. *это я*\\ *тебя учу* в (25) [Markman 2008]; [Kimmelman 2009 и др.][12]. Конструкция с квази-подлежащим обычно признается тетической, ср. (27а) [Junghanns 1997]; [Kimmelman 2009]. Однако, как кажется, альтернативой тетической структуре является категорическая структура с нейтральным порядком коммуникативных компонентов (27б).

(27) а. *Каждый почти вечер видно зарево далеких пожаров: это турки жгут болгарские деревни*\. (Пример из [Падучева 1982: 82])
б. *Каждый почти вечер видно зарево далеких пожаров: это турки/ жгут болгарские деревни*\.

Покажем, в соответствии с поставленной задачей, что вопрос с препозитивным и постпозитивным *это*$_{\text{нечп}}$ при *кто* и одноактантном предикате может считаться вопросительным аналогом утвердительной конструкции с квази-подлежащим. Напротив, остальные типы вопроса с *это*$_{\text{нечп}}$ и *кто* не допускают такой интерпретации.

---

[11] В [Kimmelman 2009] эти употребления *это* называются “focus èto-cleft” и “thetic èto-cleft”, соответственно.

[12] Здесь и далее мы помечаем акцентоносители знаками «\», «/», «\\», «//», в целом следуя нотации [Янко 2001: 36–37].

Рассмотрим примеры вопроса с *это* в составе одноактантной конструкции (28), в составе двухактантной конструкции с *кто* в позиции подлежащего (29) и двухактантной конструкции с *кто* в позиции дополнения (30).

(28) *Это кто\ пришел? <Кто\ это пришел?>*

(29) *Это кто\ варит суп? <Кто\ это варит суп?>*

(30) *Это кого\ Петя побил? <Кого\ это Петя побил?>*

Только вопрос (28) имеет формальные и содержательные признаки конструкции с *это* в роли квази-подлежащего.

Формальное сходство обеспечивается тем, что в одноактантной конструкции с *это* в роли квази-подлежащего акцентоносителем всегда выступает подлежащее — так же, как и в вопросительной одноактантной конструкции[13]:

(31) *<Послышалось звяканье входной двери.> Это Петя\ пришел.*

Содержательное сходство проявляется в том, что реплики (28) и (31) допустимы в одном и том же контексте. Так, на звяканье входной двери можно отреагировать вопросом *Это кто\ пришел*? (или: *Кто\ это пришел?*), а можно — сообщением *Это пришел Петя* (или: *Это Петя\ пришел*). Это свидетельствует о параллелизме коммуникативных структур утвердительной и вопросительной конструкций: вторая, подобно первой, должна интерпретироваться как тетическая. В самом деле, (28) — это вопрос к событию в целом (в терминологии [Янко 2001]); последовательность *кто пришел* в его составе вся соответствует неизвестному вопроса. Показательно, что *это* не сочетается с вопросом *Что случилось?*, который сам по себе является тетическим [Ibid.: 131]. Ср. (32а) и (32б):

(32) а. ?*Это что случилось?*
б. ОК*Это что упало?*

В (32а) слово *это* маркирует вопрос как относящийся к событию в целом, поэтому оно здесь избыточно. Между тем в (32б) избыточности нет, поскольку вопрос без *это* (*Что упало?*) не является вопросом к событию в целом[14].

С тетичностью вопросительных одноактантных конструкций связано и поведение частицы *еще*, обсуждавшееся в **разделе 3**, а именно, тот факт, что в сфере действия *еще* в контексте *это* оказывается целая ситуация. В самом деле, в вопросе без *это* (*Кто еще пришел?*) частица *еще* модифицирует подлежащее, а сказуемое остается вне сферы действия аддитивной семантики *еще* (ср. порождаемую *еще* презумпцию: ‘наряду с тем, о ком спрашивает говорящий,

[13] В вопросе с вопросительным словом и в утвердительном предложении используются разные типы понижающихся акцентов — ИК-2 и ИК-1 соответственно (по Е. А. Брызгуновой [Шведова и др. 1980: 97]). В нашей упрощенной нотации это различие не отражено.

[14] Несколько бóльшая приемлемость постпозитивного *это* в этом контексте (ср. *Что это случилось?*) может объясняться тем, что здесь оказывается востребована постпозитивная *это*-частица, выражающая недоумение (см. раздел 2).

пришел кто-то другой', но не 'кто-то сделал что-то другое'). Между тем в вопросе с *это* (*Это кто еще пришел*?) местоимение *это* как раз оказывается той именной группой («квази-подлежащим»), которую модифицирует частица *еще*, а собственно подлежащее (*кто*) входит в единую коммуникативную составляющую («квази-сказуемое») с глаголом. Поскольку *это* отсылает к ситуации (дейктически или анафорически), то и сферой действия *еще* оказывается ситуация.

В отличие от одноактантной вопросительной конструкции, вопросы (29) и (30) не имеют сходства с утвердительной конструкцией, содержащей *это* в роли квази-подлежащего.

С формальной точки зрения, в вопросе с *кто* в позиции подлежащего (29) акцентоносителем выступает подлежащее. Между тем в утвердительной конструкции с квази-подлежащим акцентоносителем должно стать дополнение при тетической структуре (33а) (см. о правилах выбора акцентоносителя в [Янко 2001: 188 ff.]); при категорической структуре (33б) в конструкции с квази-подлежащим два акцентоносителя.

(33) а. *С кухни доносятся странные звуки. Это Петя варит суп*\.
б. *С кухни доносятся странные звуки. Это Петя/ варит суп*\.

В вопросе с *кто* в позиции прямого дополнения (30) акцентоносителем выступает прямое дополнение. Это отвечает акцентному оформлению конструкции с квази-подлежащим (ср. (33а)). Однако в вопросе отсутствует возможность альтернативного оформления по образцу категорической структуры (ср. (33б)).

Содержательное различие между двухактантной вопросительной конструкцией и утвердительной конструкцией с квази-подлежащим состоит в том, что первая, в отличие от второй, обычно не может интерпретироваться как тетическое предложение, т. е. как вопрос к событию в целом. Так, доносящийся с кухни грохот уместно прокомментировать утвердительным высказыванием с квази-подлежащим *Это Петя Васю*\ *побил*, и неуместно — вопросом: *Это кого*\ *Петя побил?*, потому что *Петя побил* в нем составит известное вопроса, что не отвечает требованиям контекста. Вероятно, дело в том, что чем больше актантов, тем проблематичнее их объединение в нечленимую коммуникативную структуру, учитывая что обычно частный вопрос характеризуется расчлененной коммуникативной структурой[15].

Итак, только одноактантная вопросительная конструкция с *кто* обнаруживает формальное и содержательное сходство с утвердительной конструкцией,

[15] Сближение двухактантной утвердительной и вопросительной конструкций происходит при условии, что информация, соответствующая в (30) известному вопроса, активирована в предтексте. Ср. уместность утверждения *А это// Петя Васю*\\ *побил* и вопроса *А это*\/ *кого*\\ *Петя побил?* в ответ на доносящийся из кухни грохот в случае, когда известно, что Петя раньше уже кого-то побил. Слово *это* здесь сближается с местоимением (ср. его ударность) несмотря на наличие второго актанта. Для сравнения, в трехактантной конструкции *это*, по-видимому, не бывает местоимением ни при каких условиях. Ср. сомнительность ударного *это* в ?*А это*\/ *кто*\\ *тебе ее подарил?*

содержащей *это* в роли квази-подлежащего. Это подтверждает наше предположение о том, что только в такой конструкции *это* сближается с местоимением (но ср. **сноску 15**), и согласуется с данными о сочетаемости *это* с *еще* в контексте *кто*.

5. *Это* на шкале «частица-местоимение»: заключение

Рассмотренные данные требуют различать четыре разновидности слова *это* в частном вопросе.

Во-первых, различаются препозитивная и постпозитивная частица. Можно думать при этом, что на шкале «частица-местоимение» препозитивная частица ближе к полюсу «местоимение», чем постпозитивная. На это указывают наличие у постпозитивной частицы прагматического компонента значения (‘недоумение’) и более выраженная анафоричность в значении препозитивной частицы: такая частица, как мы стремились показать, в большей степени зависима от характеристик предтекста (ср. требование эксплицитного указания на истинность положения дел, стимулировавшего вопрос, **раздел 2.1**; чувствительность к активации невопросительных актантов, **раздел 3**).

Во-вторых, в вопросе со словом *кто* при одноактантном предикате следует выделять местоименное *это*$_{\text{нечп}}$, функционально близкое слову *это* в роли квази-подлежащего. Такое *это* отличается по свойствам и от *это*$_{\text{чп}}$, и от *это*$_{\text{нечп}}$ в роли частицы. От *это*$_{\text{чп}}$ местоименное *это*$_{\text{нечп}}$ тривиальным образом отличается тем, что не соответствует члену предложения (другое возможное отличие, касающееся употребления *это*$_{\text{чп}}$ и местоименного *это*$_{\text{нечп}}$ в косвенном вопросе, не было установлено достоверно, см. **раздел 2.2**). От *это*$_{\text{нечп}}$ в роли частицы местоименное *это*$_{\text{нечп}}$ отличается тем, что только последнее обладает формальным и содержательным сходством с *это* в роли квази-подлежащего и сочетается с частицей *еще*.

Наконец, в-третьих, отдельную единицу представляет собой *это*$_{\text{чп}}$.

Выделенные разновидности слова *это* могут быть следующим образом ранжированы по степени убывания местоименности:

(34) *это*$_{\text{чп}}$ > *это*$_{\text{нечп}}$ при одноактантном предикате > препозитивное *это*$_{\text{нечп}}$ при двух- и трехактантном предикате > постпозитивное *это*$_{\text{нечп}}$ при двух- и трехактантном предикате.

Разновидности *это*, представленные в (34), реализуются при вопросительном слове *кто*. При вопросительных словах, не имеющих статуса актанта (*почему, когда, где* и др.), реализуются только три разновидности *это* — местоимение и две частицы:

(35) *это*$_{\text{чп}}$ > препозитивное *это*$_{\text{нечп}}$ > постпозитивное *это*$_{\text{нечп}}$.

Вопрос со словом *что* ведет себя, по-видимому, так же, как вопрос с *кто*. Обоснование этого предположения остается за рамками работы.

References

1. *Boguslavskij I. M.* (1996), Scope of lexical units [Sfera dejstvija leksicheskih edinits], Shkola "Jazyki Russkoj Kul'tury", Moscow.
2. *Green G. M.* (1976), Main Clause Phenomena in Subordinate Clauses, Language, 52(2), pp. 382–397.
3. *Janko T. E.* (2001), Communicative strategies of Russian speech [Kommunikativnye strategii russkoj rechi], Jazyki Slavjanskoj Kul'tury, Moscow.
4. *Junghanns U.* (1997), On the so-called eto-cleft construction, Annual Workshop on Slavic Linguistics, The Indiana meeting of 1996, pp. 166–190.
5. *Kimmelman V.* (2009), On the interpretation of èto in so-called èto-clefts, Studies in Formal Slavic Phonology, Morphology, Syntax, Semantics and Information Structure: Proceedings of FDSL 7, Leipzig 2007, pp. 319–329.
6. *Markman V. G.* (2008), Pronominal copula constructions are what? Reduced specificational pseudo-clefts, Proceedings of the 26th West Coast Conference on Formal Linguistics, Berkeley, pp. 366–374.
7. *Paducheva E. V.* (1982), The meaning and syntactic characteristics of the word ETO [Znachenie i sintaksicheskie svojstva slova ETO], Problems of Structural Linguistics 1980 [Problemy strukturnoj lingvistiki 1980], Nauka, Moscow, pp. 76–90.
8. *Paducheva E. V.* (2009), Modality through the prism of deixis [Modal'nost' skvoz' prizmu dejksisa], Articles of different years [Statji raznyh let], Jazyki Slavjanskih Kul'tur, Moscow, pp. 463–476.
9. *Paducheva E. V.* (2016), Demonstrative pronouns [Ukazatel'nye mestoimenija], Towards a corpus description of Russian grammar [Materialy dlja proekta korpusnogo opisanija russkoj grammatiki], available at: (http://rusgram.ru), Manuscript, Moscow.
10. *Shvedova N. Ju. et al.* (1980), Russian grammar: vol. 1 [Russkaja grammatika: t. 1], Nauka, Moscow.

TENSE AND LAX BODY PARTS IN THE RUSSIAN DEICTIC GESTURES: THE CASE OF INDEX FINGER POINTING

Pereverzeva S. I. (P_Sveta@hotmail.com)
National Research University Higher School of Economics, Moscow, Russia

The article regards the way in which the deictic gestures with the active index finger are executed in Russian body language and focuses on the role of the tension of the index finger (slightly curved vs. extended). Using the data retrieved from the Russian Multimedia Corpus, we discover the dependency between the tension of the index finger and the tension of the arm, which is engaged in executing the deictic gestures. We also reveal correlations between the tension of the index finger and (a) the primary / secondary reference to the pointed object, (b) the closest and the farthest distance between the speaker and the pointed object. We examine the difference in meaning and usage of the deictic gestures with the slightly curved vs. extended index finger. We argue that the choice between these types of pointing may be influenced both by physical and pragmatic factors.

Key words: deictic gestures, index finger, tension, physical factors, pragmatic, iconicity

НАПРЯЖЁННЫЕ И РАССЛАБЛЕННЫЕ ЧАСТИ ТЕЛА В РУССКИХ УКАЗАТЕЛЬНЫХ ЖЕСТАХ: УКАЗАНИЕ СЛЕГКА СОГНУТЫМ ПАЛЬЦЕМ

Переверзева С. И. (P_Sveta@hotmail.com)
Национальный исследовательский университет «Высшая школа экономики», Москва, Россия

Introduction

0.1. Previous works

The tension of body parts in the process of gesturing has been studied in various ways. Most attention has been paid to the role of hand tension in the segmentation of nonverbal units, particularly to discriminating fluidly connected gestures [Harling & Edwards 1997] and gesture phases [Wachsmuth & Kopp 2002]; [Bressem & Ladewig 2011]. [Povinelli & Davis 1994] compared the resting state of the index finger of humans and chimpanzees to account for the evolution of the pointing gesture. The general dichotomy 'tenseness'—'laxness' of bodily behavior was outlined by [Puppel 2018].

However, the studies which examined the correlation between the tension of body parts and semantic or pragmatic features of a pointing gesture are unknown to us, except for the ones carried out by E. A. Grishina on the basis of Russian deictic gestures. These studies were accumulated in the book [Grishina 2017]. Basing on the idea of compositionality of gestures, Grishina [ibid., 57–60] lists five parameters characterizing the form of the deictic gestures (formal parameters), two of which are discrete (i.e. with qualitative values) and three are gradual (i.e. with quantitative values).

Discrete parameters:

(a) **configuration** of the pointing hand, which takes the values "index finger" and "open hand";

(b) **orientation** of the hand, which takes the values "vertical", "supine" and "prone".

Gradual parameters:

(c) **hand tension** in open hand pointing, which takes the values "tense hand" (all fingers extended) and "lax hand" (fingers lax and slightly bent);

(d) **arm tension**, which takes the values "extended arm" and "half-bent arm";

(e) **formedness of the fingers' combination** in index-finger pointing, which takes the values "tight combination", "half-formed combination" and "loose combination"[1].

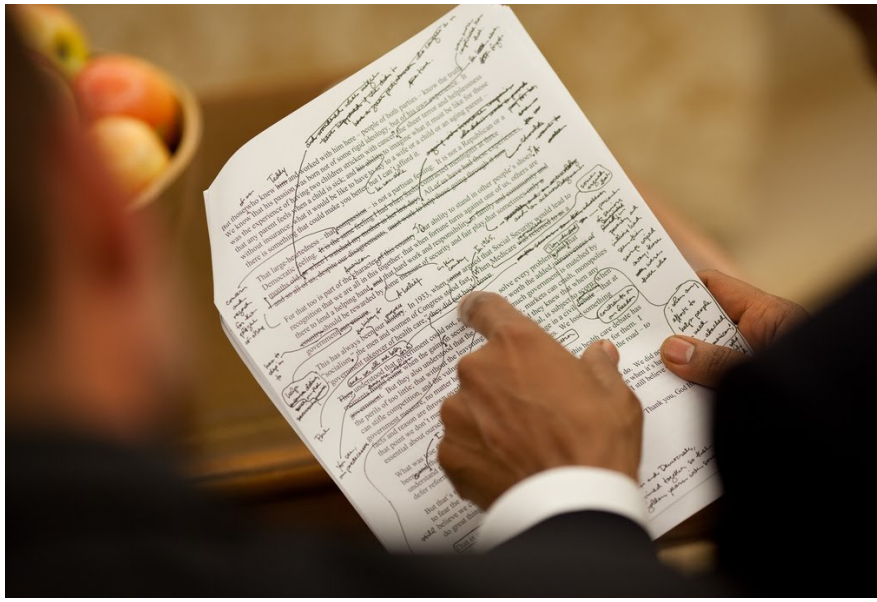
According to Grishina, each of these parameters has its own linguistic meaning, and the book is dedicated to discovering the most significant correlations between the physical parameters of a gesture and its meaning.

0.2. Object of the study

Accepting altogether the idea of gestures' compositionality and the list of parameters in question, we assume that one gradual parameter is still missing in this list. It is the **tension of the index finger** (IF), which can be thoroughly extended ("tense"), or slightly curved ("lax", ½ IF) in executing the deictic gestures.

[1] The parameters (a) and (c) are in complementary distribution, as they are innate to pointing with different body parts.

Deictic gestures with extended or ½ IF can be found both in everyday communication and in the movies (particularly in those which are presented in the Multimedia Russian Corpus—MURCO, http://www.ruscorpora.ru/search-murco.html). Also, the examples from fiction[2] show that the both Russian and English-speaking writers realize that the deictic gestures with ½ IF exist, as they directly name these gestures in their texts, cf. *Что можно посоветовать голодному двухлетнему человеку, когда он за словом в карман не лезет и тычет в материнское лицо* ***толстым полусогнутым пальцем*** ["What can one advise to a hungry two-years-old man, who never has to search for words and pokes his **thick half-bent finger** into his mother's face"] (K. Surikova); (2) *...Маленький этот старичок с мышиными глазками тыкал в собеседника* ***полусогнутым пальцем*** *так же, как и в молодости* ["...This little old man with mouse-like eyes poked **the half-bent finger** into the interlocutor just like he did when he was young"] (V. Likhonosov); (3) *I glanced at a map spread out on the desk in front of her and asked her what she was looking at. "I'm looking at these small islands in the middle of the sea." <...> Her* ***slightly bent index finger*** *pointed out a few yellow dots in the blue sea* (Xiaolu Guo). Cf. also the picture in the blog where Barak Obama points with his bent index finger at his notes:



Pic. 1: An example of the bent index finger: the act of pointing

The author of the blog comments upon it: "*I am completely enamored with that* ***slightly curved index finger***" (https://thealchemistskitchen.blogspot.com/2010/03/art-of-revision-president-obamas-recent.html).

[2] Examples in Russian are taken from the Russian National Corpus, http://www.ruscorpora.ru/; their translation into English is ours.

Surprising as it may seem, the deictic gesture with ½ IF, as far as we know, was left out by the researchers of Russian everyday gestures. This gesture has not even been regarded as a variant of a standard deictic gesture with an extended finger. In [Grishina 2017], where the significant configurations of the pointing hand are studied quite elaborately, it cannot be found in the set of manual deictic gestures [Grishina 2017: 55], nor is it mentioned in the works by G. E. Kreydlin—see, first of all, [Kreydlin 2002] and the conference papers [2007], [2008] dedicated to the deictic gestures of academic lecturers.

What has been said above is also largely true for the English studies of everyday gestures, cf. the set of English and Neapolitan deictic gestures in [Kendon 2004: 206] and a special collection of papers on pointing [Kita & Planck 2003]. Neither of these sources discusses the ½ IF in the phase of the stroke. The only exceptions may be the works on sign languages, see, e.g., the M. A. theses on the Egyptian sign language, where ½ IF is mentioned as a special variant of the gesture meaning "to understand" [Fan 2014: 21], or a thesis proposal on sign language gesture recognition, which mentions that the difference between the lexemes D, G, X and 1 in the American Sign Language lies in the degree of extension of the index finger [McNeil 2017: 28–29].

0.3. Goals of the study

In the present study we aim to

(a) introduce a gestural parameter "tension of the index finger" as a significant characteristic of lax vs. tense deictic gestures;
(b) reveal its relations with the other parameters of body parts which execute lax and tense deictic gestures according to [Grishina 2017], namely "arm tension" and "tension of the fingers' combination";
(c) reveal its relations with those semantic and pragmatic gestural parameters, which are bound with "arm tension" and "tension of the fingers' combination";
(d) discover other factors which are likely to cause the usage of the extended IF or ½ IF in everyday Russian gestures.

0.4. Data and method

Our study is based on the data retrieved from MURCO, which include

(a) clips from the deep annotated (containing the annotation of gestures) subcorpus of MURCO, where the deictic index finger gestures are marked;
(b) non-annotated clips from the movie subcorpus of MURCO, which contain the words *zdes'* 'here', *tam* 'there', *vot* 'here is...', *von* 'there is...', *eto* 'this is...'.
(c) non-annotated clips from MURCO which serve as video illustrations to the chapter "Russian deictic gestures" of the book [Grishina 2017].

Having examined more than 800 clips altogether, we have sorted out about 50 instances of ½ IF and about 100 instances of extended IF. The collection of the clips with IF gestures can be found at https://yadi.sk/d/iVCpsLcI8R18kA.

Following the methodology adopted in [Grishina 2017], we apply the χ^2 test to examine the dependencies between our data.

1. Tense vs. lax body parts in the Russian deictic gestures

Measuring the degree of dependency between the five parameters of deictic gestures (configuration, orientation, tension of the hand, arm tension and formedness of the fingers' combination), [Grishina 2017: 73, 76–78] uses the χ^2 test and discovers the following relations (in this paper each one is supplied with the values of χ^2 and p; the value of p is displayed in exponential notation):

Strong relations:

- configuration—orientation (χ^2 = 514.02; p = 2.4−112)—for the open-hand deixis, the supine hand is preferred, whereas for the index-finger deixis, the prone and vertical orientation is preferred;
- arm tension—hand tension (χ^2 = 60.36; p = 7.89−15)—the extended arm is generally accompanied by a tense hand, and a half-bent arm—with a lax hand;
- arm tension—orientation (χ^2 = 39.5; p = 2.65−09)—the extended arm tends to appear together with a prone hand, and a half-bent arm—with a supine or vertical hand;
- formedness of fingers' combination—orientation (χ^2 = 29.66; p = 5.75−06)—the loose combination is associated with a supine hand, and a tight combination—with a prone hand. The half-formed combination and the vertical hand show no preferences;

Weak relations:

- hand tension—orientation (χ^2 = 9.88; p = .007)—the tense hand is generally vertical, other values show no preferences;
- arm tension—formedness of fingers' combination (χ^2 = 7.47; p = .024)—the extended arm tends to prefer tight combination, other values show no preferences;

No relation: configuration—arm tension (χ^2 = 0.46; p = .5).

E. A. Grishina also attempts to find correlations between the parameters given and the pragmatic features of the deictic gestures. She shows that the tension of hand and arm, as well as the formedness of fingers' combination, tend to distinguish the primary and the secondary reference in the same way as the Russian collocations *vot / von X* ('Here / There is X', primary reference in activating contexts) vs. *eto X* ('This is X', secondary reference in anaphoric contexts) do.[3] The tense hand, the extended arm and the tight combination of fingers are typical for the activating contexts; the lax body parts generally appear in the anaphoric contexts [Grishina 2017: 104]. The pragmatic feature of the closest and the farthest distance between the speaker and the pointed object (represented by the words *tut* 'here' and *tam* 'there') demonstrates a weak correlation with hand tension: the tense hand is likely for *tam*, the lax hand—for *tut* [ibid., 94].

[3] In case of the primary reference, the speaker (= the gesturer) draws the listener's attention to some object, and in case of the secondary reference, the speaker refers to the object which has already been introduced to the listener.

In our study we measure the degree of dependency between the five parameters mentioned, on the one hand, and the parameter "tension of the index finger", on the other hand. Our study yields the following results:

1.1. Correlations between the tension of the index finger and other formal parameters of the deictic gestures

Table 1. Correlation between the tension of the index finger and the arm tension

Index finger \ Arm	Extended	Half-bent
Extended	34	60
½	*9*[4]	39

$\chi^2 = 4.86$; $p = .003$

Table 2. Correlation between the tension of the index finger and the formedness of fingers' combination

Index finger \ Fingers' combination	Tight	Loose (including half-formed)
Extended	21	73
½	11	36

$\chi^2 = 0.02$; $p = .089$

Table 3. Correlation between the tension of the index finger and the hand orientation

Index finger \ Orientation	Vertical	Prone[5]
Extended	48	47
½	21	27

$\chi^2 = 0.59$; $p = .044$

The parameter "hand tension" is not relevant for index-finger pointing, and the parameter "hand configuration" evidently takes the value "index finger". Thus, we regard the correlations with only three parameters out of the five proposed by Grishina.

[4] In the tables we follow the font design adopted in [Grishina 2017]: if the number is larger than the theoretically expected result, it is marked in bold and placed upon the dark background; if the number is smaller than the theoretically expected result, it is marked in italics and underlined.

[5] The supine orientation of the hand has not been found in our data.

The data given shows that there is a weak relation between the tension of the IF and the arm tension: the ½ IF is not likely for the deictic gestures executed with the extended arm. Thus, we can conclude that generally **there is a direct dependency between the degree of tension of all body parts, which take an active part in the execution a deictic gesture (arm, hand, index finger)**.

Some suggestions which may account for the instances of ½ IF in deictic gestures with a tense arm, are given further in paragraph 2.

According to our data, the tension of the IF is independent both of the hand orientation and of the formedness of fingers' combination. However, we assume that the dependencies between these parameters may still exist, and the reason for why they are not revealed in our study is a rather small number of gestures with the IF that we have examined. Indeed, if we search for correlations between the parameters that Grishina proved to be interrelated, our data will show that there is a strong correlation between the arm tension and the hand orientation, which agrees with Grishina's conclusion that the extended arm is typical for the prone hand. The difference is that, in our case, $p = 3.14-05$; thus, the exponent is less (-05 instead of -09), and the correlation is weaker. The other two pairs of parameters regarded by Grishina, in which relations are even weaker (the value of the exponent is -06 for the pair "formedness of fingers' combination—orientation" and -02 for the pair "formedness of fingers' combination—arm tension"), will appear to be independent in our data ($p = 0.42$ and 0.41 respectively).

1.2. Correlations between the tension of the index finger and pragmatic features of the deictic gestures

Table 4. Correlation between the tension of the index finger and the distance between the speaker and the pointed object

Distance from the speaker / Index finger	closest (*tut* 'here')	farthest (*tam* 'there')
Extended	3	30
½	**4**	4

$\chi^2 = 7.61$; p = .006

Table 5. Correlation between the tension of the index finger and the activating / anaphoric contexts of usage of the deictic gesture

Contexts of usage / Index finger	Activating (*vot*, *von X* 'Here / There is X')	Anaphoric (*eto X* 'This is X')
Extended	16	10
½	*2*	**9**

$\chi^2 = 5.82$; p = .002

Tables 4 and 5 show that the two pragmatic parameters in question correlate with the ½ IF in the same way as they do with the lax arm and hand, according to Grishina's study; thus, we may suggest that **the deictic gestures, which are accompanied by the words *tut* 'here' and *eto X* 'This is X', are likely to be executed by a lax body part,** and **this can be any body part which is engaged in forming the gesture (the index finger, the hand, or the arm)**. In case of the *tut*-contexts, the ½ IF is an iconic sign—its shortness reflects a short distance between the speaker and the pointed object.

However, unlike the tense hand and arm, the extended IF is not sensitive to the pragmatic parameters—it can be freely used in any of these contexts. The reason for this is probably that the gesture with the extended IF is a standard variant of pointing, whereas the gesture with the ½ IF is a marked one, as it is used less often.

2. Factors which govern the usage of the ½ IF in Russian deictic gestures

In this paragraph, we will shortly describe the factors which have not been mentioned above, but which may influence the usage of the ½ IF in executing a deictic gesture. These are the physical comfort, the complex trajectory of movement, the soft imperative and the hesitation of the speaker.

2.1. Physical comfort[6]

The ½ IF can be preferable if:

(a) the speaker's **palm** is oriented **towards the speaker**, i.e. if he points backwards (Pic. 2) or to himself (Pic. 3);

(b) the speaker's IF **touches the surface** of the pointed object, and the surface is **round** (e.g., a bottle, Pic. 4) or **horizontal** (e.g., a book on the table, Pic. 5).

The extended IF can be preferable if:

(a) the speaker's IF **touches the surface** of the pointed object, and the surface is **vertical** (Pic. 6);

(b) the speaker **does not touch the surface** of the pointed object (Pic. 7).

[6] The factor of physical comfort was described in [Grigor'eva, Grigor'ev, Kreydlin 2001]; [Kreydlin & Pereverzeva 2010].

Pic. 2: *У нас там мент ключ от наручников потерял…*
'We have a cop out there who has lost the key to the handcuffs…'
(«Операция С Новым Годом», oper_SNG_307)[7]

Pic. 3: *И кто почтит Лебедева? 'And who will commemorate Lebedev?'*
(«Идиот»; video illustration 5.3_3 to chapter 5 of [Grishina 2017], https://yadi.sk/d/pJmwnlk4koqB3)

[7] In the subscription to the clips taken from MURCO the number of the picture is followed by the cue and its English translation, both given in italics (the part of the cue which goes together with the gesture is underlined). After the cue, there is (in brackets) the Russian name of the movie and the name of the clip in our collection, which is the same as in MURCO.

Pic. 4: *А это хорошая водка / хоть ведро выкушай. Фрау водка.*
'And this is a good vodka. One may drink up a bucket. Frau vodka'
(«Операция С Новым Годом», oper_SNG_369)

Pic. 5: *А я уже родился в городе. Вот здесь.*
'And me, I was born in the town. Right here'
(«Про уродов и людей», urody_041)

Pic. 6: *Вот здесь. Вот. 'Here. Right here'*
(«Адъютант его превосходительства», adjutant_0517)

Pic. 7: *И вот здесь. 'And right here'*
(«Гардемарины, вперёд!», gardemarin_317)

2.2. Complex trajectory of movement

½ IF can mark a complex trajectory of movement—in particular, it can be used for pointing at something beyond a high barrier. Two clips (Pic. 8 and 9) are remarkable from that point of view:

Pic 8: — *Скажи, Левиус, где моя родина?* — *Вон там*.
'— Tell me, Levius, where is my home? — Over there'
(«*И* на камнях растут деревья», i_na_kamniah_190)

Pic. 9: *Ты здесь купаешься? 'Do you take a bath here?'*
(«Подкидыш», podkidysh_051)

In Pic. 8, Levius points to the mountains, having in mind that the home of Kuksha, who asks him the question, is not in the mountains, but to the other side of them (Kuksha is Ilmen Slav), and the ½ IF here means that to reach Kuksha's home one should get over the mountains and not just go up to the top.

In Pic. 9, the little girl also executes the gesture with the ½ IF, maybe because the side of the bathtub is a high enough barrier for her (it is remarkable that the grown-up, who answers the girl's question, points downwards at the bathtub with an extended IF, as the sides are not that high for him).

In both cases, the ½ IF is an iconic representation of the trajectory of movement to the destination—it is not a straight line, as in the case of the extended IF, but a curve which is caused by the necessity to overcome a barrier.

2.3. Soft imperative

The notion of soft imperative component ('to ask for something') as opposed to hard imperative component ('suppression') is discussed in detail in the book [Grishina 2017: 67–68]. In our data, we have come across the deictic gestures where the ½ IF was used to soften the request, cf. **Pic. 10**, where the speaker emphasizes his respect for the addressee:

Pic. 10: *…Если к этому наброску именно Вы приложите ну хоть частицу Вашего тонкого ума и высокого таланта… 'If it were you who would apply a small piece of your fine mind and your high talent to this draft…'* («Чокнутые», choknutie_140)

The usage of the ½ IF to convey this communicative intention looks quite natural: the extended IF symbolizes a vector, which is associated with suppressing the addressee [Grishina 2017: 70], whereas the ½ IF looks less like a vector.

2.4. Hesitation of the speaker

½ IF can mark the lack of the speaker's self-confidence, his perplexity and humbleness. Just as in the previous case, the speaker nullifies the idea of suppressing the addressee by using the ½ IF (**Pic. 11**):

Pic 11: <Semen Semenovich, rendered speechless with fear, is gesticulating instead of speaking> («Бриллиантовая рука», bril_ruka_154)

3. Conclusion

Our study shows that Russian deictic gestures with the slightly curved index finger are specific variants of standard deictic gestures executed with the fully extended index finger. The usage of the curved index finger is caused by a series of factors, both physical and pragmatic. The problem of comparing the strength of these factors, as well as checking the dependencies between the tension of the index finger and hand orientation, or between the former and the formedness of the fingers' configuration, should be resolved in further studies basing on more numerous examples.

References

1. *Bressem J., Ladewig S. H.* (2011), Rethinking gesture phases: Articulatory features of gestural movement? Semiotica 184–1/4 (2011), pp. 53–91.
2. *Fan R. C.* (2014), Verb Agreement, Negation, and Aspectual Marking in Egyptian Sign Language, M. A. thesis, available at: https://repositories.lib.utexas.edu/handle/2152/28287.
3. *Grigor'eva S. A., Grigor'ev N. V., Kreydlin G. E.* (2001), The dictionary of Russian gestures [Slovar' yazyka russkih zhestov], Wiener Slawistischer Almanach, Moskva–Wiena.

4. *Grishina E. A.* (2017), Russian gestures from a linguistic perspective [Russkaya zhestikulyatsiya s lingvisticheskoy tochki zreniya]. Yazyki slavyanskoy kultury, Moskva.
5. *Harling P. A., Edwards A. D. N.* (1997), Hand Tension as a Gesture Segmentation Cue, Progress in Gestural Interaction: Proceedings of Gesture Workshop '96, Springer, pp. 75–87.
6. *Kendon A.* (2004), Visible action as utterance, Cambridge University Press, Cambridge.
7. *Kita S., Planck M.* (2003), Pointing: Where Language, Culture, and Cognition Meet, Taylor & Francis Group.
8. *Kreydlin G. E.* (2002), Nonverbal semiotics [Neverbal'naya semiotika]. Novoe literaturnoe obozrenie, Moskva.
9. *Kreydlin G. E.* (2007), Mechanisms of interaction between verbal and nonverbal in a dialog. IIb. Deictic gestures and their types [Mehanizmy vzaimodeystviya verbal'nyh i neverbal'nyh edinits v dialoge. IIb. Deykticheskie zhesty i ih tipy], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2007" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2007"], Moskva, pp. 320–327.
10. *Kreydlin G. E.* (2008), Mechanisms of interaction between verbal and nonverbal in a dialog. IIb. Deictic gestures and speech acts [Mehanizmy vzaimodeystviya verbal'nyh i neverbal'nyh edinits v dialoge. IIb. Deykticheskie zhesty i rechevye akty], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2008" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2008"], Moskva, pp. 248–253.
11. *Kreydlin G. E., Pereverzeva S. I.* (2010), Body parts and their names in Russian: biological and semiotic pairs of body parts [Chasti tela ii h imena v russkom yazyke: biologicheskaya i semioticheskaya parnost' chastey tela], Papers and reports of the II International conference "Russian language and literature in the international educational environment: actual matters and perspectives" [Doklady i soobscheniya II Mezhdunarodnoy konferentsii "Russkiy yazyk v mezhdunarodnom obrazovatel'nom prostranstve: sovremennoe sostoyanie i perspektivy"], Vol. II, Granada, pp. 2064–2069.
12. *McNeil S. H.* (2017), Sign language static gesture recognition using leap motion, Thesis proposal, available at: https://pdfs.semanticscholar.org/c87e/39986b03108f200c9855566cfced3a188626.pdf
13. *Povinelli D. J., Davis D. R.* (1994), Differences between chimpanzees (Pan troglodytes) and humans (Homo sapiens) in the resting state of the index finger: Implications for pointing, Journal of Comparative Psychology, 108, pp. 134–139.
14. *Puppel J.* (2018), Human body-gesture capacity: the tense—lax distinction: a preliminary outline of research work, Scripta Neophilologica Posnaniensia, vol. XVIII, pp. 73–81.
15. *Wachsmuth I., Kopp S.* (2002), Lifelike gesture synthesis and timing for conversational agents, Gesture and Sign Languages in Human-Computer Interaction, GW 2001, LNAI 2298. Springer, pp. 120–133.

AN ANATOMY OF A LIE: DISCOURSE PATTERNS IN ULTIMATE DECEPTION DATASET

Pisarevskaya D. (dinabpr@gmail.com)
Institute for Systems Analysis FRC CSC RAS, Moscow, Russia

Galitsky B. (boris.galitsky@oracle.com)
Higher School of Economics, Moscow, Russia and Oracle Corp, Redwood Shores CA, USA

We propose a hypothesis that a deception in text should be visible from its discourse structure. The problem of deception detection is then formulated as classification of a discourse tree of this text, according to the Rhetorical Structure Theory. This discourse tree (DT) is extended by the speech acts expressions attached as the labels for the edges. We employ what we call an ultimate deception dataset: a set of customer complaints for English, that includes descriptions of problems customers experienced with certain businesses. It contains about 2,400 complaints about banks and provides clear ground truth, based on available factual knowledge in the financial domain. The complaints are written by non-professional writers. We conduct experiments to explore correlation between implicit cues of the rhetorical structure of texts and how truthful/deceptive are these texts. The results show that a deception in text can be detected reliably enough to assure industrial applications. Automated detection of text with misrepresentations such as fake reviews is an important task for online reputation management.

Keywords: Customer Complaints, Rhetorical Structure Theory, Discourse analysis, Deception detection, Fake Reviews

1. Introduction

It has been discovered that a lot of forms of human intellectual and communication activity are associated with certain discourse structures. Rhetorical Structure Theory (RST) [1] is a good means to express correlation between such form of activity and its representation in how associated thoughts are organized in text. Rhetorical Structure Theory presents a hierarchical, connected structure of a text as a discourse tree, with rhetorical relations between its parts. The smallest text spans are called elementary discourse units (EDUs). In communicative discourse trees (CDTs), the labels for communicative actions (CAs) (VerbNet expressions for verbs) are added to the discourse tree edges to show which speech acts are attached to which rhetorical relations; this structure helps to understand argumentation [2].

Logical Argumentation needs a certain combination of rhetorical relations of *Elaboration, Contrast, Cause* and *Attribution* [3]. Persuasiveness relies on certain structures

linking *Elaboration, Attribution* and *Condition* [4]. Explanation needs to rely on certain chains of *Elaboration* relations plus *Explanation* and *Cause*. A rhetorical agreement between a question and an answer is based on specific mappings between the rhetorical relations of *Contrast, Cause, Attribution* and *Condition* between the former and the latter [5]. Discourse trees turned out to be helpful to form a dialogue and to build dialogue from text, in order to better understand the structure of texts.

In this paper, we study rhetorical structure correlated with certain forms of verbal activity, namely we focus on deception in texts of various genres such as news articles, customer reviews and customer complaints. We intend to discover the distinct features of discourse trees associated with deception. Some of such features can be observed as a result of manual analysis, but most of such features are concealed and need to be tackled by a data-driven approach, so we adjust our customer complaints dataset tagged to detect improper argumentation patterns and invalid claims to serve as a training/test dataset for detection of deceptions.

Research on automated deception detection in written texts is focused on classifying if a narrative is truthful or deceptive. Even if an exhaustive factual information / ontology for a domain is available, it is still hard to perform fact-checking in texts since substantially deep text understanding is necessary and text representation via a logic form is required. It is much more difficult to assess truthfulness when such ontology does not exist, as even manual deception detection, in order to collect datasets for machine learning, as a biased and subjective task. The main difficulty is to detect deception where factual knowledge is not available to a degree sufficient to computationally establish the truth. This situation is typical in the real world, from intuitive choice of product based on reviews to judges' verdicts. It is impossible to establish the truth based on known facts, so decisions are based on implicit cues such as the way people explain what they have done and provide arguments for why they have done so.

While detecting misrepresentation in writing, it is important to differentiate between different categories of writers. Professional writers are frequently good at misrepresenting, and they do not include cues for what might be a lie. Conversely, a content written by non-professional writers is often authentic in how it indicates the thought patterns of the writer where the traces of a lie and hints for how it is motivated can be found.

That's why we analyze how misrepresentation occurs in both professional writing and user generated content (and provide examples of different genres: customer complaints and news stories). Due to this reason, we also provide the ground truth dataset that contains texts written by non-professional writers (bank customers). We also evaluate our classifier, trained on the new dataset, in the domain of business correspondence of non-professional writers such as Enron dataset.

We focus on deception in reviews of products and services as a special case. Automated detection of fake reviews is important for online reputation management tasks. Since fake reviews dataset is available, this is a good domain to evaluate our general domain-independent deception detection algorithm. Fake reviews are deception, but they are artificial since their purpose is not to do a misrepresentation to achieve an agent goal. Usually, this goal is associated with a desired action of another agent who is the addressee of the text that includes this misrepresentation (that is a main scenario of why people lie in the real world). Instead, in the domain of reviews, its subgenre—fake

complaints—are written on demand to manipulate public opinion, that is not an usual purpose of misrepresentation in interaction between people expressed in text. They are written with a definite objective, in order to get a better service after the complaint. Therefore, we believe that customer complaints could be the most adequate data source to explore the linguistic correlates of deception and train a classifier.

In customer complaints, complainants frequently write that they have been provided a misrepresentation by a customer support personnel. At the same time, it might be possible that the complaints are in turn lying about what was said to them by their opponents. It is hard to determine, who is lying: customer support or the complaint author himself; however, the very fact that a given complain arose usually means that there is a misrepresentation associated with the text of the given complaint. That is why the complaints are a valuable systematic source of data on deception.

To train a truth vs lie detection classified, one needs a corpora with defined ground truth. It is needed for classification tasks solving and exploring the links between implicit cues of rhetorical structure of texts and how truthful/deceptive are these texts.

The first contribution of this paper is to investigate how discourse features can be used for deception detection. The second contribution is to present the new ultimate deception dataset of bank customer complaints, it contains ground truth, is written by non-professional writers and can be used for deception detection in written texts.

The research was done for English. The paper is organized as follows. Firstly we show examples of misrepresentation in reviews and news stories, in order to highlight how it is presented in the discourse structure of texts of different genres, in both professional writing and user generated content (**Sections 2**, **3**). **Section 4** examines the existing datasets for deceptive reviews detection, it also presents briefly the main methods for deceptive texts detection, in general. In **Section 5**, the new dataset of customer complaints, with clear ground truth, is provided. In **Section 6**, we describe the deception detection methods, namely how communicative discourse trees construction and Tree Kernel learning can be applied in a system for classification of genuine/deceptive texts. **Section 7** consists of first evaluation results of the classification methods, based on CDTs construction and Tree Kernel learning, on the new provided dataset, accompanied by the results on the 'gold standard' dataset of genuine/fake reviews and on the dataset from the real world. **Section 8** contains conclusions.

2. Example of Misrepresentations in User-Generated Content

We provide some examples of misrepresentation in texts of different genres, in order to show how it is emphasized in the discourse structure of texts. Regarding possible misrepresentation in the user-generated content, the following example from customer complaints can be provided (1). We highlight the statement determined by the authors of this paper to be a deception in both text and its discourse tree. The statement is deceptive based on its factuality.

(1) *'I have accounts with them for almost 10 years, I hated it their customer service! Worst one ever. I don't know what's their problems, I'm not recommending their services and banking to anybody, I stopped using their credit cards already! The only*

reason I can't close my accounts with them, it could drop my credit score. I will not close my credit cards, but I'm not definitely using them so they can't make money from on us! I just had conversation with a supervisor from California called Steve he and his representative didn't even understand my situation, which was not common at all, basically didn't want to help me!'

The author of this complaint does not provide a single argument backing up his claim. And the author's statement that his credit history can be negatively affected by his closing an account is a misrepresentation.

We show the text split into elementary discourse units as done by discourse parser [6]. What do we see in the discourse tree for this text? We show important (non-default) rhetorical relations in bold and highlight the verbs with the role of communicative actions which are an important addition to the rhetorical relations.

```
elaboration (LeftToRight)
 elaboration (LeftToRight)
  attribution (LeftToRight)
   TEXT:I have accounts with them for almost 10 years,
   TEXT:I hated it their customer service !
  TEXT:Worst one ever.
 elaboration (LeftToRight)
  elaboration (LeftToRight)
   explanation (LeftToRight)
    attribution (LeftToRight)
     cause (LeftToRight)
      attribution (RightToLeft)
       TEXT:I do not know
       TEXT:what is their problems,
      TEXT:I'm not recommending their services and banking to anybody,
     TEXT:I stopped using their credit cards already !
    attribution (RightToLeft)
     TEXT:The only reason I can not close my accounts with them,
     TEXT:it could drop my credit score.
   contrast (RightToLeft)
    TEXT:I will not close my credit cards,
    enablement (LeftToRight)
     TEXT:but I'm not definitely using them
     TEXT:so they can not make money from on us !
  elaboration (LeftToRight)
   TEXT:I just had conversation
   same-unit
    elaboration (LeftToRight)
     TEXT:with a supervisor from California called Steve, he and his representative did not
          even understand my situation,
     TEXT:which was not common at all,
    TEXT:basically did not want to help me !
```

Figure 1. A communicative discourse tree for the user-generated text example

There is an unusual chain of rhetorical relations explanation-attribution-cause-attribution-attribution. It is a suspicious explanation pattern on its own. Unsurprisingly, the atom statement for the last attribution (which is the basis of this explanation, highlighted in **Figure 1**) turns out to be false.

3. Example of Misrepresentations in Professional Writing

For comparison with misrepresentation in texts written by non-professional writers, we show misrepresentation examples in news stories. In our first example, the objective of the author is to attack a claim that the Syrian government used chemical weapon in the spring of 2018 (2, **Figure 2**). An acceptable proof would be to share a certain observation, associated from the standpoint of peers, with the absence of a chemical attack. For example, if it is possible to demonstrate that the time of the alleged chemical attack coincided with the time of a very strong rain, that would be a convincing way to attack this claim. However, since no such observation was identified, the source, Russia Today, resorted to plotting a complex mental states expressing how the claim was communicated, which agents reacted which way for this communication. It is rather hard to verify most statements about the mental states of involved parties. We show the text split into EDUs as done by [6] discourse parser:

(2) *[Whatever the Douma residents,][who had first-hand experience of the shooting of the water][dousing after chemical attack video,][have to say,] [their words simply do not fit into the narrative][allowed in the West,] [analysts told RT.] [Footage of screaming bewildered civilians and children] [being doused with water,][presumably to decontaminate them,][was a key part in convincing Western audiences][that a chemical attack happened in Douma.] [Russia brought the people][seen in the video][to Brussels,][where they told anyone][interested in listening][that the scene was staged.] [Their testimonies, however, were swiftly branded as bizarre and underwhelming and even an obscene masquerade][staged by Russians.] [They refuse to see this as evidence,][obviously pending][what the OPCW team is going to come up with in Douma], [Middle East expert Ammar Waqqaf said in an interview with RT.] [The alleged chemical incident,][without any investigation, has already become a solid fact in the West,][which the US, Britain and France based their retaliatory strike on.]*

This article (RussiaToday 2018) does not really find counter-evidence for the claim of the chemical attack it attempts to defeat. Instead, the text says that the opponents are not interested in observing this counter-evidence. The main statement of this article is that a certain agent "disallows" a particular kind of evidence attacking the main claim, rather than providing and backing up this evidence. Instead of defeating a chemical attack claim, the article builds a complex mental states conflict between the residents, Russian agents taking them to Brussels, the West and a Middle East expert. That's why we consider this example as misrepresentation.

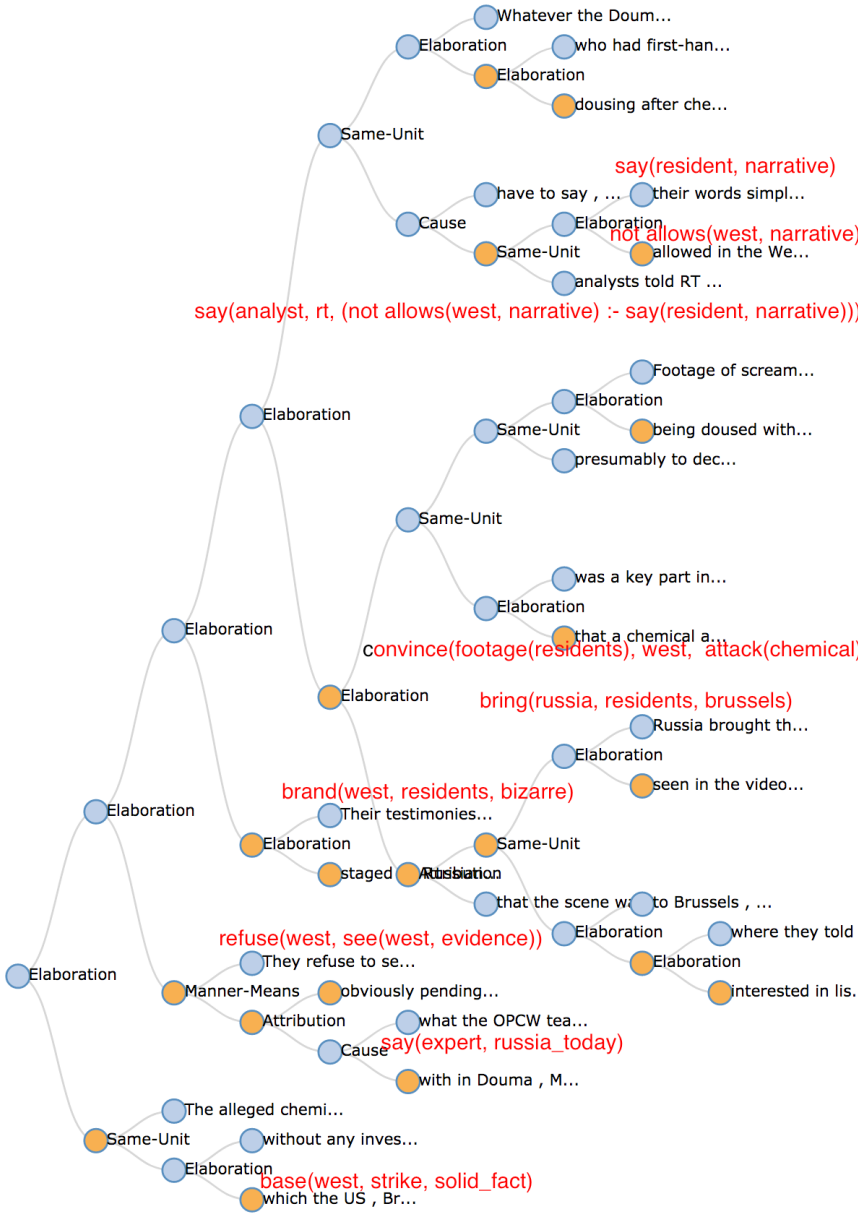


Figure 2. CDT for the chemical attack claim. An author attempts to substitute a desired valid argumentation chain by a fairly sophisticated mental states expressed by CA

Our other example of controversial news is a Trump-Russia link acquisition (BBC 2018, 3, **Figure 3**). For a long time it was unable to confirm the claim, so the story is repeated over and over again to maintain a reader's expectation that it would be instantiated one day. There is neither confirmation nor rejection that the dossier exists, and the goal of the author is to make the audience believe that such dossier does exist neither providing evidence nor misrepresenting events. To achieve this goal, the author can attach a number of hypothetical statements about the existing dossier to a variety of mental states to impress the reader in the authenticity and validity of the topic.

(3) *In January 2017, a secret dossier was leaked to the press. It had been compiled by a former British intelligence official and Russia expert, Christopher Steele, who had been paid to investigate Mr Trump's ties to Russia.*

The dossier alleged Moscow had compromising material on Mr Trump, including claims he was once recorded with prostitutes at a Moscow hotel during a 2013 trip for one of his Miss Universe pageants. Mr Trump emphatically denies this.

The file purported to show financial and personal links between Mr Trump, his advisers and Moscow. It also suggested the Kremlin had cultivated Mr Trump for years before he ran for president.

Mr Trump dismissed the dossier, arguing its contents were based largely on unnamed sources. It was later reported that Mr Steele's report was funded as opposition research by the Clinton campaign and Democratic National Committee.

Fusion GPS, the Washington-based firm that was hired to commission the dossier, had previously been paid via a conservative website to dig up dirt on Mr Trump.

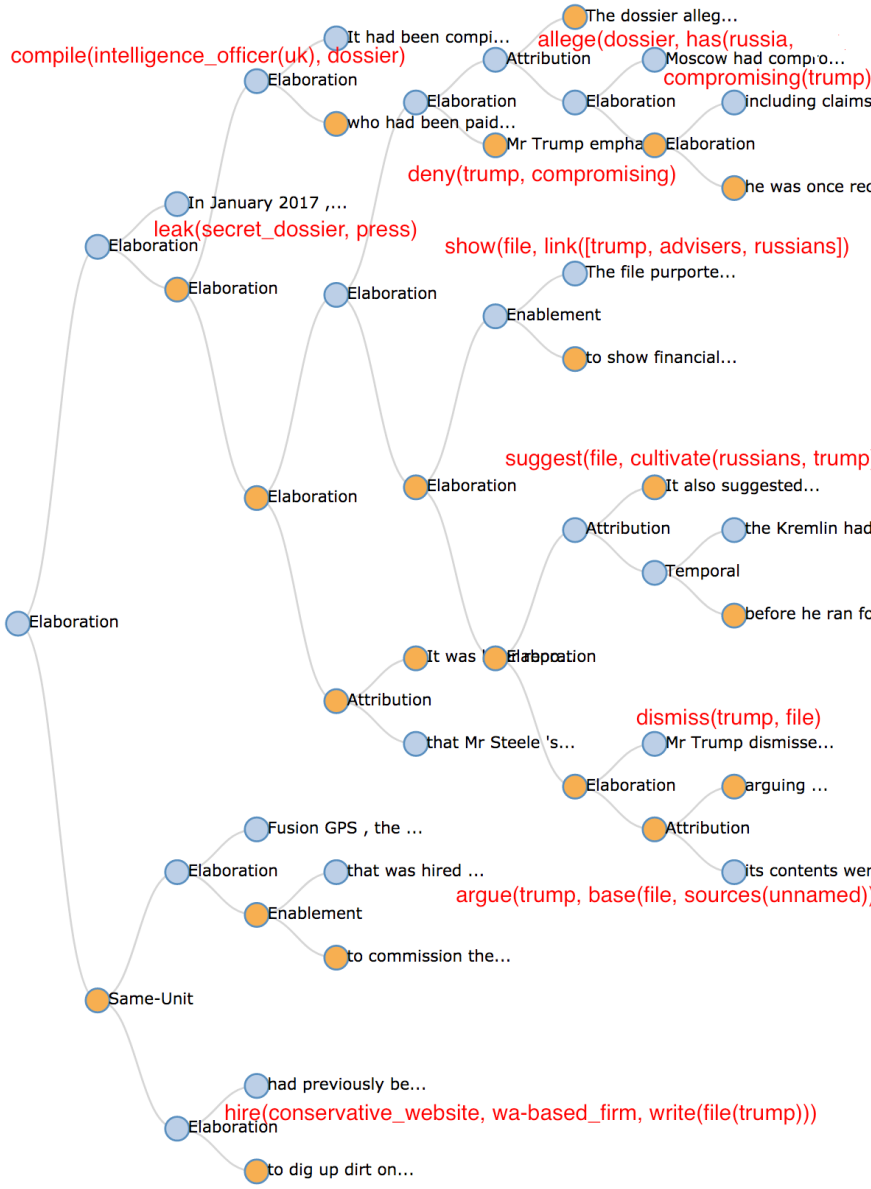


Figure 3. CDT for an attempt to prove something where an evidence is absent so the facts are "wrapped" into complex mental states as expressed by communicative actions

4. Background and Related Work on Deception Datasets

As customer complaints are a subgenre of reviews, we pay the main attention to the existing truthful/deceptive reviews datasets. Deceptive product reviews can be referred to as deceptive opinion spam: fictitious opinions that have been deliberately written to sound authentic, in order to deceive the reader [11]. Spammers write fake reviews to promote or demote target products. They are deliberately written in order to sound authentic, and it is difficult to recognize them manually: human average accuracy is merely 57.3% [11].

Automated deception detection for reviews faces the lack of 'gold standard' corpora with verified examples of deceptive uses of language. Besides this, intentionally written (e.g. by crowdsourcing) texts are distinct from genuinely produced texts. Hence, such artificial texts classified as deceptive by human annotators are not necessarily totally deceptive.

The release of two 'gold standard' datasets (available at http://myleott.com/) allowed for applying supervised learning methods, taking stylistic, syntactic and lexical features into consideration [12], [11], [13], [14]. Hotels reviews were chosen for the datasets, because it was suggested that deception rates among travel reviews is reasonably small. The latter dataset includes, among other reviews, crowdsourced generation of deceptive reviews. It contains 400 truthful positive reviews from TripAdvisor; 400 deceptive positive reviews from Mechanical Turk; 400 truthful negative reviews from reviews websites; 400 deceptive negative reviews from Mechanical Turk.

Later researchers tried to overcome the lack of large realistic datasets on different topics and domains. For example, Yao et al. [15] apply a data collection method based on social network analysis to quickly identify deceptive and truthful online reviews from Amazon. The dataset contains more than 10,000 deceptive reviews in diverse product domains. The problem of the mentioned 'gold standard' datasets is that the fake reviews were not taken from genuinely written ordinary reviews and manually classified as fake. Instead, whey were written on demand by the Amazon Mechanical Turk workers, hence they are not indicative of deception [16]. However, they are accepted as 'gold standard' datasets for this research field. Rules used in [12] to create ground truth datasets were used in later projects, such as in [17].

The real-life Amazon dataset [18] contains reviews from Amazon.com (crawled in 2006) which is large and covers a very wide range of products. It was used, for example, in Sun et al. [19], namely, three domains: Consumer Electronics, Software, and Sports. The metadata in this dataset provides only helpfulness votes of the reviews.

In cases where there was no certain knowledge of the ground truth, different ways to collect reviews corpora, relying on other features, were used. For example, in [14] the DeRev corpus of books reviews, originally posted on Amazon, was collected using definite pre-defined deception clues, Book reviews in the corpus are marked as clearly fake, possibly fake, and possibly genuine. The corpus is constituted by 6,819 instances whose 236 were labeled with the higher degree of confidence and are considered as the 'gold standard'.

In [20], two publicly available Yelp datasets were presented. They are labeled with respect to the Yelps classification in recommended and not recommended reviews. Mukherjee et al. [21] found that the Yelp spam filter primarily relies on linguistic,

behavioral, and social networking features. Classification provided by Yelp has been also used in many previous works before as a ground truth, where recommended reviews correspond to genuine reviews, and not recommended reviews correspond to fake ones, so these labels can be trusted. The Yelp NYC dataset contains reviews of restaurants located in New York City (359,052 reviews; 10.27% are fake); the Zip dataset is larger, since it contains businesses located in contiguous regions of the U.S. (608,598 reviews; 13.22% are fake).

Big Amazon dataset is annotated with compliant/non-compliant labels. It has many different topics: from electronics and books to office products (https://s3.amazonaws.com/amazon-reviews-pds/readme.html). It contains labels about star rating, helpful vote, total votes, verified purchase. That could be used for making decisions.

Hence, the existing recent datasets rely on external factors provided by their source, such as review's rating, number of votes, social networking features of review's author, metadata features etc. They are not annotated manually. So, despite the presence of different corpora, lack of corpora with exact ground truth can be understood as a bottleneck in deception detection of online reviews and similar text genres.

For fake reviews detection, language features and behavioral features are usually used, as in [22], [23]. The impact of different language features on deception detection, in general, was studied in [24], [25]. In recent years, big amounts of news stories with misinformation caused by political reasons [26] led to the specific attention to fake news detection studies for English. Several new datasets were proposed, as in [27], [28], [29]. In [30], the combined approach, based on language features, was suggested: there are linguistic (n-gram), credibility-related (capitalization, punctuation, pronoun use, sentiment polarity), and semantic (embeddings and DBPedia data) features. Close approach based on a set of various language features was suggested in [31] (ngrams, punctuation, psycholinguistic features, readability, syntax) and [32] (stylistic, complexity, psychological features). Deep learning approaches were used in [28], [33]. Source and web page features were added in [34], [35]. As to language features, unlike lexical, syntactic and semantic features, discourse features are less used due to the complexity of the approach. Despite this, automated fake news detection, based on simple discourse features, was studied in [36] and is included in the proposed methods for deception detection in written texts. Hence, we decided to examine if more complex discourse features could be useful for automated deception detection in case of reviews and complaints.

5. Description of the Training Dataset

We introduce the ultimate deception dataset. It contains customer complaints—emotionally charged texts which are very similar to reviews and include descriptions of problems they experienced with certain businesses. Raw complaints in English were collected from PlanetFeedback.com for a number of banks submitted in 2006–2010. The dataset consists of 2,746 complaints totally. 400 complaints were manually tagged with respect to the parameters related to argumentation and validity of text: perceived complaint validity; argumentation validity; presence of specific argumentation patterns; detectable misrepresentation. Here, validity of information

is connected with validity of arguments. The dataset contains texts with direct truth confirmation based on manual annotation. It contains authentic data: both truthful and deceptive reviews were taken from spontaneously written customers' texts. Among the annotated 400 complaints, 163 contain a deception.

This dataset includes more emotionally-charged complaints in comparison with other argument mining datasets, such as [37], [38], [39]. For a given topic such as insufficient funds fee, this dataset provides many distinct ways of argumentation that this fee is unfair. Authors attempt to provide as strong argumentation as possible to back up their claims and strengthen their case.

If a complaint is not truthful, it is usually invalid: either a customer complains out of a bad mood or wants to get a compensation. However, if the complaint is truthful it can easily be invalid, especially when arguments are flawed. When an untruthful complaint has valid argumentation patterns, it is hard for an annotator to properly assign it as valid or invalid, without the guidelines. So, according to the guidelines for the manual tagging of the dataset, a complaint was considered as valid if a judge believed that the main complaint claim is truthful under the assumption that a complainant is making truthful statement. Valid complaint needs to include proper discourse and acceptable argumentation patterns. Following this approach, a complaint is marked as truthful if a judge cannot defeat it, using commonsense knowledge, available factual knowledge about a domain or implicit, indirect cues. Inconsistencies detected by a judge also indicate that the complaint author is deceiving. Mentioning multiple unusual, very rarely occurring claims also indicate that the complaint author is deceiving. The judge does not have to be able to prove that the complainant is lying: judge's intuition is sufficient to tag a complaint as untruthful. We suggest that one can provide a valid argumentation and also provide a false statement in a single sentence: 'Rule is like this <correct rule> and I followed it, making <false statement>. Conversely, one can be truthful but provide an invalid argumentation pattern "I set this account for direct deposit and sent a check out of it <truthful statement>, as my HR manager suggested <should not have followed advice from not a specialist in banking>. Therefore validity (of argumentation patterns) and truthfulness are correlated.

Initial set of 400 complaints was tagged by the authors of the paper as experts. After that, three annotators worked with this dataset, having a set of definitions and applying them. Then precision and recall were measured by matching the tags done by the authors as the 'gold standard', after that the set of definitions was edited and elaborated. In the further work, the Krippendorff's alpha measure (for three annotators) was applied as inter-annotator agreement measurement, and it exceeds 80%. Complaints reveal shady practice of banks during the financial crisis of 2007—for instance, manipulating an order of transactions to charge a highest possible amount of non-sufficient fund fees. As it is possible to know, retrospectively and based on facts, the established ground truth, we suggest that the annotators can find out, with high confidence, what information in texts is deceptive. So the dataset would provide ground truth.

The rest complaints were auto-tagged based on the model trained on this 400 set. Then they have been partially manually evaluated. The accuracy of auto tagging exceeds 75%, so these labeled complaints can be also used for the classifiers training.

Customer complaints can be considered as a subgenre of reviews in general, but despite this complaints have much more significance for well-being of customers

in comparison with customer reviews. Furthermore, customer complaints have much more significance for well-being of customers in comparison with customer reviews. Therefore, tagged customer complaints have much more importance associated with truth/deception than customer reviews. Since reviews are associated with opinions which can be random and complaints with customers doing their best to achieve their goals, both the truth and a lie is much more meaningful and serious in comparison with review datasets.

Complaints usually have a simple motivational structure; they are written with an obvious goal. Most complainants are faced with a strong deviation between what they expected from a service, what they received and how it was communicated. Most complaint authors report incompetence, flawed policies, ignorance, indifference to customer needs from the customer service personnel. The authors are frequently exhausted communicative means available to them, confused, seeking recommendation from other users and advising others on avoiding particular financial service. The focus of a complaint is a proof that the proponent is right and the opponent is wrong, as well as resolution proposal and a desired outcome.

6. Detecting Deception via Communicative Discourse Trees

In the Rhetorical Structure Theory [1], [7], discourse is understood as a hierarchical system of discourse units of different size, where smaller discourse units can be successively incorporated into larger ones. Discourse unites can be combined into a higher unit in case there is a rhetorical (discourse) relation of a certain type between them, e.g. *Concession*, *Elaboration*. One of the discourse units is the nucleus (more important), while the other is a satellite (contains the additional information). An elementary discourse unit (EDU) usually corresponds to a clause.

Two RST parsers constructing discourse tree (DT) from paragraphs of text are available at the moment. We used the tool provided by [6], [8]. After that, we build CDTs involving VerbNet.

Argumentation analysis needs a systematic approach to learn associated discourse structures. The features of CDTs could be represented in a numerical space so that argumentation detection can be conducted; however, structural information on DTs would not be leveraged. Also, features of argumentation can potentially be measured in terms of maximal common sub-DTs, but such nearest neighbor learning is computationally intensive and too sensitive to errors in DT construction. Therefore, a CDT-kernel learning approach is selected which applies a support vector machine (SVM) learning to the feature space of all sub-CDTs of the CDT for a given text where an argument is being detected.

Tree Kernel (TK) learning for strings, parse trees and parse thickets is a well-established research area nowadays. The CD-TK counts the number of common subtrees as the discourse similarity measure between two DTs. In this study, we extend the TK definition for the CDT, augmenting DT kernel by the information on CAs. TK-based approaches are not very sensitive to errors in parsing (syntactic and rhetorical) because erroneous sub-trees are mostly random and will unlikely be common among different elements of a training set.

A CDT can be represented by a vector V of integer counts of each sub-tree type (without taking into account its ancestors):

$$V(T) = (\#\ of\ subtrees\ of\ type\ 1, \ldots, \#\ of\ subtrees\ of\ type\ I, \ldots, \#\ of\ subtrees\ of\ type\ n).$$

Given two tree segments CDT_1 and CDT_2, the tree kernel function is defined:

$$K\ (CDT_1, CDT_2) = < V(CDT_1), V\ (CDT_2) > = \Sigma_i V(CDT_1)[i], V(\mathrm{CDT}_1)[i] = \Sigma n_1 \Sigma n_2\ \Sigma_i I_i(n_1) \times I_i(n_2),$$

where $n_1 \in N_1$, $n_2 \in N_2$ and N_1 and N_2 are the sets of all nodes in CDT_1 and CDT_2, respectively; $I_i(n)$ is the indicator function:

$$I_i(n) = \{1\ if\ a\ subtree\ of\ type\ i\ occurs\ with\ a\ root\ at\ a\ node;\ 0\ otherwise\}.$$

Further details for using TK for paragraph-level and discourse analysis are available in [9].

Only the arcs of the same type of rhetorical relations (presentation relation, such as antithesis, subject matter relation, such as condition, and multinuclear relation, such as List) can be matched when computing common sub-trees. We use *N* for a nucleus or situations presented by this nucleus, and *S* for a satellite or situations presented by this satellite. Situations are propositions, completed actions or actions in progress, and communicative actions and states (including beliefs, desires, approve, explain, reconcile and others). Hence we have the following expression for RST-based generalization '^' for two texts $text_1$ and $text_2$:

$$text_1\ \hat{}\ text_2 = \cup_{i,j}(rstRelation_{1i}, (\ldots, \ldots)\ \hat{}\ rstRelation_{2j}(\ldots, \ldots)),$$

where $i \in (RST\ relations\ \text{in}\ text_1)$, $j \in (RST\ relations\ \text{in}\ text_2)$. Further, for a pair of RST relations their generalization looks as follows:

$$rstRelation_1(N_1, S_1)\ \hat{}\ rstRelation_2\ (N_2, S_2) = (rstRelation_1 \hat{} rstRelation_2)\ (N_1 \hat{} N_2, S_1 \hat{} S_2).$$

We define CA as a function of the form verb (agent, subject, cause), where verb characterizes some type of interaction between involved agents (e.g., explain, confirm, remind, disagree, deny, etc.), subject refers to the information transmitted or object described, and cause refers to the motivation or explanation for the subject. To handle meaning of words expressing the subjects of CAs, we apply word2vec models [10].

For EDUs as labels for terminal nodes only the phrase structure is retained. The terminal nodes are labeled with the sequence of phrase types instead of parse tree fragments.

We combined Stanford NLP parsing, coreference resolution tool, entity extraction, DT construction (discourse parser), VerbNet and Tree Kernel builder into one system.

The system is available at https://github.com/bgalitsky/relevance-based-on-parse-trees with the more detailed description. It can be used for similar tasks.

For EDUs as labels for terminal nodes only the phrase structure is retained: we suppose to label the terminal nodes with the sequence of phrase types instead of parse tree fragments. For the evaluation purpose Tree Kernel builder tool [5] was used. These discourse trees features are given to the classifiers.

7. Evaluation Results

We first train the deception detection model on our ultimate deception dataset. For the initial and automatically derived datasets, we show the accuracies of training (grayed) row and testing, averaging through 5x cross-validation. For the bottom three datasets, we only tested the obtained model. For genuine reviews, 380 cases of deception were detected which were false positives, assuming that review writers do not lie (Table 1).

Table 1: Datasets, evaluation settings and recognition accuracies for deception detection

Dataset	Deception	No deception	Precision	Recall	F1 score
Manually tagged complaints	163	237	91	85	88
			83	81	82
Automatically tagged based on initial classifier	1,132	1,615	78	75	76
			69	71	70
Genuine reviews	580	3,420	83	100	91
Fake reviews	414	286	100	59	74
Enron	27	10,000	85	0.1 (estimated)	0.2

We explored whether fake opinionated text have a similar rhetorical structure to text with deception, and genuine reviews have similar rhetoric structure to texts without deception. We took the 'gold standard' reviews dataset: fake reviews and genuine reviews [11], [12] (Table 1).

In [11], [12] authors addressed the problem of detection of opinion spam: obvious instances that are easily identified by a human reader, including advertisements, questions, and other irrelevant or non-opinionated texts. The authors investigated a more implicit type of opinion spam such as deceptive opinion spam, ones that have been deliberately written to sound authentic, in order to deceive the reader. Fake reviews were written by Amazon Mechanical Turk workers. The instructions asked the workers to assume that they are employed by a hotel's marketing department, and to pretend that they are asked to write a fake review (as if they were a customer) to be posted on a travel review website; additionally, the review needs to sound realistic and portray the hotel in a positive light. A request for negative reviews was done analogously.

Although our SVM TK system did not achieve [11], [12] performance of 90% on their data, the task of detection of fake review texts as the ones including deception was performed at 74% accuracy by the classifier.

We suggest that the system could be applied to different text genres (written by non-professional writers), so it could be the universal text classification system for deception, the same which extracts arguments and assesses sentiments polarity. Hence, we run the following evaluation experiment in order to start checking this point.

To assess the deception detection in a real world deception-neutral environment, we ran our detector again the business communication dataset of Enron [40], using

it as the evaluation dataset. This dataset represents neither user-generated content since this is work-related correspondence, not professional writing since the email authors are employees of an organization with various roles. Naturally, deception is concealed, and we do not know what was actually happening in the company and among its business partners. However, a small number of interesting email have been discovered which have a peculiar logical structure and might well be a misrepresentation. Annotators looked at them manually to understand if they were similar to misrepresentation, although we did not have ground truth here. They could not be sure if reviews with tricky patterns similar to misrepresentation were really misrepresentation, but the detector could identify possible reviews with misrepresentation, that were also identified as the 'suspicious' ones (containing possible misrepresentation) by human annotators. Precision turned out to be high and recall extremely low since only a small fraction of deception emails has been discovered. The resultant 0.2% F-score is not an indication of recognition accuracy but instead of our available estimate of the classes in the Enron dataset.

We do not know the actual proportion of emails with misrepresentation in Enron dataset but all detected cases are important since a misrepresentation is uncovered. Recall is not as important for this task as precision: we want to avoid false positives: once an email is classified as the one with deception we would expect to manually confirm it.

We now zoom into the deception detection methodology for the most adequate case, the set of 2,747 automatically tagged complaints (Table 2).

Table 2: Classification accuracy for the baseline and the approach being proposed for deception detection

Method	Precision	Recall	F1 score
Keyword-based	56	53	54
Naïve Bayes	61	63	62
SVM-TK over parse trees and DTs	67	69	68
SVM-TK over parse trees and DTs labeled with CAs	69	71	70

One can see that keyword-based and Naive Bayes classifier perform slightly better than random, since deception manifests itself at the discourse level, not the syntactic one. Then we observe that proceeding to machine learning of DTs delivers 8% gain in classification accuracy.

A deep learning approach could be potentially applied to our structured representation. However, based on our experience with discourse-level data that the amount and quality of data contributes significantly more to the overall accuracy of a classifier, we believe experiments with the same data but different machine learning framework would be redundant.

8. Conclusions

An extensive corpus of literature on RST parsers does not address the issue of how the resultant DT will be employed in practical NLP systems. RST parsers are mostly evaluated with respect to agreement with the test set annotated by humans rather than its expressiveness of the features of interest. In this work we focused on interpretation of DT and explored ways to represent them in a form indicative of a conflict rather than neutral enumeration of facts.

In several previous papers about SVM TK and discourse, it was observed that using SVM TK, one can differentiate between a broad range of text styles, genres and abstract types. These classes of texts are important for a broad spectrum of applications of recommendation and security systems, from finance to data loss prevention domains. Each text style and genre has its inherent rhetorical structure which is leveraged and automatically learned. Since the correlation between text style and text vocabulary is rather low, traditional classification approaches which only take into account keyword statistics information could lack the accuracy in the complex cases.

We showed that deception detection methodology based on rhetorical structure of texts, being applied to various text genres—news texts, online reviews, customer complaints, business communication texts—seems promising and needs to be investigated further. Next steps for proving the hypothesis of a deception being visible from a text's discourse structure should be done. Here, further experiments based on the presented ultimate deception dataset of bank customer complaints should be held. This dataset is in the initial stage now and is still being developed. In the future studies, the whole complaint dataset should be manually annotated. The recognition method will be applied to a bigger annotated dataset part. Results obtained on this dataset should be also compared with other results obtained on 'gold standard' datasets. For a bigger dataset training, we could also apply deep learning models. We will also focus on more experiments for precision improvements, as reducing the number of false positives is mostly important for deception detection task. We also plan to run further experiments on different text genres, to check if the universal text classification system for deception, based on discourse features, could be universal. Both truthfulness and validity are recognized reasonably well which is a value for Customer Relation Management systems and could be useful in different NLP tasks that are based on online reviews analysis.

Acknowledgements

This paper is partially supported by Russian Foundation for Basic Research (project No. 17-29-07033).

References

1. *William C. Mann and Sandra A. Thompson* (1988), Rhetorical structure theory: Toward a functional theory of text organization. Text-Interdisciplinary Journal for the Study of Discourse, 8(3), pp. 243–281.
2. *Boris Galitsky, Dmitry Ilvovsky, and Dina Pisarevskaya* (2018), Argumentation in text: Discourse structure matters, CICLing 2018 (unpublished).
3. *Grasso Floriana* (2003), Characterizing Rhetoric Argumentation, PhD Thesis HERIOT-WATT UNIVERSITY.
4. *B. Galitsky, D. Ilvovsky, and T. Makhalova* (2019), Enabling a Bot with Understanding Argumentation and Providing Arguments, Developing Enterprise Chatbots, Springer—Cham, Switzerland.
5. *B. Galitsky. Rhetorical Agreement: Maintaining Cohesive Conversations* (2019), Developing Enterprise Chatbots, Springer—Cham, Switzerland.
6. *S. Joty, G. Carenini, R. T. Ng, and Y. Mehdad* (2013), Combining intra-and multisentential rhetorical parsing for document-level discourse analysis, ACL (1), pp. 486–496.
7. *M. Taboada, W. C. Mann* (2006), Applications of rhetorical structure theory, Discourse Studies, 8(4), pp. 567–588.
8. *M. Surdeanu, T. Hicks, and M. A. Valenzuela-Escarcega* (2015), Two Practical Rhetorical Structure Theory Parsers, NAACL HLT.
9. *Galitsky B.* (2017), Improving relevance in a content pipeline via syntactic generalization, Engineering Applications of Artificial Intelligence 58, pp. 1-26..
10. *Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, Armand Joulin* (2017), Advances in Pre-Training Distributed Word Representations, LREC 2018.
11. *Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock* (2011), Finding deceptive opinion spam by any stretch of the imagination, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, volume 1, pp. 309–319.
12. *Myle Ott, Claire Cardie, and Jeffrey T. Hancock* (2013), Negative deceptive opinion spam, NAACLHLT 2013, Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 497–501.
13. *S. Feng, R. Banerjee, and Y. Choi* (2012), Syntactic stylometry for deception detection, ACL 12, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pp. 171–175.
14. *Tommaso Fornaciari and Massimo Poesio* (2014), Identifying fake amazon reviews as learning from crowds, Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 279–287.
15. *Wenlin Yao, Zeyu Dai, Ruihong Huang, and James Caverlee* (2017), Online deception detection refueled by real world data collection, Proceedings of Recent Advances in Natural Language Processing, pp. 793–802.
16. *A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance* (2013), Fake review detection: classification and analysis of real and pseudo reviews. tech. rep. uic-cs-2013-03. University of Illinois at Chicago.

17. *Z. Hai, P. Zhao, P. Cheng, P. Yang, X.-L. Li, and G. Li* (2016), Deceptive review spam detection via exploiting task relatedness and unlabeled data, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1817–1826.
18. *Nitin Jindal and Bing Liu* (2008), Opinion spam and analysis, Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM '08). ACM, New York, NY, USA, pp. 219–230.
19. *Chengai Sun, Qiaolin Du, and Gang Tian* (2016), Exploiting product related review features for fake review detection, Mathematical Problems in Engineering.
20. *S. Rayana and L. Akoglu* (2015), Collective opinion spam detection: Bridging review networks and metadata, Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 985–994.
21. *A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance* (2013), What yelp fake review filter might be doing?, Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media.
22. *J. Fontanarava, G. Pasi, and M. Viviani* (2017), Feature Analysis for Fake Review Detection through Supervised Classification, 2017 International Conference on Data Science and Advanced Analytics, pp. 658–666.
23. *S. Mukherjee* (2017), Probabilistic Graphical Models for Credibility Analysis in Evolving Online Communities. Doctor Thesis.
24. *E. Fitzpatrick, J. Bachenko, and T. Fornaciari* (2015), Automatic Detection of Verbal Deception. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers.
25. *P. Rosso and L. Cagnina* (2017), Deception Detection and Opinion Spam, A Practical Guide to Sentiment Analysis, Cambria, E., Das, D., Bandyopadhyay, S., Feraco, A. (Eds.), Socio-Affective Computing, vol. 5, Springer-Verlag, pp. 155–171.
26. *H. Allcott and M. Gentzkow* (2017), Social Media and Fake News in the 2016 Election., Journal of Economic Perspectives, Vol. 31–2, pp. 211–236.
27. *Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi* (2017), Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2931–2937.
28. *William Yang Wang* (2017), "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 422–426.
29. *Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu* (2018), FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media, available at: https://arxiv.org/abs/1809.01286.
30. *Momchil Hardalov, Ivan Koychev and Preslav Nakov* (2016), In Search of Credible News, Artificial Intelligence: Methodology, Systems, and Applications. AIMSA 2016. Lecture Notes in Computer Science, vol 9883. Springer, Cham, pp. 172–180.
31. *Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea* (2018), Automatic Detection of Fake News, Proceedings of the 27th International Conference on Computational Linguistics, pp. 3391–3401.

32. *Benjamin D. Horne, and Sibel Adali* (2017), This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News, available at: https://arxiv.org/abs/1703.09398.
33. *Natali Ruchansky, Sungyong Seo, and Yan Liu* (2017), CSI: A Hybrid Deep Model for Fake News Detection, CIKM'17 Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 797–806.
34. *Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov* (2018), Predicting Factuality of Reporting and Bias of News Media Sources, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3528–3539.
35. *Benjamin D. Horne, William Dron, Sara Khedr, Sibel Adalı* (2018), Assessing the News Landscape: A Multi-Module Toolkit for Evaluating the Credibility of News, WWW 2018, April 23–27, 2018, Lyon, France, pp. 235–238.
36. *Rubin, V. L. Conroy, N. J. and Chen Y. C.* (2015), Towards News Verification: Deception Detection Methods for News Discourse, Proceedings of the Hawaii International Conference on System Sciences (HICSS48) Symposium on Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium, January 5–8, 11 pages.
37. *Christian Stab and Iryna Gurevych* (2017), Recognizing insufficiently supported arguments in argumentative essays, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017), pp. 980–990.
38. *Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn A. Walker* (2016), Internet argument corpus 2.0: An sql schema for dialogic social media and the corpora to go with it, Language Resources and Evaluation Conference.
39. *Shereen Oraby, Lena Reed, Ryan Compton, Ellen Riloff, Marilyn Walker, and Steve Whittaker* (2015), And thats a fact: Distinguishing factual and emotional argumentation in online dialogue, NAACL HLT 2015 2nd Workshop on Argumentation Mining.
40. *Cohen W. W.* (2019), Enron Email Dataset, available at: https://www.cs.cmu.edu/~./enron/.

ПРОСОДИЯ И ГРАММАТИКА ПРЕДИКАТИВНОГО СОЧИНЕНИЯ: КОНСТРУКЦИИ С СОЮЗОМ *И* ПО ДАННЫМ ПРОСОДИЧЕСКИ РАЗМЕЧЕННОГО КОРПУСА

Подлесская В. И. (vi_podlesskaya@il-rggu.ru)

Российский государственный гуманитарный университет, Москва, Россия

Ключевые слова: сложное предложение, русский язык, корпус, устная речь, просодия

PROSODY AND GRAMMAR OF CLAUSAL AND VP COORDINATION: THE RUSSIAN CONJUNCTION *I (AND)* VIEWED THROUGH THE PRISM OF PROSODICALLY ANNOTATED CORPUS DATA

Podlesskaya V. I. (vi_podlesskaya@il-rggu.ru)

Russian State University for the Humanities, Moscow, Russia

The paper focuses on Russian constructions with clauses or VPs combined by means of the conjunction *I* 'and'. Prosodically, the construction may come up in two forms: (a) integrated, i.e.—as a single illocution with the first clause pronounced with a rising pitch that projects discourse continuation, and (b) disintegrated, i.e. as two separate illocutions with the first clause pronounced with a falling pitch that projects no continuation. Basing on the data from the Prosodically Annotated Corpus of Spoken Russian, prosody and grammar of coordinate constructions with the conjunction I 'and' were analyzed qualitatively and quantitatively. The results show that coordinated clauses and VPs are more frequent than coordinated NPs and other types of groups; in spoken narratives, coordinated clauses are more frequent than VPs, while in written narratives, coordinated VPs are more frequent than clauses; coordinated clauses and VPs more often come up as prosodically integrated than as prosodically disintegrated; the rate of integrated constructions is higher in coordinated VPs than in coordinated clauses.

Key words: clause combining, Russian, corpus, natural discourse, prosody

1. Постановка вопроса

Предметом данного исследования являются русские сочинительные конструкции с союзом *И*, в которых конъюнктами являются предикативные единицы — клаузы или их вершины, что в традиционной терминологии примерно соответствует сложносочиненным предложениям и предложениям с однородными сказуемыми[1]. Грамматике и семантике русского предикативного сочинения посвящена большая литература, в том числе, ставшие уже классическими монографические сочинения [Санников 2008], [Урысон 2011]; с корпусных позиций этот феномен проанализирован в соответствующих разделах «Русской корпусной грамматики» [Апресян, Пекелис 2012], [Пекелис 2013]. Однако просодия таких конструкций остается слабо изученной. Цель данной работы частично восполнить этот пробел.

Прежде всего, нас будет интересовать степень просодической интеграции компонентов конструкции и ее возможная корреляция со степенью синтаксической интеграции. В нейтральном случае предикативное сочинение обычно демонстрирует симптомы просодической интеграции: первый конъюнкт (клауза или глагольная группа) реализуется как просодически незавершенный, симптомом просодической незавершенности в русском может быть, в частности, подъем тона в главном фразовом акценте. Так, естественное произнесение (1) — с подъемом тона на акцентном центре первой из сочиненных глагольных групп, на словоформе *отдохну*. На письме такая реализация конструкции нотируется пунктуационно — с помощью запятой при сочинении клауз и с помощью отсутствия пунктуационного знака при сочинении предикатных вершин:

(1) *Я /отдохну и схожу в \магазин.*

Подъем тона является в русском языке одним из стандартных маркеров просодической незавершенности (хотя и не единственно возможным): эти маркеры формируют у адресата ожидание продолжения, т. е. если использовать терминологию, принятую в рамках анализа бытового диалога, можно говорить о том, что просодия является просодическим проектором [Auer 2002]. При этом никаких лексических или грамматических «проекторов» первый компонент конструкции с сочиненными предикатами может не содержать. Это делает возможным альтернативную просодическую реализацию, при которой первый конъюнкт произносится как просодически завершенный — с падением тона в главном фразовом акценте, и, тем самым, конструкция оказывается просодически дезинтегрированной, так происходит, например, при парцелляции второго конъюнкта. На письме такое произнесение имитируется с помощью точки перед *И* и написания *И* с заглавной буквы:

(1’) *Я \отдохну. И схожу в \магазин.*

[1] Исследование выполнено при поддержке РНФ, грант №17-18-01184. Выражаю свою искреннюю признательность анонимным рецензентам, высказавшим проницательные замечания, которые позволили — как я надеюсь — уточнить грамматический, семантический и прагматический статус рассмотренных в работе конструкций.

В литературе такое написание используется иногда как стилистический прием, чтобы продемонстрировать пошаговое развертывание дискурса, при котором повышается значимость каждого отдельного шага. См. два примера из «Денискиных рассказов» В. Драгунского, появляющихся в контексте философского перечисления «что я люблю» и «чего не люблю». В (2) имитируется просодическая завершенность при сочинении клауз, в (3) — при сочинении глагольных групп:

(2) *Я люблю ходить в зоопарк! Там чудесные слоны. И есть один слонёнок.*

(3) *Не люблю, когда ребята задаются. И очень не люблю, когда порежусь, вдобавок — мазать палец йодом.*

В данной работе, опираясь на корпусные данные, я попытаюсь ответить на следующие вопросы:

- Каков арсенал просодических конфигураций, используемых в предикативном сочинении с *И*?
- Какие из этих конфигураций могут использоваться в качестве проекторов в разворачивающейся структуре дискурса?
- Как просодические проекторы могут согласовываться (или не согласовываться!) с лексико-грамматическими?
- Различается ли употребление конструкций с *И* в устном и письменном дискурсе?

Я использую материал корпуса «Веселые истории из жизни» электронной коллекции [Spokencorpora 2018]. Корпус содержит 40 устных монологов (аудиофайлы с синхронизированными просодически размеченными транскриптами, респонденты от 18 до 60 лет, около 10 000 словоупотреблений), плюс письменные версии этих же рассказов, самостоятельно записанные авторами спустя несколько дней после записи устной версии (около 7000 словоупотреблений). Исследуемый корпус предоставляет уникальный материал для сравнения дискурсивных стратегий — в том числе, стратегий предикативного сочинения — в устной и письменной речи именно потому, что включает устные и письменные версии рассказов одного и того же говорящего, основанные на одном и том же сюжете. Были проанализированы все вхождения *И*, обнаруженные в корпусе: 341 вхождение в устном подкорпусе (34,1 на 1000 слов, далее в тексте — примеры с индексом FS-Sp) и 281 вхождение в письменном подкорпусе (40,1 на 1000 слов, примеры с индексом FS-Wr). Дальнейшее изложение будет строиться следующим образом: в **разделе 2** будут качественно и количественно проанализированы основные структурные типы конструкций с *И*, представленные в корпусе, и место конструкций с предикативным сочинением в общей номенклатуре употреблений *И*; в **разделе 3** будут рассмотрены качественно и количественно основные просодические формы реализации предикативного сочинения в устном подкорпусе; в **разделе 4** мы сравним распределение просодически интегрированных и просодически дезинтегрированных конструкций с предикативным сочинением в устном подкорпусе с распределением их пунктуационно маркированных аналогов в письменном подкорпусе и подведем итоги исследования.

2. Номенклатура и частотность употреблений *И* в корпусе

Все задокументированные употребления *И* в корпусе могут быть подразделены на три больших структурных класса: предикативное сочинение, непредикативное сочинение и не собственно сочинительные употребления *И*.

2.1. Предикативное сочинение

В этот класс входят следующие случаи: сочинение клауз (4) и более крупных фрагментов текста (групп клауз), см. *И* в строке 42 в (5), сочинение финитных глагольных групп (6), (7), сочинение инфинитных глагольных групп, в т. ч., инфинитивов (8) и деепричастий (9), а также сочинение глагольной группы с группой другого типа (10):

(4) FS_13-f-sp[2]
32. Я напишу об этом в /газете,
33. и все \узнают!»

(5) FS_18-f-sp
40. ··· Ну и так ти́хонько-о ээ ” ээ ” мм /заговорили,
41. что типа в-вроде как \полиция приехала.
42. ···· И мы /смотрим,
43. что люди значит стали перебегать из нашего поезда в \противоположный.

(6) FS_09-f-sp
54. В общем мы с папой вот это вот всё круто нагрузили ·· на заднее /сиденье,
55. и вместе с /торто́м поехали в Новые \Черёмушки,

(7) FS_09-f-sp
98. Ты спустись /вниз¡,
99. и помоги \вещи поднять¡»

(8) FS_29-m-sp
38. ··· эээ достали /–колоночки,,,
39. стали слушать ·· /Шопена,
40. и \закусывать.<<

(9) FS_14-f-wr
Народ сходил с ума, обмахиваясь платочками и попивая воду.

(10) FS_38-m-wr
Конструкция у него очень простая и состоит из двух бачков, а также насоса для смыва.

[2] Об используемой системе дискурсивной транскрипции см. [Кибрик, Подлесская (ред.) 2009], [Spokencorpora 2018]. Индекс примера содержит отсылку к ярлыку текста в составе корпуса. В примерах из устного подкорпуса указаны также номера строк в транскрипте текста.

2.2. Непредикативное сочинение

Наряду с предикативным сочинением, являющимся основным предметом данного исследования, в корпусе, естественным образом, обнаруживаются и случаи, когда *И* связывает непредикативные единицы. В этот класс входят случаи сочинения именных (11), (12), адъективных (13) и адвербиальных (14) групп:

(11) FS_30-m-wr
Полезное общение перемежалось искромётным юмором
и зажигательными шутками.

(12) FS_37-f-wr
где всегда по весне образуется огромная лужа из воды и талого снега

(13) FS_25-m-sp
6. и была у меня /знакомая одна,
7. ·· ты /знаешь,
8. барышня ·· очень такая /полная,
9. ·· и честно говоря \страшная.

(14) FS_28-f-sp
24. то есть за эту сумму можно было туда-обратно и ещё раз /туда слетать,

2.3. Не собственно сочинительные употребления *И*

Эта группа употреблений представлена тремя подгруппами: частично идиоматизированные и лексикализованные употребления, употребления *И* в качестве частицы, и употребления, «деформированные» речевым сбоем.

К подгруппе лексикализованных условно отнесены сочетания полнозначной группы с аппроксиматором (*и прочее, и так далее*), сочетания *И* с другими союзами (*но и, да и*), эмфатические вопросы, вводимые *И* (*И куда это ты?*), идиоматизированные финализирующие реплики (*И всё!)*. Сюда же отнесены и конструкции с так называемым вторичным сочинением, в которых с помощью *И* присоединяется эмфатически выделенный атрибут:

(15) FS_03-m-sp
16. «Ну \что ж такое,
17. ···/щи,
18. и ·· без \сметаны!»

Употребления *И* в качестве частицы используются для выражения аддитивного, верификативного и ряда других значений:

(16) FS_04-f -wr
Но и это его не развеселило

(17) FS_18-f-wr
хотя за нами никто и не гнался.

Наконец, особняком стоят конструкции, в которых употребление *И* было сопряжено с речевым сбоем. Они представлены, естественно, только в устном подкорпусе:

(18) FS_36-m-sp
14. ··· они побратались с /–Вовой,,,
15. после чего ==
16. и он пригласил их ··· эээ к себе в /кабинет,

В **Таблицах 1**, **2** представлены количественные данные о распределении структурных классов сочиненных групп в корпусе.

Из приведенных в **Таблицах 1**, **2** данных следует несколько примечательных выводов:

Первое. Союз *И* и в устном, и в письменном подкорпусе используется, прежде всего, для выражения предикативного сочинения: предикативное сочинение обнаруживается в письменном корпусе примерно в пять раз чаще непредикативного, а в устном — почти в десять (!) раз чаще. Соответственно, в письменном тексте несколько выше доля сочиненных непредикативных — именных, адвербиальных и адъективных — групп, которая, впрочем, в обоих подкорпусах удивительно мала.

Таблица 1. Распределение сочиненных групп (N=341) в устном подкорпусе

	Предикативное сочинение	**Непредикативное сочинение**	**Несочинительные употребления *И***
Число эпизодов (доля от общего числа употреблений *И* в подкорпусе)	270 (79,18%)	28 (8,21%)	43 (12,61%)
	в том числе:	*в том числе:*	*в том числе:*
	клаузы 176 (51,61%)	NP 24 (7,04%)	лексикализованные употребления 20 (5,86%)
	VP 94 (27,57%)	Adj\Adv 4 (1,17%)	частицы 8 (2,35%)
			сбои 15 (4,40%)

Таблица 2. Распределение сочиненных групп (N=281) в письменном подкорпусе

	Предикативное сочинение	Непредикативное сочинение	Несочинительные употребления *И*
Число эпизодов	198 (70,46%)	43 (15,3%)	40 (14,24%)
(доля от общего числа употреблений *И* в подкорпусе)	*в том числе:*	*в том числе:*	*в том числе:*
	клаузы 82 (29,18%)	NP 32 (11,39%)	лексикализованные употребления 26 (9,26%)
	VP 116 (41,28%)	Adj\Adv 11 (3,91%)	частицы 14 (4,98%)

Второе. Внутри предикативного сочинения по-разному распределено сочинение клауз и сочинение глагольных групп: в устном подкорпусе *И* используется преимущественно для сочинения клауз (и более крупных фрагментов текста), эти употребления составляют примерно половину всех вхождений *И* в подкорпусе и 65,18% (176 из 270) случаев предикативного сочинения в подкорпусе; а в письменном подкорпусе большинство употреблений составляют случаи сочинения глагольных групп: 41,28% от всех вхождений *И* в подкорпусе и 58,59% (116 из 198) случаев предикативного сочинения в подкорпусе[3].

Если считать условно, что степень синтаксической интеграции у сочиненных глагольных групп выше чем у сочиненных клауз (хотя бы потому, что для первых обязательно сочинительное сокращение), то можно сказать, что в зоне предикативного сочинения в устном тексте *И* обслуживает чаще конструкции с меньшим уровнем синтаксической интеграции, а в письменном тексте — наоборот. Посмотрим теперь, какие приемы просодической интеграции обнаруживаются в устном подкорпусе.

3. Просодические формы реализации предикативного сочинения в устном подкорпусе

Обнаруженные в корпусе просодические способы реализации предикативного сочинения были разбиты на три класса: (1) конструкции, в которых первый конъюнкт реализуется как просодически незавершенный (просодически интегрированные конструкции); (2) конструкции, в которых первый конъюнкт

[3] Один из анонимных рецензентов предложил возможное объяснение этого феномена в духе [Brown, Levinson 1987:115]: рецензент полагает, что, согласно теории вежливости, говорящий часто стремится представить новое суждение как вытекающее из предыдущего — отсюда более высокая частотность таких неспециализированных средств когезии, как английское *then* или русское И. Это, в целом, очень привлекательная гипотеза, однако, учитывая, что наш массив составляют элицитированные монологи, она нуждается в эмпирической проверке на таких жанрах дискурса, где более выражено взаимодействие локуторов, и, в первую очередь, на материале неподготовленных диалогов.

реализуется как просодически завершенный (просодически дезинтегрированные конструкции); (3) конструкции, в которых противопоставление по просодической завершенности первого конъюнкта нейтрализовано или не верифицируемо (спорные случаи). Рассмотрим эти классы подробнее.

3.1. Просодически интегрированные конструкции

Как и следовало ожидать, в нашем материале широко представлены прототипические бинарные конструкции с *И*, компоненты которых иллокутивно однородны, т.е. имеют одинаковую иллокутивную силу (преимущественно, сообщение, так как мы имеем дело с коллекцией нарративов), и первый компонент которых реализуется как просодически незавершенный. Дефолтный способ маркирования просодической незавершенности в таких конструкциях «прототипический русский подъем» с падением на заударных, если они есть [Янко 2008:31], т.е. по типу ИК-3 в терминологии интонационных конструкций [Брызгунова 1982], ср. подъем на слове *снотворное* в строке 81 в (19) с сочиненными клаузами и на слове *пропишем* в строке 43 в (20) с сочиненными глагольными группами:

(19) FS_11-m-sp
81. \Вкалывают ээ /снотв<u>о</u>рное,
82. /и-и я \улет<u>а</u>ю.

(20) FS_11-m-sp
43. ·· мы тебя сразу ·· /проп<u>и</u>шем,
44. и \прооп<u>е</u>р<u>и</u>руем.

Незавершенность перед И может также маркироваться конфигурацией типа ИК-4 — падением на ударном слоге с последующим подъемом на заударных или непосредственно на ударном слоге, если заударных нет. Эта конфигурация в русском языке, по наблюдениям [Т. Е. Янко 2008: 33, 200–225], связана со значением «рассказа по порядку», ср. движение тона на слове *рожу* в (21):

(21) FS_39-f-sp
60. она ему состроила \/р<u>о</u><u>ж</u>у,
61. {СМЕХ} ··· и гордо пошла к \эск<u>а</u>латору.

К числу просодически интегрированных конструкций можно отнести и такие, в которых между компонентами, связанными сочинительным отношением, имеется вставка, обычно уточняющая значение первого компонента. Эта вставка часто произносится как парентеза — в более узком частотном диапазоне, со сниженным уровнем громкости — и может завершаться нисходящим акцентом, но семантически сферой действия просодической незавершенности является не материал вставки, а именно фрагмент, вводимый союзом *И*. Так, в следующем примере просодическим проектором является восходящий акцент на слове *вышла* в строке 8, именно он предсказывает появление строки 12, связанной сочинительным отношением со строкой 8; во вставке — последовательность из парентетической клаузы (строки 9–10) и дискурсивного маркера *вот*:

(22) FS_27-f-sp
8. на балкон вот помню один раз /вышла,
9. (Вернее не /помню,
10. мне потом \мама рассказала.)
11. ·· –вот,
12. и меня эти соседи стали \выспрашивать всё это.

Возможна и конфигурация с так называемым «нефинальным» падением, [Кибрик, Подлесская (ред.): 152–155], т. е. падением не в самый низкий для данного говорящего уровень. Такое падение выступает в двух вариантах — адаптивном и деклинационном. Функция адаптивного нефинального падения как маркера просодической незавершенности — показать, что данное движение тона адаптируется к подъему тона в следующей коммуникативно-просодической составляющей, чтобы интегрировать текущую составляющую с последующей. Именно такая функция в (23) у падения на слове *открылись* в строке 50 — оно адаптировано к подъему в главном акценте следующей строки, на слове *закрываться*. Подъем на слове *закрываться* вызван внешними дискурсивными причинами — он маркирует адвербиальное обстоятельство. Благодаря адаптивному падению в строке 50, строки 50–51 интегрируются в качестве обстоятельства времени к строке 52:

(23) FS_39-f-sp
50. когда уже двери ··· можно сказать эээ \открылись,
51. и собирались /закрываться,
52. ·· она схватила с этого парня /–шапку,,,

Функция деклинационного нефинального падения — показать, что каждый из конъюнктов обладает некоторой коммуникативно-просодической автономностью, однако они интегрируются в единую цепочку за счет того, что каждое следующее падение в акцентном центре конъюнкта осуществляется в более низкий уровень, пока последний не будет реализован в самый низкий уровень для данного говорящего (так называемый «уровень точки»). См. (24) , где строка 13 реализуется с подготовительным падением в уровень примерно 120 Hz на слове *сметана*, а строка 14 — с финальным падением на слове *щи* в уровень примерно 85 Hz:

(24) FS_03-m-sp
10. А /я-ah || ·· я с= || см= || /смотрю,
11. у /них там ·· /рядом —
12. ·· на \стойке,
13. — стоит \сметана,
14. ·· и /явно её ·· нужно класть в \щи.

Таковы, в самом кратком описании, основные просодические паттерны, которые используются в интегрированных конструкциях с предикативным сочинением. Перейдем теперь к дезинтегрированным конструкциям.

3.2. Просодически дезинтегрированные конструкции

Просодическая завершенность перед *И* стандартно маркируется падением тона в главном фразовом акценте первого компонента по типу ИК-1 в терминологии интонационных конструкций [Брызгунова 1982], [Янко 2008].

(25) FS_39-f-sp
8. чувствовала себя ·· очень /–уверенно,,,
9. очень ·· /–красивой,,,
10. ··· всё было \замечательно.
11. ·· И вот однажды она ехала в \метро.

Условно в эту же группу «инициальных *И*» отнесены случаи, когда первый конъюнкт не расположен контактно, а находится очень далеко в предтексте или вообще отсутствует и реконструируется слушающим исходя из конситуации. Таковы *И*, которыми открывается реплика в диалоге, или *И*, которыми открывается цитируемый фрагмент после авторской ремарки. Препозитивная авторская ремарка, как и независимое сообщение, обычно реализуется с падением тона в главном фразовом акценте. Это создает уникальную дискурсивную коллизию: просодическая реализация не проецирует продолжения, но в то же время, наличие предиката речи с незаполненной валентностью, напротив, предполагает дальнейшее развертывание дискурса. Такой конфликт просодии и лексико-грамматической формы конвенционально нотируется в письменной речи с помощью двоеточия; двоеточие используется в таких случаях и в транскриптах устного подкорпуса:

(26) FS_13-f-sp
52. ··· А /потом ··ну такая /пауза,
53. /тишина,
54. и он \говорит:
55. ··· «И /ещё ·· к \тому же ·· она /землячка \Медведева!»

Одиночные дискурсивные маркеры, которые регулируют продвижение дискурсивной последовательности — как просодически неавтономные, так и просодически автономные — мы считаем прозрачными для просодической завершенности (т. е. их наличие и просодическое оформление не берутся в расчет при оценке статуса *И*). Так, в следующем примере считается, что *И* следует после просодически завершенной клаузы, хотя в строке 40 союз И расположен не в абсолютном начале строки, а после неакцентированной (просодически неавтономной) частицы *ну*:

(27) FS_18-f-sp
39. ·· Потом поезд \остановился.
40. ··· Ну и так ти́хонько-о ээ ” ээ ” мм /заговорили,
41. что типа в-вроде как \полиция приехала.

Аналогичным образом, в (28) *И* также считается инициальным, т. к. строка 4 — просодически завершенная, а строка 5 — регуляторное просодически автономное *вот*, которое реализуется с нефинальным падением:

(28) FS_25-m-sp
1. В общем ⋯ /де̲ло было,
2. ⋯ \Фё̲дор@
3. когда я ещё ·· \выпива̲л,
4. \периоди̲чески.
5. ⋯ \Во̲т,
6. и была у меня /знако̲мая одна,

Таковы, кратко, основные просодические паттерны, которые используются в дезинтегрированных конструкциях с предикативным сочинением. Перейдем теперь случаям, в которых противопоставление по просодической завершенности первого конъюнкта нейтрализовано или не верифицируемо.

3.3. Спорные случаи

Первую группу проблемных случаев составляют контексты, в которых фрагмент, предшествующий *И*, маркирован акцентом типа ИК-6 по Е. А. Брызгуновой — с подъемом на ударном слоге, за которым не следует падения на заударных. Эта конфигурация, которую условно можно назвать интонацией многоточия, особенно при растянутом ударном слоге, выражает, согласно Т. Е. Янко, значение имитации ментальной деятельности (припоминание, недоумение), однако при этом широко используется и для выражения незавершенности при описании череды событий, открытого списка [Янко 2008: 109–113, 166–167]. Получая на вход фрагмент, оформленный таким образом, слушающий допускает, что возможно продолжение, но оно жестко не проецируется. Так, в следующем примере, строки 6–7 с повтором одного и того же глагола призваны передать идею растянувшейся во времени вечеринки по случаю встречи Нового года. Главные фразовые акценты в этих строках реализуются как ИК-6. В принципе, ни лексико-грамматическое, ни просодическое оформление этих строк не проецируют их потенциальную связь со следующим фрагментом, однако и не противоречат такой связи. В той системе нотации, которая используется в корпусах [Spokencorpora 2018], такого рода омонимия дискурсивного статуса делает допустимыми два варианта нотации в строке 7 в позиции перед *И*, ср. (29а) — знак «…» плюс заглавная буква, т. е. завершенность перед инициальным *И*, открывающим новую иллокуцию:

(29а) FS_29-m-sp
6. ···· ну мы встречали его встречали-встречали-/–встреча̲ли,,,
7. ⋯ встречали дня наверное /–тр̲и…
8. ·· И потом у нас родилась революционная \иде̲я:
9. что нужно поехать в \Н̲ижний.

и (29б) — знак «,,,» (ослабленное, внутрииллокутивное многоточие — такое же, как в строке 6) плюс строчная буква, т. е. незавершенность перед *И* в составе единой иллокуции:

(29б) FS_29-m-sp
6. ···· ну мы встречали его встречали-встречали-/–встреча̲ли,,,
7. ⋯ встречали дня наверное /–тр̲и,,,

8. ·· и потом у нас родилась революционная \идея:
9. что нужно поехать в \Нижний.

Таким образом, в контекстах подобного рода противопоставление по просодической завершенности оказывается нейтрализованным.

Вторая группа проблемных случаев связана со сменой типа иллокуции на границе перед *И*. Сочетаемость иллокутивных значений вопроса, директива, обращения, других частных типов иллокуций с дискурсивной незавершенностью пока плохо изучена, поэтому не всегда удается однозначно квалифицировать наблюдаемые просодические паттерны как завершенные или как незавершенные. Так, в следующем примере строка 33 реализуется с подъемом в главном фразовом акценте на слове *можете* по типу ИК3. Этот подъем отвечает за иллокутивное значение этой строки, представляющей собой вопрос. Однако этот вопрос может быть как завершающим элементом иллокутивной цепочки — и тогда трансктипт будет иметь вид (30а) с инициальным *И* в строке 34, так и незавершающим — тогда более адекватным будет вариант транскрипта (30б), с «внутрииллокутивным» *И*:

(30а) FS_38-m-sp
32. ··· Я говорю
33. «Ну а вы не /можете сказать?»
34. И они говорят
35. «–Нет,
36. мы не \можем сказать.

(30б) FS_38-m-sp
32. ··· Я говорю
33. «Ну а вы не /можете сказать?»,
34. и они говорят
35. «–Нет,
36. мы не \можем сказать.

Аргументированный выбор между этими вариантами будет возможен лишь после тщательного изучения возможных просодических отличий между ИК3 вопроса и ИК3 незавершенности, в том числе, по таким параметрам, как диапазон, тайминг, характер возврата к начальному уровню частоты основного тона после падения на заударных и проч., а также условий совмещения значений вопроса и незавершенности в едином просодическом паттерне. В противном случае противопоставление по (не)завершенности в такого рода контекстах придется признать неверифицируемым.

Таковы, в самом сжатом изложении, ситуации, в которых фрагмент, предшествующий *И*, может быть охарактеризован как просодически завершенный, как просодически незавершенный или однозначная характеризация его просодического статуса наталкивается на определенные трудности. Обратимся теперь к некоторым общим количественным наблюдениям.

3.4. Просодическая реализация конструкций с предикативным сочинением: количественные данные

Три способа просодической реализации предикативного сочинения — с просодической завершенностью перед *И*, с просодической незавершенностью перед *И* и спорные случаи — по-разному распределены в конструкциях с сочиненными клаузами и в конструкциях с сочиненными глагольными группами. Эти различия отражены в **Таблице 3**:

Таблица 3. Просодическая (не)завершенность перед *И* при предикативном сочинении в устном подкорпусе «Веселых историй из жизни»

Тип конъюнкта	**И, всего**	**просодическая завершенность перед *И*** (доля от общего числа)		**просодическая НЕзавершенность перед *И*** (доля от общего числа)		**Спорная просодия** (доля от общего числа)	
Клауза	176	**48**	(27,27%)	**96**	(54,55%)	**32**	(18,18%)
VP	94	**8**	(8,51%)	**71**	(75,53%)	**15**	(15,96%)
Суммарно клаузы и VP	270	**56**	(20,74%)	**167**	(61,85%)	**47**	(17,41%)

Как видно из **Таблицы 3**, при предикативном сочинении — как клауз, так и глагольных групп — перед союзом *И* просодическая завершенность встречается значительно реже, чем просодическая незавершенность. Существенно однако, что при сочинении клауз незавершенность встречается примерно в два раза чаще завершенности, а при сочинении глагольных групп — в девять (!) раз чаще. Таким образом, в зоне предикативного сочинения просодическая интеграция оказывается полностью скоррелирована с синтаксической интеграцией.

Сравним теперь полученные данные с тем, как распределены пунктуационные аналоги (не)завершенности в конструкциях с предикативным сочинением в письменных версиях тех же рассказов.

4. Просодия предикативного сочинения и ее пунктуационные аналоги в письменном тексте

В письменных версиях рассказов решения о знаках пунктуации принимали сами испытуемые, поэтому есть основания считать аналогом просодической незавершенности в конструкциях с сочинением клауз — запятую с последующим строчным *И*, а в конструкциях с сочинением глагольных групп — отсутствие знака с последующим строчным *И*. Аналогом просодической завершенности будем считать точку (вопросительный знак, двоеточие, восклицательный знак, кавычки) — с последующим заглавным *И*. Распределение этих вариантов в письменном подкорпусе представлено в **Таблице 4**.

Таблица 4. Пунктуационные аналоги просодической (не)завершенности перед *И* при предикативном сочинении в письменном подкорпусе «Веселых историй из жизни»

Тип конъюнкта	*И*, **всего**	**Точка (или другой «иллокутивный» знак) плюс заглавная** *И* доля от общего числа *И* с данным типом конъюнкта		**Запятая (или отсутствие знака) + строчная** *и* доля от общего числа *И* с данным типом конъюнкта	
клауза	82	27	(32,93 %)	55	(67,07 %)
VP	116	2	(1,72 %)	114	(98,28 %)
Суммарно клаузы и VP	198	29	(14,65 %)	169	(85,35 %)

Как видно из Таблицы 4, пунктуационная незавершенность перед *И* (строчная *и*) встречается многократно чаще, чем пунктуационная завершенность (заглавная *И*). При этом в конструкциях с сочиненными клаузами сохраняется та же пропорция, что для просодической (не)завершенности в устном подкорпусе: незавершенных в два раза больше, чем завершенных. Если же речь идет о сочинении глагольных групп, то тут тенденция радикализируется: доля инициальных *И* при сочинении глагольных групп составляет всего 1,72 %, тогда как доля неинициальных , соответственно — 98,28 %. Можно предположить, что в письменном тексте авторы следуют нормативам прескриптивной пунктуации и отказываются от других опций — например, от парцелляции второго конъюнкта, тогда как в устном тексте, они свободнее оперируют просодическиой возможностью дезинтегрировать сочинительную конструкцию.

5. Итоги

Подведем краткие итоги. Проанализированные корпусные данные демонстрируют, что:

- функция предикативного сочинения является основной для союза *И* как в устном, так и в письменном тексте, причем преобладание предикативного сочинения над непредикативным особенно выражено в устных текстах;
- внутри предикативного сочинения в устном тексте выше доля сочинения клауз, а в письменном тексте преобладает сочинение глагольных групп;
- конструкции с предикативным сочинением значительно чаще реализуются как просодически интегрированные, чем как просодически дезинтегрированные, причем эта диспропорция особенно выражена в конструкциях с сочиненными глагольными группами, т. е. налицо корреляция синтаксической и коммуникативно-просодической интеграции;
- распределение по пунктуационной (не)завершенности конструкций с сочинением клауз в письменной речи дублирует распределение по просодической незавершенности в устной речи, а при сочинении глагольных групп доля пунктуационной дезинтеграции оказывается ничтожно малой.

Разумеется, обследованный корпус имеет небольшой объем для полновесных количественных выводов, однако он дает эмпирическую основу для качественного анализа, позволяющего сложить общий портрет союза *И* из разнообразия его лексико-граммматических и просодических проявлений. Дальнейшее расширение просодически размеченного массива данных позволит усилить предложенную функциональную аргументацию и повысить статистическую валидность материала.

Литература

1. *Апресян В. Ю. Пекелис О. Ю.* (2012) Сочинительные союзы. Материалы для проекта корпусного описания русской грамматики (http://rusgram.ru). На правах рукописи. М.
2. *Брызгунова Е. А.* (1982) Интонация, Русская грамматика, том 1, М.: Наука, 103–118.
3. *Кибрик А. А., Подлесская В. И.* (ред.) (2009) Рассказы о сновидениях: Корпусное исследование устного русского дискурса. М.: ЯСК.
4. *Пекелис О. Ю.* (2013) Сочинение Материалы для проекта корпусного описания русской грамматики (http://rusgram.ru). На правах рукописи. М.
5. *Санников В. З.* (2008) Русский синтаксис в семантико-прагматическом пространстве. М.: ЯСК.
6. *Урысон Е. В.* (2011) Опыт описания семантики союзов. М.: ЯСК.
7. *Янко Т. Е.* (2008) Интонационные стратегии русской речи в сопоставительном аспекте. Москва: Языки славянских культур.
8. *Auer, Peter* (2002) Projection in interaction and projection in grammar.” // Interaction and linguistic structures # 33
9. *Brown, Penelope and Stephen C. Levinson.* (1987). Politeness: Some universals in language usage. Cambridge: Cambridge University Press.
10. *Bryzgunova E. A.* (1982) Intonation [Intonatsiya], Russian Grammar [Russkaja grammatika]. Vol. 1, Nauka, Moscow, pp. 98–118
11. *Janko T. E.* (2008) Intonacionnye strategii russkoj rechi v tipologicheskom aspekte [Intonational strategies in spoken Russian from a comparative perspective]. Moskva: Jazyki Slavjanskix Kul’tur.
12. *Kibrik A. A., Podlesskaya V. I. [Eds.]* (2009) Rasskazy o snovidenijax: korpusnoe issledovanie usntogo russkogo diskursa [Night Dream Stories: A corpus study of spoken Russian discourse]. Moscow: Jazyki Slavjanskix Kul’tur.
13. *Sannikov, Vladimir* (1989) Russkie sočinitel’nye konstrukcii [Russian coordinate constructions]. Moscow: Nauka.
14. *Spokencorpora* (2018) Prosodically Annotated Corpus of Spoken Russian (PrACSRuss). Pilot version. Online: http://spokencorpora.ru.

САМОИСПРАВЛЕНИЯ ГОВОРЯЩЕГО В РУССКОМ МОНОЛОГИЧЕСКОМ И ДИАЛОГИЧЕСКОМ ДИСКУРСЕ: ОПЫТ КОРПУСНОГО ИССЛЕДОВАНИЯ

Подлесская В. И. (vi_podlesskaya@il-rggu.ru),
Коротаев Н. А. (n_korotaev@hotmail.com),
Мазурина С. И. (svet.mazurina95@gmail.com)
РГГУ, Москва, Россия

A CORPUS STUDY OF SELF-REPAIRS IN RUSSIAN MONOLOGUES AND DIALOGUES

Podlesskaya V. I. (vi_podlesskaya@il-rggu.ru),
Korotaev N. A. (n_korotaev@hotmail.com),
Mazurina S. I. (svet.mazurina95@gmail.com)
RSUH, Moscow, Russia

Self-initiated and other-initiated self-repairs (N=632) were investigated in a subcorpus (1 h 14 min) extracted from the multichannel corpus "Russian Pear Chats and Stories". The subcorpus consists of three communication sessions where participants retell and discuss the "Pear stories" film, hence each session contains both monologue and dialogue discourse parts. The overall rates of self-repairs and the distribution of their particular types were compared in monologues and dialogues. The results show that while, overall, speakers tend to repair more often in conversational than in retelling parts, particular types of repairs are distributed differently, e. g. (a) repetitions and restarts have higher rates in conversational parts, while corrections appear more often in retellings; (b) in retellings, reparandum and reparans appear more often within the same discourse unit, while in conversational parts, they tend to appear in separate discourse units.

Key words: disfluencies, speech repairs, spoken discourse, monologues, conversation

1. Постановка вопроса

В неподготовленной устной речи говорящий неизбежно сталкивается с необходимостью *самоисправлений (коррекций)*, т. е. с необходимостью повторять, изменять или отменять фрагменты дискурса, которые оказываются полностью или частично не соответствующими той речевой задаче, которую говорящий перед собой ставит[1]. С расширением возможностей электронного документирования и инструментального анализа устной речи феномен самоисправления говорящего становится объектом систематического исследования, в том числе, с использованием корпусных данных. Начиная со ставших уже классическими работ [Shriberg 1994], [Eklund 2004], изучаются разные аспекты этого явления: семантика и прагматика [Wilkinson & Weatherall 2011], [Ginzburg et al. 2014], просодия [Moniz et al. 2012], возможности автоматического извлечения самоисправлений из больших массивов текстов с использованием машинного обучения [Feng et al. 2018] и др. Изначально привлекались почти исключительно данные английского языка, но постепенно круг исследуемых языков расширяется, ср. [Fox et al. 2017] (английский, немецкий, иврит, японский, корейский, персидский, финский, бикольский, индонезийский, сочиапамский чинантекский), [Laakso & Sorjonen 2010] (финский), [Maruyama & Sano 2006] (японский), [Zhang & Chan 2013] (китайский) и др. Имеющиеся немногочисленные публикации, посвященные самоисправлениям в русском языке [Podlesskaya 2015], опираются на данные монологического нарративного дискурса, между тем материал других языков показывает, что самоисправления в диалоге и монологе могут существенно различаться и по форме, и по локализации, и по частотности. Отсюда — задача настоящей работы: на основе корпусных данных сравнить номенклатуру и частотность самоисправлений в русском монологическом и диалогическом дискурсе и выяснить, чувствительны ли к коммуникативному режиму дискурса частотность самоисправлений в целом и частотность отдельных их типов. Изложение будет строиться следующим образом. В разделе 2 кратко представлен корпус, послуживший эмпирической базой нашего исследования. В разделе 3 будет описана номенклатура обнаруженных в корпусе самоисправлений и принципы их аннотирования. Раздел 4 посвящен количественному анализу самоисправлений. В разделе 5 подводятся итоги.

2. Материал исследования

В работе используется материал мультиканального корпуса «Рассказы и разговоры о грушах» (www.multidiscourse.ru) Корпус состоит из отдельных коммуникативных эпизодов (сессий, или, условно, «записей»). В каждой сессии участвует четыре человека — Рассказчик (Narrator), Комментатор (Commentator), Пересказчик (Reteller) и Слушатель (Listener). Рассказчик и Комментатор вначале смотрят стимульный материал — так называемый «Фильм о грушах» (см. http://www.linguistics.ucsb.edu/faculty/chafe/pearfilm.htm); Пересказчик

[1] Работа выполнена при финансовой поддержке гранта РФФИ №19-012-00626

и Слушатель фильма не видят. Далее, Рассказчик рассказывает Пересказчику содержание фильма в режиме монолога. Затем наступает интерактивный этап, в течение которого Комментатор дополняет или уточняет рассказ, а Пересказчик задает вопросы обоим собеседникам, видевшим фильм. После этого появляется Слушатель, и Пересказчик пересказывает ему фильм, опять в режиме монолога. Таким образом, дизайн эксперимента мотивирует всех участников к полноценной и осмысленной коммуникации, подробнее см. [Кибрик 2018]. В каждой записи имеются как монологические, так и диалогические фрагменты дискурса. Для анализа самоисправлений нами были отобраны три записи (ниже в примерах им соответствуют коды записи 04, 16 и 22) общей продолжительностью 1 час 14 мин. Внутри каждой из трех записей были проанализированы партии Рассказчика и Пересказчика, поскольку Рассказчик и Пересказчик участвуют и в диалогических, и в монологических частях сессии, а Комментатор — только в диалогах. Таким образом, мы получили возможность сравнивать номенклатуру и частотность самоисправлений у шести локуторов в двух коммуникативных режимах: монологическом и диалогическом. Общее количество зарегистрированных самоисправлений в исследованном массиве — 632.

3. Основные классы самоисправлений

Все самоисправления были проаннотированы с учетом параметров, позволяющих с разных точек зрения оценить урон, который наносит самоисправление структурной и просодической когерентности текста. Иными словами, перечисляемые ниже признаки самоисправлений направлены на то, чтобы операционализировать характер и степень отклонений от плавного, «идеального» [Clark & Clark 1977] порождения речи.

3.1. Нарушение vs. сохранение когерентности: онлайн- и офлайн-стратегии самоисправления

Этот базовый параметр разграничивает коррекции, которые нарушают когерентость дискурса, и те, которые происходят в рамках плавного речепорождения. Онлайн-стратегия предполагает трехчастную структуру самоисправления [Levelt 1993], [Shriberg 1994]: «репарандум (фрагмент, подлежащий исправлению) / точка прерывания / репаранс (откорректированный коррелят)». В точке прерывания возможны заполненные паузы, лексические маркеры хезитации и прочие сигналы сбоя, ср. в (1) строку N-vE007, где репарандумом является слово *сзади*, которое забраковано говорящим и заменено на (репаранс) *позади*, и строку N-vE010, где репарандум — оборванное *яблоневы=*, в точке прерывания возникает лексический маркер обнаружения ошибки *ой нет*, а далее следует репаранс *грушóвое*:

(1) Pears16N[2]

N-vE007	Сзади[a] || (0.21) /позади /холма \лес.
pN-008	(0.54)
N-vE008	И (0.36) п-перед (0.27) (ə 0.40) (0.18) /хóлмом,
N-vE009	на первом /плане,
pN-009	(0.53)
N-vE010	(ə 0.22) яблоневы= ==
N-vE011	\ой,
N-vE012	\нет,
pN-010	(0.33)
N-vE013	грушóвое \дерево.

При онлайн-стратегии плавное развертывание дискурса прерывается, наличие прерывания показывает говорящему, что репарандум забракован, его следует устранить из текста, чтобы получить правильный с точки зрения говорящего фрагмент. Эта стратегия сопряжена с речевым сбоем, т. е. с нарушением лексико-грамматической и/или просодической когерентности дискурса, в том числе, с обрывом текущего фрагмента. Однако самоисправление может и не сопровождаться нарушением когерентности. Если говорящий предпочтет офлайн-стратегию, то исправление ошибки откладывается до завершения текущего речевого отрезка, который оказывается цельнооформленным и структурно, и просодически. Офлайн-стратегия не предполагает нарушений плавного развертывания речи, самоперебива и таких внешних сигналов сбоя, как например, обрывы слов. Так, в (2) говорящим постфактум была обнаружена и исправлена оговорка: вместо изначально задуманного ‘поставил корзину на багажник’ в C-vE084 ‘корзина’ была ошибочно оформлена как локативная группа. При этом исправление в строке C-vE086 не демонстрирует никаких следов нарушения плавного развертывания речевого потока (в строке C-vE085 имеется, правда, хезитационный маркер *ну*, он не задействован в самоисправлении, но может косвенно свидетельствовать о том, что говорящий уже заметил ошибку и вступает в фазу поиска выхода из этой ситуации). Репаранс этого офлайн-самоисправления снабжен дополнительным эксплицитным маркером коррекции *в смысле*:

---

[2] Индекс примера содержит отсылку к номеру записи и статусу локутора в составе корпуса (N — Рассказчик, R — Пересказчик, C — Комментатор); номера строк соответствуют нумерации строк в транскрипте записи. Об используемой системе вокальной транскрипции см. [Коротаев 2019]. Для понимания нотации в приводимых примерах достаточно знать, что текст разбивается на строки, каждая из которых соответствует элементарной дискурсивной единице (ЭДЕ) — минимальному кванту дискурса, целостность которого задается единым рематическим акцентом; движение тона в акцентированном слове указывается перед словом иконически с помощью косых черт; ударный слог в слове — носителе фразового (рематического) акцента подчеркивается; значимые движения тона в заударных и предударных слогах обозначаются иконически с помощью вертикальных стрелок; незавершенность открытого списка обозначается в транскрипте многоточием на границе иллокуции и знаком «,,,» (три запятых) внутри иллокуции; точка прерывания при речевом сбое маркируется знаком «==» на границе ЭДЕ и знаком «||» внутри ЭДЕ. Продолжительность пауз указана в скобках с точностью до сотой доли секунды.

(2) Pears04C

C-vE084	Ну он же на \корзину?!,
pC-063	(0.13)
C-vE085	(? 0.35) ну –/поставил её.
pC-064	(0.45)
C-vE086	На \/багажник в_смысле.

3.2. Изоморфизм репарандума и репаранса

Изоморфизм между репарандумом и репарансом является важнейшим средством поддержания связности дискурса при коррекции. Чем меньше сходство между репарандумом и репарансом, тем грубее нарушение. Изоморфизм проявляется в семантической близости, сходстве грамматического оформления и синтаксической функции. По признаку «наличие vs. отсутствие изоморфизма между репарандумом и репарансом» мы выделяем три класса самоисправлений: повторы, модификации и отмены (ср. близкое разграничение repetition, repair и restart, введенное в [Shriberg 1994][3]). Повтор — это случай полного изоморфизма между репарандумом и репарансом, ср. (3), где фрагментированное *В следующ=* после абсолютной и заполненной пауз повторено полностью в составе репаранса:

(3) Pears16N

N-vE015	В следующ= || (0.22) (ə 0.32) в следующем \↑кадре (0.43) показывается-а (0.21) (ɐ 0.29) (0.39) /человек,

На другом полюсе противопоставления по изоморфизму — отмены, т. е. такой тип сбоев, при котором говорящий в принципе отказывается от забракованного фрагмента, изменяет исходный замысел и переходит к новому, когерентно построенному эпизоду. Фактически, в неизоморфных онлайн-коррекциях репарандум есть, а репаранса — нет. Так, в (4) в строке R-vE310 *по пути* (репарандум) Пересказчик приступает к рассказу о встрече, которая, согласно сюжету, должна произойти дальше, но вспоминает, что забыл упомянуть некоторые детали, отказывается от уже начатого и переключается на описание шляпы и «звуков»:

(4) Pears16R

R-vE309	Едет /вперёд,
pR-174	(0.74)
R-vE310	(ɐ 0.49) по пути-и ==
R-vE311	\да,
R-vE312	у него слетает /–шляпа,,,
pR-175	(0.74)
R-vE313	(ɐ 0.32) (*Ну \да.
R-vE314	–Ещё вот его вот /звуки идут на заднем /фоне,

[3] Это тройственное деление считается общепризнанным и в той или иной версии эксплуатируется практически во всех работах про самоисправления последних десятилетий, ср., например, «нулевые, частичные и полные отмены» в [Богданова-Бегларян 2013].

Между повторами, демонстрирующими полный изоморфизм между репарандумом и репарансом, и отменами, демонстрирующими отсутствие изоморфизма, располагается третий класс самоисправлений — модификации, демонстрирующий частичный изоморфизм. Репарандум и репаранс при модификациях могут быть изоморфны по различным признакам. Это может быть замена лексемы на близкую по смыслу в той же грамматической форме, ср. (5), или замена грамматической формы одной и той же лексемы, ср. смена рода в (6):

(5) Pears16N

N-vE107	Он засматривается (ə 0.09) на эту девоч= || (0.14) на эту /↑девушку,

(6) Pears22N

N-vE156	если \один-н || (ɥ 0.26) (ˀ 0.11) \одну корзину —
N-vE157	(полностью \полную,)
N-vE158	— /увезли,
N-vE159	почему другая полностью \пустая?

При модификации может происходить расширение репарандума, в том числе, за счет присоединения дополнительного фрагмента слева, ср. понятие insertion repair в [Wilkinson & Weatherall 2011]. Таково в (7) присоединение фрагмента *абсолютно* перед отрицанием:

(7) Pears04N

N-vE083	его нь= || абсолютно не \↑видит,

К классу модификаций мы относим и коррекции, при которых в качестве репарандума выступают так называемые маркеры препаративной подстановки, или плейсхолдеры [Podlesskaya 2015] — выражения чаще всего местоименного типа, которые говорящий использует в качестве временного заместителя того фрагмента, который не может своевременно встроить в дискурс. В примере ниже плейсхолдер *такая* изоморфен репарансу *багажник* по падежу, числу и синтаксической функции (подлежащее); в то же время при коррекции сменился род — во всей видимости, кандидатом на это место могло быть слово *подставка*; трудность поиска сигнализируется и уточняющими ремарками в последующих строках R-vE278 — R-vE279:

(8) Pears16R

R-vE277	(ɐ 0.39) \спереди у него такая || (0.83) \/↑багажник,
pR-162	(0.26)
R-vE278	(это — называется,)
pR-163	(0.25)
R-vE279	(ɐ 0.36) (\решётчатый,)

3.3. Линейный диапазон коррекций

Этот параметр регулирует утяжеление сбоя в зависимости от того, насколько далеко друг от друга расположены репарандум и репаранс в линейной развертке текста; естественно, он релевантен только для повторов и модификаций, так как

при отменах репаранса как такового нет. Наименее травматичными являются контактные коррекции — такие, при которых репаранс следует за репарандумом непосредственно, или между ними есть только пауза (абсолютная или заполненная): таковы самоисправления в (3), (6), (7) выше или строка N-vE007 в (1). Далее следуют коррекции «средней» дистанции, при которых репаранс отодвинут дискурсивным маркером, как строках N-vE010 — N-vE013 в (1), и «дальней» дистанции, в которых между репарандумом и репарансом имеются слова с пропозициональным значением, ср. (9), где конкурирующие *кульмина=* и *развязка* отделены несколькими словами:

(9) Pears22R

R-vE246	ну в /общем видимо это кульмина= ==
R-vE247	(ə 0.14) в_смысле \да,
R-vE248	уже-е \развязка /картины,

3.4. Структурный диапазон коррекций

Этот параметр регулирует утяжеление сбоя в зависимости от того, насколько далеко друг от друга расположены репарандум и репаранс в структуре дискурса. Как и линейный диапазон, структурный диапазон значим только для повторов и модификаций. По этому параметру наименее травматичными являются «микрокоррекции», в которых и репарандум, и репаранс расположены внутри одной ЭДЕ, как в (3) или (6). Далее следуют «однотактные макрокоррекции», в которых репарандум и репаранс расположены в соседних ЭДЕ:

(10) Pears16R

R-vE430	что груш н= ==
R-vE431	что корзины \нет,

Наконец, наиболее травматичные коррекции по этому параметру, условно «многотактные макрокоррекции», — те, в которых затрагиваются блоки, большие, чем пара последовательных ЭДЕ. В этих случаях исправление может требовать пространного репаранса, как в (11), где говорящему не удается в один ход завершить строку N-vE019; там предположительно могло было бы быть *как он снимает груши*, но вместо этого предлагается более дробное развертывание последовательности событий в кадре — сначала ‘руки’, потом ‘весь’:

(11) Pears16N

N-vE019	(ɐ 0.13) сначала (ə 0.12) (0.15) показывается как он с-с= ==
N-vE020	(ə 0.20) только –/↑руки его,
N-vE021	как они-и (0.11) просто –снимают,
N-vE022	потом и весь –/сам,

Многотактные макрокоррекции возникают и тогда, когда говорящий преждевременно приступает к некоторой ЭДЕ, но вынужден вернуться на шаг назад, достроить предыдущий фрагмент, а затем вновь вернуться к текущему. Так, в (12) в строке N-vE046 описываются действия козы, строка N-vE047 — это преждевременная попытка перейти к описанию действий ее хозяина — сказать, предположительно, *но он не даёт ей съесть груши*; далее говорящий

возвращается в строке N-vE048 к запоздалой интерпретации поведения козы и только потом возобновляет линию рассказа:

(12) Pears16N

N-vE045	— проходит (0.11) (ə 0.19) (0.06) \человек с-с /козо̲й,
N-vN005	(ɥ 0.42)
N-vE046	она-а (0.14) так пытливо смотрит на-а эти /гру̲ши,
N-vE047	но-о он н-нь= ==
pN-033	(0.15)
N-vE048	видимо ʔона хочет их /съе̲сть,
N-vE049	но-оʔ (0.05) человек не \даё̲т.

3.5. Объем репарандума

По этому параметру менее травматичны коррекции, при которых забракованный фрагмент условно отнесен нами к классу «мелких» — он не содержит ни одного полнозначного слова, т. е. это может быть фрагмент полнозначного слова , как *яблоневы*= в (1), или служебные слова — изолированно или в сочетании с фрагментами полнозначных, как *В следующ*= в (3). Более травматичны «крупные» коррекции, в которых забракованный фрагмент включает хотя бы одно полнозначное слово, ср. *что груш н*= в (10).

3.6. Способ инициации самоисправления

По этому параметру разграничиваются коррекции, где инициатором самоисправления является сам говорящий, испытывающий трудности речепроизводства, и коррекции, которые вызваны вмешательством собеседника, условно — «внутренние» и «внешние» (ср. противопоставление self-initiated и other-intitiated repair, выработанное в рамках анализа бытового диалога, i. a. [Schegloff et al 1977]). Все примеры, которые приводились нами до сих пор, демонстрировали внутренние коррекции. Внешние коррекции можно считать более травматичными, поскольку они затрагивают когерентность всего дискурсивного пространства, а не только партию единичного локутора.

Внешние коррекции часто связаны с тем, что собеседник вступает в коммуникацию, не дожидаясь окончания реплики текущего локутора, в связи с чем текущий локутор бывает вынужден свою реплику оборвать. В корпусе «Рассказы и разговоры о грушах» имеется особый способ представления разметки коммуникативного взаимодействия — так называемая партитурная запись, которая позволяет увидеть взаимную привязку всех коммуникативных событий ко временно́й шкале. В примере (13) представлен редуцированный фрагмент партитурной записи эпизода, где Комментатор начинает строку C-vE134 в момент 413,13 сек от начала записи, но еще до ее завершения, в момент 413,51 сек, вступает Пересказчик (R-vE010). Комментатор обрывает строку C-vE134 уже «внутри» речи Пересказчика, в момент 413,78 сек, и больше уже к ее содержанию не возвращается. Обрыв строки по причине вмешательства собеседника нотируется в транскрипте двойным волнистым знаком равенства (≈≈):

(13) Pears22C_R

	411.04	411.40	C-vN038	(ц 0.35)		
	411.40	412.81	C-vE132	И он на них так /смо̲трит подозрительно,		
	412.81	413.13	C-vE133	(\вслед̲,)		
	413.13		C-vE134	что типа они ≈≈	R-vE010	А /мальчик (ə 0.52) —
		413.78				
		414.59				
(0.23)	414.59	414.82				
	414.82	415.84			R-vE011	(ʔ 0.10) (*тот который* на \велосипеде,)
	415.84	416.75			R-vE012	— увёз /–корзи̲ну,
	416.75	418.01			R-vE013	в которой больше / всего̲ было груш?
(0.36)	418.01	418.38				
	418.38	418.64	C-vN039	{sm 0.27}		
	418.64		C-vE135	/Ну \да̲,		

Однако вмешательство собеседника может приводить и к более серьезной перестройке. Так, в примере (14) строка R-vE300 оборвана, поскольку Комментатор (с некоторым опозданием) замечает, что Пересказчик в строке R-vE299 дал ошибочное название фрукту (*яблоко* вместо *грушу*), и поправляет того (строка C-vE238), причем R-vE300 и C-vE238 собеседники произносят практически одновременно, временны́е координаты этих реплик различаются на сотые доли секунды. Дальше в партии Пересказчика следует комплексная коррекция — сначала в строке R-vE301 он исправляет название фрукта, а затем в R-vE302 возвращается к исполнению брошенной строки *смотрит*:

(14) Pears16C_R

	1170.00	1173.26			R-vE299	Значит он берёт || (0.17) \хочет сначала взять \одно̲ /я̲блоко,
(0.25)	1173.26	1173.52				
	1173.52				R-vE300	смотрит ≈≈
	1173.58	1174.09	C-vE238	\↑Гру̲шу.		
		1174.15				
(0.38)	1174.15	1174.53				
	1174.53	1175.32			R-vE301	одну /гру̲шу,
	1175.32	1175.62			R-vN012	(ц 0.30)
	1175.62	1176.29			R-vE302	/смо̲трит,
	1176.29	1178.98			R-vE303	с= || что-о и= || / игнори̲руют,
	1178.98	1180.12			R-vE304	(как бы –его̲,)

Таковы, в вынужденно сжатом изложении, основные классы самоисправлений говорящего, представленные в исследованном нами подкорпусе.

В следующем разделе приведены количественные данные об общей частотности самоисправлений и частотности некоторых их отдельных классов в двух коммуникативных режимах — монологе и диалоге.

4. Количественный анализ самоисправлений

В партиях Рассказчика и Пересказчика в трех исследованных нами записях нами было зарегистрировано в общей сложности 632 коррекции, из которых 94,6% (598) составляют онлайн-коррекции. В табл. 1 показано, как соотносится общая частотность онлайн-коррекций в диалоге и монологе. В таблице приведены как абсолютные цифры для партии каждого локутора в монологе и диалоге, так и число коррекций, приведенное ко времени вокализации и числу слов в соответствующей (монологической или диалогической) части партии. Под временем вокализации понимается суммарная продолжительность всех ЭДЕ (включая и любые типы пауз и неречевых звуков внутри ЭДЕ), а также изолированных (находящихся вне ЭДЕ) заполненных пауз, смеха, прочих неречевых вокальных действий и пауз, заполненных громкими вдохами. (Таким образом, не включается во время вокализации продолжительность абсолютных пауз, располагающихся между ЭДЕ и / или другими единицами верхнего уровня сегментации; см. [Коротаев 2019]). В число слов, помимо «обычных» словоформ, также включены заполненные паузы и смех (в том случае если он не накладывается на произнесение других единиц).

Таблица 1. Распределение общего числа онлайн-коррекций по коммуникативным режимам (в абсолютных значениях и в приведении к числу слов и времени вокализации) у шести говорящих размеченного подкорпуса

Говорящий	Режим	Время вокализации, с	Число слов	Всего онлайн-коррекций	Коррекций на 100 с	Коррекций на 100 слов
04N	Монолог	290,65	731	31	10,67	4,24
	Диалог	190,41	539	36	18,91	6,68
04R	Монолог	325,06	797	29	8,92	3,64
	Диалог	299,84	876	60	20,01	6,85
16N	Монолог	277,06	730	32	11,55	4,38
	Диалог	477,87	1387	130	27,20	9,37
16R	Монолог	342,04	807	60	17,54	7,43
	Диалог	543,29	1323	115	21,17	8,69
22N	Монолог	211,59	585	29	13,71	4,96
	Диалог	209,84	675	29	13,82	4,30
22R	Монолог	293,21	796	43	14,67	5,40
	Диалог	93,37	190	4	4,28	2,11
Всего	Монолог	1739,61	4446	224	12,88	5,04
	Диалог	1814,62	4990	374	20,61	7,49

Как видим, генеральная тенденция состоит в том, что в диалогических частях записей коррекции встречаются чаще, чем в монологических. Однако индивидуальные дискурсивные предпочтения говорящих явным образом влияют на общую картину[4]. Пересказчица в записи 22 фактически уклоняется от ведения беседы, диалогическая часть ее партии многократно короче, чем у других локуторов; кроме того, если обратиться к медиафайлам и транскриптам корпуса, то мы увидим, что в ее диалоге мало пропозиционального содержания, он наполовину состоит из сигналов обратной связи типа *Угу*. Неудивительно, что в нем обнаружилось всего четыре самоисправления. Поэтому нарушения генеральной тенденции у 22R можно отнести на счет ненадежности количественных данных по этой партии. Возможно, косвенным образом это задевает и результаты по партии ее собеседника 22N, у которого частоты коррекции в монологе и диалоге различаются очень слабо.

Перейдем к анализу частот отдельных типов самоисправлений. Отметим, что мы исключили из подсчетов данные по говорящей 22R, учитывая ее слабую занятость в диалоге (см. выше). Чувствительными к коммуникативному режиму оказались два параметра: «наличие vs. отсутствие изоморфизма между репарандумом и репарансом» и «структурный диапазон коррекции». Различающиеся по признаку «наличие vs. отсутствие изоморфизма между репарандумом и репарансом» повторы, модификации и отмены по-разному распределены в монологе и диалоге — 52:89:41 в монологе, 126:80:165 в диалоге, см. **рис. 1**. Как видим, в монологе преобладают модификации, а в диалоге — повторы и отмены.

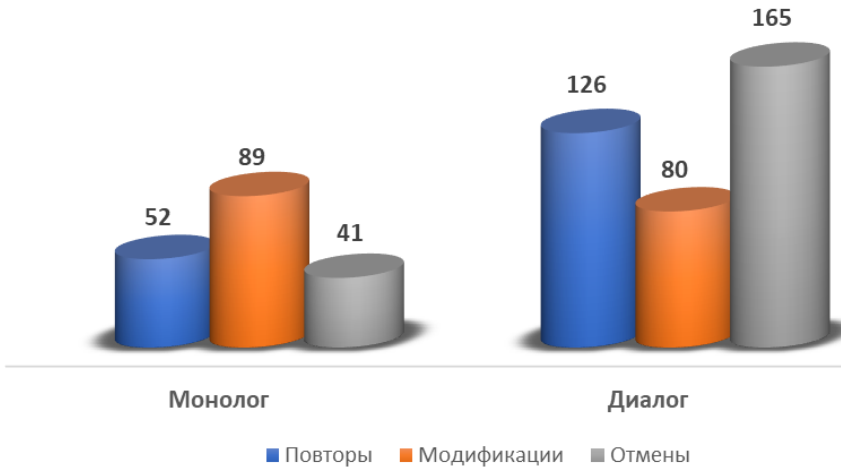


Рис. 1. Распределение онлайн-коррекций по параметру «наличие vs. отсутствие изоморфизма между репарандумом и репарансом» в двух режимах коммуникации (суммарные данные по пяти говорящим)

[4] О соотношении универсальных тенденций и индивидуальных особенностей в мультиканальной коммуникации (тоже на материале «Рассказов и разговоров о грушах») см. также [Федорова, Кибрик 2018].

Наблюдаемые различия в распределениях обладают статистической значимостью ($p < 0{,}001$ при оценке методом «хи-квадрат»). По данному параметру из общей картины несколько выбивается говорящая 04N: хотя в ее речи характер распределения типов онлайн-коррекций в монологе vs. в диалоге похож на общий случай, наблюдаемые различия не имеют статистической значимости ($p > 0{,}1$).

Различающиеся по признаку «структурный диапазон» микро- и макрокоррекции (суммарно одно- и многотактные) также по-разному распределены в монологе и диалоге — 114:68 в монологе, 175:195 в диалоге, см. **рис. 2**. Как видим, в монологах преобладают коррекции, в которых и репарандум, и репаранс размещены внутри одной ЭДЕ, в диалогах реперандум и репаранс чаще расположены в разных, иногда даже не в соседних ЭДЕ. Различия в распределении обладают статистической значимостью по критерию Фишера ($p < 0{,}001$). При этом есть и индивидуальное варьирование: у 4N и16R значимых различий по этому параметру нет.

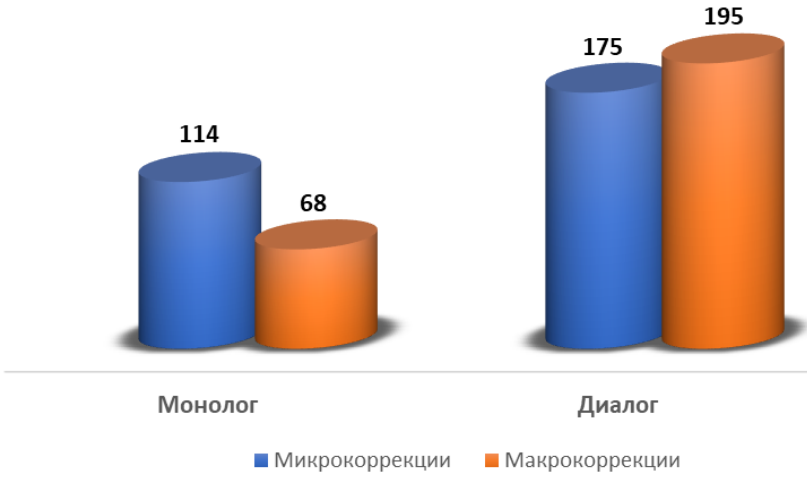


Рис. 2. Распределение онлайн-коррекций по параметру «структурный диапазон коррекции» в двух режимах коммуникации (суммарные данные по пяти говорящим)

По параметрам «линейный диапазон коррекции» и «объем забракованного фрагмента» различий между коммуникативными режимами нет. Контактные коррекции и в монологе, и в диалоге встречаются многократно чаще дистантных; распределение «контактные» : «средняя дистанция» : «длинная дистанция» в монологе — 102:26:11, в диалоге — 165:22:21. Статистически значимого различия нет (критерий «хи-квадрат», $p > 0{,}05$). Индивидуальное отличие замечено только у 4R — значимое повышение доли коррекций средней дистанции в монологе. Мелкие забракованные фрагменты встречаются незначительно чаще в монологе, а крупные — незначительно чаще в диалоге: монолог — 94:88, диалог — 170:200. Однако статистически значимого различия нет

(критерий Фишера, $p > 0{,}02$). Единственный говорящий, у которого это различие достигает статистической значимости, — 16R.

По параметру «способ инициации коррекции» сравнение не проводилось, поскольку коррекции, инициированные собеседником, по определению, представлены только в диалоге.

5. Заключение

Подведем итоги. Мы проанализировали три коммуникативных эпизода (сессии) с участием шестерых говорящих, включающих как монологические (пересказы самостоятельно просмотренного видеофильма или пересказы сюжета видеофильма со слов собеседника), так и диалогические фрагменты (обсуждение фильма). В этом массиве была произведена разметка самоисправлений говорящего, которая учитывала следующие параметры:

- приводит ли самоисправление к нарушению структурной и просодической плавности разворачиваемого дискурса (онлайн- vs. офлайн-коррекции);
- сходны ли по форме и функции забракованный фрагмент, или репарандум, и его исправленный коррелят, или репаранс (изоморфизм репарандума и репаранса);
- далеко ли отстоят репарандум и репаранс в линейной развертке текста (линейный диапазон коррекции);
- происходит ли самоисправление в одной ЭДЕ или затрагивает две и больше ЭДЕ (структурный диапазон коррекции);
- включает ли забракованный фрагмент хотя бы одно полнозначное слово (объем репарандума);
- инициировано ли самоисправление самим говорящим или оно вызвано вмешательством собеседника (способ инициации коррекции).

Анализ общей частотности коррекций и частотности их отдельных типов показал следующее:

1. В целом самоисправления чувствительны к коммуникативному режиму: их общая частотность выше в диалоге, чем в монологе.
2. К коммуникативному режиму чувствительно распределение коррекций по признаку «изоморфизм репарандума и репаранса» (в диалогах больше отмен и повторов, в монологах больше модификаций), а также распределение по признаку «структурный диапазон коррекции» (в монологах больше коррекций в пределах ЭДЕ, в диалогах коррекции чаще выходят за пределы ЭДЕ).
3. Чувствительность к коммуникативному режиму по признаку «линейный диапазон коррекции» и «объем репарандума» не прослеживается.

Эти предварительные количественные данные позволяют предположить, что такие параметры коррекций, как степень изоморфности репарандума и репаранса и структурный диапазон коррекции, т. е. степень ее компактности при размещении в иерархической структуре дискурса, в большей степени ориентированы на эффективность взаимодействия с собеседником. В то же время,

по-видимому, такие параметры коррекций, как объем репарандума и его линейное расстояние от репаранса, в большей степени ориентированы на внутренние процессы говорящего, связанные с проблемами речепорождения, в том числе, с проблемами выбора адекватных способов вербализации исходного речевого замысла. Разумеется, полученные данные должны быть в дальнейшем проверены на расширенной выборке, тем более что обнаруженные нами тенденции демонстрируют заметную индивидуальную вариативность.

Литература

1. *Богданова-Бегларян Н. В.* (ред.) (2013), Звуковой корпус как материал для анализа русской речи. Коллективная монография. Часть 1. Чтение. Пересказ. Описание. СПб.: Филологический факультет СПбГУ.
2. *Кибрик А. А.* (2018), Русский мультиканальный дискурс. Часть I. Постановка проблемы // Психологический журнал 39(1). С. 70–80.
3. *Коротаев Н. А.* (2019), «Рассказы и разговоры о грушах»: принципы вокальной аннотации. Версия 10.01.2019, http://multidiscourse.ru.
4. *Федорова О. В., Кибрик А. А.* (2018), Общее, индивидуальное и контекст в мультиканальной коммуникации // Когнитивные исследования языка 33. С. 637–645.

References

1. *Bogdanova-Beglarian N. V.* (ed.) (2013), Speech corpus as a base for analysis. Collective monograph. Part 1. Reading. Retelling. Description [Zvukovoj korpus kak material dlja analiza russkoj reči. Kollektivnaja monografija. Čast' 1. Čtenie. Pereskaz. Opisanie], Saint-Petersburg.
2. *Clark, H. H., Clark, E.* (1977), Psychology and language: An introduction to
3. *psycholinguistics, Harcourt Brace, New York.*
4. *Eklund R.* (2004), Disfluency in Swedish human-human and human-machine travel booking dialogues, Unitryck, Sweden.
5. *Fedorova O. V., Kibrik A. A.* (2018), General, singular and the context in multichannel communication [Obščee, individual'noe i kontekst v mul'tikanal'noj kommunikacii], Cognitive studies of language [Kognitivnye issledovanijz jazyka], 33, pp. 637–645.
6. *Feng Wang, Wei Chen, Zhen Yang, Qianqian Dong, Shuang Xu, Bo Xu* (2018), Semi-Supervised Disfluency Detection, Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA, August 20–26, pp. 3529–3538
7. *Fox B., Wouk F., Fincke S., Hernandez Flores W., Hayashi M., Laakso M., Maschler Yael M., Sorjonen M.-L., Uhmann S., Yang Hyun Jung* (2017), Morphological self-repair: Self-repair within the word, Studies in Language, Volume 41, Issue 3, pp. 638–659.

8. *Ginzburg J., Fernández, R., Schlangen, D.* (2014), Disfluencies as intra-utterance dialogue moves, Semantics and Pragmatics, 7, pp. 1–64.
9. *Kibrik A. A.* (2018), Russian multichannel discourse. Part I. Setting up the problem [Russkij mul'tikanal'nyj diskurs. čast' I. Postanovka problemy], Psixologiceskij žurnal, Vol. 39 (1), pp. 70–80.
10. *Korotaev N. A.* (2019), "Russian Pear Chats and Stories": Vocal annotation guide. Version 10.01.2019, available at: https://www.multidiscourse.ru/data/ann/pears_vocal%20annotation.pdf.
11. *Laakso, M., Sorjonen, M.-L.* (2010), Cut-off or particle — devices for initiating self-repair in conversation, Journal of Pragmatics, 42(4), pp. 1151–1172.
12. *Levelt, W.* (1983), Monitoring and Self-Repair in Speech, Cognition, 14, pp. 41–104.
13. *Maruyama T., Sano S.* (2006), Classification and annotation of self-repairs in Japanese spontaneous monologues, Linguistic Patterns in Spontaneous Speech, Taipei, November 2006, pp. 283–298.
14. *Moniz H., Batista F., Mata A. I., Trancoso I.* (2012), Analysis of disfluencies in a corpus of university lectures, ExLing 2012, Proceedings of 5th Tutorial and Research Workshop on Experimental Linguistics, Athens, Greece, pp. 96–99.
15. *Podlesskaya V. I.* (2015), A corpus-based study of self-repairs in Russian spoken monologues, Russian Linguistics, Vol. 39, Issue 1, pp. 63–79.
16. *Schegloff, E. A., Jefferson, G., Sacks H.* (1977), The preference for self-correction in the organization of repair in conversation, Language 53(2), 361–382.
17. *Shriberg, E.* (1994), Preliminaries to a Theory of Speech Disfluencies, University of California in Berkeley.
18. *Wilkinson, S., Weatherall, A.* (2011), Insertion repair, Research on Language & Social Interaction, 44(1), pp. 65–91.
19. *Zhang, W., Chan, A.* (2013), Self-repair in Mandarin and Cantonese: delaying the next item due in casual conversation and news interview, Chinese Discourse and Interaction: Theory and Practice, Equinox Publishing Ltd, pp. 35–57.

MEASURE CLUSTERING APPROACH TO MWE EXTRACTION[1]

Rossyaykin P. O. (petrrossyaykin@gmail.com),
Loukachevitch N. V. (louk_nat@mail.ru)
Lomonosov Moscow State University, Moscow, Russia

In this paper we present an unsupervised and resource-independent approach to the well-known task of discovery of multiword expressions (MWE) in text corpora. We experimented on extracting Russian nominal phrases (Adj-N and N-N.Gen) relevant for lexical resources (thesauri, WordNet, etc.). Our approach is based on the assumption that idiosyncrasy of MWEs can be due to different properties (morphosyntactic, semantic, pragmatic and statistical), and thus, different types of measures (statistical, context, distributional) are efficient at extracting different MWEs. We propose new context measures as well as an unsupervised method of combining measures in which we cluster vectors of ranks assigned by individual measures. The proposed method accounts for different properties of MWEs and allows surpassing both individual measures and their simple sum/product.

Key words: multiword expressions (MWEs), MWE extraction, association measures, context measures, distributional semantics, clustering, thesaurus, Russian language

ИЗВЛЕЧЕНИЕ MWE НА ОСНОВЕ КЛАСТЕРИЗАЦИИ МЕР

Россяйкин П. О. (petrrosyaykin@gmail.com),
Лукашевич Н. В. (louk_nat@mail.ru)
МГУ им. М. В. Ломоносова, Москва, Россия

Ключевые слова: устойчивые словосочетания, извлечение устойчивых словосочетаний, меры ассоциации, контекстные меры, дистрибутивная семантика, кластеризация, тезаурус, русский язык

1. Introduction

MWEs, also called collocations or multiword units (MWU), have a long history in NLP. Numerous definitions of MWE were proposed in the literature on both theoretical and computational linguistics. All of them emphasize two core features of MWEs:

[1] The reported study was partially funded by RFBR according to the research project № 18-00-01226 (18-00-01240).

1) they are 'words with spaces', i.e. sequences of graphical words not shorter than 2 words and 2) they exhibit unusual, unpredictable properties at any level of linguistic analysis. We, thus, adopt a broad definition proposed by [Baldwin & Kim 2010]:

> "Multiword expressions (MWEs) are lexical items that:
> (a) can be decomposed into multiple lexemes; and
> (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity"

This definition accounts for expressions of very different nature: idioms (*to kick the bucket*), which are semantically idiosyncratic, lexicalized expressions (*black and white television*) displaying statistical idiosyncrasy, terms (*vowel harmony*), which are usually idiosyncratic both semantically and statistically, proper names (*South Korea*), morphologically/syntactically rigid expressions (*by and large*), etc. Information on MWEs of all of these types is crucial for performance in many NLP tasks and applications: machine translation [Ren et al. 2009]; [Carpuat & Diab 2010], syntactic parsing [Korkontzelos & Manandhar 2010], word sense disambiguation [McCarthy et al. 2004], etc.

The number of MWEs in the language is comparable to that of single lexemes or even surpasses it [Jackendoff 1997]; [Sag et al. 2002]. Moreover, new MWEs appear constantly which makes manual compilation of MWE lists inefficient. This prompts the task of automatic MWE extraction (or discovery), which consists in providing a ranked list of expressions that can be either examined and refined by human experts or used in further applications as is.

The goal of this work is to elaborate a method to supplement lexical resources with MWEs. Hence, we focus on nominal phrases (Adj-N and N-N.Gen) of all semantic types mentioned. Despite being composed of multiple lexemes, they should be included in lexical resources as single entries due to 1) their correspondence to single entities on the ontological level and 2) impossibility to account for them with any regular rules of syntax and/or semantics.

We deal with Russian data; however, the methods discussed and proposed in this paper are mostly language-independent — they require only a text corpus and basic pre-processing (lemmatizing and PoS-tagging).

The structure of paper is as follows: we provide a short overview of previous work (paragraph 2), describe our corpus and the set of candidate expressions (3), introduce individual features used in our combinational method: most common lexical association measures, original context measures which yield good results on their own and two state-of-the-art distributional measures (4), introduce a new clustering-based approach to combining measures (5), provide and discuss results (6), (7).

2. Related work

Starting with the papers by [Choueka 1988] and [Church & Hanks 1990] statistical methods prevail in MWE extraction. The only information used by statistical association measures is frequency distribution of words in the corpus, in most cases number of occurrences and co-occurrences of MWEs' components (collocates). Numerous

association measures, previously used in other tasks, were adapted to MWE extraction: PMI [Church & Hanks 1990], t-score [Church et al. 1991], log-likelihood ratio [Dunning 1993] are among the most popular.

Crucially, association measures are restricted by design in their ability to discover MWEs. They take advantage of just one property of MWEs (statistical idiosyncrasy), which is irrelevant for certain expressions (e.g. *red tape* with both components being very frequent independently). Moreover, they do not take into account semantic and statistical asymmetry of most MWEs and can be biased to either rare or frequent bigrams (for further criticism of association measures see [Evert 2007: 7.1]).

Alternative approach is based on detecting non-compositionality with the help of either context measures [Nakagawa & Mori 2003], [Riedl & Biemann 2015] or distributional semantics [Lin 1998], [Padó & Lapata 2007], [Van de Cruys & Moirón 2007], [Baroni & Zamparelli 2010]. The introduction of word2vec [Mikolov et al. 2013] triggered a new surge of research in distributional semantics with word embeddings being adapted to different tasks including MWE extraction. In most cases these methods are designed for either particular syntactic patterns (phrasal verbs — [Baldwin et al. 2003], [Salehi et al. 2015]; V-N idioms — [McCarthy et al. 2007], [Senaldi et al. 2016] or lexical types of MWEs [Rodríguez-Fernández et al. 2016], [Enikeeva & Mitrofanova 2017].

Most recent papers deal with combining different methods rather than individual association or distributional measures. [Pecina & Schlesinger 2006] used hierarchical clustering to select a set of statistical and context features and different machine learning algorithms to provide ranking function. [Tsvetkov & Wintner 2011], [Buljan & Šnajder 2017] connected statistical and morphosyntactic measures in a Bayesian network. [Tutubalina & Braslavski 2016] adopted learning-to-rank methods from information retrieval.

Unsupervised approaches to combining measures are much less common. [Zakharov 2017] combined association measures by averaging ranks of MWEs obtained with the use of individual measures. [Tutubalina 2015] used clustering in 2-dimensional space with log-likelihood ratios calculated on 2 different corpora. In contrast to this method, we used one corpus and measures of different nature (statistical, context, distributional) as dimensions. We assume that such an approach allows separating MWEs of different types from free phrases.

3. Data

The corpus we experimented on was composed of news' texts from the Russian Internet published in 2011. We deleted all punctuation, lemmatized and uppercased it, PoS-tagging was used in order to obtain the initial list of candidate Adj-N and N-N bigrams. Bigrams with the observed frequency of less than 200 were excluded, resulting in the list of 37,767 candidate expressions. Given PoS-filtering and a high frequency threshold, we suppose that our dataset contained no bigrams which systemically were not actual syntactic constituents.

We used the Russian language thesaurus RuThes [Loukachevitch et al. 2014] as our gold standard. Expressions present in it were regarded as actual MWEs (9,837 in total). Our task, thus, was to provide a method which would rank these 9,837 expressions on the top of the list.

4. Individual measures

4.1. Overview

We calculated 22 statistical association measures including the most popular ones (PMI and its variants, t-score, LLR, Dice coefficient, etc.) as well as less common measures which showed good performance in previous comparative studies [Pecina 2008], [Hoang et. al 2009]. 5 asymmetric variants of MI and PMI proposed by [Hoang et al. 2009] and [Carlini 2014] were also added to our comparison.

8 context measures, 4 of which are introduced in this study (see detailed description below), were calculated to obtain a more semantics-based view on our dataset. Formulae for all 30 individual measures are presented in **Table 1**.

Table 1. Statistical association and context measures used for ranking MWE candidates

Name	Formula
frequency	$f(xy)$
PMI	$\log \frac{P(xy)}{P(x)P(y)}$
Sørensen–Dice coefficient (DC)	$\frac{2 * f(xy)}{f(x) + f(y)}$
log-likelihood ratio	$2 \sum_{x,y} f(xy) * \log \frac{p(xy)}{p(x) * p(y)}$
chi-square	$\frac{(f(xy) - \frac{f(x) * f(y)}{N})^2}{f(x) * f(y)}$
Piatetsky-Shapiro coefficient	$P(xy) - P(x) * P(y)$
t-score	$\frac{P(xy) - P(x) * P(y)}{\sqrt{\frac{P(xy)}{N}}}$
geometric mean	$\frac{f(xy)}{\sqrt{f(x) * f(y)}}$
normalized PMI	$\frac{PMI(xy)}{-\log(P(xy))}$
odds ratio	$\log \frac{(f(xy) + \frac{1}{2})(f(\bar{x}\bar{y}) + \frac{1}{2})}{(f(x\bar{y}) + \frac{1}{2})(f(\bar{x}y) + \frac{1}{2})}$
Poisson significance measure	$((P(xy) * N - f(xy) * \log(P(xy) * N) + \log(f(xy)!))/ \log N$
modified DC	$\log\big(f(xy)\big) * DC(xy)$
Confidence	$\max(P(y\|x), P(x\|y))$

Name	Formula
local PMI	$f(xy) * PMI(xy)$
augmented PMI	$\log \frac{P(xy)}{P(x\bar{y})P(\bar{x}y)}$
cubic PMI	$\log \frac{P(xy)^3}{P(x)P(y)}$
normalized MI	$\frac{\sum_{x,y} P(xy) * \log \frac{P(xy)}{P(x) * P(y)}}{-\sum_{x,y} P(xy) * \log P(xy)}$
MI/NF(0.5)	$\frac{MI}{0.5 * P(x) + 0.5 * P(y)}$
PMI/NF(0.77)	$\frac{PMI}{0.77 * P(x) + 0.23 * P(y)}$
MI/NFmax	$\frac{MI}{\max(P(x), P(y))}$
PMI/NFmax	$\frac{PMI}{\max(P(x), P(y))}$
NPMIC	$\frac{PMI(xy)}{-\log(P(x))}$
gravity count (GC)	$\log \frac{f(xy) * \lvert r(x)\rvert}{f(x)} + \log \frac{f(xy) * \lvert l(y)\rvert}{f(y)}$
modified GC	$\log(\frac{f(xy) * \lvert r(x)\rvert}{f(x)} + \frac{f(xy) * \lvert l(y)\rvert}{f(y)})$
type-LR	$\sqrt{\lvert r(x)\rvert * \lvert l(y)\rvert}$
type-FLR	$\frac{f(xy)}{typeLR(xy)}$
context intersection (CI)	$\frac{\lvert l(xy) \cap l(x)\rvert}{\lvert l(x)\rvert} * \frac{\lvert r(y) \cap r(xy)\rvert}{\lvert r(y)\rvert}$
independent CI	$\frac{\lvert l(xy) \cap l(xW)\rvert}{\lvert l(xW)\rvert} * \frac{\lvert r(Wy) \cap r(xy)\rvert}{\lvert r(Wy)\rvert}$
CI*freq	$f(xy) * CI(xy)$
ICI*log(freq)	$\log f(xy) * ICI(xy)$

Where N is the number of tokens in corpus, xy — bigram consisting of words x and y, $f(x)$ is the observed frequency of the word x, $P(x) = f(x)/N$, $\bar{x}$ stands for any word except x, $r(x)$ is a set of unique words occurring in corpus immediately to the right from the word x, $l(x)$ is a set of unique words which occur in corpus immediately to the left from the word x, $\sum_{x,y} A(x,y) = A(x,y) + A(x,\bar{y}) + A(\bar{x},y) + A(\bar{x},\bar{y})$, W stands for any word which does not form a candidate expression with an adjacent word (x in xW or y in Wy)

4.2. Context measures

In their work on automatic term recognition [Nakagawa & Mori 2003] proposed to calculate how many distinct compound nouns contain the simple noun in question as their part in a given corpus, i.e. to build sets of unique words which occur immediately to the left and to the right from the word W. Cardinalities of these sets are multiplied in the scoring function. We use this idea to model lexical rigidity (non-substitutability) of MWE components. In our measure type-LR (see **Table 1**) we take geometric mean of the number of unique words which can occur in the first and in the second position of the MWE under consideration (with the other word being fixed). Our assumption is as follows: the fewer words occur to substitute components of the given bigram, the higher its probability to be an actual MWE.

Taking into account that usual statistical association measures and context measures use different properties of MWEs, we also incorporated the observed frequency into type-LR. This modification (type-FLR) gave a significant increase in average precision (see paragraph 6 for results).

The other group of measures proposed is based on the idea of [Riedl & Biemann 2015] that MWEs tend to have single-word synonyms and, thus, contexts of MWEs of different lengths are similar to that of single words. We took another perspective on the context data comparing immediate contexts of MWEs with those of their components (see context intersection (CI) and independent context intersection (ICI) in **Table 1**). We also combined CI and ICI with either raw observed frequency or its binary logarithm, the best combinations are present in **Table 1**.

Using context and mixed measures introduced in this paragraph we achieved average precision comparable to that of the best measures included in comparison. ICI multiplied by logarithm of frequency significantly surpassed all other individual measures (see **Table 2** below).

4.3. Distributional measures

The only distributional measure we used is DFsing/DFthes proposed in [Loukachevitch & Parkhomenko 2018]. It showed extremely high average precision on the same data. We used the model trained by the authors of the original paper with word2vec [Mikolov et al. 2013] and the following parameters: vector size 200, window size 3, min_count 3 (other parameters left default). Bigrams from dataset were concatenated into single tokens using underscores ('x y' -> 'x_y') to make word-2vec able to build vectors for them. DFsing is calculated as the similarity between the phrase vector $v(xy)$ and vector of the most similar single word w; the word should be different from the phrase components.

$DFsing = \max(\cos(v(xy), v(w)))$, where w is a word from the model vocabulary distinct from x and y.

Given the task of thesauri extension [Loukachevitch & Parkhomenko 2018] also proposed modification of DFsing, which calculates the maximal cosine similarity of a phrase with the existing text entries of the basic thesaurus (RuThes in our case) and orders phrase candidates according to decreasing value of similarity with thesaurus entries (single words or phrases).

$DFthes = \max(\cos(v(xy), v(te)))$, where te is a thesaurus entry (word or phrase).

Note that it is the only measure requiring external resources and its inclusion is not crucial for our combinational method. Following [Loukachevitch & Parkhomenko 2018] we also multiplied DFsing and DFthes by binary logarithm of frequency (see **Table 2** below).

5. Clustering

The idea behind combining measures is the following: an MWE can be indistinguishable from free phrases according to, for example, its frequency properties but stick out as for context or distributional properties (or vice versa). The example of simple combination of two features is provided at **Figure 1**. Values of PMI3 and type-FLR (normalized with binary logarithm), which do not use any common data when calculated, serve as coordinates in 2-dimensional space. It is clearly visible that MWEs (red dots) and free phrases (blue dots) tend to cluster. However, if we look at the border zone of the clusters we will see that expressions of two classes are still substantially mixed and there is no hyperplane which would be able to separate them with high precision.

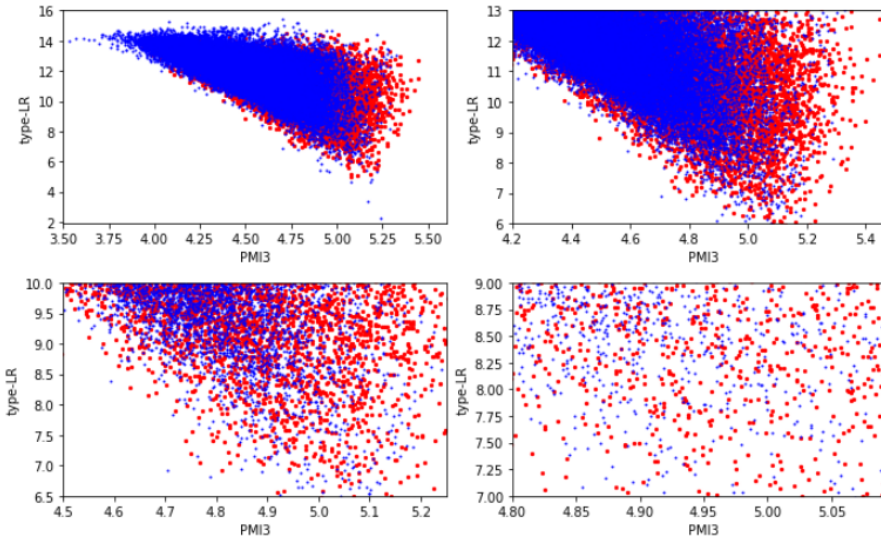


Figure 1. Distribution of expressions in 2-dimensional space with binary logarithms of values assigned by PMI³ and type-LR used as coordinates

This prompts us to increase the number of features, i.e. dimensionality of feature space. The question is how to map feature vectors to numbers (for ranking purposes) without training any classifier. **Table 3** below shows that simple sum or product of coordinates (= feature values) tends to yield deteriorating results with the increasing number of features.

The alternative approach proposed by us aims at preserving dimensionality rather than simply compressing vectors into numbers. First of all, since values assigned by different measures vary considerably, for clustering purposes they were mapped to ranks with binary logarithms[2] taken to make contribution of higher ranks more significant. As a result, every candidate expression was associated with a vector consisting of logarithms of ranks assigned by individual features. These vectors were divided into 2 clusters using implementations of k-means and agglomerative hierarchical clustering (with Ward linkage strategy) in scikit-learn library of Python. We assume that the smaller cluster corresponds to actual MWEs with the larger one consisting mostly of free phrases.

Since we are interested in ranking, rather just classifying, bigrams, we used the following centroid-oriented scoring function:

$$rank(xy) = d(xy, \vec{\mu}_0) - d(xy, \vec{\mu}_1)$$

Where $d(a,b)$ is Euclidean distance, xy is the vector of an expression 'x y', $\vec{\mu}_0$ is the centroid of the larger cluster, and $\vec{\mu}_1$ is the centroid of the smaller cluster.

6. Results

To evaluate the list rankings, we utilized uninterpolated average precision measure (AP), which achieves the maximal value (1) if all expressions of the positive class are located in the beginning of a list without any interruptions. AP at the level of k first candidates is calculated as follows:

$$AP@k = \frac{1}{m}\sum_{i=1}^{k}(r_i * (\frac{1}{i}\sum_{1\le j\le i} r_j))$$

Where $r_i = 1$ if i-th candidate belongs to the positive class, $r_i = 0$ otherwise, m is the number of elements in the positive class.

Table 2 shows the results for individual measures with the best ones being compared at Figure 2.

When combining measures we have experimented on the set of 8 measures with the highest AP. They include measures of all three types — statistical (cubic PMI and LLR), context (type-FLR and ICI*log(freq)) and distributional (DFsing, DFthes and variants multiplied by binary logarithms of frequency). We performed 2 variants of clustering (k-means and agglomerative) with the scoring function defined in paragraph 5 on all 247 feature subsets with more than 1 element. For every feature subset we also tried to combine logarithms of ranks by simply multiplying or summing them up. Table 3 shows the best results, all of them except for the first one were obtained using agglomerative clustering.

[2] Sigmoid function can be used instead as was proposed by an anonymous reviewer.

Table 2. Average precision of individual measures

measure	AP@100	AP@500	AP@1000	AP@2500
Statistical association measures (except PMI and MI variants)				
frequency	0.725	0.734	0.698	0.615
PMI	0.518	0.544	0.545	0.532
Sørensen–Dice coefficient	0.697	0.683	0.674	0.636
LLR	0.778	**0.802**	**0.780**	**0.705**
chi-square	0.699	0.704	0.693	0.657
Piatetsky-Shapiro	0.726	0.740	0.706	0.625
t-score	0.727	0.743	0.710	0.631
geometric mean	0.699	0.704	0.693	0.657
odds ratio	0.559	0.598	0.59	0.562
Poisson	0.775	0.799	0.777	0.702
modified DC	**0.827**	0.736	0.713	0.664
confidence	0.56	0.647	0.642	0.608
Symmetric variants of PMI and MI				
local PMI	0.768	0.792	0.768	0.693
augmented PMI	0.567	0.6	0.591	0.562
cubic PMI	**0.907**	**0.821**	**0.795**	**0.726**
NPMI	0.653	0.663	0.651	0.615
NMI	0.687	0.672	0.666	0.63
Asymmetric variants of PMI and MI				
MI / NF(0,5)	0.64	**0.655**	**0.652**	**0.618**
PMI / NF(0,77)	0.508	0.5	0.495	0.471
MI / NFmax	**0.641**	0.646	0.643	0.609
PMI / NFmax	0.452	0.476	0.472	0.447
$NPMI_c$	0.567	0.529	0.516	0.497
Context measures				
gravity count	0.708	0.694	0.662	0.594
modified GC	0.703	0.695	0.666	0.599
type-LR	0.521	0.563	0.553	0.529
type-FLR	0.818	0.825	0.796	0.74
CI	0.748	0.789	0.783	0.743
ICI	0.894	0.869	0.843	0.78
CI*freq	0.902	0.866	0.847	0.777
ICI*log(freq)	**0.915**	**0.879**	**0.855**	**0.789**
Distributional measures				
DFsing	0.846	0.770	0.694	0.583
DFsing*log(freq)	0.929	0.877	0.834	0.731
DFthes	**0.953**	0.853	0.823	0.759
DFthes*log(freq)	0.950	**0.910**	**0.879**	**0.807**

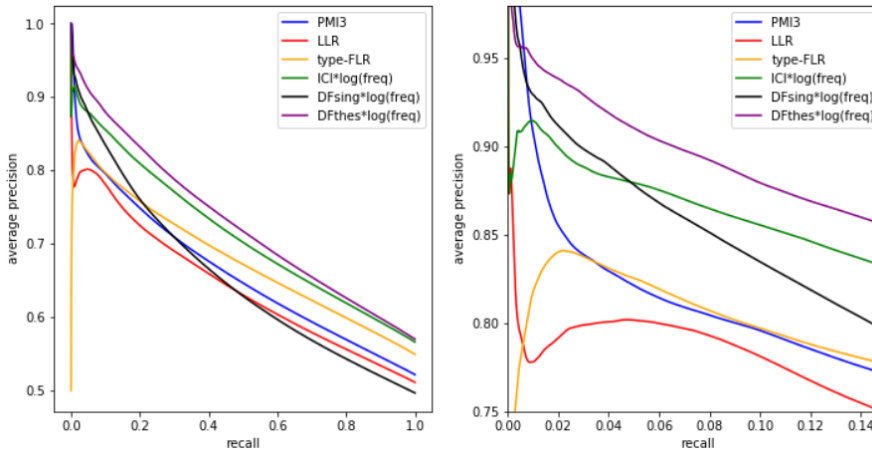


Figure 2. Average precision of the best individual measures with the recall up to 1 and 0.15

Table 3. Best variants of measures' combinations

Method	measures used	AP@100	AP@500	AP@1000	AP@2500
sum of ranks' logarithms	type-FLR, DFthes, DFsing*log(f)	0.976	0.945	**0.92**	**0.847**
agglomerative clustering of ranks' logarithms	type-FLR, ICI*log(f), DFthes, DFsing*log(f)	0.986	0.94	0.907	0.84
	LLR, type-FLR, DFsing, DFthes, DFthes*log(f)	**0.991**	**0.955**	0.917	**0.847**
	LLR, type-FLR, ICI*log(f), DFsing, DFthes, DFsing*log(f), DFthes*log(f)	0.988	0.95	0.914	0.844

7. Discussion

Although simple sum of ranks provides more consistent results, especially when using low amounts of features, the best results are achieved with clustering-based ranking function applied to larger subsets of features. Importantly, two best variants use measures of all three types. Note also that except cubic PMI all of the measures we tried to combine appear in the best setups at least twice.

It is well-known that statistical association measures tend to be biased to either rare or frequent expressions [Evert 2007], [Bouma 2009]. Use of context and distributional measures allows promoting 'unusual' expressions. Table 4 shows top-10 lists obtained with type-FLR and DFthes which turned out to be the most robust features for combining purposes appearing in all four best combinational setups (see **Table 3**). Finally, Table 5 shows top-20 list obtained with the best variant of clustering (LLR, type-FLR, DFsing, DFthes, DFthes*log(freq)). Note that there are no common expressions in these three lists.

Table 4. Top-10 bigrams extracted with type-FLR and DFthes

type-FLR		DFthes	
Едиот Ахронот 'Yedioth Ahronoth'	N	детский сад 'kindergarten'	T
заработная плата 'salary'	T	Европейский союз 'European Union'	T
правоохранительный орган 'law enforcement agency'	T	атомная электростанция 'nuclear power station'	T
Централ Партнершип 'Central Partnership'	N	атомная станция 'nuclear station'	T
точка зрения 'point of view'	T	международное сообщество 'international community'	T
рубрика Автоновости 'Autonews column'	N	мировое сообщество 'world community'	T
уголовное дело 'criminal case'	T	генеральная прокуратура 'prosecutor-general's office'	T
Ближний Восток 'Near East'	T	районный суд 'district court'	T
алкогольное опьянение 'alcohol intoxication'	T	государственный бюджет 'government budget'	T
тройская унция 'Troy ounce'	T	следственный изолятор 'detention center'	T

Table 5. Top-20 bigrams extracted with the best combinational ranking function

1–10		11–20	
ремонтные работы 'reconditioning'	T	Донецкая область 'Donetsk region'	T
Красноярский край 'Krasnoyarsk region'	T	Оренбургская область 'Orenburg region'	T
исполнительный директор 'executive director'	T	антиправительственное выступление 'antigovernment rally'	T
административная ответственность 'administrative liability'	T	избирательная кампания 'election campaign'	T
товарищеский матч 'exhibition game'	T	киевское Динамо 'Kievan Dynamo'	T
Приморский край 'Primorsky kray'	T	наркотическое средство 'narcotic substance'	T
Томская область 'Tomsk region'	T	добыча нефти 'oil extraction'	T
мобильный телефон 'mobile phone'	T	силовая структура 'uniformed service'	T
сотовый оператор 'mobile network operator'	T	общеобразовательная школа 'comprehensive school'	T
федеральный бюджет 'federal budget'	T	Иркутская область 'Irkutsk region'	T

8. Conclusion

In this paper we have introduced simple yet highly efficient unsupervised approach to extracting MWEs appropriate for lexical resources. We have shown that the choice of features is crucial for the efficiency of combinational MWE extraction. Comparing 22 statistical association measures, 8 context measures (4 of which were introduced in this paper) and 4 distributional measures we showed that ranked lists extracted by measures of different types exhibit significant variation. We also showed that the highest average precision is achieved with the help of measures which utilize both frequency and context/distributional information.

We tried out two unsupervised approaches to combining measures: simple sum or product and clustering. Dividing feature vectors of MWEs into 2 clusters and computing distances to their centroids allows incorporating larger number of measures with higher average precision. We leave for further research testing the stability of average precision given particular subset of measures and varying input data.

References

1. *Baldwin T., Bannard C., Tanaka T., Widdows D.* (2003), An empirical model of multiword expression decomposability. In Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, pages 89–96.
2. *Baldwin T., Kim S. N.* (2010), Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, Handbook of Natural Language Processing, pages 267–292. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA, 2 edition.
3. *Baroni M., Zamparelli R.* (2010), Nouns are vectors, adjectives are matrices: Representing adjective–noun constructions in semantic space. In Proceedings of the EMNLP-2010, pp. 1183–1193.
4. *Bouma G.* (2009), Normalized (pointwise) mutual information in collocation extraction. In From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009, volume Normalized, pages 31–40, Tübingen.
5. *Buljan, Šnajder J.* (2017), Combining Linguistic Features for the Detection of Croatian Multiword Expressions. Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), Valencia, 194–199.
6. *Carlini, R., Codina-Filbà, J., Wanner, L.* (2014), Improving collocation correction by ranking suggestions using linguistic knowledge. Proceedings of the 3rd Workshop on NLP for computer-assisted language learning, Uppsala, Sweden.
7. *Carpuat M., Diab M.* (2010), Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 242–245. Association for Computational Linguistics.
8. *Choueka Y.* (1988), Looking for needles in a haystack. In Proceedings of RIAO '88, pages 609–623.

9. *Church K., Hanks P.* (1990), Word Association Norms, Mutual Information, and Lexicography. In Proceedings of ACL, pages 76–83, 1989.
10. *Church K., Gale W. A., Hanks P., Hindle D.* (1991), Using statistics in lexical analysis. In Lexical Acquisition: Using On-line Resources to Build a Lexicon, pages 115–164. Lawrence Erlbaum.
11. *Van de Cruys, T., Moirón, B. V.* (2007), Semantics-based multiword expressions extraction. In: Proceedings of the Workshop on A Broader Perspective on Multiword Expression, pp. 25–32.
12. *Dunning T. E.* (1993), Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1), 61–74.
13. *Enikeeva E. V., Mitrofanova O. A.* (2017), Russian Collocation Extraction Based on Word Embeddings. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2017", Vol. 1, 2017, pp. 52–65.
14. *Evert S.* (2007), Corpora and collocations. Extended Manuscript of Chapter 58 of A. Lüdeling and M. Kytö, 2008, Corpus Linguistics. An International Handbook, Mouton de Gruyter, Berlin.
15. *Hoang H. H., Kim S. N., Kann M.* (2009), A re-examination of lexical association measures. In Proceedings of the ACL 2009 Workshop on MWEs, pages 31–39, Singapore.
16. *Jackendoff R.* (1997), The Architecture of the Language Faculty. Number 28 in Linguistic Inquiry Monographs. MIT Press, Cambridge, MA, USA. 262 p.
17. *Korkontzelos I., Manandhar S.* (2010). Can recognising multiword expressions improve shallow parsing? In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 636–644. Association for Computational Linguistics.
18. *Lin D.* (1998), Automatic retrieval and clustering of similar words. In Proceedings of COLING/ACL-98, pages 768–744, Montreal.
19. *Loukachevitch N., Dobrov B., Chetviorkin I.* (2014). "Ruthes-lite, a publicly available version of thesaurus of russian language ruthes." Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue", Bekasovo, Russia. 2014.
20. *Loukachevitch N., Parkhomenko E* (2018), Recognition of multiword expressions using word embeddings // Artificial Intelligence. RCAI 2018. — Vol. 934 of Communications in Computer and Information Science. — Springer Cham, 2018. — P. 112–124.
21. *McCarthy D., Koeling R., Weeds J., Carroll J.* (2004), Finding predominant word senses in untagged text. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, pp. 280–287.
22. *Mikolov T., Chen K., Corrado G., Dean J.* (2013), Efficient Estimation of Word Representations in Vector Space. In: Proceedings of Workshop at ICLR.
23. *Nakagawa H., Mori T.* (2003), Automatic Term Recognition based on Statistics of Compound Nouns and their Components // Terminology. — 2003. — Vol. 9, №2. — P. 201–219.

24. *Padó S., Lapata M.* (2007), Dependency-based Construction of Semantic Space Models. Computational Linguistics, 33(2):161–199.
25. *Pecina P., Schlesinger P.* (2006), Combining association measures for collocation extraction. In Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006), Sydney, Australia.
26. *Pecina P.* (2008), A Machine Learning Approach to Multiword Expression Extraction. In Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions (MWE 2008), pages 54–57, Marrakech, 2008.
27. *Ren Z., Lü Y., Cao J., Liu Q., Huang Y.* (2009), Improving statistical machine translation using domain bilingual multiword expressions. In Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications, pages 47–54. Association for Computational Linguistics.
28. *Riedl M., Biemann C.* (2015), A single word is not enough: Ranking multiword expressions using distributional semantics. In Proceedings of EMNLP 2015, pages 2430–2440, Lisbon.
29. *Rodríguez-Fernández S., Anke L., Carlini R., Wanner L.* (2016), Semantics-driven recognition of collocations using word embeddings'. In: Proceedings of the 2016 Annual Meeting of the Association for Computational Linguistics (ACL), Berlin, Germany.
30. *Sag I. A., Baldwin T., Bond F., Copestake A., Flickinger D.* (2002), Multiword expressions: A pain in the neck for NLP. In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2002), pages 1–15, Mexico City, Mexico.
31. *Salehi B., Cook P., Baldwin T.* (2015), A word embedding approach to predicting the compositionality of multiword expressions. In NAACL-HTL, pages 977–983, Denver, Colorado, 2015.
32. *Senaldi M. S. G., Lebani G. E., Lenci A.* (2016), Lexical Variability and Compositionality: Investigating Idiomaticity with Distributional Semantic Models. Proceedings of the 12th Workshop on Multiword Expressions (MWE 2016): 21–31.
33. *Tsvetkov Y., Wintner S.* (2011), Identification of multi-word expressions by combining multiple linguistic information sources. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 836–845. Association for Computational Linguistics.
34. *Tutubalina E.* (2015), Clustering-based Approach to Multiword Expression Extraction and Ranking. In NAACL-HTL, pages 39–43, Denver, Colorado, 2015.
35. *Tutubalina E. V., Braslavski P. I.* (2016), Multiple features for multiword extraction: A learning-to-rank approach// Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016", Vol. 1, 2016, pp. 782–791.
36. *Zakharov V.* (2017), Automatic Collocation Extraction: Association Measures Evaluation and Integration // Computational Linguistics and Intellectual Technologies: Papers from the Annual conference "Dialogue". Volume 1 of 2. Computational Linguistics: Practical Applications. — Moscow: RSUH, 2017. — P. 396–407.

WORD VECTOR MODELS AS AN OBJECT OF LINGUISTIC RESEARCH

Shavrina T. O. (rybolos@gmail.com)
NRU HSE, Moscow, Russia; Sberbank, Moscow, Russia

This article launches a series of studies in which popular vector word2vec models are considered not as an element of the architecture of an NLP application, but as an independent object of linguistic research. The linguist's view on the surrogate of contexts on the corpus, as which vector models can be considered, makes it possible to reveal new information about the distribution of individual semantic groups of vocabulary and new knowledge about the corpus from which these models are derived. In particular, it is shown that such layers of English and Russian vocabulary, such as the names of professions, nationalities, toponyms, personal qualities, time periods, have the greatest independence from changing the model and retain their position relative to their neighbour words—that is, they have the most stable contexts regardless of the corpus; it is shown that the vocabulary from the Swadesh list is statistically more resistant to changing the model than the frequency vocabulary is; it is shown which word2vec models for the Russian language preserve best the ontological structures in vocabulary.

Key words: word2vec, word vector model, word vectors, vector model evaluation, word2vec interpretation

ВЕКТОРНЫЕ МОДЕЛИ КАК ОБЪЕКТ ЛИНГВИСТИЧЕСКОГО ИССЛЕДОВАНИЯ

Шаврина Т. О. (rybolos@gmail.com)
НИУ ВШЭ, Москва, Россия; Сбербанк, Москва, Россия

В данной статье начата серия исследований, в которых популярные векторные модели word2vec рассматриваются не как элемент архитектуры NLP-приложения, а как самостоятельный объект лингвистического исследования. Взгляд лингвиста на суррогат контекстов, коим можно назвать такие модели, позволяет выявить новую информацию о распределении отдельных семантических групп лексики и о корпусах, на которых эти модели получены. В частности, показывается, что такие пласты английской и русской лексики, как названия профессий, национальностей, топонимы, качества личности, временные сроки, обладают наибольшей независимостью от смены модели и сохраняют свое положение относительно соседей — то есть имеют наиболее устойчивые контексты независимо от корпуса; показывается, что лексика из списка Сводеша в среднем более устойчива к смене модели, чем частотная лексика; показывается, какие модели word2vec для русского языка наилучшим образом сохраняют онтологические структуры в лексике.

Ключевые слова: word2vec, векторные модели, word vectors, evaluation

1. About word vector models

Vector word models are currently one of the main elements in architectures for language modelling and processing, showing themselves to be an effective way to convey information about the meaning and generalized contexts of individual words. From the point of view of mathematics, modern vector skip-gram and CBOW models have an indisputable advantage over other ways of vectorizing words—they simultaneously describe the distribution of words relative to each other and also take into account their sequential order.

However, the assessment of such vector models like word2vec [Mikolov et al., 2013], GloVE [Pennington et al., 2014], fasttext [Bojanowski et al., 2017] is currently hampered by the "black box" of the algorithm for obtaining them—and the quality of the models is estimated very indirectly. This study is devoted to the development of a linguistic apparatus for assessing the quality of vector models based on linguistic knowledge.

The main hypothesis on which the training word2vec models is based is "the words having the same contexts mean the same". Both Skip-gram and CBOW models provide high-quality word embeddings with this hypothesis, however, any linguist will call the resulting problems:

- there are words with similar contexts meaning the opposite—antonyms;
- there are words with different contexts meaning the same—historical synonyms, multi-word expressions, etc;
- also, well-known problems are polysemy, morphological derivatives, misprints.

These problems lead to the attention shift in the human evaluation of vector models: the vocabulary of the medium frequency, non-homonymous, unambiguous gives those beautiful examples of vector calculations (*"king"—'man' = 'queen', etc. by* [Mikolov et al., 2013]). What happens on the other groups of lexis?

For a linguistic point of view, the word vector model is a linguistic surrogate all the contexts in a corpus from which it is derived. Thus it can be considered interesting as an independent object of study, object situated in the middle of the usual division [Saussure, 1916] of synchronic and diachronic approaches: cumulative information about word behaviour in the language, obtained on the basis of all contexts over a certain (usually broad) time period, can be surprisingly accurate—examples (1) and (2) show that such a model can even accumulate extralinguistic knowledge if trained on the billionth volume of words.

(1) *5 closest words to the word 'otradnoye'* (adj, name of a Moscow metro station) *on word2vec model are 5 geographically adjacent Metro stations* (model trained on Russian National Corpus).

Semantic associates for отрадное[1] (computed on Ruscorpora and Wikipedia)

лобаново	*0.614*
петровско-разумовское	*0.585*
романово	*0.577*
глухово	*0.574*
лукино	*0.572*

[1] https://rusvectores.org/en/ruwikiruscorpora_upos_skipgram_300_2_2019/отрадное_ADJ/

(2) *10 closest words to the word 'shabolovskaya'* (adj, name of a station on crossing Metro lines) *on word2vec model are geographically adjacent Metro stations, street names and names of crossing metro lines* (model trained on news corpus).

Semantic associates for шаболовская[2] (ADJ) computed on news corpus

шаболовский$_{ADJ}$	*0.59*
щёлковская$_{ADJ}$	*0.53*
серпуховской$_{ADJ}$	*0.51*
-радиальный$_{ADJ}$	*0.49*
таганско$_{ADJ}$	*0.49*
198-ть$_{ADJ}$	*0.49*
добрынинская$_{ADJ}$	*0.49*
филетовый$_{ADJ}$	*0.48*
подбельский$_{ADJ}$	*0.47*
калужско-рижский$_{ADJ}$	*0.47*

Hereinafter, all results will be presented on RusVectores project [Kutuzov, Andreev, 2015], [Kutuzov, Kuzmenko, 2017] models—all skip-gram, with lemmatization and pos-tagging, trained on 1) news corpus[3], 2) Russian National Corpus[4] and Wikipedia, 3) Taiga corpus[5] and 4) Aranea corpus[6]. Results in English are computed on a sister project—WebVectors[7]

2. Word vector model evaluation

Vector models are of great interest in connection with the mediated material they represent—for the needs of corpora comparison and assessment, for analyzing the nature of lexis. Knowledge of the "normal" and "anomalous" behaviour of lexis on the corpora would allow a much more accurate assessment of the quality of the obtained model vectors.

However, the quality assessment of vector models is still fairly superficial—this is either enumerating all possible models and choosing one that showed the best result in a particular architecture and specific task [Kutuzov, 2015], or an assessment on a small set of individual pairs of words with human assessment of their

---

[2] https://rusvectores.org/en/news_upos_skipgram_300_5_2019/шаболовская_ADJ/

[3] News: news stream from 1500 primarily Russian-language news sites, model: http://vectors.nlpl.eu/repository/11/184.zip

[4] Full Russian National Corpus http://ruscorpora.ru/en/, model https://rusvectores.org/static/models/rusvectores4/RNC/ruscorpora_upos_skipgram_300_5_2018.vec.gz

[5] **Taiga:** open and structured Russian web corpus https://tatianashavrina.github.io/taiga_site/, model https://rusvectores.org/static/models/rusvectores4/taiga/taiga_upos_skipgram_300_2_2018.vec.gz

[6] **Araneum Russicum Maximum:** large web corpus of Russian http://ella.juls.savba.sk/aranea_about, model https://rusvectores.org/static/models/rusvectores4/araneum/araneum_upos_skipgram_300_2_2018.vec.gz

[7] http://vectors.nlpl.eu/explore/embeddings/en/about/

distance (completely subjective)—SimLex999 [Hill et al. 2015] and Google Analogy [Mikolov 2013]. Several significant studies [Tsvetkov et al. 2016], [Vulich et al. 2017] have already shown that the quality of vector models for the English language is unstable and depends on many factors, and for an independent assessment of models, a new methodological apparatus is needed.

The evaluation problem grows like a snowball—in 2018, the first studies devoted to obtaining the best vector models were published, claiming universality for all words and sentences in a language—BERT [Devlin et al. 2018], ELMo [Peters et al. 2018], and OpenAI architecture [Radford et al. 2018]. The main trend in NLP remains—we search for an effective way to vectorize words and whole texts, but to evaluate model effectiveness, a new approach and a new level of understanding of the resulting models despite the corpus features is missing. Next, we consider a series of experiments devoted to the study of the lexis behaviour in word2vec models and the linguistic interpretation of the quality of word vectors—the preservation of known ontological relationships, most stable vocabulary groups, and so on.

3. The behaviour of lexis in word vector models

In accordance with the first hypothesis about the lexis behaviour in word2vec models, it was decided to check the Swadesh list [Swadesh, 1950]—words from a manually compiled list that are considered chronologically the most stable in the language. Words from the Swadesh list do have interesting characteristics from the point of view of vector models—they denote the basic concept—relatives, animals, main action verbs, colorus, numbers, etc., and have a frequency above the average, that is, have enough contexts in any corpus. Hypothetically, on vector models, such vocabulary should be stable relative to its neighbours.

3.1. Experiment 1

Swadesh list was obtained for Russian and English in its fullest form (200 words), then only those words that were found on all models in concern were left—these are 173 words for Russian and 160 words for English since stop words are removed from the models before training[8].

Then, for each of the words in the list, the share of the word neighbours always presented regardless of the model was calculated—in the window of the 10 closest ones, as well as the 20, 50, 100, 200 and 300 nearest neighbours. For the Russian language, the models RNC + wiki, Taiga, Aranea were used, and for English—BNC, Wiki, Gigaword.

For comparison, random words of a general dictionary of models were also taken, and, separately, random words with a high frequency (top 2000). For the Russian language, frequencies were taken from [Lyashevskaya, Sharoff, 2009], for English [Kilgarriff, 1997] served as material.

Thus, it was obtained 15 samples for each language (3 types of words—Swadesh, frequent and random x 5 amounts of the nearest neighbours)—words and

[8] The full list can be found in the repository https://github.com/TatianaShavrina/wordvector_metrics.

corresponding numbers from 0 to 1, denoting % of the stable neighbours. A statistical Mann—Whitney U-test [Mann, Whitney, 1947] was used to evaluate the differences between two independent samples based on the level of any trait measured quantitatively (simple non-parametric criterion).

On each triple of samples (Swadesh, frequent words, random words), a test was conducted with an alternative hypothesis that the values in the second sample were larger. The obtained result for each window of the nearest neighbours is the same:

1. words from Swadesh's list have a higher percentage of stored neighbours than random frequency words from the top 2000;
2. words from Swadesh's list have a higher percentage of saved neighbours than random words of a language;
3. frequency words from the top 2000 have a greater percentage of saved neighbours than just random words of the language[9].

The p-value for all such tests clearly shows that the values in Swadesh's samples are significantly larger than values in frequency word lists; frequency word values are in turn larger than values in random word lists.

(3) *for 100 nearest word neighbours for English:*
fr = frequent, sv = svodesh, rn = random
rn ≤ sw
annwhitneyuResult(statistic = 8932.0, pvalue = 4.4298335409745345e − 11)
fr ≤ sw
MannwhitneyuResult(statistic = 13363.0, pvalue = 0.04262714406973201)
rn ≤ fr
MannwhitneyuResult(statistic = 9962.0, pvalue = 3.771709496130687e − 08)

(4) *for 100 nearest word neighbours for Russian:*
fr = frequent, sv = svodesh, rn = random
rn ≤ sw
MannwhitneyuResult(statistic = 8931.0, pvalue = 2.4298335409745345e − 11)
fr ≤ sw
MannwhitneyuResult(statistic = 13363.0, pvalue = 0.05262714406973201)
rn ≤ fr
MannwhitneyuResult(statistic = 7344.0, pvalue = 1.6367105050242702e − 11)

3.2. Experiment 2

Further, it was decided to scale up the previous experiment for the entire vocabulary of the existing models and conduct a test on the most stable words model, sorting them all one by one.

The intersection of dictionaries of all models was obtained, then for each word from the list, the number of stable neighbours was calculated- in the window of the

[9] More complete numbers can be found https://github.com/TatianaShavrina/wordvector_metrics.

100 nearest neighbours. The list has been sorted by percentage of saved neighbours, remaining the same regardless of model—to measure that the intersection of the list of N nearest neighbours of the word was used on the entire list of models.

Thus, 2 interesting results were obtained at once—at the top of the list, we get the most stable words, which, regardless of the corpus source, keep their neighbours, and at the bottom—the most unstable ones. It is curious that the semantically given top of the list is grouped into distinct semantic groups:

- nouns denoting the personal qualities of a person,

(5) *Russian:*

находчивость_NOUN	*(resourcefulness_NOUN)*	*0.2781 neighbours saved*
радушие_NOUN	*(welcome_NOUN)*	*0.2670*
аккуратность_NOUN	*(accuracy_NOUN)*	*0.2626*
идеализм_NOUN	*(idealism_NOUN)*	*0.2542*

- emotions,

(6) *Russian:*

неприязнь_NOUN	*(hostility_NOUN)*	*0.3059 neighbours saved*
недоверие_NOUN	*(distrust_NOUN)*	*0.2832*
восхищение_NOUN	*(admiration_NOUN)*	*0.2528*
негодование_NOUN	*(resentment_NOUN)*	*0.2473*

- nationalities,

(7) *Russian:*

итальянец_NOUN	*Italian_NOUN*	*0.2558*
ирландец_NOUN	*Irish_NOUN*	*0.2690*
узбек_NOUN	*Uzbek_NOUN*	*0.2389*

- professions,

(8) *Russian:*

скрипач_NOUN	*violinist_NOUN*	*0.2193*
палеонтолог_NOUN	*paleontologist_NOUN*	*0.2179*
филолог_NOUN	*philologist_NOUN*	*0.2391*
географ_NOUN	*geographer_NOUN*	*0.2320*

- toponyms,

(9) *Russian:*

казах_NOUN	*Kazakh_NOUN*	*0.2444*
нижегородский_ADJ	*Nizhny Novgorod_ADJ*	*0.2432*
бразилия_PROPN	*Brazil_PROPN*	*0.2350*
испанский_ADJ	*spanish_ADJ*	*0.2337*

- term adjectives.

(10) *Russian:*

двухлетний_ADJ	*two-year_ADJ*	*0.2428*
четырехмесячный_ADJ	*four month_ADJ*	*0.2278*
трехдневный_ADJ	*three-day_ADJ*	*0.2240*
шестимесячный_ADJ	*six month_ADJ*	*0.2198*

Results are stable for Russian and English (see appendix 1 and appendix 2 correspondingly). Only a few words they are knocked out of a list and can not be assigned to any group: these are 'pregnancy' (0.1386), 'whale' (0.1268), 'intercourse' (0.1226), 'waste' (0.1208) for English, 'неразбериха' ('confusion', 0.2431) 'материализм' ('materialism', 0.2228), 'коррупция' ('corruption', 0.2193) for Russian. There are practically no verbs in the top of the list, for both Russian and English they have too diverse contexts. All the above-mentioned semantic categories were postulated while analyzing the list, the reverse statement that all the words of these categories on average have more stable contexts is not proven because of the difficulty of demarcating these categories.

The most unstable group of words is:

- proper names

(11) *Russian:*

Неклюдов_PROPN	*Neklyudov PROPN*	*0*
Свинцов_PROPN	*Svintsov_PROPN*	*0*
Софронов_PROPN	*Sofronov_PROPN*	*0*
Робсон_PROPN	*Robson_PROPN*	*0*

Having the most inconsistent contexts and low frequency, the proper names—surnames, full names occupy the bottom of the list for both Russian and English.[10]

It is noteworthy that these results partially reproduce the results of clustering in the work [Zobnin, 2017], where groups of proper names, toponyms and other semantic categories are also distinguished.

4. First steps to a linguistic assessment of models

Learning more about the standard properties of a wide list of lexemes in a language, we can more accurately assess both the adequacy of specific models for applied problems and the perspective of their potential improvement.

In the next experiment, we will show how the most popular models for the Russian language retain ontological relations in the vector space. The ontology of RuWordNet [Loukachevitch, Lashevich, 2016], containing more than 300 thousand pairs of words connected by relationships, was taken as a bank of such relations:

POS-synonymy, antonym, cause, domain, entailment, hypernym, hyponym, instance hypernym and instance hyponym, part holonym, part meronym.

4 popular word2vec models for Russian—based on News, Aranea, RNC+wiki, Taiga—were studied on the subject of 1) the presence of words in the dictionary, 2) % of the preservation of connections between words—the "presence of a word in the list of N closest neighbours". N is 10, 20, 50, 100. Multi-word expressions are also included in the test—see **Table 1**.

[10] See full lists at https://github.com/TatianaShavrina/wordvector_metrics.

Table 1: Experimental data examined

child_word	parent_ word	relation	has_ in_10	has_ in_20	has_ in_50	has_ in_100
рабочий, работник физического труда (worker)	каменщик (mason)	hypernym	FALSE	FALSE	TRUE	TRUE
промышленность (industry)	каменщик (mason)	domain	FALSE	FALSE	FALSE	FALSE

We have 3 values for each word2vec model—"False"—both words presented in a model, but no relation found, 'OOV'—out of vocabulary, one of the words is not presented in a model, 'True'—both words presented in a model, relation established through N nearest words.

The results are surprising in some ways: first, all the metrics turned out to be quite low. Synonymy and antonymy, so beautifully illustrated with examples of original articles, generally stop reproducing for most of the vocabulary. Secondly, the best quality is shown by the model obtained on the largest corpus, Aranea (internet-crawled data), while the model of the Russian National Corpus and Wikipedia shows results below average. The results are also reproduced for the 100 nearest neighbour words (table 2). However, a model trained on the Russian National Corpus and Wikipedia has one of the most comprehensive dictionaries—the number 'not in vocabulary' in it is the smallest in almost all relationships (shown in bold).

Table 2: Remaining % of ontological relations on popular word2vec models, Russian. 100 nearest neighbours

relation	value	taiga	me	news	aranea	mean
antonymy	FALSE	73.05	57.47	75.65	45.56	62.93
antonymy	**TRUE**	25.11	37.77	16.78	**48.70**	32.09
antonymy	OOV	**1.84**	4.76	7.58	5.74	4.98
cause	FALSE	68.44	55.15	78.24	31.23	58.26
cause	**TRUE**	19.44	**41.03**	10.13	15.28	21.47
cause	OOV	12.13	**3.82**	11.63	53.49	20.27
domain	FALSE	96.91	92.73	96.41	90.00	94.01
domain	**TRUE**	1.75	5.80	3.19	**8.13**	4.72
domain	OOV	1.34	1.47	**0.41**	1.87	1.27
entailment	FALSE	91.92	80.13	89.32	64.44	81.45
entailment	**TRUE**	4.46	**17.73**	6.78	12.81	10.45
entailment	OOV	3.62	**2.14**	3.90	22.75	8.10
hypemym	FALSE	88.95	82.18	87.64	61.34	80.03
hypemym	**TRUE**	6.81	14.33	7.73	**19.71**	12.15
hypemym	OOV	4.24	**3.49**	4.63	18.95	7.83
hyponym	FALSE	81.57	76.49	79.87	63.20	75.28
hyponym	**TRUE**	7.87	13.40	7.18	**17.75**	11.55
hyponym	OOV	10.57	**10.11**	12.95	19.06	13.17

The lowest quality is shown by popular vector models when conveying relationships like instance hyponymy and domain—it is possible that low quality, among other factors, can be explained by a low frequency of individual occurrences and their absence in the model dictionary. Also, the hypothesis that such relations as hyponymy, hypernymy and domain should be expressed by nearest neighbours can be too simplifiying, as they are hierarchical relations that cannot be extracted from embeddings directly by cosine similarity, unlike pairwise-equivalent synonymy relations.

Relationships of antonymy are fairly well preserved (48% on the best model, Aranea), cause (41% on a best model, Aranea), hypernym, part holonym и part meronym (20% each on Aranea)—but such quality can be considered rather low. Nonetheless, in a similar experiment for the English language [Rogers et al. 2018] skip-gram models show lower quality—synonyms—0.447, antonyms—0.144, hyponyms—0.038, other relations—0.013 of ontological relations.

1) a modern amount of data is still not enough—we need at least an order of magnitude more data to get a large number of contexts for low-frequency words and multi-word expressions, which can be distinguished in a large number in any language;
2) the efficiency of the vectors obtained is far from ideal—between words that are obviously close to the claimed hypothesis: synonyms and antonyms, as well as part-whole and class-subclass relations—the proportion of the saved relations is low.

5. Further work and discussion

Within the framework of the initiated methodology, it is planned to further study the distributional lexis behaviour, and based on the results obtained, it is planned to develop metrics that allow obtaining a more complete interpretation of vector models.

The author would like to start a discussion on whether vector models can be used as a tool for a full-fledged linguistic lexical study on big corpora: potentially, such areas of study could be:

- assessment of corporal context biases, corpus thematic focus
- assessment of the sufficiency of the presented contexts of basic vocabulary in the corpus
- search for the most universal vocabulary groups that preserve the structure of relations among themselves regardless of the corpus and model
- the formation of a clearer picture of the set of mandatory properties that characterize a representative corpus of a language.

6. Conclusion

In this paper, as a result of experiments conducted on popular word2vec models for Russian and English, it was shown that the most stable lexical groups, having most uniform contexts, independent from the corpus, are:

- adjectives denoting the personal qualities of a person,
- nationalities,

- professions,
- toponyms,
- term adjectives.

At the same time, the most unstable group are proper names—as the rarest and context-dependent.

It has been established that words from Swadesh list (for English and Russian) are more resistant to a change of model and retain their closest neighbours regardless of the model much more often than words from the frequency vocabulary, as well as more often than random words.

For the Russian language, an experiment was conducted to assess the residual number of semantic and ontological links between known pairs of words and the quality of models was estimated on the basis of this number of relations remaining in the model.

All the data and code for this paper are available on github[11]—we welcome other authors to contribute word2vec metrics and evaluate their models.

7. Acknowledgement

The author is sincerely grateful to Olga Lyashevskaya and Serge Sharoff who prompted the author to think about the need for a different methodology for evaluating vector models, to Andrei Kutuzov for providing additional information about RusVectores models, and to Natalia Lukashevich for providing materials from the RuWordNet project.

References

1. *Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean* (2013). Efficient estimation of word representations in vector space. CoRR, abs/1301.3781.
2. *Jeffrey Pennington, Richard Socher, and Christopher D. Manning* (2014). GloVe: Global Vectors for Word Representation.
3. *P. Bojanowski*, E. Grave*, A. Joulin, T. Mikolov* (2017) Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, volume 5, 2017, issn 2307–387X, pp. 135–146.
4. *Ferdinand de Saussure* (1916) Cours de linguistique générale, ed. C. Bally and A. Sechehaye, with the collaboration of A. Riedlinger, Lausanne and Paris: Payot; trans. W. Baskin, Course in General Linguistics, Glasgow: Fontana/Collins, 1977.
5. *Kutuzov A., Kuzmenko E.* (2017) WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In: Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science, vol 661. Springer, Cham.

[11] https://github.com/TatianaShavrina/wordvector_metrics.

6. *Kutuzov, A., Andreev, I.* (2015) Texts in, meaning out: neural language models in semantic similarity task for Russian. In: Proceedings of the Dialog Conference, Moscow, RGGU.
7. *Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.* (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv e-prints
8. *Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer* (2018). Deep contextualized word representations. In NAACL.
9. *Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever* (2018). Improving language understanding with unsupervised learning. Technical report, OpenAI.
10. *Swadesh, Morris.* (1950). "Salish Internal Relationships." International Journal of American Linguistics, Vol. 16, 157–167.
11. *О. Н. Ляшевская, С. А. Шаров,* (2009). Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник.
12. *Loukachevitch N., Lashevich G.* (2016) Multiword expressions in Russian Thesauri RuThes and RuWordNet. Proceedings of the AINL FRUCT 2016, pp. 66–71.
13. *Kilgarriff, A.* (1997) Putting Frequencies in the Dictionary. International Journal of Lexicography 10 (2). Pp 135–155.
14. *Mann, Henry B.; Whitney, Donald R.* (1947). "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other". Annals of Mathematical Statistics. 18 (1): 50–60. doi:10.1214/aoms/1177730491. MR 0022058. Zbl 0041.26103.
15. *Felix Hill, Roi Reichart, and Anna Korhonen* (2016). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. Computational Linguistics
16. *Rogers, Anna, Shashwath Hosur Ananthakrishna, and Anna Rumshisky* (2018). What's in your embedding, and how it predicts task performance. In Proceedings of the 27th International Conference on Computational Linguistics, pages 2690–2703.
17. *Yulia Tsvetkov, Manaal Faruqui, and Chris Dyer* (2016). Correlation-based intrinsic evaluation of word vector representations. CoRR abs/1606.06710.
18. *Ivan Vulic, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen* (2016). Hyperlex: A large-scale evaluation of graded lexical entailment. CoRR,abs/1608.02117.
19. *A. Zobnin* (2017) "Rotations and interpretability of word embeddings: the case of the russian language," arXiv preprint arXiv:1707.04662.

Appendix

Appendix 1 and 2: Top 100 most stable words for English and Russian

N top	intersection	word English	intersection	word Russian
1	0.2658959538	six-month_ADJ	0.3058823529	неприязнь_NOUN
2	0.2388888889	three-week_ADJ	0.2832369942	недоверие_NOUN
3	0.2131147541	two-week_ADJ	0.2781065089	находчивость_NOUN
4	0.1935483871	eight-year_ADJ	0.269005848	ирландец_NOUN
5	0.1917098446	six-week_ADJ	0.2670454545	радушие_NOUN
6	0.1917098446	two-year_ADJ	0.2625698324	аккуратность_NOUN
7	0.1904761905	four-month_ADJ	0.2558139535	итальянец_NOUN
8	0.1808510638	three-month_ADJ	0.2542372881	идеализм_NOUN
9	0.1804123711	extremely_ADV	0.2528089888	изобретательность_NOUN
10	0.175879397	Uganda_PROPN	0.2528089888	восхищение_NOUN
11	0.1693121693	unease_NOUN	0.25	самоотверженность_NOUN
12	0.1691542289	Malawi_PROPN	0.2472527473	негодование_NOUN
13	0.1675126904	four-week_ADJ	0.2471264368	добросовестность_NOUN
14	0.16	Tanzania_PROPN	0.2458100559	расторопность_NOUN
15	0.1592039801	seven-day_ADJ	0.2445652174	невероятный_ADJ
16	0.158974359	resentment_NOUN	0.2444444444	казах_NOUN
17	0.1565656566	disappointment_NOUN	0.2432432432	нижегородский_ADJ
18	0.1534653465	Botswana_PROPN	0.2430939227	неразбериха_NOUN
19	0.1507537688	astonishing_ADJ	0.2427745665	двухлетний_ADJ
20	0.1477832512	immense_ADJ	0.2391304348	филолог_NOUN
21	0.1469194313	Zambia_PROPN	0.2388888889	узбек_NOUN
22	0.1464646465	Guyana_PROPN	0.2369942197	десятилетний_ADJ
23	0.1464646465	homosexual_ADJ	0.2362637363	недовольство_NOUN
24	0.1463414634	incompetence_NOUN	0.2349726776	бразилия_PROPN
25	0.1435897436	violin_NOUN	0.2349726776	сметка_NOUN
26	0.1428571429	Mozambique_PROPN	0.2346368715	неодобрение_NOUN
27	0.1428571429	five-day_ADJ	0.2336956522	испанский_ADJ
28	0.1407035176	ten-year_ADJ	0.2329545455	смекалка_NOUN
29	0.14	inaccurate_ADJ	0.2320441989	географ_NOUN
30	0.14	three-year_ADJ	0.2316384181	настойчивость_NOUN
31	0.1386138614	pregnancy_NOUN	0.2316384181	грузин_NOUN
32	0.1379310345	five-week_ADJ	0.2315789474	румыния_PROPN
33	0.1359223301	incredible_ADJ	0.2311827957	вологодский_ADJ
34	0.1359223301	Grenada_PROPN	0.2307692308	оплошность_NOUN
35	0.1359223301	tedious_ADJ	0.2295081967	ирландский_ADJ
36	0.1355140187	amazing_ADJ	0.2287234043	омский_ADJ
37	0.1355140187	Kenya_PROPN	0.2287234043	финн_NOUN
38	0.1346153846	enormous_ADJ	0.2277777778	дружелюбие_NOUN
39	0.1346153846	shocked_ADJ	0.2277777778	четырехмесячный_ADJ
40	0.1343283582	disquiet_NOUN	0.2272727273	лицемерие_NOUN
41	0.1339712919	Lesotho_PROPN	0.227027027	саратовский_ADJ
42	0.1333333333	Sierra::Leone_PROPN	0.226519337	азербайджанец_NOUN
43	0.1320754717	dismay_NOUN	0.226519337	католик_NOUN
44	0.1320754717	Zimbabwe_PROPN	0.2263157895	венгрия_PROPN
45	0.1317073171	greatly_ADV	0.2252747253	спокойствие_NOUN
46	0.1306532663	appalling_ADJ	0.2252747253	пунктуальность_NOUN
47	0.1280788177	alarmed_ADJ	0.2247191011	бесполезность_NOUN

N top	intersection	word English	intersection	word Russian
48	0.1279620853	remarkable_ADJ	0.2247191011	некомпетентность_NOUN
49	0.1267605634	whale_NOUN	0.2245989305	ростовский_ADJ
50	0.1261682243	incredibly_ADV	0.2240437158	трехдневный_ADJ
51	0.125	Antigua_PROPN	0.2240437158	деликатность_NOUN
52	0.125	teenager_NOUN	0.222826087	чудовищный_ADJ
53	0.1237623762	honesty_NOUN	0.222826087	материализм_NOUN
54	0.1231527094	Nigeria_PROPN	0.222826087	индус_NOUN
55	0.1231527094	ankle_NOUN	0.2216216216	трехнедельный_ADJ
56	0.1227272727	biologist_NOUN	0.2215909091	американец_NOUN
57	0.1226415094	intercourse_NOUN	0.2204301075	недоумение_NOUN
58	0.1225490196	Dominica_PROPN	0.2204301075	австралия_PROPN
59	0.1218274112	frustration_NOUN	0.2197802198	шестимесячный_ADJ
60	0.1213592233	underwear_NOUN	0.2192513369	скрипач_NOUN
61	0.1209302326	Barbados_PROPN	0.2192513369	коррупция_NOUN
62	0.1207729469	waste_NOUN	0.217877095	палеонтолог_NOUN
63	0.1207729469	trumpet_NOUN	0.2173913043	неясность_NOUN
64	0.1201923077	generosity_NOUN	0.2173913043	неимоверный_ADJ
65	0.1196172249	clarinet_NOUN	0.2162162162	пакистан_PROPN
66	0.119266055	conspiracy_NOUN	0.2159090909	биолог_NOUN
67	0.119266055	whisky_NOUN	0.2157894737	недельный_ADJ
68	0.1188118812	cello_NOUN	0.2154696133	антисемитский_ADJ
69	0.1184834123	courage_NOUN	0.2150537634	венгерский_ADJ
70	0.117370892	sex_NOUN	0.2142857143	предусмотрительность_NOUN
71	0.117370892	surgeon_NOUN	0.2131147541	презрение_NOUN
72	0.1170731707	pear_NOUN	0.2131147541	усидчивость_NOUN
73	0.1165048544	nine-year_ADJ	0.2131147541	дотошный_ADJ
74	0.1165048544	ten-day_ADJ	0.2124352332	возмущение_NOUN
75	0.1165048544	Nairobi_PROPN	0.2116402116	таджик_NOUN
76	0.1162790698	flute_NOUN	0.2116402116	ирландия_PROPN
77	0.1162790698	headache_NOUN	0.2111111111	этнограф_NOUN
78	0.1157407407	uncle_NOUN	0.2111111111	сноровка_NOUN
79	0.1153846154	craftsman_NOUN	0.2108108108	геолог_NOUN
80	0.1148325359	sadness_NOUN	0.2105263158	армянин_NOUN
81	0.1148325359	weather_NOUN	0.2099447514	выразительность_NOUN
82	0.1142857143	t-shirt_NOUN	0.2099447514	ангола_PROPN
83	0.1141552511	marvellous_ADJ	0.2099447514	православие_NOUN
84	0.1141552511	frustrating_ADJ	0.2099447514	плечистый_ADJ
85	0.1136363636	biology_NOUN	0.2096774194	ярославский_ADJ
86	0.1132075472	despair_NOUN	0.2096774194	тщательность_NOUN
87	0.1132075472	consternation_NOUN	0.2096774194	симпатия_NOUN
88	0.112745098	concerto_NOUN	0.2096774194	грузинский_ADJ
89	0.1126760563	sexual_ADJ	0.2094240838	пермский_ADJ
90	0.1126760563	perseverance_NOUN	0.2087912088	дагестанский_ADJ
91	0.1126760563	husband_NOUN	0.2087912088	голландец_NOUN
92	0.1126760563	inventiveness_NOUN	0.2087912088	беспокойство_NOUN
93	0.1126760563	arduous_ADJ	0.2085561497	зависть_NOUN
94	0.1126760563	false_ADJ	0.2085561497	злорадство_NOUN
95	0.1126760563	homosexuality_NOUN	0.2085561497	невероятно_ADV
96	0.1121495327	tuna_NOUN	0.2085561497	томский_ADJ
97	0.1121495327	frequently_ADV	0.2078651685	никчемность_NOUN
98	0.1121495327	rivalry_NOUN	0.2076502732	коренастый_ADJ
99	0.1116504854	shirt_NOUN	0.2076502732	воронежский_ADJ

ПЕРЕДАЧА ЦЕРКОВНОСЛАВЯНСКОГО ТЕКСТА СРЕДСТВАМИ ГРАЖДАНСКОЙ ГРАФИКИ: МОЖНО ЛИ ПОЛУЧИТЬ ЕЕ ПРИ ПОМОЩИ ФОРМАЛЬНОЙ ПРОЦЕДУРЫ?[1]

Шмелев А. Д. (shmelev.alexei@gmail.com)

Московский педагогический государственный университет; Институт русского языка им. В. В. Виноградова РАН; Православный Свято-Тихоновский гуманитарный университет

CHURCH SLAVONIC TEXT IN THE RUSSIAN SCRIPT: CAN ONE USE ANY FORMAL PROCEDURE TO GET IT?

Shmelev A. D. (shmelev.alexei@gmail.com)

Moscow Pedagogical State University, Moscow, Russia; Vinogradov Institute of Russian Language, Russian Academy of Sciences, Moscow, Russia; St Tikhon's Orthodox University

The paper discusses the problem of rendering Church Slavonic text in the modern Russian script, which is a common practice at present. The relevant procedure would include the following stages: spelling out words with titla, replacing the letter-based denotation of numerical values with Arabic numerals, replacing characters that are absent from the Russian alphabet with characters with the same phonetic value, removing breathings, replacing different accent marks with a unified stress accent. Certain semantic and grammatical information will be lost in the resulting text while the sound will be kept. In other words, the resulting text may be regarded as a practical transcription of the original text. At the next point, the procedure should aim at replacing the original punctuation with the common Russian punctuation (within certain limits) and at the capitalization of certain words (the latter task might require a system of determining co-reference links). The need for a system of automatic punctuation (when the input is a written text) and a system of automatic resolution of referential ambiguity poses challenges to computational linguistics.

Key words: New Church Slavonic, procedure, Russian alphabet, phonetic value, practical transcription, punctuation, co-reference

[1] Раздел, посвященный пунктуации в церковнославянском тексте, переданном средствами гражданской графики, написан в рамках научно-исследовательского проекта РФФИ № 18-012-00778 «Теоретические основания кодификации русской пунктуации».

1. Вступительные замечания

В подавляющем большинстве современных изданий церковнославянских текстов, предназначенных для употребления мирянами, используется русская гражданская графика. Такая запись менее информативна, нежели исходная запись, осуществленная посредством церковнославянской графики, так что исходную запись без привлечения дополнительной информации по ней восстановить невозможно, однако предполагается, что в ней содержатся все необходимые данные для чтения вслух церковнославянского текста в соответствии с современной нормой.

Способ представления церковнославянского текста средствами гражданской графики, как правило, одинаков во всех изданиях, хотя небольшое варьирование имеет место (об этом речь пойдет ниже). Возникает вопрос: какими правилами перехода от церковнославянской записи к записи средствами гражданской графики руководствуются издатели и возможна ли формализация этих правил, чтобы запись средствами гражданской графики, максимально близкая к той, что представлена в реально существующих изданиях, получалась алгоритмически или почти алгоритмически? На входе в такой алгоритм должна поступать оригинальная церковнославянская запись, а на выходе должна быть запись церковнославянского текста в гражданской графике, или церковнославянская гражданская запись (далее — ЦСГЗ).

Практическая потребность в таком алгоритме представляется очевидной: на его основе может быть создана программа автоматического перевода церковнославянских текстов в ЦСГЗ (в настоящее время перевод производится вручную). Кроме того, ряд преобразований, производимых таким алгоритмом, полезно учитывать для усовершенствования системы поиска в церковнославянском подкорпусе «Национального корпуса русского языка» (ЦСНКРЯ) в «модернизированной» орфографии.

Необходимо подчеркнуть, что ЦСГЗ — это вовсе не та передача церковнославянского текста, которая иногда используется в поисковых системах и, как правило, предполагает обратимость. Речь идет именно о реально существующих изданиях церковнославянских текстов (молитвословах, Псалтири и т. д.)[2].

Правила передачи должны включать в себя два раздела: правила побуквенной передачи и правила расстановки прописных букв и знаков препинания. Побуквенная передача должна предполагать возможность однозначного прочтения получившегося текста в соответствии с общими правилами чтения. Правила расстановки прописных букв и знаков препинания, вообще говоря, могли бы считаться факультативными; однако, поскольку реальная ЦСГЗ, как правило, включает в себя прописные буквы, а также (чаще всего) некоторые знаки препинания, отсутствовавшие в оригинальном тексте, алгоритм, ориентированный на то, чтобы получить текст, максимально приближенный к реально существующим

[2] Можно сказать, что ЦСГЗ представляет собою не транслитерацию, а практическую транскрипцию церковнославянского текста.

текстам в ЦСГЗ, должен стремиться к расстановке прописных букв и знаков препинания, соответствующим тому, с чем мы имеем дело в этих текстах[3].

В последующем изложении правила побуквенной передачи и правила расстановки прописных букв и знаков препинания будут рассмотрены отдельно.

2. Побуквенная передача церковнославянского текста

Как уже говорилось, главное требование, которое предъявляется к побуквенной передаче церковнославянского текста в рамках ЦСГЗ, — это возможность его однозначного чтения вслух в соответствии с нормами современного церковнославянского произношения (в пределах сегментной фонетики[4]). Надо сказать, что это требование легко выполнимо. Если мы имеем полностью акцентуированный текст в рамках ЦСГЗ, то нетрудно составить алгоритм, позволяющий перейти от буквенной записи к его фонетической транскрипции. Кстати, эта задача решается значительно легче, чем для современного русского языка. Даже акцентуированный современный русский текст с расставленными точками над *ё* часто содержит элементы, для которых правильное прочтение требует дополнительной информации (при этом большинство реальных русских текстов не содержат знаков ударения и точек над *ё*). Так, буква *г* читается как [г] в слове *много* и как [в] в словоформе *иного*, как [в] в слове *сегодня* и как [г] в названии города *Сеговия*; в дискурсивном слове *конечно* буква *ч* читается как [ш], а в кратком прилагательном *конечно* (*Это множество конечно*) — как [ч']. В церковнославянском таких случаев практически нет; немногочисленные случаи, когда возможны отклонения от стандартных правил чтения, обычно являются зонами варьирования и допускают чтение в соответствии со стандартными правилами. Так, при стандартном чтении *с* перед *м* читается как [с] (*смоко́вница*, *осмы́й*), но в таких словах, как *кафи́сма* или в имени *Исма́ил* может читаться и как [з]; при стандартном чтении слова, начинающиеся на *е* читаются с неслоговым [и] в начале (*Ева́нгелие*), но в таких именах собственных, как *Едо́м*, *Есфи́рь*, *Еми́лия*, *Емману́ил* возможно произношение на месте начальной буквы *е* нейотированного [э] (в слове *Едо́м*, кроме того, возможно произношение твердого согласного перед ударным *е*)[5]. Однако произношение указанных слов в соответствии со стандартными правилами чтения также возможно, а иногда даже считается предпочтительным. Помимо этого, можно упомянуть возможность вариативного чтения буквы *г* — а именно, она

---

[3] Проблема слитного, раздельного и дефисного написания (одна из самых трудных проблем современной русской орфографии) не существует для ЦСГЗ: дефисное написание в ней вообще не используется, а выбор слитного или раздельного написания определяется написанием в оригинальном церковнославянском тексте.

[4] В настоящее время активно изучается интонационное оформление церковнославянского текста при его чтении вслух [Прохватилова 1999]; [Янко 2010]; [Ianko 2011]; однако ЦСГЗ не ориентирована на передачу особенностей интонации, так что этот аспект церковнославянского произношения может игнорироваться.

[5] Нейотированное [э] может читаться на месте *е* не после согласной и в слове *а́ер* ‘воздух’ и в некоторых других (крайне немногочисленных) случаях.

может читаться как взрывной или как фрикативно-придыхательный звук. Иногда делается попытка ввести в правила чтение дополнительное распределение двух *г*, однако следование дополнительному распределению на практике встречается крайне редко. Все эти особенности церковнославянской орфоэпии не находят отражения в реально существующих изданиях текстов ЦСГЗ, поэтому при решении вопроса о возможности строгой процедуры получения ЦСГЗ можно не придавать им значения.

Как же может строиться процедура побуквенной передачи церковнославянского текста средствами гражданской графики? На первом шаге должно происходить раскрытие слов под титлами: они заменяются на полную буквенную запись[6]. Морфемы, которые пишутся в оригинальных церковнославянских текстах под титлами, задаются списком с указанием их расшифровок. При этом, если над частью, находившейся вне титла, в исходной записи стоял знак ударения, то он сохраняется над той же буквою; если же знака ударения вне титла не было, то он ставится над некоторой буквою в «раскрываемой» части (над какой именно — указывается в исходном списке). Информация о том, что в слове произошло раскрытие титла, на этом этапе сохраняется (слово помечается определенным образом), поскольку данная информация в дальнейшем окажется существенной для расстановки прописных букв.

На следующем шаге буквы приобретают «гражданский» облик и одновременно происходит унификация одинаково произносимых букв: *є* и «ять» (ѣ) заменяются на *е*[7], *«зело» (ѕ)* заменяется на *з*, *ї* «десятеричное» заменяется на *и*, «от» (ѿ) заменяется на *от*, «ук» в обоих начертаниях (диграф *оу* и лигатура) заменяется на *у*, «омега» в обоих начертаниях (*ѡ* и *ѽ*) заменяется на *о*, *я* в обоих начертаниях («юс малый» *ѧ* и лигатура) заменяется на *я*, «кси» (*ѯ*) заменяется на *кс*, «пси» (*ѱ*) заменяется на *пс*, «фита» (*ѳ*) заменяется на *ф*. Для «ижицы» (*ѵ*) замена осуществляется в зависимости от наличия/отсутствия над этой буквой надстрочного знака: при отсутствии надстрочного знака «ижица» (это возможно только после букв *а* и *е*) заменяется на *в*, а при наличии — на *и*. Кроме того, осуществляется еще ряд контекстно обусловленных замен: «ер» (ъ) на конце слов просто элиминируется, «еры» (ы) после шипящих заменяется на *и*[8], *сочетание гк* — на *нк* (напр., в слове *синкли́т*). Заметим, что постулировать замену *гг* на *нг* для слов *а́нгел* (и слов с первой частью *ангел-*, напр. *ангелоподо́бный*), *а́нгельский*, *архан́гел*, *Евáнгелие*, *евáнгельский* нет никакой необходимости: эти слова в оригинальных церковнославянских текстах принято писать под титлами, так что

---

[6] А цифровая запись чисел, которая, как известно, в церковнославянском осуществляется при помощи букв под титлами, переводится в запись арабскими цифрами.

[7] Здесь могут оказаться релевантными различия между региональными вариантами церковнославянского языка: «ять» при чтении церковнославянских текстов на Западной Украине и часто в Польше (где среди православных много этнических украинцев) читается как [и]; соответственно, в текстах ЦСГЗ, изданных на Западной Украине, «ять» иногда передается как *і*.

[8] Это преобразование, необходимое для получения ЦСГЗ, привычной для подавляющего большинства пользователей, не учитывается системой поиска в ЦСНКРЯ (даже в «модернизированной» орфографии).

сочетание *нг* должно появиться уже на этапе раскрытия титл в соответствии с заранее заданным списком. Напротив того, в слове *а́ггел*, обозначающем злого духа (в противопоставлении св. ангелу) сохраняется сочетание *гг*, что соответствует традиции читать в этом слове звук [г’][9]. Соответственно, различаются:

(1) *…я́ко А́нгелом Свои́м запове́сть о тебе́… (Пс. 90, 11)*

(2) *…се́ А́з посыла́ю А́нгела Моего́ пред лице́м Твои́м… (Мк. 1, 2)*

с одной стороны, и:

(3) *…дадеся́ ми́ па́костник пло́ти, а́ггел сатани́н, да ми́ па́кости де́ет… (2 Кор. 12, 7)*

(4) *Отрица́еши ли ся сатаны́, и все́х де́л его́, и все́х а́ггел его́…? (из последования крещения)*

— с другой. Бывают и случаи, когда *а́нгел* и *а́ггел* встречаются в пределах одной фразы, как в следующем примере из словаря [Кравецкий, Плетнева 2016: 83], представленном здесь в соответствии с нормами ЦСГЗ:

(5) *…Михаи́л и А́нгели его́ бра́нь сотвори́ша со змие́м, и змиі́й бра́ся и а́ггели его́…*

(Заметим в скобках, что, помимо слова *а́ггел* как обозначения злого духа, сочетание *гг* встречается в календарном имени *Агге́й* и читается оно опять-таки как [г’].)

Наконец, на последнем этапе побуквенной передачи церковнославянского текста происходит обработка надстрочных знаков. Знак «звательце», исторически восходящий к знаку придыхания, просто элиминируется, а знаки ударения: острого, тяжелого и облегченного — заменяются единым знаком ударения, принятым в современной гражданской графике, как это видно из приведенных выше примеров. (Знаки титл были элиминированы еще на этапе раскрытия титл.)

Здесь, правда, надо сделать одну оговорку. В реально существующих изданиях ЦСГЗ знаки ударения часто не ставятся над односложными словами, а в некоторых изданиях они оставляются только в тех словах, которые, по мнению издателей, могут вызвать затруднения у современного русскоязычного читателя. Эту практику нельзя одобрить. Решение, согласно которому то или иное слово считается «не вызывающим затруднений», не только не представимо алгоритмически, но по существу своему оказывается крайне субъективным. Но даже и снятие ударений над односложными словами, легко представимое алгоритмически, может привести к неточностям при чтении вслух текста ЦСГЗ. При таком решении стираются различия между безударными служебными односложными словами (предлогами и частицами) и односложными автосемантичными словами. Между тем отсутствие ударения на предлоге или частице может казаться вовсе не очевидным русскоязычному читателю. Так, в сочетании *по су́ху* (из ирмоса первой песни ряда канонов, в частности широко употребительного в молитвенной практике покаянного

[9] Сам собою вспоминается неподражаемый диалог из романа Михаила Булгакова «Белая гвардия»: *Он уехал в царство антихриста в Москву, чтобы подать сигнал и полчища ангелов вести на этот Город в наказание за грехи его обитателей… — Это вы большевиков ангелами? Согласен.*

канона, а также, напр., канона молебного при разлучении души от тела: *Яко по су́ху пешеше́ствовав Изра́иль, по бе́здне стопа́ми, гони́теля фараóна ви́дя потопля́ема…*) в церковнославянском тексте предлог *по* не несет на себе ударения. Однако, если ударение снимается в односложных словах, русскоязычный читатель с большой долей вероятности может контаминировать это сочетание с русским наречием *посуху*, в котором ударение стоит как раз на первом слоге, и произносить предлог *по* с ударением. Поэтому разумно предпочесть практику, в соответствии с которой знак ударения сохраняется и над односложными словами.

В результате проведенных преобразований мы потеряем значительную часть информации, относящейся к семантическим и грамматическим свойствам словоформ. Дело в том, что одинаково читающиеся буквы и разные знаки ударения используются в церковнославянском языке для разграничения на письме омофонов, т. е. словоформ, совпадающих в произношении, но различающихся по семантике или грамматическим характеристикам. Преобразования приведут к тому, что омофоны превратятся в полные омонимы. Так, перестанут различаться разные понимания слов *язы́к* (*γλώσσα* и *ἔθνος*), *ми́р* (*εἰρήνη* и *κόσμος*), форм *ра́б* (И. ед. и Р. мн.), *творя́щим* (Тв. ед. и Дат. мн.) и т. д. Однако фонетическая информация (то, как текст должен звучать при чтении вслух) потеряна не будет, а это и есть критерий «правильности» ЦСГЗ.

В отношении буквенного состава словоформ в реальных текстах ЦСГЗ вариативность крайне незначительна. Так, имя пророка Илии может передаваться двояко: в одних изданиях как *Илиа́*, а в других — как *Илия́*. Второй способ точнее передает произношение, но потребует внесения дополнительного преобразования в процедуру. Разумеется, решение вопроса, какой способ передачи предпочесть, не входит в компетенцию автора статьи.

3. Знаки препинания и прописные буквы в ЦСГЗ

3.1. Знаки препинания в ЦСГЗ

Церковнославянская пунктуация по целому ряду параметров отличается от принятой современной русской пунктуации (напр., в отношении расстановки запятых). Пунктуация в реально существующих изданиях ЦСГЗ, как правило, представляет собою компромисс между нормами церковнославянской и современной русской пунктуации, причем не одинаковый для разных изданий. Разработка процедуры расстановки знаков препинания в ЦСГЗ в значительной мере зависит от того, что мы хотим получить на выходе: текст, в котором знаки препинания максимально соответствуют пунктуации в оригинальном церковнославянском тексте, или текст, который максимально удовлетворяет нормам современной русской пунктуации. При этом мы не располагаем строгой процедурой и для автоматической расстановки знаков препинания в текстах на русском языке.

Начнем с шагов, которые заведомо необходимы при переходе к ЦСГЗ. Следует осуществить автоматическую замену тех знаков, которые имеют в церковнославянском и русском языке разное средство выражения. Так, на первом же

этапе следует заменить церковнославянский вопросительный знак (;) русским (?), а «малую точку» оригинального текста[10] — на точку с запятой ЦСГЗ.

Некоторые издания этими изменениями и ограничиваются. Однако в большинстве случаев пунктуация оригинального текста еще несколько модифицируется.

Для оригинальных церковнославянских текстов чрезвычайно характерен такой знак, как двоеточие (:). При этом функции двоеточия могут быть различны.

Во-первых, двоеточие может выступать как знак того, что для того или иного возгласа или молитвы приведены лишь первые слова. В современных русских текстах сходную функцию может выполнять многоточие. Во многих изданиях ЦСГЗ двоеточие в этой функции и заменяется на многоточие. Так, вместо *Отче наш:* в издании ЦСГЗ может быть напечатано *Отче наш…* (при этом предполагается, что молитва в этом месте должна читаться целиком). Однако в конце таких сочетаний, как *Слава: И ныне:*, а также *Слава, и ныне:* в большинстве изданий ЦСГЗ сохраняется двоеточие (читается *Слава Отцу́, и Сы́ну, и Свято́му Ду́ху, и ны́не, и при́сно, и во ве́ки веко́в, ами́нь*).

Случаи, когда двоеточие выполняет именно эту функцию и, соответственно, уместна его замена на многоточие, распадаются на два класса. Это некоторый набор первых слов наиболее часто повторяющихся возгласов и молитв (как в вышеприведенных примерах), который можно задать списком, а также случаи, когда только что приведенную молитву положено повторить, напр.:

> *Ми́лостию Бо́гови послу́жим, я́ко же Мари́а на ве́чери, и не стя́жим сребролю́бия, я́ко Иу́да: да всегда́ со Христо́м Бо́гом бу́дем.Ми́лостию Бо́гови послу́жим:*

Так этот текст мог бы выглядеть при сохранении двоеточия в качестве знака сокращения. В большинстве современных изданиях ЦСГЗ в таких случаях используется другой прием: после текста молитвы, повторенного один раз, в скобках указывается, что его следует повторить дважды.

Во-вторых, двоеточие используется сходно с двоеточием в современных русских текстах (напр., вводит прямую речь или используется в бессоюзном предложении, в котором вторая часть поясняет первую). Такое двоеточие сохраняется в изданиях ЦСГЗ.

В-третьих, двоеточие иногда используется в качестве показателя смыслового или интонационного членения текста подобно тому, как в современных русских текстах используются точка с запятой или запятая. Такое двоеточие в ЦСГЗ иногда заменяется на запятую или точку с запятой, однако сформулировать четкие формальные критерии, по которым этот случай можно было бы отличить от предыдущего, а также правила, позволяющие осуществить выбор между точкой с запятой и запятой, пока не удается. Здесь необходимы дополнительные исследования.

При этом ЦСГЗ использует не все возможности современной русской пунктуации, в частности не все знаки препинания. В ЦСГЗ не встречаются так

---

[10] «Малая точка» в церковнославянском тексте внешне ничем не отличается от «обычной» точки. Ее формальный признак — после нее предложение продолжается с маленькой буквы.

называемые парные знаки: кавычки и скобки (точнее, скобки могут использоваться для комментариев, касающихся чтения текста, напр. в них может помещаться слово *дважды*). Восклицательный знак (*!*) представляет собою редкость для оригинальных церковнославянских текстов, однако, если он в них встречается, то, как правило, сохраняется и в ЦСГЗ; это можно указать в алгоритме. Напр.:

(6) *О коли́ких бла́г непа́мятлив бы́л еси́!*

Впрочем, некоторые издания ЦСГЗ используют восклицательные знаки значительно чаще, в том числе в тех случаях, когда в оригинальном церковнославянском тексте никакого восклицательного знака не было, напр. при возгласе *Прему́дрость!* Ср. также:

(7) *Ели́цы оглаше́ннии, изыди́те! Оглаше́ннии, изыди́те! Ели́цы оглаше́ннии, изыди́те!*

Это представляется субъективным решением издателей и едва ли может быть предусмотрено алгоритмом.

Знак тире (–) не используется в оригинальных церковнославянских текстах, и появляется лишь в немногих изданиях ЦСГЗ, да и то довольно редко. Напр.:

(8) *…И́же спасе́ние ро́ду челове́ческому ниспосла́вый — Единоро́днаго Сы́на Твоего́…*

В целом введение в текст тире тоже представляется сомнительным решением, тем более что не удается сформулировать строгие и формальные правила, определяющие, когда это следует делать.

3.2. Прописные буквы в ЦСГЗ

Как известно, прописные буквы используются в русском письме в двух различных функциях [Шмелев 2017: 698–699], и эти же функции они выполняют и в ЦСГЗ.

Первая функция (позиционная) заключается в том, чтобы маркировать определенные позиции в тексте. Так, с прописной буквы принято начинать первое слово в каждом самостоятельном предложении. Вторая функция (выделительная) заключается в том, чтобы выделять определенные слова, независимо от строения текста. Обычно, когда говорят о правилах употребления прописных букв, имеют в виду их использование в выделительной функции.

В современном церковнославянском языке в церковнославянской записи прописные буквы используются лишь в позиционной, но не в выделительной функции. Очевидно, что использование прописных букв в позиционной функции будет сохраняться и в ЦСГЗ.

Употребление прописных букв в выделительной функции в каждом языке регулируется в первую очередь традицией. Употребление прописных букв в выделительной функции в ЦСГЗ отчасти соответствуют употреблению прописных букв в современном русском языке, принятому в церковной печати. Соответствующие

правила детально описаны в специальной инструкции для изданий Московской Патриархии [Редакционно-издательское… 2015]. Однако эти правила не являются полностью алгоритмизуемыми.

Понятно, что с прописной буквы пишутся имена собственные в узком смысле слова (в первую очередь — личные имена и географические наименования). Встает вопрос: как установить, что то или иное слово представляет собою имя собственное?

Для церковнославянского языка самое простое решение могло бы состоять в том, чтобы задать списком все имена собственные, которые могут встретиться в церковнославянских текстах, тем более что новые имена собственные появляются в церковнославянских текстах весьма редко; обычно используются перифрастические именования, напр. Ленин может обозначаться как *Русѝ Православ́ной погубѝтелей путевóждь*, Сталин — как *всея́ страны́ тира́н безбóжный и мучѝтель немилосе́рдый*, Ежов — как *приспе́шник мучѝтеля кровáвый* (примеры из одного доклада на конференции по православной гимнографии в ИРЯ РАН в 2014).

Однако, во-первых, указанное правило не абсолютно: изредка в текстах на церковнославянском языке все же появляются имена собственные, не включенные в существующие списки. Так, о мученике, расстрелянном на Бутовском полигоне, может быть сказано:

(9) *У́тру глубоку́ вскра́й ро́ва Бу́товскаго, а́дски зия́юща, по ру́ку свя́занный приведе́н и во главу́ устрелéн…*

При этом слово *Бу́товский* отсутствует в словаре [Кравецкий, Плетнева 2016], который, вообще говоря, составители стремились сделать как можно более полным. Тем самым оно имеет мало шансов быть включенным в заданный список имен собственных.

Во-вторых, некоторые слова могут выступать и в качестве имен собственных, и в качестве имен нарицательных, напр. *А́вгуст* (имя римского императора) и *а́вгуст* (название месяца).

Помимо имен собственных в узком смысле, с прописной буквы в ЦСГЗ пишутся некоторые другие слова, напр. все существительные, обозначающие Бога. В частности, это относится к самому слову *Бо́г*, если в оригинальном церковнославянском тексте оно писалось под титлом; если же слово *бо́г* писалось без титла, то оно относится к языческому божеству и должно писаться со строчной буквы.

С прописной буквы должны писаться не только существительные, обозначающие Бога, но и кореферентные этим существительным личные местоимения. Здесь мы упираемся в то, что мы не располагаем строгой процедурой, однозначно устанавливающей кореферентность (в настоящее время все процедуры установления кореферентных связей в русском тексте носят вероятностный характер). Лишь при наличии такой процедуры правила написания личных местоимений с прописной буквы можно было бы сделать полностью алгоритмичными.

4. Заключительное замечание

Итак, побуквенная передача церковнославянского текста средствами гражданской графики может быть осуществлена на основе формальной процедуры. Однако для внесения в полученную запись пунктуации и прописных букв в выделительной функции необходимо располагать дополнительными средствами, а именно — автоматической системой расстановки знаков препинания (на основе синтаксического анализа?) и системой автоматического установления кореферентности личного местоимения и существительного. Именно здесь ощущается недостаточность лингвистических знаний, хотя современная компьютерная лингвистика делает важные шаги в обоих указанных направлениях. Что касается до полноты исходного списка собственных имен, то это относится скорее к энциклопедической составляющей процедуры перехода. В настоящее время придется ограничиться побуквенным переходом, частичной русификацией пунктуации, а в отношении прописных букв автоматически поставить их в бесспорных случаях и отметить спорные, чтобы осуществить выбор «прописная или строчная?» в режиме ручного редактирования.

References

1. *Editorial and Publishing Design of Church Publications: Handbook of Author and Publisher* (2015) [Redaktsionno-izdatel'skoe oformlenie tserkovnykh pechatnykh izdanii]. Moscow Patriarchate Publ., Moscow.
2. *Kravetskii A. G., Pletneva A. A.* (ed.) (2016), Big Dictionary of Contemporary Church Slavonic [Bol'shoi slovar' tserkovnoslavianskogo iazyka Novogo vremeni], vol. 1, Slovari XXI veka, Moscow.
3. *Prokhvatilova O. A.* (1999), Orthodox Sermon and Prayer as a Modern Speech Phenomenon [Pravoslavnaia propoved' i molitva kak fenomen sovremennoi zvuchashchei rechi]. Volgograd St. Univ. Publ., Volgograd.
4. *Shmelev A. D.* (2017) Capital letters in church and secular press [Propisnye bukvy v tserkovnoi I svetskoi pechati], Sretensky Collection [Sretenskii sbornik], 7–8, 698–734.
5. *Yanko T. E.* (2010) Prosody of sentences with no illocutionary force [Prosodiia predlozhenii so "sniatoi" illokutivnoi siloi]. Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" [Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog"], Issue 9, Moscow, pp. 609–621.
6. *Yanko T. E.* (2011) Accent placement principles in Russian. Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue", Issue 10, Moscow, pp. 712–724.

Литература

1. *Кравецкий А. Г., Плетнева А. А.* (ред.) (2016) Большой словарь церковнославянского языка Нового времени. Т. 1. А–Б. Москва: Словари XXI века.
2. *Прохватилова О. А.* (1999) Православная проповедь и молитва как феномен современной звучащей речи. Волгоград: Издательство ВолГУ, 362 с.
3. *Редакционно-издательское оформление церковных печатных изданий:* справочник автора и издателя. М.: Издательство Московской Патриархии Русской Православной Церкви, 2015. 208 стр.
4. *Шмелев А. Д.* (2017) Прописные буквы в светской и церковной печати // Сретенский сборник. № 7–8. С. 698–734.
5. *Янко Т. Е.* (2010) Просодия предложений со «снятой» иллокутивной силой // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2010 по компьютерной лингвистике и ее приложениям. М.: Издательство РГГУ. С. 609–621.
6. *Yanko T. E.* (2011) Accent placement principles in Russian. Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue", Issue 10, Moscow, pp. 712–724.

AGRR-2019: AUTOMATIC GAPPING RESOLUTION FOR RUSSIAN

Smurov I. M., **Ponomareva M.**
ABBYY, Moscow, Russia

Shavrina T. O.
NRU HSE, Sberbank, Moscow, Russia

Droganova K.
Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic

The 2019 Shared Task on Automatic Gapping Resolution for Russian (AGRR-2019) aims to tackle non-trivial linguistic phenomenon, gapping, that occurs in coordinated structures and elides a repeated predicate, typically from the second clause.

In this paper we define the task and evaluation metrics, provide detailed information on data preparation, annotation schemes and methodology, analyze the results and describe different approaches of the participating solutions.

Key words: shared task, ellipsis, gapping, gapping resolution, Russian

1. Introduction

During the last two decades, just a few works have been dealing with ellipsis detection and resolution and almost exclusively for English. Most of these works address VP-ellipsis, which refers to the omission of a verb phrase whose meaning can be reconstructed from the context [Johnson 2001], for instance, in "Mary loves flowers. John does too" [Hardt 1997]; [Nielsen 2004]; [Lappin 2005]; [McShane and Babkin 2016]. [Anand and Hardt 2016] concentrate on sluicing, which refers to reduced interrogative clauses [Merchant 2001], for instance, in "Mary loves those flowers. I want to know why". [Schuster et al. 2018] and [Droganova and Zeman 2017] focus on gapping (i.e., an omission of a repeated predicate which can be understood from context [Ross 1970]).

To the best of our knowledge, there has been only one attempt to organize a shared task on ellipsis detection and resolution, specifically the shared task dedicated to VP-ellipsis detection and resolution for English, which was one of the SemEval-2010 tasks[1]. Unfortunately, the results of this shared task are not available.

Ellipsis exists in the majority of languages [Merchant 2001]. However, according to [Testelets 2011], a single rule that motivates elliptical constructions cannot

[1] Task 4, descripion avaliable at http://semeval2.fbk.eu/semeval2.php?location=tasks#T14.

be defined even within one language. In addition to the adversity of the construction itself, the phenomenon is naturally rare, thus research was conducted so far on rather small amount of data, not exceeding several hundreds of sentences; with the exception of [Anand and Hardt 2016], whose dataset consists of 4,100 sluicing examples from The New York Times subset of the Gigaword Corpus.

AGRR-2019 aims at detection and resolution of gapping constructions for Russian. For the purpose of the shared task we defined the task and evaluation metrics and developed a gapping dataset for Russian that consists of 7.5k sentences with gapping (as well as 15k relevant negative examples) and comprises data of various genres: news, fiction, social media and technical texts. We hope that the proposed methodology and dataset will encourage further development and regular comparison of systems for gapping detection and resolution.

2. Data

2.1. Linguistic Description

In this work we use the following terminology for gapping elements. We call the pronounced elements of the gapped clause **remnants**. Parallel elements found in full clause that are similar to remnants both semantically and syntactically are called **remnant correlates**. The missing material is called **the gap** [Coppock 2001].

Traditionally gapping is defined as an omission of a repeating predicate in non-initial composed and subordinate clauses where both remnants to the left and to the right remain expressed.

(1) Один имел силу солнца, другой — луны.
one had power sun other moon
'One had the power of the Sun, the other (had the power of) the Moon'

However Russian language allows a broader interpretation, thus it is important to mention the cases that were selected for the shared task and included into the gapping dataset for Russian.

The cases where the second remnant is missing and the second clause contains just one remnant are called stripping and can be considered a special case of gapping [Merchant 2016]. Canonical examples of stripping are limited to a small number of constructions (2)–(4). According to the [Hankamer and Sag 1976] who introduced the termin: "Stripping is a rule that deletes everything in a clause under identity with corresponding parts of a preceding clause except for one constituent (and sometimes a clause-initial adverb or negative)."

(2) The man stole the car after midnight, **but not** the diamonds. [Merchant 2016]

(3) Abby can speak passable Dutch, and Ben, **too**. [Wurmbrand 2013]

(4) Все мы любим Мамбу и Сережа **тоже**.
All we love Mamba and Serezha too
'All of us love Mamba, and Serezha ~~loves it~~ too'

Such examples were not included in the corpus. The set of constructions for Russian that implement stripping seems to be broader than for English. Therefore we encountered wide variety of examples that go beyond the canonical examples. Examples (5) and (6) illustrate the cases when arguments of the elided verb do not fully correspond to the arguments of the pronounced verb, thus some of the arguments of the elided verb (highlighted in bold) do not have correlates. We consider such examples gapping with one remnant and include them in the corpus.

(5) Добавляем муку, крахмал и разрыхлитель, а **в конце** — сметану.
add flour starch and baking.powder and in end sour.cream
'We add flour, starch and baking powder, and at the end ~~we add~~ sour cream.'

(6) Рост цен составил 11,9 процента (**за 2009 год** — 4,4 процента)
growth prices amounted.to 11.9 percent in 2009 year 4.4 percent
'The prices growth amounted to 11.9 percent (in 2009 ~~it amounted to~~ 4.4 percent)'

Elements remaining after predicate omission can be of different nature. Consider the following examples where remnants are predicates (7), preposition phrases (8), adverbs (9), adjectives (10) possibly with their dependents.

(7) Одно может вдохновлять, а другое вгонять в тоску.
one can inspire and other put in melancholy
'One thing can inspire and the other ~~can~~ put you in a melancholy mood.'

(8) Советую вам поменьше думать о проблемах, и побольше
Recommend you less think about problems and more
об их решении.
about their solution
'I recommend you to think less about problems, and ~~think~~ more about solving them.'

(9) Вначале они играли интересно, потом прескучно.
at.first they played interesting.ADV after extremely.boring.ADV
'At first they played interesting, then ~~they played~~ extremely boring.'

(10) Сердце ее было слишком чистым, чувства слишком искренними.
heart her was too pure feelings too sincere
'Her heart was too pure and her feelings ~~were~~ too sincere.'

While collecting the corpus, one of our main goals was to make it diverse. Along with grammatical diversity briefly described above, we intended to make the corpus heterogeneous both lexically and topically. We discuss how different genres contribute to the corpus in the next section.

2.2. Obtaining the Data

Reasonable amount of data is crucial to train a system utilizing machine learning techniques. At the same time, gapping is a relatively rare syntactic phenomenon: according to our data, no more than 5 sentences out of 10,000 contain gapping. Furthermore, annotation is a laborious process and existing corpora do not exceed several hundred examples. Thus, for the purpose of the shares task our priority was to collect

as much data as possible. For this reason we opted to validate automatically obtained markup instead of annotating sentences from scratch.

Compreno parser [Anisimovich et al. 2012] was used to provide syntactic analysis for several millions of sentences. This parser includes a template-based module for gapping detection [Bogdanov et al. 2012] which allowed us to identify sentences with gapping elements. Such sentences were selected and automatically annotated using bracket markup (see subsection **Dataset Format**).

Over 22,500 sentences were shared among 11 assessors. Assessors were asked to evaluate the automatically obtained annotations, classifying each sentence into one of the following classes:

[0] no gapping, no markup is needed;
[1] correctly annotated;
[2] incorrectly annotated;
[3] difficult to analyse.

Each sentence was evaluated by two assessors. If both assessors considered a sentence class 1, it was added to the corpus as a positive example.

Since the markup was only evaluated without correcting it, we managed to collect a reasonably large corpus in a relatively short time.

To serve the training purpose, the corpus has to include negative examples. We considered two types of negative examples to select more relevant sentences. The first type comprises problematic negative sentences on which Compreno parser false-positively predicted gapping (labeled with 0 by both assessors). Introducing negative examples of this type supposedly would allow a system to improve upon the results of the source parser. The second type comprises sentences not shorter than 6 words that contain dash or comma, and a verb. We made the negative class twice as large as the positive one.

We intended to produce a corpus comprising a variety of genres. The main part of the corpus consist of fiction, technical texts and news. We deliberately added texts from social media and balanced their proportion in both positive and negative classes, so they form 25% of the corpus.

All obtained sentences were split in development set and training set in proportion 1:5. For the final submission, the participants were allowed to train their systems on training set and development set jointly.

The annotation of the test set was evaluated by the organizers: it contains ten times less examples than joined training and development sets with the same distribution of genres and the same ratio of positive to negative classes.

In addition, we released optional training materials, that comprise 115,563 examples of noisy data with the same proportion of positive and negative examples. The annotation was obtained automatically by Compreno without further manual validation.

Table 1: Number of examples by class; vk stands for social media texts

		0		1		sum
dev	vk	670	2,760	326	1,382	20,548
	other	2,090		1,056		
train	vk	2,860	10,864	1,366	5,542	
	other	8,004		4,176		
test	vk	343	1,365	185	680	2,045
	other	1,022		495		
sum		14,989		7,604		22,593

2.3. Dataset Format

We have two versions of annotation schemata. The first schema provides human-readable format useful for analysing and evaluating the annotation. Square brackets are utilized to mark all gapping elements (whole NP, VP, PP etc. for remnants and their correlates and the predicate controlling the gap). The gap is marked with **V**. The syntactic head of the predicate that corresponds to the ellided predicate is marked with **cV**.

- V—the gap
- cV—the head of the VP that controls the gap
- R1—the first remnant
- cR1—correlate of the first remnant
- R2—the second remnant
- cR2—correlate of the second remnant

The sentence (10) would have the following bracket annotation (11).

(11) [$_{cR1}$ Сердце ее] [$_{cV}$ было] [$_{cR2}$ слишком чистым], [$_{R2}$ чувства] [$_{V}$] [$_{R2}$ слишком искренними].
heart her was too pure feelings too sincere
'Her heart was too pure and her feelings (were) too sincere.'

While the bracket format is convenient for human analysis, it is less suitable as input for automatic systems. Thus we utilize the alternative format: information concerning every sentence is represented by 8 columns. The first column contains plain text, which serves as input for automatic systems. The second column contains 0 or 1 depending on the presence of gapping. The rest of the columns correspond to gapping elements (cV, cR1, cR2, V, R1, R2) and contain character offsets for annotation borders for each gapping element if it is present in the sentence. Consider an example (12).

(12) **text**
Сердце ее было слишком чистым, чувства слишком искренними.

class	**cV**	**cR1**	**cR2**	**V**	**R1**	**R2**
1	10:14	0:9	15:30	39:39	31:38	39:57

2.4. Assessment Analysis

In this section we provide analysis of the examples that were labeled as class [2] or [3] by the assessors. Tables 2 and 3 show the confusion matrices of assessors' marks.

Table 2. Assessment analysis for the subcorpus of technical and fiction texts

	0	1	2	3
0	1,533	138	129	136
1	240	5,301	1,021	237
2	213	451	1,600	281
3	307	177	117	108

Out of 11,989 sentences 44% were considered correctly annotated and 13% were unanimously considered to have no gapping.

Table 3. Assessment analysis for the social media subcorpus

	0	1	2	3
0	1,817	232	174	118
1	154	1,900	142	46
2	75	130	360	21
3	139	53	36	25

Out of 5,422 sentences 35% were considered correctly annotated and 34% were unanimously considered to have no gapping.

The annotators classified slightly more than half of the automatically annotated examples as correctly annotated or having no gapping at all. Out of the rest of examples the most interesting are the examples unanimously attributed to class 2—incorrect annotation of sentence with gapping—and class 3—problematic sentences that are difficult to analyse.

Let us illustrate cases frequently encountered in these two classes. All sentences are given with automatic annotation, which has errors that show the bias of the source system and the corpus.

The following cases are common for class 2:

- Gapping with more than two remnants

(13) В Виннице больше оставаться было нельзя, [$_{cR1}$ семья] [$_{cR2}$ самолётом] [$_{cV}$ отправилась] **в Россию**, а [$_{R1}$ я] [$_{V}$] [$_{R2}$ поездом **на восток**].
in Vinnitsa longer stay was impossible family plane travaled to Russia and I train to east
'It was impossible to stay any longer in Vinnitsa, and the family traveled by plane to Russia, while I took a train to the east.'

Among other cases listed below, this is the only case that always gets erroneous annotation due to the limitations of the rule-based algorithm for gapping detection in Compreno.

- Lack of markup in some of multiple clauses that contain a gap

(14) [$_{cR1}$ В Петербурге] делами [$_{cV}$ ведал] старший сын [$_{cR2}$ Фёдор],
in St.Petersburg business involved eldest son Fedor
[$_{R1}$ в Казани] —[$_{V}$] [$_{R2}$ Иван], **в Ростове и Рыбинске — Дмитрий,**
in Kazan Ivan in Rostov and Rybinsk Dmitry
в **Самаре** — **Михаил**.
in Samara Mikhail.
'In St. Petersburg the eldest son Fedor was involved in business, in Kazan—Ivan, in Rostov and Rybinsk—Dmitry, in Samara—Mikhail.'

- Particular type of gapping when the correlate clause semantically generalizes over instances described in following clauses

(15) [$_{cR1}$ Два Ангела] [$_{cV}$ уселись] на плечах: один— [$_{cR2}$ на левом],
two angels sat on shoulders one on left
а [$_{R1}$ второй] —[$_{V}$] [$_{R2}$ на правом].
and second on right
'Two Angels sat on their shoulders: the first set on the left and the second on the right.'

This type of gapping is not limited to semantic relations between the clauses. The main clause may lack the correlates of some remnants, e.g. *в правую руку, в левую руку* in (16).

(16) [$_{cV}$ Возьмите] лист [$_{cR1}$ бумаги] и два карандаша разного цвета:
take piece paper and two pencils different colour
один [$_{cR2}$ в правую руку], [$_{R1}$ другой] —[$_{V}$] [$_{R2}$ в левую].
one in right hand another in left
'Take a piece of paper and two different coloured pencils: one in the right hand, the other in the left.'

- The correlates may remain unmarked in case of coordinated predicates in the full clause.

(17) **Ты** продолжала молчать и оценивающе [$_{cV}$ смотрела]
you kept be.silent and appraisingly looked
[$_{cR2}$ на меня], а [$_{R1}$ я] [$_{V}$] [$_{R2}$ на тебя].
at me and I at you
'You kept silent and were looking at me appraisingly, while I ~~was looking~~ at you.'

- Incorrectly predicted boundaries of gapping elements. In (18) the unknown word *Суне* may be the reason for the erroneous prediction.

(18) Тётя Яна [$_{cV}$ купила] [$_{cR1}$ **своей**] [$_{cR2}$ **Суне** сказки], а [$_{R1}$ себе] [$_{V}$] [$_{R2}$ прописи].
Aunt Yana bought her Suna fairy.tales and for.herself copybook
'Aunt Yana bought a book of fairy tales for Suna and a copybook for herself.'

- When gapping appears deeper in the syntactic tree, the main clause of the whole sentence may be erroneously predicted as correlate.

(19) Поэтому-то [$_{cR2}$ Евангелие] и [$_{cV}$ советует] нам благословлять,
That.is.why Gospel PART advises us to.bless
а не проклинать, так как **благословение** **приносит** **благо**,
and not to.curse because blessing brings good
а [$_{R1}$ проклятье] — [$_{V}$] [$_{R2}$ беду и несчастье].
and curse misfortune and grief
'That is why the Gospel advises us not to curse but to bless, because blessing brings good, and curse ~~brings~~ misfortune and grief.'

- The pair of a remnant and its correlate are missing in annotation

(20) [$_{cR1}$ Кто-то из нас] [$_{cV}$ выживает] **благодаря**, а [$_{R1}$ кто-то] **вопреки**.
Somebody of us survive due.to and somebody despite.of
'Some of us survive due to something, and some, despite of something.'

- The analysis that is syntactically possible, but semantically doubtful and causes incorrect sentence interpretation. In (21) the correlate of R1 is erroneously detected due to morphological homonymy of *слова мысли* (it is interpreted as NomPl), thus the correlate of the predicate is predicted incorrectly as well.

(21) [$_{cR1}$ Евангелие] [$_{cV}$ призывает] человека **привести** свои **дела**
Gospel encourages person to.bring their deeds
[$_{cR2}$ в соответствие со словами], [$_{R1}$ слова] [$_{V}$] [$_{R2}$ в соответствие
into.line with words words into.line
с мыслями], а [$_{R1}$ мысли] — [$_{V}$] [$_{R2}$ в соответствие со Словом Божиим].
with thoughts and thoughts into.line with word of.God
'The Gospel encourages a person to bring their deeds into line with words, ~~to bring~~ their words into line with thoughts, and ~~to bring~~ their thoughts into line with the Word of God.'

- The parser may miss some remnants in coordinated clauses that contain a gap. In this case some remnants may be erroneously merged together and form one remnant instead of two that would correspond to different correlates.

(22) [$_{cV}$ Нарезать] [$_{cR1}$ лук и шампиньоны **полукольцами**, **куриное** **филе**]
to.slice onion and champignons half.moons chicken fillet
[$_{cR2}$ кубиками], а [$_{R1}$ картофель] [$_{V}$] [$_{R2}$ полосками].
cubes and potatoes sticks
'Slice the onion and champignons into half moons, ~~dice~~ the chicken fillet into cubes, and ~~cut~~ the potatoes into sticks.'

- Coordinated correlates or remnants are not predicted as an entire gapping element

(23) [$_{cR1}$ Раньше] я [$_{cV}$ хотела] [$_{cR2}$ любви] **и** **замуж**, а [$_{R1}$ сейчас]
previously I wanted love.NOUN and married.ADV and now
[$_{R2}$ кожанку и джип].
leather.jacket and jeep
'Previously, I wanted love and to get married, and now ~~I want~~ leather jacket and jeep.'

Both assessors considered approximately 1% of all the automatically annotated examples problematic. The cases where the markup was inapplicable rather then wrong or the assessors could not mark an example as lacking any kind of gap are the following:

- Canonical stripping (with *тоже, нет* etc.)

(24) Пронумеруйте [$_{cR1}$ такты], а то [$_{cV}$ глаза] [$_{cV}$ могут] сместиться,
number.IMP bars because eyes may shift
а [$_{R1}$ цифры] — **нет**!
and numbers not
'Number the bar lines, because the eyes may shift, but not the numbers'

(25) [$_{cR1}$ Мертвых] [$_{cR2}$ мы] охотно [$_{cV}$ принимаем] сюда, но [$_{R1}$ живых] — [$_{V}$]
dead.NOUN we gladly accept here but living.NOUN
[$_{R2}$ **дудки**]!
no.EXCLAM
'We gladly accept the dead here, but the living—not on your life'

- Stripping with less typical markers

(26) Они [$_{cV}$ оказывают] психологическую поддержку [$_{cR2}$ жертвам
they provide psychological support victims
землятресений], **в особенности** [$_{V}$] [$_{R2}$ детям].
earthquakes especially children
'They provide psychological support to earthquake victims, especially to children'

- Conjunction rather then gapping

(27) [$_{cR2}$ Рисовать рисунок] [$_{cV}$ надо] на кальке, а затем [$_{V}$]
to.draw picture should on tracing.paper and then
[$_{R2}$ вырезать как показано на картинке].
cut.out as is.shown on figure
'One should draw the picture on tracing paper and then cut it out as the figure shows.'

(28) Меня [$_{cV}$ попросили] [$_{cR1}$ привезти] вас [$_{cR2}$ сюда], а [$_{R1}$ самому поехать]
me asked bring you here and myself go
[$_{V}$] [$_{R2}$ куда-нибудь еще].
somewhere else
'I was asked to bring you here and to go somewhere else myself.'

3. Shared Task Set-Up

Training data were released on January 26th 2019, automatic noisy data were released a week after. Participants had approximately a month to create solutions (system submissions were due February 23rd) and the results were announced on March 5, 2019.

Further details on the task schedule, evaluation, and results are available on the task web site at: https://github.com/dialogue-evaluation/AGRR-2019.

3.1. Shared Task

We offered participants two tracks concerning different technological limitations:

1. Closed track—an open-source track, convenient for research groups and student teams. Participants of the track were allowed to train their models only on open-access data (open source dictionaries, word embeddings, open universal embedders, open parsing systems, etc.). To verify the results, participants placed their code and models on github making it publicly available both for organizers and other teams.
2. Open track—no restriction on data and systems used; recommended for participants from industry who present their products. Track participants were allowed to bring any data for learning beyond the provided data and use their own commercial programs. Github sharing was not required.

All the systems participating in the shared task have chosen closed (open-source) track. All the models are publicly available on participants' github (links can be found at AGRR github page).

The participants were offered 3 different gapping tasks:

1. Binary presence-absence classification—for every sentence, participating systems must decide if there is a gapping construction in it.
2. Gap resolution—for every sentence with gapping, participating systems must predict the position of the elided predicate and the pronounced predicate in the antecedent clause.
3. Full annotation—for every sentence with gapping, participating systems must predict the linear position of the elided predicate and positions of its remnants in the clause with the gap, as well as the positions of remnant correlates and pronounced predicate in the antecedent clause.

3.2. Metrics

For the binary classification task we have decided to use standard metrics: precision, recall and f-measure (the participants' submissions were ranked according to the latter one). For the tasks 2 and 3, we have decided to avoid using standard metrics that require gold-standard tokenization. Our main motivation was to allow participants to use any available syntactic parser (since tokenization is often a part of a syntactic parsing pipeline, choosing any particular tokenization could have potentially made some parsers less suitable for the shared task than others). Given this reasoning, for gap resolution and full annotation tasks we have chosen symbolwise f-measure as the main metric. More specifically:

- true-negative samples for binary classification task do not affect total f-measure;
- for true-positive samples, symbolwise f-measure is obtained for each relevant gapping element separately, thus generating 6 numbers for full annotation task and 2 numbers for gap resolution task (if the evaluated sentence is either false-positive or false-negative, all the generated numbers are equal to 0);
- the obtained f-measures are macro-averaged on the whole corpus.

For instance, if the gold standard offset for particular gapping element is 10:15 and the prediction is 8:14, we have 4 true positive chars, 1 false negative char and 2 false positive chars, and the resulting f-measure equals 0.727.

It should be noted that the evaluation results on task 1 are always greater or equal to the results on tasks 2 and 3 (and while f-measure on task 2 may theoretically be lower than f-measure on task 3, the former is normally expected to be higher than the latter). This feature correlates with the hierarchy of the tasks: each subsequent task requires solving the previous ones (i. e. tasks 2 and 3 have nonzero annotations only on the sentences with gapping and task 3 provides richer annotation than task 2).

3.3. Results

Research groups from various Russian universities (MIPT, MSU, HSE, IITP, NSU), participants from two IT companies and several independent researchers have taken part in the competition, making 9 teams in total.

Binary classification and gap resolution tasks were equally popular among the participants (all teams have submitted solutions for the tasks); all teams but one have also participated in full annotation task. Final results are shown in Table 4 (sorted by gap resolution score). The implemented solutions are described in detail in the next section.

Table 4. The official results of the AGRR-2019 shared task

	binary			gap resolution	full
team	precision	recall	f-measure	f-measure	f-measure
fit_predict	0.969	0.95	**0.959**	**0.905**	**0.892**
EXO	0.899	0.964	**0.931**	**0.815**	**0.786**
Koziev Ilya	0.774	0.903	0.834	**0.677**	**0.647**
Derise	0.801	0.906	**0.850**	0.665	0.622
Meanotek	0.891	0.781	0.832	0.635	0.514
МГУ-DeepPavlov	0.934	0.644	0.762	0.601	0.587
vlad	0.778	0.915	0.841	0.574	
MorphoBabushka	0.763	0.619	0.683	0.466	0.440
nsu-ai	0.485	0.123	0.196	0.037	0.036

Some participants have submitted their solutions after the deadline (but before the release of the test data). These solutions were not scored alongside the official results of the AGRR-2019 shared task. These results are given in a separate table (see Table 5).

Table 5. Results of after-deadline submissions

	binary			gap resolution	full
team	precision	recall	f-measure	f-measure	f-measure
МГУ-DeepPavlov	0.973	0.646	0.776	0.617	0.599
МГУ-DeepPavlov	0.898	0.934	0.916		
МГУ-DeepPavlov[2]	0.97	0.712	0.821	0.658	0.653
EXO	0.946	0.946	**0.946**	**0.859**	**0.836**
Meanotek	0.815	0.939	0.872	0.727	0.688

3.4. Methods

All participants but one (МГУ-DeepPavlov) have reduced gap resolution and full annotation tasks to sequence labeling task. The most fruitful approaches were to enhance standard BLSTM-CRF architecture [Lample et al. 2016]; [Ma and Hovy 2016], to pretrain LSTM-based language model or to use transformer-based solutions [Vaswani et al. 2017]; [Devlin et al. 2018].

The methods used are summarized in Table 6.

Table 6. Methods of the AGRR participants

team	architecture	token features	sequence labeler	additional features
fit_predict	Trasformer (BERT)	BERT (pretrained)	Custom FSA-based postprocesser	Joint model resolving both full annotation and binary classification; Noisy data (not validated by assessors) was used.
EXO	BLSTM + MultiHead self-attention	BERT (pretrained)	NCRF++ (n-best CRF implemented in [Yang and Zhang 2018])	Joint model resolving both full annotation and binary classification.
Koziev	BLSTM	Word2vec + CharRNN(CNN)	CRF	Separate models for binary classification and full annotation tasks.
Derise	BiGRU, Transformer	fastText	None	Separate models for binary classification (BiGRU) and full annotation (Transformer) tasks.
Meanotek	2 layer LSTM	Character-level LM (LSTM)	None	Full annotation task model was trained; Binary classification is resolved with heuristics on full annotation.

[2] These results were further improved two weeks after the end of AGRR-2019, when all the gold answers and the systems of the other participants were available. We do not consider these results relevant for the shared task and thus do not to include them into this paper.

team	architecture	token features	sequence labeler	additional features
МГУ-Deep-Pavlov	Rule-based; BLSTM model (submitted after deadline)	ELMo, UDPipe, Morphological features	Not sequence labeling approach	Rule-based system (scored system) BLSTM model (submitted after deadline) uses dot-product similarity to determine if a pair of tokens are a particular pair of gapping elements (cV and V, V and R1 etc)
vlad	ULMFit + linear decoder	ULMFit (pretrained)	None	Separate models for binary classification (MLP) and full annotation (linear decoder) tasks.
Morpho-Babushka	BERT	BERT (pretrained) + Pymorphy2	None	Separate models for binary classification and full annotation tasks.
nsu-ai	BERT	BERT (pretrained)	None	Joint model, separate outputs for class (one per sentence) and each gapping element label (one per token).

Most participating systems did not use any token-level features other than word embeddings, character-level embeddings, or language model embeddings [Peters et al. 2018]; [Devlin et al. 2018]; [Howard and Ruder 2018]. Of particular note is that neither of the top-scoring systems use morphological or syntactic features. While it may be theorized that using such features could yield some improvements, we suppose that language model embeddings (especially when coupled with self-attention like in the top two systems) contain most syntactic information relevant for ellipsis resolution.

4. Conclusion

In this paper we have introduced Automatic Gapping Resolution for Russian (AGRR-2019), the first shared task centered on gapping. We have outlined the design of the dataset used for the shared task and provided a brief assessment analysis. We have defined three tasks on gapping detection and resolution as well as evaluation metrics. Finally, we have presented the official results of the shared task.

The two most important features of our dataset are its diversity and size. Russian language allows rather broad interpretation of gapping (see section Linguistic Description for details). Furthermore, we were able to increase the diversity of our corpus not only by varying its genre composition, but also by including a substantial social media component (see details in section Obtaining the Data). The size of our corpus (7.5k sentences with gapping and 15k relevant negative sentences) is sufficient for successful gapping resolution with ML-methods (as shown in Results). To the best of our knowledge no other publicly available dataset contains a comparable amount of gapping examples.

The task attracted considerable attention from a number of researchers, but only nine teams have submitted their solutions. Nevertheless the participants have demonstrated that gapping can be successfully resolved using sequence-labeling techniques

(the best solution has achieved 0.96 in gapping classification task, 0.91 in gap resolution task and 0.89 in full annotation task). A surprising observation is that rich morphological and syntactic features are not necessary to achieve satisfactory results on gapping resolution.

We hope that AGRR-2019 has provided useful insights both for researchers interested in improving parsing quality and those who study theoretical aspects of gapping. We believe that it is a small step towards fully resolved ellipsis.

5. Acknowledgments

Authors are thankful to Alexey Bogdanov for help with linguistic analysis and useful advices.

The work was partially supported by the GA UK grant 794417.

References

1. *Anand P. and Hardt D.* (2016). Antecedent selection for sluicing: Structure and content. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1234–1243.
2. *Anisimovich K. V., Druzhkin K. Ju., Minlos F. R., Petrova M. A., Selegey V. P. and Zuev K. A.* (2012). Syntactic and semantic parser based on abbyy compreno linguistic technologies. In Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"], Bekasovo, Vol. 2, pp. 90–103.
3. *Bogdanov A.* (2012). Description of gapping in a system of automatic translation. In Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"], Bekasovo, Vol. 2, pp. 61–70.
4. *Coppock E.* (2001). Gapping: In defense of deletion. In Proceedings of the Chicago Linguistics Society, Vol. 13, pp. 133–148.
5. *Devlin J., Chang M. W., Lee K. and Toutanova K.* (2018). Bert: Pretraining of deep bidirectional transformers for language understanding. In arXiv preprint arXiv:1810.0480.
6. *Droganova K. and Zeman D.* (2017). Elliptic Constructions: Spotting Patterns in UD Treebanks. In Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017), 48–57.
7. *Hankamer J. and Sag I.* (1976). Deep and surface anaphora. In Linguistic Inquiry, 7:391–426.
8. *Hardt D.* (1997). An empirical approach to VP ellipsis, Computational Linguistics, MIT Press, Vol. 23(4), pp. 525–541.
9. *Howard J. and Ruder S.* (2018). Universal language model fine-tuning for text classification. In Association for Computational Linguistics
10. *Johnson K.* (2001). What VP ellipsis can do, and what it can't, but not why, Citeseer

11. *Lample G., Ballesteros M., Subramanian S., Kawakami K. and Dyer C.* (2016). Neural Architectures for Named Entity Recognition. In NAACL-HLT.
12. *Lappin S.* (2005). A sequenced model of anaphora and ellipsis resolution, Anaphora Processing: Linguistic, Cognitive, and Computational Modelling. Amsterdam: John Benjamins, pp. 3–16.
13. *Ma X. and Hovy E.* (2016). End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics.
14. *McShane M. and Babkin P.* (2016). Detection and resolution of verb phrase ellipsis, LiLT (Linguistic Issues in Language Technology), Vol. 13.
15. *Merchant J.* (2001). The syntax of silence: Sluicing, islands, and the theory of ellipsis, Oxford University Press on Demand.
16. *Merchant J.* (2016). Ellipsis: A survey of analytical approaches. University of Chicago, Chicago, IL.
17. *Nielsen L. A.* (2004) Verb phrase ellipsis detection using automatically parsed text. In Proceedings of the 20th international conference on Computational Linguistics, Association for Computational Linguistics, p. 1093.
18. *Peters M. E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K. and Zettlemoyer L.* (2018). Deep contextualized word representations. In Proceedings of NAACL.
19. *Ross J. R.* (1970). Gapping and the order of constituents. In Manfred Bierwisch and Karl Erich Heidolph, editors, Progress in linguistics: A collection of paper, De Gruyter, 43:249–259.
20. *Schuster S., Nivre J. and Manning C.* (2018). Sentences with Gapping: Parsing and Reconstructing Elided Predicates, arXiv preprint arXiv:1804.06922.
21. *Testelets Ya. G.* (2011). Ellipsis in Russian: Theory versus Description. Typology of Morphosyntactic Parameters [Ellipsis v russkom yazyke: teoreticheskiĭ i opisatel'nyĭ podkhody], Typology of Morphosyntactic parameters, MSUH, pp. 1–6.
22. *Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł. and Polosukhin I.* (2017) Attention is all you need. In Advances in Neural Information Processing Systems, pp. 5998–6008.
23. *Wurmbrand S.* (2013). Stripping and topless complements. Ms., University of Connecticut.
24. *Yang J. and Zhang Y.* (2018). NCRF++: An Open-source Neural Sequence Labeling Toolkit. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics.

PHRASE-BASED ATTENTIONAL TRANSFORMER FOR HEADLINE GENERATION

Sokolov A. M. (sokolov.andrej.m@gmail.com)
SPBU, Saint-Petersburg, Russia

Nowadays the task of selecting key information from large amount of text data is becoming more and more relevant. This article proposes a model of deep neural network with phrase-based attentional mechanism used for automatic generation of news headlines. The proposed architecture achieves a new state-of-the-art on the RIA news dataset.

Key words: text summarization, headline generation, Russian language, neural networks, self-attention

PHRASE-BASED ATTENTIONAL TRANSFORMER ДЛЯ ГЕНЕРАЦИИ НОВОСТНЫХ ЗАГОЛОВКОВ

Соколов А. М. (sokolov.andrej.m@gmail.com)
СПбГУ, Санкт-Петербург, Россия

В настоящее время задача выделения ключевой информации из больших объемов текстовых данных становится все более и более востребованной. В данной статье предлагается модель глубокой нейронной сети с фразовым механизмом внимания, применяемой для автоматической генерации новостных заголовков. Предложенная архитектура достигает наилучших на данный момент результатов на наборе новостей РИА.

Ключевые слова: автоматическое реферирование текстов, генерация заголовков, нейронные сети, механизм внимания

1. Introduction

Name of the task of generating news headlines is pretty self-explanatory: having a news text you need to generate a short title for it that reflects an essence of the news. This problem is a special case of the *abstractive summarization*. In contrast to the *extractive summarization*, where it is sufficient to select the most important words or sentences from the text, in the abstractive summarization we can use paraphrasing or words not contained in the original text.

The rapid development of recurrent networks and language models shakes-up research of abstractive summarization methods. Transformer architecture [17] became an excellent replacement of RNN and allowed us to train deep networks faster without loss of quality. A key part of the Transformer is an attention mechanism. In classic version of the Transformer, attention allows to model and recognize connections between individual tokens, ignoring connections between phrases directly. An architecture used in this article is based on *Phrase Based Attentions* [10], which allows us to fix the described drawback.

RIA Dataset proposed in [4] has been used for training. The dataset consists of approximately one million headline-news pairs of the Russian news agency "Rossiya Segodnya".

Section 2 describes the architecture of the model used. **Section 3** discusses the dataset, data preprocessing pipeline and training in more detail. **Section 4** and **5** describe the experiments and results respectively. **Section 6** is devoted to the analysis of offered approach shortcomings and reflections on ways of its solution. **Section 7** provides a brief overview of abstract summarization methods proposed by various researchers. In the last section you can find conclusions and arguments on the work done.

2. System description

Denote $\mathbf{x}$, $\mathbf{y}$ as sequences of the news text tokens vector representations $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and news header $\mathbf{y} = (y_1, y_2, \ldots, y_m)$. We will use a statistical language model $p(\mathbf{y}|\mathbf{x}, \theta) = \prod_{i=1}^{m} p(y_i|y_{i-1}, \ldots, y_1, \mathbf{x}, \theta)$ to generate header, where θ— model parameters. It is proposed to use the classical Transformer architecture [17] with *heterogeneous* attention [10] as language model parameterization. Unlike recurrent neural networks, which have been a state-of-the-art approach to NLP problems for a long time, the use of the Transformer allows us to learn more effectively. The self-attention mechanism affords to calculate hidden representations of sequences in parallel, while RNN hidden state h_t can be obtained only after calculating the previous state h_{t-1}. One modification applied to the original transformers in our article is *heterogeneous* attention. This type of attention extends receptive fields of the model adding an ability to directly model relationships between tokens and phrases. This effect is achieved through the use combination of 1, 2 kernel convolutions applied to the input sequences of attention blocks. Then this convolved sequences are concatenated and used as input of multi-head attention. In the interest of space, we omit the details and send an interested reader to the original articles [10], [17].

3. Data and training

3.1. Dataset

We use RIA dataset[1] for train. It contains 1,003,869 news with written headings. On average, header consists of 10 words, and text of news consists of 316 words. A subset of 20,000 examples of this dataset was reserved for testing proposes. The remaining part is used for training.

3.2. Preprocessing

First of all, the entire text of the dataset was reduced to lowercase, all html tags and their contents were removed. In order to simplify model training we use only the first 3 news sentences. It is acceptable due to the fact that the main essence of news contains in the first few sentences. We also limit the maximum number of tokens processed to 150.

The next step in data preparation is to split the text into tokens. It is proposed to use *Byte Pair Encoding* [15] as a tokenization method for both news text and header. This approach is currently a state-of-the-art tokenization method for NLP problems. BPE solves a problem of out-of-vocabulary words and works well with morphologically rich languages. We also use *word2vec* [8] as a vector representation of tokens.

3.3. Training

Using a given model parameterization $p(\mathbf{y}|\mathbf{x}, \theta)$, we will minimize the *negative log likelihood* function $NLL(\theta) = -\log \mathcal{L}(\theta) = -\sum_i \log p(\mathbf{y}_i|\mathbf{x}_i, \theta)$ during training, where $(\mathbf{x}_i, \mathbf{y}_i)$ are training pairs of the dataset.

Models were trained using *Adam* [5] optimizer with parameters $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$ with learning rate $lr = 10^{-4}$. Each iteration of the optimizer uses a batch of data of size equal to 16. Training continues for 10 epochs.

3.4. Inference and evaluation

During inference we will use a *beam-search* algorithm. Unlike a greedy approach, where a token with the maximum probability is selected for each decoding step, *beam-search* uses m most likely independent header generations, trying to maximize resulting likelihood.

The most popular summarization quality metric is ROUGE [6]. We use ROUGE-1 (unigram), ROUGE-2 (bigram) and ROUGE-l (longest common subsequence) version, scored by F-measure.

[1] The dataset is available at https://vk.cc/8W0l5P

4. Experiments

In this article[2], *BPE* and *word2vec* were trained on the training part of the dataset used. The generated vocabulary contains 30,000 tokens, the dimension of word vector representations is 300. Decoding beam-search size equal to 5.

First Sentence: The simplest and most naive approach to generating headlines. The first sentence of the text is used as a summarization. This approach is used with the assumption that the main point is contained in the beginning of the news.

RNN: We will use a classic seq2seq [16] architecture with attention as a baseline. Encoder and decoder are five-layer bidirectional GRU [2] with hidden size equal to 500. A dropout with probability equal to 0.1 after each layer is used for regularization. As attention we use attention via dot-product [7].

Universal Transformer: For comparison, we use the results of Gavrilov et al. [4], based on Universal Transformer. They used 4 layers in encoder and decoder with 8 heads of attention.

Vanilla Transformer: In this case, we will use the classic Transformer architecture with default settings for both encoder and decoder: 6 layers, 8 attention heads, model hidden size is 512, position-wise block hidden size is 2048, dropout equal to 0.1.

Phrase Based Attentional Transformer: PBA transformer settings are very similar to the classic Transformer, except for the attention block. It uses the *heterogeneous* approach: one and two kernel convolution applied to key, value and query. Next, we concatenate this convolved sequence representations and calculate scaled dot-product attention.

5. Results

Table 1: Evaluation results on RIA dataset

Model	ROUGE-1-f	ROUGE-2-f	ROUGE-l-f
First Sentence	24.08	10.57	16.70
RNN	37.98	20.51	35.36
Universal Transformer	39.75	22.15	36.81
Vanilla Transformer	42.42	25.06	39.50
PBA Transformer	**42.96**	**25.43**	**40.02**

[2] The source code for all experiments is available at https://github.com/gooppe/deep-summarization-toolkit

Table 2: PBA Trabsformer generation sample

Text: к 2016 году 20 % школ должны быть доступными для обучения инвалидов, в настоящее время этот показатель составляет чуть больше 2 %, сообщил министр труда и социальной защиты максим топилин на заседании правительства рф. «к 2016 году 20 % школ, не коррекционных, а обычных, должны быть приведены в доступный вид для обучения инвалидов, сегодня этот показатель на начало реализации программы (по доступной среде для инвалидов) составляет 2,5 %», — сказал топилин. по его словам, увеличение в 10 раз — это неплохой показатель, хотя в дальнейшем доступными для обучения инвалидов должны быть все школы. **Original summary**: топилин: к 2016 году 20 % школ должны быть доступными для инвалидов **Generated summary**: к 2016 году 20 % школ должны быть доступны для инвалидов — минтруд
Text: бригада сахалинского бассейнового аварийно-спасательного управления из-за непогоды приостановила работы на аварийном судне мр-150–289 у южного побережья сахалина, сообщили риа новости в главном управлении мчс рф по региону. судно мр-150–289 шло в портпункт озерск корсаковского района. но из-за поломки в системе теплоподачи судно зашло в бывший портпункт новиково. **Original summary**: спасатели из-за непогоды приостановили работы на подтопленном судне **Generated summary**: спасатели приостановили работу на аварийном судне у сахалина

As you can see from Table 1, PBA Transformer shows the best result. The classical recurrent sequence-to-sequence approach is slow. A recurrent network needs much more time to achieve the quality of Transformers. Vanilla Transformer has better results than Universal Transformer presumably due to greater depths. PBA transformer shows interesting results. Qualitatively, it has a small increase, but its ability to generate abbreviations seems to be quite interesting. Table 2 shows example of PBA Transformer generation sample.

Unfortunately, this model is not perfect. Sometimes it makes mistakes, such as usage of incorrect forms of words, confuses with key figures or repeating them. However, the model almost always highlights relevant information, which is inspiring.

6. System and error analysis

In this section, we would like to draw reader's attention to one important problem that appears during testing of trained models on another datasets. The essence of the problem is that a model trained on the dataset of one news Agency cannot be applied to generate news headlines from another source. It seems that data structure should be the same for different news Agencies, but alas, models confuse and generate bad

headlines. During testing on different datasets, we made sure that naive use of the first news sentence as a title shows results better than generations of trained models. This effect can not be considered as overfitting, because on large test subsets from the same dataset, the model achieves good metric estimates. Firstly, this problem may occure due to strong variability in a style of news and headlines. Each Agency uses its own writing style, and model is strongly attached to it. Secondly, there may be shifts in news domains, because of them model focuses on some specific topics and can not cover all areas of text news. Third, it is hard to find supervised summarization dataset, that covers all aspects of human life. There will be always some aspect missing from the training dataset that will be processed with difficulties by proposed model. More formally, it is impossible to use proposed model on different datasets due to Distribution Assumption: there is one probability distribution D that governs both training and testing examples.

The first thing that comes to mind for solving this problem is the use of pre-trained language models on large corpora [3], [11], [12]. These models parametrize language prior distribution. Language model can be fine-tuned on a specific dataset, directly to solve summarization problem. It stands to reason that having some General language representation we can more accurately distinguish information from texts that are not even present in the task-specific dataset. This approach can be used as a baseline for further research.

7. Related work

The task of abstract summarization and the task of generating news headlines in particular deserve much attention of researchers. So, Rush et al. [13] were the first, who proposed to use a deep, fully connected neural network with an attention mechanism as a language model for generating news headlines. Later, approaches based on recurrent neural networks were proposed: Chopra et al. [1] suggested to use the classic recurrent sequence-to-sequence architecture with attention. Other researchers tried to adapt such approaches to the specifics of automatic summarization. Nallapati et al. [9] offered several ideas, potentially improving previously proposed models: large vocabulary trick, hieracical attention and copy-from-text approach. [See et al. 2017] developed the idea of copying some text from the original and proposed to use the Pointer-Generator Network for modelling rare or unseen words. [Gavrilov et al. 2019] applied Universal Transformer with BPE tokenization and offered a new Russian dataset for the research of automatic referencing methods.

8. Conclusion and future work

Under this article, an attempt was made to use a modified Transformer with a phrase based attention mechanism. This modification has improved the quality of the base model and achieved a new state-of-the-art in the task of headline generation on the RIA dataset. Experiments with testing of trained models on another datasets have led us to the problem of model dependence on the news Agency. In the future the results can be improved by using language models or unsupervised approaches.

References

1. *Chopra, S. et al.:* Abstractive sentence summarization with attentive recurrent neural networks. In: Proceedings of the 2016 conference of the north American chapter of the association for computational linguistics: Human language technologies. pp. 93–98 Association for Computational Linguistics, San Diego, California (2016).
2. *Chung, J. et al.:* Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR. abs/1412.3555, (2014).
3. *Devlin, J. et al.:* BERT: pre-training of deep bidirectional transformers for language understanding. CoRR. abs/1810.04805, (2018).
4. *Gavrilov, D. et al.:* Self-attentive model for headline generation. In: Proceedings of the 41st european conference on information retrieval. (2019).
5. *Kingma, D. P., Ba, J.:* Adam: A method for stochastic optimization. CoRR. abs/1412.6980, (2015).
6. *Lin, C.-Y.:* ROUGE: A package for automatic evaluation of summaries. In: ACL 2004. (2004).
7. *Luong, T. et al.:* Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 conference on empirical methods in natural language processing. pp. 1412–1421 Association for Computational Linguistics, Lisbon, Portugal (2015).
8. *Mikolov, T. et al.:* Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th international conference on neural information processing systems - volume 2. pp. 3111–3119 Curran Associates Inc., USA (2013).
9. *Nallapati, R. et al.:* Sequence-to-sequence rnns for text summarization. CoRR. abs/1602.06023, (2016).
10. *Nguyen, P. X., Joty, S.:* Phrase-based attentions. CoRR. abs/1810.03444, (2018).
11. *Peters, M. et al.:* Deep contextualized word representations. In: Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers). pp. 2227–2237 Association for Computational Linguistics, New Orleans, Louisiana (2018).
12. *Radford, A. et al.:* Improving language understanding by generative pre-training. In: Technical report, openai. (2019).
13. *Rush, A. M. et al.:* A neural attention model for abstractive sentence summarization. In: Proceedings of the 2015 conference on empirical methods in natural language processing. pp. 379–389 Association for Computational Linguistics, Lisbon, Portugal (2015).
14. *See, A. et al.:* Get to the point: Summarization with pointer-generator networks. CoRR. abs/1704.04368, (2017).
15. *Sennrich, R. et al.:* Neural machine translation of rare words with subword units. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers). pp. 1715–1725 Association for Computational Linguistics, Berlin, Germany (2016).
16. *Sutskever, I. et al.:* Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. pp. 3104–3112 (2014).
17. *Vaswani, A. et al.:* Attention is all you need. CoRR. abs/1706.03762, (2017).

FILLING THE GAPS WITH RULES AND NETWORKS[1]

Sorokin A. A. (alexey.sorokin@list.ru)

Moscow Institute of Physics and Technology, Neural Networks and Deep Learning Lab, Dolgoprudny, Russia; Moscow State University, Faculty of Mathematics and Mechanics, Moscow, Russia

In this paper we describe rule-based and neural approaches to gapping resolution task for Russian language. Our study was conducted on the material of AGRR-2019 Shared Task. We demonstrate that neural model definitively outperforms the rule-based one even when only 2000 annotated sentences are available. The rule-based model took the 6th place in AGRR-2019 competition (2nd in terms of precision), while the neural one was better than the second-ranked system[2].

Keywords: Gapping, ellipsis, automatic gapping resolution, neural networks

ПРАВИЛОВЫЕ И НЕЙРОННЫЕ МОДЕЛИ ДЛЯ ГЭППИНГА

Сорокин А. А. (alexey.sorokin@list.ru)

Московский Физико-технический Институт, Лаборатория нейронных систем и глубокого обучения, Долгопрудный, Россия; Московский Государственный Университет, механико-математический факультет, Москва, Россия

Данная работа посвящена автоматическому распознаванию гэппинга. Мы показываем, что нейросетевой подход к этой задаче более эффективен в сравнении с правиловым, в том числе на обучающей выборке небольшого размера. Наша нейросетевая модель показала качество выше второго результата в соревновании AGRR-2019[3], в то время как правиловая модель заняла шестое место, показав при этом вторую точность.

Ключевые слова: гэппинг, автоматическое распознавание гэппинга, эллипсис, семантический парсинг

[1] The research was conducted under support of National Technological Initiative Foundation and Sberbank of Russia. Project identifier 0000000007417F630002.

[2] The neural model was submitted two weeks after the end of AGRR-2019 competition when all the gold answers and the systems of other participants were available.

[3] Данный результат был получен после завершения соревнования.

1. Introduction

In linguistics, gapping is a type of ellipsis that occurs in the non-initial conjuncts of coordinate structures [11], [2]. The elided material usually includes the finite verb as well as some of its dependents. For example, in the sentence *Маша любит чай, а Петя кофе.* (*Mary likes tee and Peter coffee*), the elided segment consists of the main verb *любит* (*likes*). Identifying the presence/absence of gaps and their resolution is important in Natural Language Understanding. For example, consider the sentence

(1) *Президентом Ирака стал Бакр, а вице-президентом — Саддам Хуссейн.*
Bakr became the president of Iraq and Saddam Hussein the vice-president.

To extract the semantic structure of this sentence and transform it, e. g., to Wikidata-like triples object-relation-subject, a system must restore the missing verb *стал* (*became*). The extracted triples can be useful, for example, for question answering or information retrieval.

Gapping has attracted high attention in theoretical linguistics [11], [8], [9], however, there are only a few works that investigate gapping in computational literature [14], [5], [6], [7], [15]. Moreover, none of this works address this problem in modern NLP paradigm, where large amounts either of labeled or unlabeled data are utilized. The main obstacle is the lack of datasets: even large existing treebanks contain not more than several hundreds of gapping examples. The only exception is Automatic Gapping Resolution for Russian Shared Task (AGRR-2019) [16] https://github.com/dialogue-evaluation/AGRR-2019, which provides more than 16,000 sentences in total for training and development. Our system was submitted to participate in this competition and the present work describes the model, as well as the results of its application to the dataset.

The paper consists of the following parts: **Section 2** describes the task and the dataset. **Section 3** describes the pipeline and our initial rule-based model. **Section 4** describes the neural model and **Section 5** is devoted to its training. **Section 6** is devoted to data pre- and postprocessing, **Section 7** measures the quality of our models as well as their individual parts. In **Section 8** we conclude with the directions for future work.

2. Task description

AGRR-2019 organizers postulate gapping resolution task as follows: given a raw sentence

(2) *Президентом Ирака стал Бакр, а вице-президентом — Саддам Хуссейн.*
Bakr became the president of Iraq and Saddam Hussein the vice-president.

detect:

1. Whether the sentence contains a gap (binary presence-absence task in organizers' terms).
2. [item:pos] The position of this gap. The annotation standards located it immediately to the left of the first symbol of right core argument *Саддам Хуссейн*. Naively it seems more natural to treat the hyphen as such position, however, the hyphen is optional. Therefore the organizers' solution is more consistent, though less obvious from the first glance.

3. The position of the predicate, corresponding to the gap (gap resolution task).
4. [item:core] The core arguments of the elided predicate: *вице-президентом* (*vice-president*+Ins) and *Саддам Хуссейн* (*Saddam Hussein*+Nom).
5. [item:core-main] The core arguments of the main predicate, corresponding to the orphaned dependents of the elided one: *Президентом Ирака* (*president*+Ins *of Iraq*+Gen) and *Бакр* (*Bakr*+Nom).

Alltogether these subtasks comprise the full resolution track. The participants were allowed to solve the entire task or simply detect the presence/absence of the gap. Note that a sentence may contain several gaps, corresponding to the same main predicate, such as.

(3) *У двоих были черные волосы, у одного — светлые, а у четвертого — каштановые.*
Two of them had black hair, one blond and the fourth — brown.

The dataset also contains several examples with only one orphaned dependent, such as

(4) *Судья посмотрела на свои часы, затем — на меня.*
The judge looked at her watch and then — at me.

The characteristics of the dataset are given in Table 1. Note that the number of gapped sentences and their relative frequency significantly exceeds the corresponding parameters of existing general-purpose corpora, such as Universal Dependencies (e. g. [5], Table 1).

Table 1: Statistics of gap sentences in the dataset

	Train	Development	Test
Total	16,407	4,143	2,046
With gaps	5,542	1,382	680
Multiple gaps	369	90	47
Single orphan gaps	174	27	17

3. Rule-based approach

As introduced in [12], the gapping relations in UD 2.x treebanks is treated via promotion (see also [5]). The highest node in "obliqueness hierarchy"[4] is promoted to be the head of the clause containing the gap, while all other core dependents of the elided predicate are attached to it via the orphan relation. It results in the dependency tree below on Figure 1.

[4] nsubj > obj > iobj > obl > advmod > csubj > xcomp > ccomp > advcl

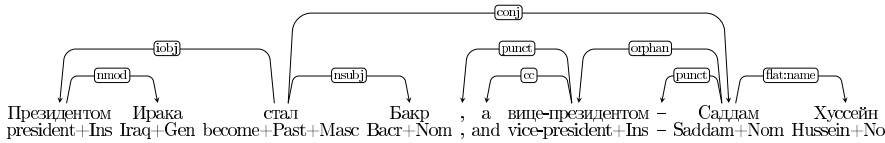


Figure 1: A dependency tree for a sentence, containing gap

Given the golden dependency tree, the rule-based pipeline may work as following:

1. Find the orphan dependency and extract the subtrees attached to its left and right edges as elided predicate dependents. If required, remove the punctuation and conjuction between two clauses.
2. Label the first word of the right subtree as gap position.
3. Start from the head of the found orphan relation and follow the dependency edges upward until a verb node is reached. Label this node as the main verb.
4. Find in the subtree of the main verb the two nodes that better match the core dependents found on the first step using their morphology, syntactic roles and semantics (e. g. embeddings).

Provided the syntactic tree is correct, the first three stages are performed algorithmically. The last task was studied in [15] for English, where it was solved by finding the cheapest alignment between the arguments in the full and gapped clause, where the cost of the alignment was based on similarities between the phrases being aligned as well as on the monotonicity of the alignment. The cost of individual alignment links used GloVe similarities between words being aligned and their part-of-speech tags. This method achieved a relatively high quality of remnant attachment with precision and recall of 87% (see **Table 5** in [15]) on golden parses for English language. However, in the real world scenario with automatically generated parse trees using state-of-the-art dependency parsers, the recall falled to 38% and even the precision to 65%.

We met the same problems as in even to a greater extent. For example, in the development set of the dataset our parser found only 338 orphan relations which is less than one quarter of the number of gapped sentences[5]. Even if the model reconstructs an approximately correct parse tree in terms of its topology (Unlabeled Attachment Score), it may fail to label the orphan edge correctly. It is frequently confused with nsubj, obl or nmod, amod relations, for example, in the sentence on **Figure 2** nsubj was predicted instead of orphan.

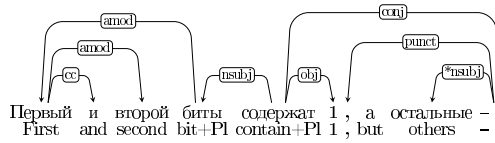


Figure 2: Wrong dependency label in automatical parse tree of gapped sequence

[5] We used the UDPipe [18] dependency parser trained on ru-syntagrus corpus together with DeepPavlov morphological tagger, based on [10] and [17].

Some of these parsing errors can be overcome using other clues except the orphan relation, such as the presence of hyphen or "…, *а* …" construction together with the verb-noun conj relation between the main clause and the promoted gap dependent, which is unlikely to occur in other conditions. Therefore we developed a complicated rule-based system, which finds the gap position using the potentially incorrect dependency tree. If the tree was correct, a system would find the gap predicates by picking the shortest edge that covers the gap position detected on the previous stage. However, parsing errors make the rules even more complicated since the parser is prone to establish local dependencies. For example, consider the sentence

(5) *Мне поставили пять, а брату моей подруги детства — четыре* (*They gave (the mark) five to me and to the brother of my childhood friend+Fem — four*)

and its correct and predicted parses on **Figure 3**.

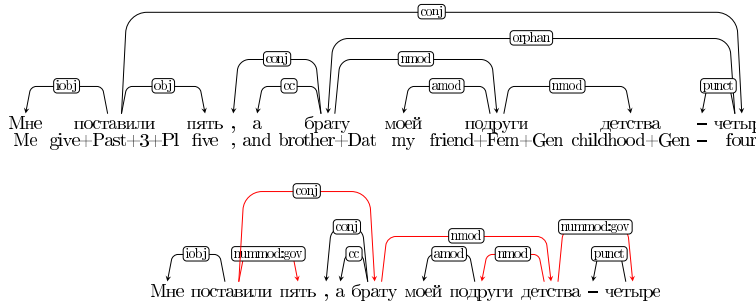


Figure 3: Local attachment error in automatically obtained dependency parse tree (bottom)

Following the obliqueness hierarchy, the gold parse tree the second direct object *четыре* "*four*" is attached to the head of the main clause and the indirect object *брату brother*+Dat — to it via orphan relation. However, the indirect object is much closer to the main verb, therefore the parser selects it as the head of the second clause. Moreover, both numeral objects obtain a wrong dependency label nummod:gov; the structure of the noun phrase *моей подруги детства* "*my*+Gen+Fem *childhood*+Gen *friend*+Gen+Fem" is also incorrect. In this case a system finds the pattern "… , а … " and the conj relation between verb and noun, but fails to restore the second predicate of the gap.

To find the remnants of the gap arguments we utilize the morphological information. Namely, we try to find the descendants of the main verb that have the same part-of-speech and case as the arguments. If only a single word satisfies these constraints, we return this word as remnant. If there are multiple remnants, we rank them using their depth in the tree, their location (to the left or to the right of the main verb) and other features. The hierarchy of features was tuned by hand on the training set.

The final part of the model transforms the constituent heads discovered on the previous stages to the correspondings constituents, which are returned as spans. We use a syntactic parser for this purpose simply considering all dependents of a constituent

head as its subtree. We also write some rules to remove the punctuation and function words on constituent boundaries and to deal with format mismatch between UD and the competition (e.g., competition and UD guidelines differently treat numeral constructions such as *десять лет* or fixed prepositional phrases).

4. Neural model

Rule-based models usually suffer from data variability and vagueness. Consequently, we decided to design a neural network that can automatically detect gapping after training on labeled data. We solve the task by stages, using separate networks for gap location, predicate location and remnant matching, although the architecture of all the networks is the same. The networks on latter stages of the pipeline use as inputs (some of) the outputs produced by the previous stages. We describe in details the network for gap prediction and only note the differences for models that perform predicate location and remnant matching.

4.1. Network for gap location

We solve the following task: given the tokenized sentence $w_1, \ldots, w_m$, the verb position i and word position j, predict whether the verb w_i was elided in position j. The network structure can be described as following:

1. Take as input a sequence of pretrained word embeddings $e_1, \ldots, e_m$.
2. Pass this embeddings through two independent bidirectional LSTMs to obtain sequences of context vectors $g_1, \ldots, g_m$ and $h_1, \ldots, h_m$.
3. Calculate similarity scores between g_i and h_j, a natural way to do it is to measure their dot product $s_{ij} = \langle g_i, h_j \rangle$.
4. Pass the similarities through a sigmoid layer $\sigma_{ij} = \sigma(s_{ij})$.
5. Label as gaps for verb w_i all the positions j, such that $\sigma_{ij} > \frac{1}{2}$.

We did not use the softmax activation because there can be several gaps for a single verb in a sentence. After preliminary experiments we replaced the attention-like similarity calculation by an Infersent-like [3] dense layer using the formula:

$$s_{ij} = \langle w, [g_i, h_j, g_i - h_j, g_i \odot h_j] \rangle + b,$$

where w and b are trainable vectors, $\odot$ denotes element-wise product and $[\,\cdot\,]$ concatenation. The intuition behind is that g-embedding of the main verb g_i should match the h-embedding of the gap position h_j.

4.2. Network modifications

For the task of gap argument location we use as inputs the pair of verb and gap positions (i and k respectively). When calculating the similarity scores s^{arg}_{ikj} for the argument in position j we simply use a dense layer $s^{arg}_{ikj} = \langle w^{arg}, [g_i, g_k, h_j] \rangle + b^{arg}$

Since there can be only one left argument for a given gap, we return the position of the highest score σ_{ikj} provided $\sigma_{ikj} > \frac{1}{2}$. We also select only those k that satisfy the

inequality $i < j < k$. For the right argument we train an analogous model, using the restriction $j > k$ during decoding phase.

For the problem of remnant matching we pass as inputs the verb index i and corresponding gap predicate index k. Since the predicate of the gapped verb usually resemble its remnant we use again the Infersent-like formula:

$$s^{rem}_{ikj} = \langle w^{rem}, [g_i, g_k, h_j, g_k - h_j, g_k \odot h_j] \rangle + b^{rem}$$

In this case also only the word with the highest score is returned.

Summarizing, given the training instance for a single sentence, which includes the main verb V, its predicates (remnants) R_l and R_r and the gap triples of the form $\langle G_i, P_{i,l}, P_{i,r} \rangle$, the prediction pipeline is expressed in Table 2.

Table 2: Inputs and outputs for different pipeline phases

Pipeline phase	Input	Output
Gap location	V	$G_1, \ldots, G_k$
Gap predicate location	v, G_i	$P_{i,1}, P_{i,r}$
Remnant location	$v, P_{i,l}$	R_l
	$v, P_{i,r}$	R_r

When the heads are found, we recover the complete remnant and predicate spans using the same dependency-based procedure as in the rule-based system.

Theoretically, the information about word morphology and syntax can be useful to detect its gapping status. Our network is enough flexible for this task: one may concatenate the embeddings of word morphological tag and dependency type to the pretrained word embedding and train the corresponding embedding matrix together with the network.

5. Model training

5.1. Loss function

Our loss function consists of two components: L_p, penalizing false positives, and L_r, preventing from false negatives. Both components use standard cross-entropy loss and sum over all verbs in all sentences of the training data. Using notation of the previous section, the loss for a single verb in position i is

$$\begin{aligned} L_p &= -\sum_{j \in \overline{I_g}} \log(1 - s_{ij}), \\ L_r &= -\sum_{j \in I_g} \log s_{ij}, \end{aligned}$$

where I_g is the set of all answers (gaps) for a given input (verb). Actually, $L = L_p + L_r$ equals to standard cross-entropy. Since recall is more important than

precision at the early stages of the pipeline, we penalize false negatives higher than false positives by weighting L_r with additional multiple α.

When doing argument prediction and remnant matching we output only the position with the highest matching score s_{ij}. Therefore we must ensure during training, that the score of the correct word in position j_g is maximal among all words. In this case we add auxiliary loss

$$L_{aux} = -\alpha(\log s_{ij_g} - \log \max_j s_{ij}),$$

which is zero only when the correct word has the highest score. It is also weighted by recall weight α, which was selected to be 2 in our experiments.

5.2. Model parameters

We used ELMo embeddings [13] as input for our model, using the implementation from DeepPavlov library [1]. We took the first layer of ElMo network, since it is known to better reflect morphological and syntactic properties that are important for gapping. The size of these embeddings was 1024. Bidirectional LSTMs for sentence processing contained 192 units in each direction.

We collected the inputs in mini-batches, a single batch contains all input-output pairs for 8 sentences (there can be several verbs in a sentence, therefore actual size of the mini-batch is larger). The input data is partitioned in 3/1 proportion to test and development sets. The network was trained for 5 epochs, if F1-score did not improve for 2 epochs, the training was stopped. Training was performed using Adam optimizer with default settings. Our network is implemented using Keras[6].

6. Model application

6.1. Input and output format

Training data consists of raw text sentences together with their annotation. The annotation consists of two parts, the first is the binary label (0/1) indicating whether a gap occurs in the current sentence. For the sentences containing the gap the second part generally contains 6 pairs of numbers, as shown on **Figure 4**. The first pair of numbers refers to the main predicate (typically, a single verb), and the second and the third to its core arguments; the fourth one contains the empty span of the gap, the two remaining ones label the positions of gap predicates. All offsets are given in characters.

Некоторые торговцы отлично продают один вид товаров или услуг, а другие - другой.
1 27:34 0:18 35:39 74:74 65:71 74:80

Figure 4: A typical example of input data sentence

[6] https://github.com/AlexeySorokin/Gapping

For some sentences, the number of input spans differ: if the gapped verb has only one argument remaining, such as *поперек* the one on Figure 5.1, then there are only only 4 spans present. On the contrary, if more the sentence contains $r > 1$ gaps, as in Figure 5.2, (*домишко* (*казался*) *дворцом*, *дома* (*казались*) *небоскребами*) and the gapped verb has k arguments, then the total number of input spans is $r(k + 1)$ (typically k equals 2).

Сосиску разрезаем вдоль на две половинки, затем поперек.
1 8:17 18:23 48:48 48:51

Я смотрел, и каждая улочка казалась мне изящной аллеей, каждый домишко — дворцом, а двухэтажные дома — небоскребами.
1 27:35 13:26 40:54 73:73 103:103 56:70 84:100 73:80 103:115

Figure 5: Other examples of input data sentences

To be evaluated, a system should provide the output of the same format as the input given. If the system does not predict the span, the corresponding column is left empty.

6.2. Data pre- and postprocessing

Since both our systems operate on the level of subtree heads, not the character-based spans, we need to convert the input data to appropriate format before applying the model and transform it back when submitting the output. Input conversion is required on two stages: to prepare training data and to evaluate system output on validation data.

First, character-level spans are converted to word-level spans using the NLTK tokenizer (we also tried the UDPipe one but found it to perform worse). Second, for each span the subtree head is determined using the automatically obtained parse tree. We select as possible heads all words in the span whose parent lies outside this span. If the parse tree is correct, there is only one such head. If multiple heads are returned due to parsing errors, we exclude the instance during training or return None during validation.

During model evaluation we need to convert its predictions back to the competition format. Subtree heads are transformed to word-level spans using the procedure described in Section 3 and word-level spans are converted to character-level to produce the final answer. We note that this procedure is prone to errors due to incorrect parsing and (potentially) tokenization.

7. Results and discussion

We present the evaluation of the entire system using the official evalution script [7] as well as our own metrics. We pay more attention to our own metrics since they allow to evaluate separate stages of the pipeline.

[7] https://github.com/dialogue-evaluation/AGRR-2019/blob/master/agrr_metrics.py

7.1. Individual models evaluation

Our gapping model essentially consists of model for 3 individual subtasks: gap location, gapped predicates location and remnants matching. Additionally, to predict argument spans, the stage of span prediction is required. We present scores for each of the stages separately, passing the gold input to them (see Table 2 for the list of inputs and outputs for each phase of the pipeline). Since our models return subtree heads, their predictions are judged against the subtree heads extracted from correct spans, when our input preprocessor fails to extract such span, any output is considered as invalid. We evaluate the model on the level of individual inputs, not sentences. We collect all input-output tuples and calculate the number of true positives (present both in gold and predicted answers), false positives and false negatives. For all the tasks except gap location we also count the number of partially correct answers, which is the number of input-output pairs for which more than one half of tuple elements was predicted correctly. These partial matches are added to the number of true positives with weight 0.5 when calculating precision, recall and F1-measure. The metrics are reported using the official test set and the gold answers on it. We compare 3 models: the rule-based one, the full neural model and the ensemble of 3 neural models.

Table 3: Quality of individual pipeline stages on test set of AGRR-2019 competition

Stage	Model	TP	FP	FN	partial	Precision	Recall	F1
Gap location	Rule-based	556	219	180	0	71.74	75.54	73.59
	Neural (single)	663	50	73	0	92.99	90.08	91.51
	Neural (ensemble)	1397	86	103	0	**94.20**	**93.13**	**93.66**
Predicate location	Rule-based	615	2	54	67	94.81	88.11	91.34
	Neural (single)	649	0	13	74	94.88	93.21	94.04
	Neural (ensemble)	1378	0	10	112	**96.24**	**95.60**	**95.92**
Remnant matching	Rule-based	475	0	208	53	94.98	68.14	79.35
	Neural (single)	557	3	106	73	93.76	80.64	86.71
	Neural (ensemble)	576	4	107	53	**95.18**	**81.86**	**88.02**
Span prediction	Rule-based	520	0	116	100	91.94	77.45	84.07

Table 3 demonstrates that rule-based model definitely looses to the neural one. Moreover, the only rule-based component of the neural pipeline, span prediction, occurs to be the weakest part of it. To measure relative impact of different pipeline components, we score the output after each stage of the pipeline.

Table 4 supports the conclusion made from individual stages evaluation: neural model is significantly stronger than the rule-based one. Though their overall performance is comparable in terms of precision, the recall of the neural model is definitely higher. Note that after remnant matching stage the precision of rule-based model goes up because this phase eliminates some arguments of the gapped verb that were incorrectly predicted on the previous stages and do not match any arguments of the main verb in the sentence. It is also the only stage, where individual rule-based model

is comparable with the neural one at least in precision terms, however it is achieved at the expense of significant decrease of recall.

Table 4: Quality after each stage of the pipeline on test set of AGRR-2019 competition

Stage	Model	TP	FP	FN	partial	Precision	Recall	F1
Gap location	Rule-based	556	219	180	0	71.74	75.54	73.59
	Neural (single)	663	50	73	0	92.99	90.08	91.51
	Neural (ensemble)	672	45	64	0	**93.72**	**91.30**	**92.50**
Predicate location	Rule-based	537	170	131	68	73.68	77.58	75.58
	Neural (single)	598	31	69	69	90.62	85.94	88.21
	Neural (ensemble)	617	27	62	57	**92.08**	**87.70**	**89.84**
Remnant matching	Rule-based	398	34	266	72	86.11	58.97	70.00
	Neural (single)	527	22	87	122	87.63	79.89	83.58
	Neural (ensemble)	561	19	75	100	**89.85**	**83.02**	**86.30**
Span prediction	Rule-based	322	45	277	137	77.48	53.06	62.98
	Neural (single)	403	38	103	230	77.20	70.38	73.63
	Neural (ensemble)	436	38	94	206	**79.77**	**74.67**	**76.13**

7.2. AGRR-2019 evaluation metrics

In Table 5 we score different stages of the pipeline using the official evaluation metrics of the competition. The organizers evaluated the performance both in term of binary detection of gapping in a sentence, as well as complete gapping resolution, which uses character-wise precision and recall. Since full resolution can be performed only when all the answers are available, intermediate stages are evaluated using only binary metrics.

Table 5: Official evaluation metrics of AGRR-2019 competition after each stage of the pipeline on development (left) and test (right) set

Stage	Model	Binary						Full resolution	
		Precision		Recall		F1			
Gap location	Rule-based	80.6	80.6	82.9	82.5	81.7	81.5	—	—
	Neural (single)	96.1	96.2	93.7	92.5	94.9	94.3	—	—
	Neural (ensemble)	*97.2*	*97.0*	**94.6**	**92.9**	**95.9**	**95.0**	—	—
Predicate location	Rule-based	80.6	80.6	82.9	82.5	81.7	81.5	—	—
	Neural (single)	96.3	96.3	93.3	91.8	94.7	94.0	—	—
	Neural (ensemble)	*97.2*	*97.0*	*94.4*	*92.2*	*95.8*	*94.6*	—	—
Remnants matching	Rule-based	93.1	93.4	63.7	64.6	75.6	76.3	—	—
	Neural (single)	97.4	97.3	91.2	89.1	94.2	93.0	—	—
	Neural (ensemble)	**98.0**	**97.9**	*92.0*	*90.1*	*94.9*	*93.9*	—	—
Span prediction	Rule-based	93.1	93.4	63.7	64.6	75.6	76.3	59.3	60.2
	Neural (single)	97.4	97.3	91.2	89.1	94.2	93.0	87.5	85.3
	Neural (ensemble)	**98.0**	**97.9**	*91.4*	*90.1*	*94.9*	*93.9*	**89.1**	**87.1**

Almost always scores decrease between subsequent stages of the pipeline. The only exception is the rule-based model, where gaps and predicates are located using the same model, therefore there accuracies coincide. On the stage of remnants matching the precision of rule-based model grows up since most incorrectly predicted predicates are not matched with any remnant and are therefore rejected. However, for a significant fraction of correct predicates matches are also not found which deteriorates the overall performance.

The first thing to note is solid performance of our model in terms of precision even without ensembling. Its recall is also rather high: the best scores achieved by AGRR-2019 participants [8] during evaluation were 95.9% for binary gap detection and 89.2% for full resolution, so we are about two percents behind[9].

Our comparison demonstrates the clear superiority of neural models in all phases of the pipeline. However, the training set provided by the organizers of AGRR-2019 was rather large and requires huge amount of manual effort in its collection. It is natural to ask whether neural models can be used in low-resource setting. We selected first 1,600 sentences of the training data for training and 400 sentences of development data for tuning and trained an ensemble of 3 models on this smaller dataset. The results in comparison with our large model is given are given in Table 6.

Table 6: Official evaluation metrics of AGRR-2019 for neural model trained on smaller dataset on development (left) and test (right) set

Stage	Model	Binary						Full resolution	
		Precision		Recall		F1			
Gap location	Rule-based	80.6	80.6	82.9	82.5	81.7	81.5	—	—
	Small neural (ensemble)	96.0	*98.2*	86.2	81.2	90.8	88.9	—	—
	Neural (ensemble)	*97.2*	97.0	**94.6**	**92.9**	**95.9**	**95.0**	—	—
Predicate location	Rule-based	80.6	80.6	82.9	82.5	81.7	81.5	—	—
	Small neural (ensemble)	96.0	**98.2**	85.9	81.2	90.7	88.9	—	—
	Neural (ensemble)	*97.2*	97.0	*94.4*	*92.2*	*95.8*	*94.6*	—	—
Remnants matching	Rule-based	93.1	93.4	63.7	64.6	75.6	76.3	—	—
	Small neural (ensemble)	97.4	**98.6**	76.7	71.5	85.8	82.9	—	—
	Neural (ensemble)	**98.0**	*97.9*	*92.0*	*90.1*	*94.9*	*93.9*	—	—
Span prediction	Rule-based	93.1	93.4	63.7	64.6	75.6	76.3	59.3	60.2
	Small neural (ensemble)	97.4	**98.6**	76.7	71.5	85.8	82.9	73.9	69.4
	Neural (ensemble)	**98.0**	*97.9*	*91.4*	*90.1*	*94.9*	*93.9*	**89.1**	**87.1**

We observe that models trained on smaller sample of data do not loose in precision, however, recall significantly decreases at all stages of the pipeline. Nevertheless, they still outperform the rule-based model. We observed severe overfitting on small datasets, which means that several parameters of the model (e. g., interlayer dropout

[8] https://github.com/dialogue-evaluation/AGRR-2019

[9] We note again that only our rule-based system was submitted during the competition, therefore other participants could potentially improve their scores as well.

or number of recurrent units) must be altered. The recall weight should be also adjusted even further than it is in the basic model. Nevertheless, even datasets of medium size allow to train neural gap detectors that outperform rule-based recognizers.

8. Conclusions

We have designed a high-quality neural model for gap resolution for Russian language, whose quality achieves 93% for binary gap detection and 89% for full gap resolution. The model is based on ELMo embeddings and recurrent neural networks. The are at least three directions for future research: the first is to apply the model to other languages, such as English, Spanish or Czech. Since model architecture is language independent, the main obstacle can be relative lack of data. Though our model shows solid performance for only 2,000 training sentences available, they are much lower than the scores of the model trained on larger datasets. Probably, curated generation of training data (especially negative examples) can make the model more robust. Another problem is the usage of ELMo embeddings which are not available for many languages and whose learning is time- and resource-consuming. The simplest solution is to replace ELMo with newer multilingual BERT [4] which demonstrated high performance in other works of AGRR-2019 competition.

Another problem to investigate is the role of morphological and syntactic information. We found that it does not have stable effect on the full dataset, however, with smaller data its significance can be higher. The main improvement can be achieved in the weakest part of the model: the detection of constituent bounds using a syntactic parser. The simplest way to enhance performance is to retrain the parser with more gapped sentences and thus increase its ability to parse such sentences. Alternatively, one may directly solve the task of constituent bound detection without reducing to syntactic parsing. We leave these questions for future research.

Acknowledgements

The author thanks the organizers of AGRR-2019 competition for giving the opportunity to participate and for their helpful cooperation during the Shared Task. I am also very grateful to the stuff of MIPT Neural Networks laboratory for warm and inspiring atmosphere during the work on the problem. I deeply thank the anonymous reviewers whose comments and suggestions helped to improve the paper.

References

1. *Burtsev M. et al.:* DeepPavlov: Open-Source Library for Dialogue Systems (2018), Proceedings of ACL 2018, System Demonstrations, Melbourne, Australia, pp. 122–127.
2. *Carnie A.:* Syntax: A generative introduction (2006), Blackwell.

3. *Conneau A. et al.:* Supervised learning of universal sentence representations from natural language inference data (2017), arXiv preprint arXiv:1705.02364, available at https://arxiv.org/pdf/1705.02364.pdf.
4. *Devlin J. et al.:* Bert: Pre-training of deep bidirectional transformers for language understanding (2018), arXiv preprint arXiv:1810.04805, available at https://arxiv.org/pdf/1810.04805.pdf.
5. *Droganova K., Zeman D.:* Elliptic Constructions: Spotting Patterns in UD Treebanks (2017), Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017), Gothenburg, Sweden, pp. 48–57.
6. *Droganova K. et al.:* Parse Me if You Can: Artificial Treebanks for Parsing Experiments on Elliptical Constructions (2018), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018), Miyazaki, Japan, pp. 1845–1852.
7. *Droganova K. et al.:* Mind the Gap: Data Enrichment in Dependency Parsing of Elliptical Constructions (2018), Proceedings of the Second Workshop on Universal Dependencies (UDW 2018), Brussels, Belgium, pp. 47–54.
8. *Jackendoff R. S.:* Gapping and related rules (1971), Linguistic inquiry, Vol. 2., 1., pp. 21–35.
9. *Johnson K.:* Gapping (2014), manuscript, available at http://people.umass.edu/kbj/homepage/Content/gapping.pdf.
10. *Heigold G., Neumann G., van Genabith J.:* An extensive empirical evaluation of character-based morphological tagging for 14 languages (2015), Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers., Vol. 1., pp. 505–513.
11. *Lakoff G., Ross J. R.:* Gapping and the order of constituents (1970), Progress in linguistics: A collection of papers., Vol. 43., p. 249.
12. *Nivre J. et al.:* (2016), Universal Dependencies v1: A Multilingual Treebank Collection, Language Resources and Evaluation (LREC), Portoroz, pp. 1659–1666.
13. *Peters M. E. et al.:* Deep contextualized word representations (2018), arXiv preprint arXiv:1802.05365.
14. *Schuster S., Lamm M., Manning C. D.:* Gapping Constructions in Universal Dependencies v2 (2017), Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017), Gothenburg, Sweden, pp. 123–132.
15. *Schuster S., Nivre J., Manning C. D.:* Sentences with Gapping: Parsing and Reconstructing Elided Predicates (2018), arXiv preprint arXiv:1804.06922, available at https://arxiv.org/pdf/1804.06922.pdf.
16. *Smurov I. et al.:* AGRR 2019: Automatic Gapping Resolution for Russian (2019), International conference on computational linguistics "Dialogue", to appear.
17. *Sorokin A.:* Improving neural morphological Tagging using Language Models (2018), International conference on computational linguistics "Dialogue"., Vol 1., pp. 707–720.
18. *Straka M.:* UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task (2018), Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies., Brussels, Belgium, pp. 197–207.

MORPHOLOGICAL PARSING OF LOW-RESOURCE LANGUAGES[1]

Sorokin A. A. (alexey.sorokin@list.ru)

Moscow Institute of Physics and Technology, Neural Networks and Deep Learning Lab, Dolgoprudny, Russia;
Moscow State University, Faculty of Mathematics and Mechanics, Moscow, Russia

It this paper we study morphological parsing and lemmatization on the material of Evenk and Selkup language. We compare basic neural models with their extensions that attempt to utilize additional linguistic information from the training data. We show that the augmented model does not improve over the baseline even decreasing performance for the task of lemmatization. We hypothesize that to be helpful additional information should be extracted from external resources, if available, not the corpus itself.

Keywords: morphological parsing, lemmatization, low-resource languages, morphological guesser

МОРФОЛОГИЧЕСКИЙ АНАЛИЗ МАЛОРЕСУРСНЫХ ЯЗЫКОВ

Сорокин А. А. (alexey.sorokin@list.ru)

Московский Физико-технический Институт, Лаборатория нейронных систем и глубокого обучения, Долгопрудный, Россия; Московский Государственный Университет, механико-математический факультет, Москва, Россия

Данная работа посвящена морфологическому анализу и лемматизации для эвенкийского и селькупского языка на материале соревнования LowResourceEval-2019 для малоресурсных языков. Мы сравниваем базовую нейронную модель с её расширениями, использующими лингвистическую информацию (морфологический словарь, извлечённый из корпуса), и показываем, что они не ведут к улучшению качества. Наша гипотеза состоит в том, что дополнительная информация

[1] The research was conducted under support of National Technological Initiative Foundation and Sberbank of Russia. Project identifier 0000000007417F630002.

должна извлекаться не из обучающего корпуса, а из внешних источников, в противном случае ту же самую информацию более эффективно извлекает сама нейронная сеть.

Keywords: морфологический анализ, малоресурсные языки, лемматизация, морфологические словари

1. Introduction

In recent years neural networks have dramatically improved the quality of natural language processing, especially in semantically-oriented tasks. One of the key advantages of neural models is their ability to extract knowledge from large amounts of unlabeled data, for example in the form of word embeddings or language models, or efficiently utilize patterns in raw data that are too complex or vague to be captured with handcrafted features. However, the applicability of neural approaches in low-resource setting is not that obvious. The main obstacle is the inclination of neural networks to overfit, especially on small datasets.

In the field of morphological tagging neural network models clearly outperform earlier approaches based on conditional random fields or local classifiers [2]. As discussed in [9], the key reason is the importance of word-level information which is readily captured by character-level embeddings, in contrast to tag-level interactions which were in the focus of hidden Markov models or conditional random fields. Several editions of CONLL shared tasks [13], [12] have demonstrated that neural networks are equally more efficient in high-resource or low-resource setting. Therefore we consider neural networks as a default choice for morphological analyzer without any need for further discussion.

What have to be discussed is the choice of information passed to the network. Usually neural models are trained on raw tokenized texts. On the contrary, earlier approaches to morphological tagging heavily relied on external morphological dictionaries and other resources. For example, the whole task of morphological tagging was treated as disambiguation, which is, the selection of the correct label from the ones presented in the dictionary. This approach is rarely applied in neural paradigm since it contradicts the main idea of neural NLP: make the model to learn arbitrarily complex patterns from data and do not impose restrictions on their form by external constraints. However, the studies for Russian language [1], [9] demonstrated, that passing the output of external morphological analyzer as additional input of the network improves tagging accuracy even in high-resource setting. Therefore we expected the benefits to be even higher in case when little data is given.

We tested our hypothesis on two datasets of LowResourceEval-2019 competition[2] [5] for Evenk and Selkup languages, solving the tasks of morphological tagging and lemmatization. Since the datasets were equipped with gold morpheme segmentation and most of the morphemes were found to correspond with morphological features, our strategy was restrict the set of potential tags given possible segmentations and pass these tags as additional inputs to the model (the approach successfully applied

[2] https://lowresource-lang-eval.github.io

in the studies mentioned above). However, our hypothesis failed, since none of the complex models was able to outperform the basic ones[3].

The structure of our paper is the following: in **Section 2** we present the architecture of our basic model, **Section 3** describes the feature extraction process, **Section 4** is devoted to data and experiments description, in **Section 5** we present the results and discuss them. We conclude in **Section 6** with the directions for future work.

2. Model architecture

2.1. Morphological tagging

Our basic model is the implementation of Heigold's character-based network [2]. For the sake of completeness we briefly describe the architecture below. Note that similar approach was also pursued in the work [4] on neural language modelling. The model consists of two subnetworks: the first transforms the words to their vector representations, the second uses the obtained embeddings to predict morphological labels.

1. Each character is encoded as a 1-hot row vector with n_c dimensions, n_c being the number of characters. Thus the word is represented by a sequence of L such vectors $x_{i_1}, \ldots, x_{i_L}$, which is a matrix X with L rows and n_c columns with exactly one unit in each row.
2. This matrix is multiplied by a matrix U of size $n_c \times n_e$, producing a sequence $X' = XU$ of L embeddings $x'_{i_1}, \ldots, x'_{i_L}$. X_{i_j} is the i_j-th column of the embedding matrix U, which is a dense representation of i_j-th character in the alphabet.
3. X' is passed through parallel convolutional layers with different window size $w_1, \ldots, w_K$ and filters number $f_1, \ldots, f_K$. After this step K vectors of dimensions $f_1, \ldots, f_K$ are accociated with each position of the word. Roughly speaking, k-th of these vectors contains information of useful ngrams of length w_k around current position.
4. All the vectors from the previous step are concatenated, producing a vector of length $F = \sum_j f_j$ for each symbol of the word. A word is now a matrix with L rows and F columns.
5. A maximum over each row is taken via max-pooling layer, finally encoding a word as a vector h' of fixed dimension F.
6. Several highway layers [10] are applied to this vector. Highway layer performs the transformation $h = s \odot g(Vh') + (1 - s) \odot h'$, where V is a square matrix with V rows, g a non-linear function and $\odot$ denotes coordinate-wise product. The idea is both to produce useful combinations of features using one-layer perceprtron output $g(Vh')$ and keep relevant dimensions of h' at the same time. The contributions of both components are balanced using s vector, which is obtained by another one-layer perceptron with sigmoid activation: $s = \sigma(Sh')$.

[3] Nonetheless, our models took the first place on all tasks where we participated

The second component of the network transforms the obtained sequence of word vectors $h_1, \ldots, h_n$ into n propability distributions $\pi_1, \ldots, \pi_n$, π_j being probabilities of tags for j-th word in the sentence. First, two LSTMs are applied[4], the first processing the sentence from left to right and the second from right to left. The first produces vectors $\vec{y}_1, \ldots, \vec{y}_n$ and the second produces $\overleftarrow{y}_n, \ldots, \overleftarrow{y}_1$, thus each word is encoded by two vectors $\vec{y}_i, \overleftarrow{y}_i \in \mathbb{R}^{n_y}$. These vectors are multiplied by a projection matrix W with n_t rows and n_y columns, n_t being the number of tags. Applying softmax layer produces the required probability distribution:

$$\begin{aligned} y_i &= [\vec{y}_i, \overleftarrow{y}_i] \text{ (concatenation)}, \\ z_i &= W y_i \\ \pi_{ij} &= \frac{e^{z_{ij}}}{\sum\limits_k e^{z_{ik}}} \end{aligned}$$

In [2] this architecture is proved to be successful for languages of different morphological structure even with only several thousands of tagged sentences available for training. It is also flexible enough to encode additional linguistic information i. e. from a morphological dictionary. This information is encoded in a vector form, for example, using a 0/1 vector of size n_t where nonzero elements correspond to the positions of possible dictionary tags. Such a vector $z_i^{\mathbf{feat}}$ is concatenated to the output z_i of bidirectional LSTM[5].

2.2. Lemmatization

In general, the task of lemmatization (the recovery of word normal form given the inflected one) is a string-to-string transduction problem. As demonstrated in several works [7] on string inflection, such a problem should be solved by sequence-to-sequence (seq2seq) neural networks. However, in most cases the transduction changes the material only on word edges (certainly, it is wrong for Semitic languages or for stem vowel alternations in German or Spanish), therefore it can be described using finite amount of information. This reduces the transduction problem "reconstruct the basic form letter by letter" to a classification problem "guess the transformation pattern" which can be solved even without neural networks. Such an approach is used, for example, in [11]. However, in the languages of the current study the problem becomes even simpler: in Evenk the inflected form is always formed by attaching several (possibly zero) letters to the end of the word. Consequently, the lemma is always an initial segment of the word under consideration. Therefore to perform lemmatization one suffices to predict the end position of the stem. We model it by predicting a probability distribution $\mathbf{p}_w = [p_1, \ldots, p_{|w|}]$ over word positions where p_i is the probability that word stem ends after its i-th letter. Then the end of the initial word form is the maximum of this distribution.

[4] We omit the equations, interested reader may consult [4].

[5] we tested other ways to append this information, but this showed the higher performance.

For Selkup the pattern is slightly more complex: inflection also includes infixation, which is the insertion of morphemes inside the stem, though the percentage of such cases is rather low. For example, the inflected form of *amrsat* "*bowl*" is *amĩrsat*'. Hence, the stem is no longer a prefix of the word, but its (possibly discontinuous) subsequence. We model this by predicting two vectors $\mathbf{p}_w$ and $\mathbf{q}_w$: the first has the same meaning as for Evenk, while $q_i \in [0; 1]$ is the probability of i-th letter to be the result of epenthese, which implies that it is not present in the initial form.

Summarizing, the network architecture is the following:

1. Each symbol is encoded as a 0/1-vector of size n_t, which is transformed to a dense vector by multiplying an embedding matrix.
2. As in the tagging model, we pass the embeddings throw several convolutional layers. Each layer contains filters of different width, whose outputs are collected together in a single vector. To facilitate learning we insert batch normalization [3] between consecutive layers[6].
3. Each positional output h_i of the convolutional layer is multiplied by a trainable vector w to obtain a number $s_i = \langle w, h_i \rangle$.
4. The vector $\mathbf{s}$ of obtained scores is passed through a softmax layer to get the final probability distribution $\mathbf{p} = softmax(\mathbf{s})$:

$$p_i = \frac{e^{s_i}}{\sum\limits_{j=1}^{|w|} e^{s_j}}$$

When we additionally predict the vector q of deletion probabilities, then the probability that i-th symbol is omitted is $q_i = \sigma(\langle w_{del}, h_i \rangle)$, where σ is the sigmoid activation function. To predict the lemma we find the maximum value of p_i: $I = \text{argmax}_i p_i$ and return w[:I] as lemma. When modelling the infixation, we additionally delete all symbols in positions j such that $q_j \geq \frac{1}{2}$.

3. Additional features

Theoretically, context-dependent morphological taggers should benefit from information, available from morphological dictionaries, lexicons and/or context-free analyzers. When a dictionary is not available, the most common tool to apply is a suffix guesser, which determines possible morphological features using word suffixes. However, this method is not easily adapted to agglutinative languages since to extract a particular morpheme one may need to observe up to 8 final symbols. We selected another strategy: in addition to the morphological tags, the training data contained the morpheme segmentation of the form:

ne:jamtli ne: ja_DIM m_ACC tɬi_3SG

[6] This solution is crucial for deep convolutional network, without batch normalization the network often fails to learn at all due to gradient decay.

Some of the morphemes can be converted to morphological features, for example, **3SG** is **Number=Sing|Person=3**. This mapping can be reconstructed automatically, by calculating probabilities of morphological features which cooccur with a given morph in a training corpus. This leads to the following method of feature extraction: given a word, we extract all possible morpheme combination on its right edge, checking not only the correctness of morph segmentation, but also the validity of corresponding sequence of morpheme types. Then for each morpheme type we extract the corresponding values of morphological features. We select as possible all morphological tags whose feature values do not contradict the selected features and cooccur with them in at least 3 training examples. To prevent overfitting we randomly replace this vector by all zeros with a fixed probability to allow the model to generalize to out-of-vocabulary inputs.

In the case of lemmatization we also experimented with either adding the part of speech label as additional input during lemmatization or multitask learning approach: we trained the network to predict word part-of-speech and detect stem boundary simultaneously, sharing all the embedding and convolutional layers between them.

4. Data and experiments

In case of neural tagging we run two models: the basic one and the one augmented with possible tag information. We also test three models for lemmatization: the basic one, the one augmented with morphological tags as input and the one with guessed possible morpheme boundaries.

4.1. Model parameters

Following [2] and [9], we choose the following parameters of morphological tagger: character embeddings are of size 32, convolutional window size changes from 1 to 7, the number of filters for width w is min (200, 50w). The number of convolutional layers is 2 with 0.2 dropout between layers and highway layer following the final convolution. This yields word embeddings of final size 1100, which are passed through a bidirectional LSTM with 128 units in each direction. We also tested several other parameter combinations but found these to give higher accuracy.

The lemmatizer had two convolutional layers of with 5 and 192 filters with no dropout (we found it useless in contrast to several previous studies). Other parameters are completely determined by the algorithm.

In preliminary studies we divided the dataset to train and development subsets and found 20 epochs of training to be optimal for lemmatization and 25 epochs for morphological tagging. Therefore our final models were trained for this number of epochs without early stopping. We used Adam optimizer and batches of size 16. All networks are implemented using Keras framework, our implementation is open-source[7].

When we used the guesser, we additionally memorized all the word-lemma pairs that appear 5 or more times in the training data. Since the precision of the guesser

[7] https://github.com/AlexeySorokin/NeuralMorphoTagger1/tree/low-resource

is much lower than its recall (see Section 5 below), we restricted its output to 5 most frequent tags.

4.2. Data

We test our models on Evenk and Selkup dataset of LowResourceLangEval contest[8]. The parameters of the dataset are given in Table 1. All the datasets were converted to CONLL-U format [9] by the Shared Task organizers, they provided the tokenization as well. We consider as tags the concatenation of part-of-speech label and morphological features, for example, ADP and NOUN, are both examples of possible tags. We also tried to predict the value of each feature, e. g. case and gender, separately, but this significantly deteriorated performance.

Table 1: Dataset parameters

Language	Dataset	words	sentences	unique tags	OOV words	hapaxes
2*Evenk	train == xbby xtby	25,869	5,527	873	0	8,063
	test == xbby xtby	2,697	548	319	814	272
2*Selkup	train == xbby xtby	13,436	2,394	316	0	5,088
	test == xbby xtby	2,426	425	151	912	246

Lemmatization algorithm is trained and tested on the same datasets. To reduce overfitting we downsample frequent words: if a word occurs $n > 5$ times in the dataset we include it to the training sample only $5 + \lceil \log_2(n - 5) \rceil$ times.

We would like to note that in comparison to low-resource languages in CONLL 2018 Shared Task [12] Evenk and Selkup corpora are substantially larger, since most of low-resource corpora there do not exceed 1000 words. That implies that several questions relevant for actual low-resource parsing do not arise in our current study.

5. Results and discussion

In case of morphological tagging we compare two approaches, the basic one (BASIC) and the one using dictionary (DICTIONARY) information. Since we have no separate morphological dictionaries, all the word-tag pairs are extracted from the training data. We report accuracy both for morphological tags (which is, the percentage of words whose full morphological descriptions are predicted correctly) and sentences (the fraction of sentences where all words obtain correct morphological tags). For each metric we report two numbers: the average across 3 randomly initialized models (left) and the ensemble of these models (right).

[8] https://lowresource-lang-eval.github.io

[9] https://universaldependencies.org/format.html

Table 2: Results of morphological tagging

Model	Evenk		Selkup	
	Tag acc.	Sent acc.	Tag acc.	Sent acc.
BASIC	81.30 83.98	45.25 50.18	80.75 82.81	40.32 43.06
DICTIONARY	81.48 83.13	45.80 48.54	80.65 82.32	39.14 42.12

For lemmatization we compare 3 models: the basic one (BASIC), the tag-augmented one (TAGS) which takes gold morphological labels as additional inputs and the joint one (MULTITASK) which tries to predict these tags as an auxiliary task. As in case of tagging, we show the average accuracy across 3 runs and the accuracy of ensemble of 3 models.

Table 3: Results of lemmatization

Model	Evenk		Selkup	
	Single	Ensemble	Single	Ensemble
BASIC	91.32	93.33	88.35	89.94
TAGS	91.73	92.66	87.68	89.40
MULTITASK	90.88	91.88	86.26	87.79

5.1. Discussion

As shown in Table 2 and Table 3, the baseline network method either is on the par with linguistically informed extensions or even outperforms them. As demonstrated earlier [1], [9], in case of several other languages morphological dictionaries and guessers do improve performance (the boost is especially valuable for Russian and Pymorphy [6] analyzer). Actually, our results do not show that external morphological knowledge is useless, it only shows that dictionary cannot be extracted from the same training corpus. It can be viewed as the kind of overfitting: the dictionary information is available in training time, but the model may lack it in test phase due to OOV words. Usually dropping a fraction of dictionary inputs in training time fixes this issue at least partially, but the present study it was not the case.

To understand the phenomenon better we tested the guesser itself. As mentioned above, we restricted the output of the guesser to 5 most probable tags. This yields to the coverage of 72% for Selkup and 66% for Evenk even on the training set itself, since other hypotheses are too rare to occur between top 5 that do not contradict with morpheme segmentation. However, the coverage on the development is not much lower: 70% and 60%, which means that our morpheme guesser is able to generalize to unseen words. Hence, it is not poor coverage that causes decrease in performance. Further, omitting the top 5 variants restriction produces 20–30 variants for a word in average, which makes the guesser helpless (the choice between 30 tags is not easier than between the original 200). And the learning curve shows that dictionary-augmented model trains significantly faster on the first few iterations (which

confirms that dictionary information is used), however, achieves lower performance in the end.

Probably, the problem lies in the datasets themselves. We compare the characteristics of the datasets with other language presented in UD 2.3 corpora [8]. Evenk belongs to Tungus family, which has no other UD corpora available, while Selkup is Uralic, though it belongs to Samoyed outgroup. **Table 4** contains the characteristics of Selkup in comparison with hu_szeged corpora of Hungarian, which also belongs to Uralic family of languages, and SST corpus of Slovenian, which is Indo-European (Slavic), but whose corpora is of the same order of size[10].

Table 4: Dataset comparison

Language	Dataset	words	sentences	unique tags	OOV words	hapaxes
2*Evenk	train == xbby xtby	25,869	5,527	873	0	8,063
	test == xbby xtby	2,697	548	319	814	272
2*Selkup	train == xbby xtby	13,436	2,394	316	0	5,088
	test == xbby xtby	2,426	425	151	912	246
2*Hungarian	train == xbby xtby	20,166	910	581	0	5,883
	test == xbby xtby	10,448	449	446	3233	513
2*Slovenian	train == xbby xtby	19,473	2,078	645	0	3,021
	test == xbby xtby	10,015	1,110	506	1,631	316

We observe that the main feature of Selkup corpus is smaller length of sentences. The ratio of unique tags and corpus length is almost the same as for Hungarian and lower than for Slovenian, but larger than for typical corpora of Universal Dependencies. The percentage of out-of-vocabulary words and hapax legomena is also much larger than in Hungarian which obviously makes tagging harder and makes it more useless to memorize training set in form of dictionaries. The quantitative properties of Evenk corpora are even more extreme than of Selkup. Last, but not the least is the origin of texts appearing in corpus, while most UD corpora are collected from media and fiction, Evenk and Selkup corpora are more informal by nature and mostly consist of native speakers oral speech records, which also makes the corpus less standartized. Summarizing, our hypothesis is that the informal nature of the corpora and dialectal variation increase the proportion of out-of-vocabulary and rare words, thus making it harder to memorize corpus via dictionaries.

All these concerns apply to lemmatization as well. The only thing we would like to mention is that multitask learning both on lemmatization and part-of-speech tagging task had failed, showing inferior performance. Probably this is due to low capacity of networks we used, however, our experiments show that even gold morphological tags does not improve lemmatization accuracy which implies that part-of-speech information is practically useless for this task. That contradicts our intuition and requires further study.

[10] The corpora for comparison are chosen randomly just to show the general pattern

6. Conclusions and further work

We compared different methods of augmenting neural models with additional information for lemmatization and part-of-speech tagging of Selkup and Evenk languages. Our results show that basic neural models outperforms its extensions. We expect this to be not a general phenomena, but the feature of particular datasets. However, a wider cross-lingual study is required to reveal the factors that affect the applicability of morphological dictionaries in tasks of computational morphology. The first experiment to perform is to use not the dictionary, extracted from the training set, but independently constructed one. However, the author does not know such dictionaries for Evenk and Selkup. Another direction of study is the usage of unlabeled corpora. Such corpora were available in the shared task, but their orthography was different from the morphology corpus. Given recent success of minimally supervised neural language models, probably we can extract more from unlabeled data than from dictionaries and grammars.

However, the main problem is to find the cheapest and quickest way for field linguists to create resources which will allow high-quality morphological analysis. For example, it is questionable whether it is easier to collect an unlabeled corpus of required size or an example grammar. The author is not a practical linguist to resolve this question, however the adoption of neural networks from industrial NLP for main world languages to the low-resource studies still has to be done.

Acknowledgements

The author thanks the Shared Task for giving the opportunity to participate and for their helpful cooperation during the Shared Task. I am also very grateful to the stuff of MIPT Neural Networks laboratory for warm and inspiring atmosphere during the work on the problem.

References

1. *Anastasiev D., Gusev I., Indenbom E.* Improving part-of-speech tagging via multitask learning and character-level word representations (2018), International conference on computational linguistics "Dialogue", Moscow, Russia, Vol 1., pp. 14–28.
2. *Heigold G., Neumann G., van Genabith J.* An extensive empirical evaluation of character-based morphological tagging for 14 languages (2017), Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Valencia, Spain, Vol. 1., pp. 505–513.
3. *Ioffe S., Szegedy C.* Batch normalization: Accelerating deep network training by reducing internal covariate shift (2015), arXiv preprint arXiv:1502.03167, available at https://arxiv.org/pdf/1502.03167.pdf.
4. *Kim Y. et al.* Character-Aware Neural Language Models (2016), AAAI., pp. 2741–2749.
5. *Klyachko E. et al.* LowResourceEval-2019: a shared task on morphological analysis for low-resource languages. (2019), Proceedings of International conference on computational linguistics "Dialogue", online articles, Moscow, Russia, to appear.

6. *Korobov M.* Morphological analyzer and generator for Russian and Ukrainian languages (2015), International Conference on Analysis of Images, Social Networks and Texts, Ekaterinburg, Russia, pp. 320–332.
7. *Makarov P., Ruzsics T., Clematide S.* Align and copy: UZH at SIGMORPHON 2017 shared task for morphological reinflection (2017), arXiv preprint arXiv:1707.01355, available at https://arxiv.org/pdf/1707.01355.pdf.
8. *Nivre, J. et al.* Universal Dependencies 2.3. (2018), LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), available at http://hdl.handle.net/11234/1-2895.
9. *Sorokin A.* Improving neural morphological Tagging using Language Models (2018), International conference on computational linguistics "Dialogue", Moscow, Russia, Vol 1., pp. 707–720.
10. *Srivastava R. K., Greff K., Schmidhuber J.* Highway networks (2015), arXiv preprint arXiv:1505.00387 available at https://arxiv.org/pdf/1505.00387.pdf.
11. *Straka M., Straková J.* Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe (2017), Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies., Vancouver, Canada, pp. 88–99.
12. *Zeman D. et al.* CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies (2018), Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies., Brussels, Belgium, pp. 1–21.
13. *Zeman D. et al.* CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies (2017), Proceedings of the CoNLL 2018 Shared Task: MultilingualParsing from Raw Text to Universal Dependencies, Vancouver, Canada, pp. 1–20.

PREDICTING DEPRESSION FROM ESSAYS IN RUSSIAN

Stankevich M. A. (stankevich@isa.ru)

Artificial Intelligence Research Institute, FRC CSC RAS, Moscow, Russia

Smirnov I. V. (ivs@isa.ru)

Artificial Intelligence Research Institute, FRC CSC RAS; Peoples' Friendship University of Russia (RUDN University), Moscow, Russia

Kuznetsova Y. M. (kuzjum@yandex.ru)

Artificial Intelligence Research Institute, FRC CSC RAS, Moscow, Russia

Kiselnikova N. V. (nv.pirao@gmail.com)

Psychological Institute of Russian Academy of Education, Moscow, Russia

Enikolopov S. N. (enikolopov@mail.ru)

Department of Medical Psychology, Mental Health Research Centre, Moscow, Russia

The study is focused on the detection of depression by processing and classification of short essays written by 316 volunteers. The set of 93 essays was provided by two different teams of psychologists who asked patients with clinically confirmed depression to write short essays on the neutral topic. The other 223 essays on the same topic were written by volunteers who completed questionnaires, which are designed to reveal depression status and did not demonstrate any signs of mental illnesses. The study describes psycholinguistic and classic text features which were calculated by utilizing natural language processing tools and were used to perform on the classification task. The machine learning classification models achieved up to 73% of f1-score for the task of revealing essays written by people with depression.

Key words: depression detection, text classification, psycholinguistic features, natural language processing

ВЫЯВЛЕНИЕ ПРИЗНАКОВ ДЕПРЕССИИ У АВТОРОВ ЭССЕ НА РУССКОМ ЯЗЫКЕ

Станкевич М. А. (stankevich@isa.ru)

Институт проблем искусственного интеллекта
ФИЦ ИУ РАН, Москва, Россия

Смирнов И. В. (ivs@isa.ru)

Институт проблем искусственного интеллекта ФИЦ ИУ РАН;
Университет дружбы народов (РУДН), Москва, Россия

Кузнецова Ю. М. (kuzjum@yandex.ru)

Институт проблем искусственного интеллекта
ФИЦ ИУ РАН, Москва, Россия

Кисельникова Н. В. (nv.pirao@gmail.com)

Психологический институт РАО, Москва, Россия

Ениколопов С. Н. (enikolopov@mail.ru)

Научный центр психического здоровья
НЦПЗ РАН, Москва, Россия

Данная работа направлена на задачу выявления депрессии при помощи обработки и классификации 316 эссе. Коллекция из 93 эссе была предоставлена двумя коллективами психологов, которые попросили пациентов с клинически подтвержденной депрессией написать эссе на нейтральную тему. Остальные 223 эссе на аналогичную тему были написаны добровольцами, которые прошли стандартный опросник на выявление депрессии и не показали признаков наличия ментальных заболеваний. Исследование описывает различные психолингвистические и стандартные текстовые признаки, полученные при помощи инструментов обработки естественного языка и использованные для задачи классификации. Основанные на машинном обучении классификационные модели продемонстрировали до 73% f1-меры в задаче обнаружения эссе, написанных людьми с депрессией.

Ключевые слова: обнаружение депрессии, классификация текста, психолингвистические признаки, обработка естественного языка

1. Introduction

It is a known fact that depression is one of the leading causes of disability worldwide and it affects millions of people around the world [11]. Depression can make a significant impact on the daily lifestyle and behavior of people. At the same time, a considerable number of depression cases stay untreated or undetected [21]. It is also known that severe types of depression affect the way human thinks and, therefore, influence human ability to express thoughts in oral speech and writings [2]. The psycholinguistic investigating this impact of depression and other mental diseases on human linguistics and propose some valuable methodology on it. But manual psycholinguistic analysis requires a lot of effort and time. Development of natural language processing tools allows to partially solve this problem [20]. At the same time, machine learning methods present a lot of opportunities to reveal human psychological attributes when applied on text data, for example in social media [17]. We currently aimed to develop such methods for the Russian social media and users' writings. But for the Russian language, we are lacking background knowledge about relations between psychological attributes of the human and his text.

The main idea of the study consists in applying machine learning and natural language processing tools to perform on the task of depression detection in essay writings on the Russian language. We formed two collections of essays: 93 essays written by people with clinically diagnosed depression, and 223 essays written by volunteers who completed a psychological questionnaire to confirm they did not demonstrate any depression signs. Thus, we focused on binary classification in order to evaluate the ability of machine learning approach to detect if a text belongs to depressed or healthy subject. We present features retrieved from essays including classical text features, psycholinguistic features, n-grams, and sentiment. It is important to note that currently there are no studies devoted to the depression detection task among Russian language and the psycholinguistic features proposed in the study are not previously tested on similar tasks.

2. Related work

Linguistic Inquiry and Word Count (LIWC) is one of the most frequently used tools for automatic text analysis for researches related to psychology and psycholinguistics [13]. The main idea embodied in this tool is that the author's psychological characteristics are related to the text's quantitative parameters: the frequency of punctuation marks, words of a certain part of speech (prepositions, conjunctions, pronouns, adverbs), words of a certain lexic-semantic group (negative or positive emotions, describing cognitive processes).

The task of depression detection mostly focused on social media data. There is a lot of studies that consider the task of detection depression by analyzing social media messages. The work presented in [6] describes the classification of social media messages written by depressed and healthy users. The authors achieved 74% of accuracy with SVM classifier using the following features: social media activity, time, N-grams, postags, and features based on LIWC.

Another work observes depression detection problem as a task of detecting vocabulary related to 9 depression symptoms [24]. Authors processed messages from Twitter to indicate the presence of these symptoms in users' writings. This approach is based on the observation that users of social networks frequently write about their mental state [3]. The experiments were focused on multi-label classification and comparison of semi-supervised topic modeling over time model (ssToT) with supervised SVM and Multinomial Naive Bayes approaches based on bag-of-words features. The ssToT model yielded 68% averaged accuracy which is competitive with a fully supervised approach presented in the study.

It is important to note studies presented by CLPsych 2015 shared task competitors [4]. The shared task provided dataset which consists of text messages samples from Twitter that belongs to users with depression and PTSD. The best average precision (roughly 87%) on the depression vs control task was achieved by the method based on lexical features with tf-idf weighting [15]. Another team with a good result (86% averaged precision) proposed the method of terms clustering and formed the feature set using clusters of terms as N-grams [14].

CLPsych 2018 [9] focused on the two tasks: predicting 11 years old child's current psychological health from essays and predicting future psychological health form the same essays. Participant's submissions on the regression tasks were compared by the Pearson Product-Moment Correlation Coefficient. The best approach for both tasks used regularized linear regression with character and word-level n-gram features [5]. The second place on the first task was achieved by the team that utilized tf-idf and sentiment features with ensembles of different methods: ridge regression, SVMs, boosting, CNNs, RNNs, and feed-forward neural networks [26].

Clef/eRisk 2017 Shared Task [8] provided noisy dataset which consists of 887 Reddit user's messages collections, where 135 of the persons were identified as belonging to a risk case of depression. The best submitted classification model yielded 64% of F1-score by the team who applied an ensemble of tf-idf based classifiers on the data [22]. It is important to note that the same team reworked their models after Clef/eRisk 2017 Shared Task completion and reported 73% F1-score on the same train/test data by utilizing sophisticated linguistic metadata features (including LIWC) with logistic regression [23].

The analysis of related works reveals that it is hard to strictly compare the results of our research with others. The data and experiments design presented in these works differ from study to study and covering only the English-speaking population. The social media based datasets contain much more textual information for each person, but at the same time it much noisier than essays. But we can observe the methods and approaches that yield promising results on depression detection task. For both CLPsych 2018 and Clef/eRisk 2017 shared tasks the classic well-tuned n-gram and tf-idf based models outperform neural networks models. The specific attributes of depression mental state usually revealed through the depression related dictionaries, sentiment, and LIWC features. Although LIWC is a very popular and effective tool, there is no appropriate adaptation for the Russian language. It is also missing some psycholinguistic characteristics of the text.

3. Dataset

The dataset for the research contains two classes of texts: 93 essays written by people with depression (depression group) and 223 essays written by healthy people (control group). Depression essays were collected with the collaboration of two different teams of the psychologist who asked patients with depression disorder to write a short essay with a minimum length of 1800 characters. The control group essays were written mostly by students from different universities and different education programs (psychology, sociology, journalism, and information technology). Volunteers completed Russian adaptation of Beck Depression Inventory [1] to reveal their depression status and only essays written by persons who did not demonstrate depression signs were included in the control group (score less than 14 on the 0–63 Beck Depression Inventory scale). Thus, volunteers from the depression group did not complete the same questionnaire because their depression status was revealed by clinical experts with face to face diagnostic, which is superior to questionnaires. In the other hand, this fact forbids us from investigating this task as regression analysis. The topic of the essays is similar for both depression and control groups. We can generalize this topic as "Me and my relations with others and the world around me". Minimal age of volunteers for both groups is 18.

We should highlight two assumptions that we made around the depression group to perform classification on the dataset. First, the depressive disorder can be divided into many subtypes, and each form of depression has a different severity. Secondly, the part of the depression essays was written by patients who have already taken medications, but another part was written by untreated persons. In another hand, it is a very difficult task to collect a big number of essays from people with clinically confirmed depression. Utilizing machine learning tools require a sufficient number of training examples, which forced us to generalize all of the depression types as one and ignore the fact of medication use by depressed patients.

We present general statistics on the dataset in Table 1. It can be noted that generally, people with depression tend to write shorter texts. The mean age of the depression group is insufficiently higher than the control group. The gender distribution in the depression group is 59% females and 41% males. Gender distribution in the control group is 69% females and 31% males.

Table 1. Dataset statistics. Mean values and standard deviation

Value	Depression group	Control group
Number of essays	93	223
Age	28.05±10.67	22.82±10.01
Gender	55 Female, 38 Male	153 Female, 70 Male
Mean characters count	1883 ± 895	1994 ± 207
Mean words count	294.83 ± 145.89	317.75 ± 35.6
Mean sentence count	22.6 ± 11.28	23.11 ± 6.93

4. Methods

4.1. N-grams

Tf-idf and n-grams features are common natural language processing approach which performed well on depression detection task. Thus, we included 2 tf-idf based feature sets: unigrams and bigrams. N-grams that appeared less than in 1% of essays and more than in 90% of essays were removed from the feature sets. The result of t-SNE [10] on the bigrams data demonstrated in **Figure 1**.

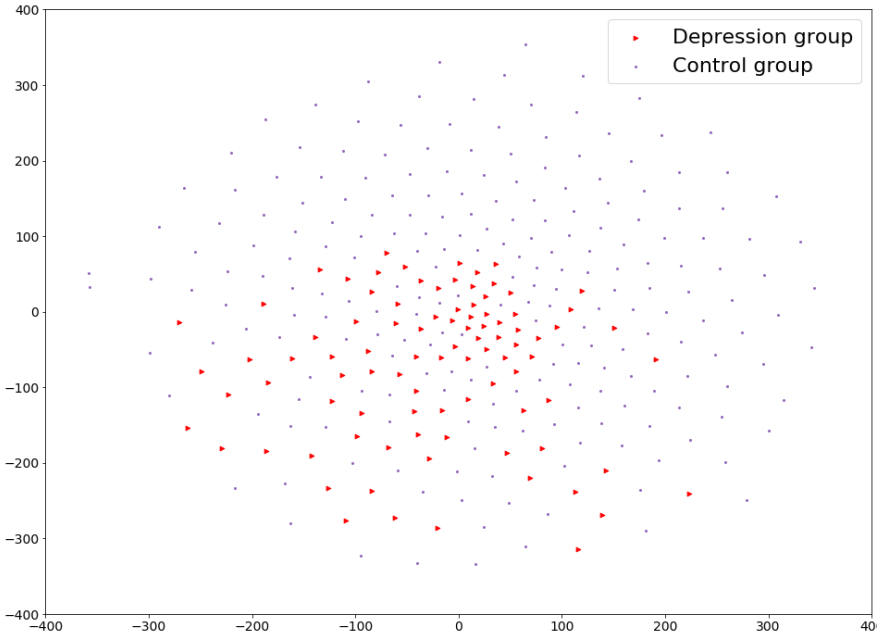


Fig. 1. Results of t-SNE applied on bigrams features

4.2. Depression markers

We annotated following feature set as Depression Markers (DM). It is containing classical text features (mean word/sentence length, POS-tags ratio) and psycholinguistic features. The part of psycholinguistic features described in following works [16], [18], [19], [25], and another part were proposed during the manual analysis of essays by psychologists and linguists. An important point is that most of these features were not previously tested on depression detection task.

To retrieve features from the essays we applied tokenization, lemmatization, and morphological analysis with MyStem. Statistics for DM presented in Table 2 (excluding POS-tags ratio). By observation of morphology related features, it can be noted that verbs and pronouns in various forms yield a lot of differences between groups. The smaller mean depth of syntax tree and mean number of words per sentence demonstrate a tendency of depression group to express themselves with shorter sentences. The *(N verbs) / (N adjectives)* ratio, which is also known as Trager coefficient, is usually differ from 1 among people with higher mental stress, which is also differ in our report among depression group [19].

Table 2. Mean+std for depression markers in depression and control group

Description	Depression group	Control group
Mean number of words per sentence	12.66±3.63	14.6±3.81
Mean number of characters per word	5.01±0.32	5.06±0.29
Lexicon: (N unique words) / (N words)	0.56±0.07	0.53±0.05
Average syntax tree depth	4.96±1.24	5.41±1.28
(N verbs) / (N adjectives)	1.36±0.45	1.11±0.30
(N verbs) / (N nouns)	0.5±0.12	0.5±0.08
(N participles) / (N sentences)	0.11±0.08	0.16±0.11
(N conjunctions + N prepositions) / (N sentences)	2.65±0.89	3.15±1.01
(N infinitives) / (N verbs)	0.23±0.07	0.28±0.08
(N singular first person past tense verbs) / (N verbs)	0.19±0.10	0.13±0.10
(N past tense verbs) / (N verbs)	0.69±0.09	0.62±0.09
(N first person verbs) / (N verbs)	0.2±0.11	0.15±0.10
(N third person verbs) / (N verbs)	0.18±0.08	0.25±0.10
(N first person pronouns) / (N pronouns)	0.56±0.14	0.45±0.15
(N singular first person pronouns) / (N pronouns)	0.53±0.15	0.35±0.19
(N plural first person pronouns) / (N pronouns)	0.01±0.02	0.08±0.08
(N words with wrong spelling) / (N sentences)	0.11±0.13	0.09±0.12
Sentiment rate	−1.31±6.36	3.91±6.62

4.3. Sentiment

Another valuable feature was computed with the help of Linis Crowd word sentiments dictionary [7]. The dictionary provides information about words estimated values of positive (1, 2), negative (−1, −2) or neutral (0) sentiment. We calculated the sentiment rate of each document by matching words from the essay with dictionary values and then summing it up. As demonstrated

in Figure 2, sentiment rates for each group greatly vary and have been included in the feature set. This value was included in the DM set.

5. Result of experiments

To perform classification, we utilized scikit-learn [12] implementation of random forest and SVM algorithms. Overall, we evaluate 5 different sets of features: depression markers (*DM*) that was described in **section 4.2**, tf-idf model computed on unigrams, tf-idf model computed on bigrams, and combination of n-gram models with depression markers. The classification report represented as averaged result of 5-fold cross-validation on the data (**Table 3**). Recall, precision, and F1-score calculated for the class of depression.

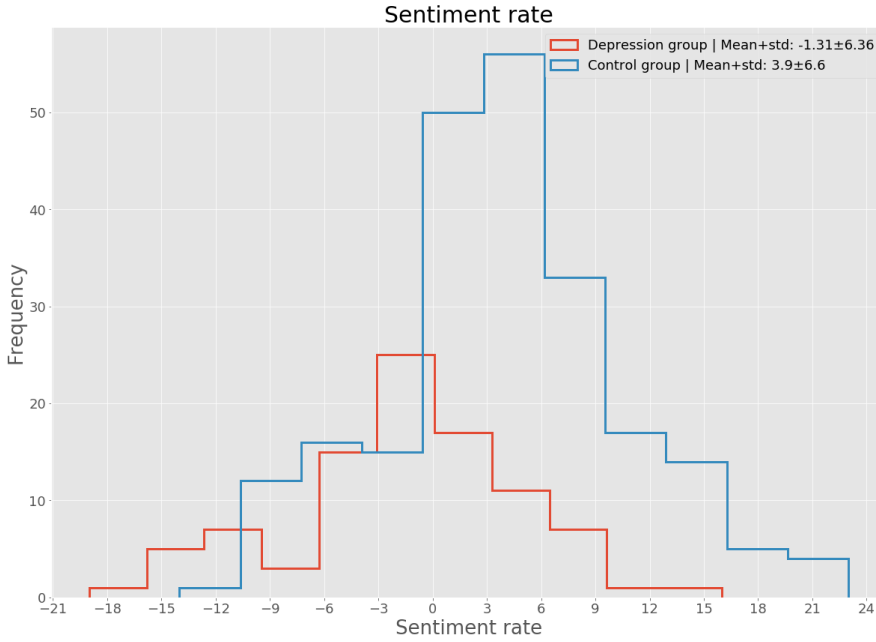


Fig. 2. Sentiment rate for depression and control groups

Depression markers model achieved best score in terms of overall classification performance on the data with 84% accuracy. N-grams based model also performed well with a F1-score around 70%. The combination of all models yielded best result for the task of depression essays classification with 73% F1-score. We relate the high values of standard deviation to the small number of samples in the data. The SVM based models are also yield best performance with bigrams features.

As it was mentioned before, it is hard to provide strict comparison with similar works, since the data format and language is different than in other studies. In terms of experiments design, studies related to the Clef/eRisk 2017 is the closest ones. The best reported F1-score for depression class on Clef/eRisk 2017 data is also 73% [23], which is close to F1-score in our experiments.

Table 3. Classification results

Feature set	Recall, %	Precision, %	F1, %	Accuracy, %
Random Forest				
DM	65.53 ± 8.31	**77.52** ± 4.91	70.65 ± 5.39	**84.16** ± 2.15
Unigrams	69.83 ± 9.36	70.87 ± 7.31	69.69 ± 4.51	82.27 ± 2.40
Unigrams + DM	72.01 ± 10.03	70.7 ± 10.18	70.48 ± 5.36	82.28 ± 3.40
Bigrams	**76.26** ± 7.37	70.42 ± 5.69	72.72 ± 2.44	83.21 ± 1.77
Bigrams + DM	74.18 ± 3.11	72.12 ± 4.16	**73.01** ± 2.11	83.85 ± 1.46
SVM				
DM	78.66 ± 14.27	52.78 ± 5.63	62.96 ± 8.12	73.11 ± 5.44
Unigrams	49.59 ± 16.11	69.04 ± 5.80	56.60 ± 12.67	78.81 ± 4.29
Unigrams + DM	72.28 ± 15.01	61.00 ± 11.09	66.05 ± 12.51	78.21 ± 8.17
Bigrams	64.67 ± 11.74	72.42 ± 8.89	68.01 ± 9.38	82.29 ± 5.04
Bigrams + DM	65.76 ± 11.54	69.16 ± 7.05	67.20 ± 8.76	81.35 ± 4.62

6. Conclusion

The study evaluates the ability of machine learning models and several types of feature sets to perform classification on essays in Russian written by depressed and healthy peoples. The depression markers that was described in the paper, as well as standard NLP approaches like unigrams and bigrams, demonstrated good performance on the data. The Bigrams+DM feature set achieved the best results for the task of revealing depression essays with 73% F1-score. It was discovered that applying word sentiment dictionaries as Linis Crowd is suitable for the depression detection task. We considering this study as a first step in the machine learning based depression detection from texts in Russian.

We currently looking forward to investigating the ability of word embeddings and neural networks models to identify depression in human writings. The dataset possibly will become public-available for research purposes in the fully anonymized format. As a general idea for future work, we planning to apply depression detection methods on Russian-speaking social networks.

Acknowledgments

The publication has been prepared with the support of the “RUDN University Program 5-100” and funded by RFBR according to the research projects №17-29-02225 and №17-29-02305.

References

1. *Beck, A. T., Steer, R. A., & Carbin, M. G.* (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. Clinical psychology review, 8(1), 77–100.
2. *Bucci, W., & Freedman, N.* (1981). The language of depression. Bulletin of the Menninger Clinic, 45(4), 334.
3. *Coppersmith, G., Dredze, M., Harman, C., & Hollingshead, K.* (2015). From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (pp. 1–10).
4. *Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M.* (2015). CLPsych 2015 shared task: Depression and PTSD on Twitter. In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (pp. 31–39).
5. *Çöltekin, Ç., & Rama, T.* (2018). Tubingen-Oslo system: Linear regression works the best at Predicting Current and Future Psychological Health from Childhood Essays in the CLPsych 2018 Shared Task. arXiv preprint arXiv:1809.04838.
6. *De Choudhury, M., Counts, S., & Horvitz, E.* (2013, May). Social media as a measurement tool of depression in populations. In Proceedings of the 5th Annual ACM Web Science Conference (pp. 47–56). ACM.
7. *Koltsova, O. Y., Alexeeva, S., & Kolcov, S.* (2016). An opinion word lexicon and a training dataset for russian sentiment analysis of social media. Computational Linguistics and Intellectual Technologies: Materials of DIALOGUE 2016 (Moscow), 277–287.
8. *Losada, D. E., & Crestani, F.* (2016, September). A test collection for research on depression and language use. In International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 28–39). Springer, Cham.
9. *Lynn, V., Goodman, A., Niederhoffer, K., Loveys, K., Resnik, P., & Schwartz, H. A.* (2018). Clpsych 2018 shared task: Predicting current and future psychological health from childhood essays. In Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic (pp. 37–46).
10. *Maaten, L. V. D., & Hinton, G.* (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(Nov), 2579–2605.
11. *Moussavi, S., Chatterji, S., Verdes, E., Tandon, A., Patel, V., & Ustun, B.* (2007). Depression, chronic diseases, and decrements in health: results from the World Health Surveys. The Lancet, 370(9590), 851–858.
12. *Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J.* (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825–2830.
13. *Pennebaker, J. W., Francis, M. E., & Booth, R. J.* (2001). Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates, 71(2001), 2001.
14. *Preoţiuc-Pietro, D., Sap, M., Schwartz, H. A., & Ungar, L.* (2015). Mental illness detection at the World Well-Being Project for the CLPsych 2015 shared task. In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (pp. 40–45).

15. *Resnik, P., Armstrong, W., Claudino, L., & Nguyen, T.* (2015). The University of Maryland CLPsych 2015 shared task system. In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (pp. 54–60).
16. *Samokhvalov V. P.* (2002), Psychiatry [Psihiatriya], Phoenix, Rostov-on-Don.
17. *Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... & Ungar, L. H.* (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one, 8(9), e73791.
18. *Shvedovsky E. F., Zvereva N. V.* (2015). Studies of speech disorders in schizophrenia. History and state of-the-art [Исследование речевых нарушений при шизофрении. История и современное состояние проблемы]. Psychological Science and Education, 20(2), 78–92.
19. *Smirnova, D. A.* (2010). Clinical and psycholinguistic characteristics of mild depression [Клинические и психолингвистические характеристики легких депрессий] (Doctoral dissertation, Moscow Research Institute of Psychiatry).
20. *Tausczik, Y. R., & Pennebaker, J. W.* (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of language and social psychology, 29(1), 24–54.
21. *Thornicroft, G., Chatterji, S., Evans-Lacko, S., Gruber, M., Sampson, N., Aguilar-Gaxiola, S., ... & Bruffaerts, R.* (2017). Undertreatment of people with major depressive disorder in 21 countries. The British Journal of Psychiatry, 210(2), 119–124.
22. *Trotzek, M., Koitka, S., & Friedrich, C. M.* (2017, September). Linguistic Metadata Augmented Classifiers at the CLEF 2017 Task for Early Detection of Depression. In CLEF (Working Notes).
23. *Trotzek, M., Koitka, S., & Friedrich, C. M.* (2018). Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. IEEE Transactions on Knowledge and Data Engineering.
24. *Yazdavar, A. H., Al-Olimat, H. S., Ebrahimi, M., Bajaj, G., Banerjee, T., Thirunarayan, K., ... & Sheth, A.* (2017, July). Semi-supervised approach to monitoring clinical depressive symptoms in social media. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 (pp. 1191–1198). ACM.
25. *Zagorovskaya, O. V., Litvinova, O. A., & Litvinova, T. A.* (2016). Identify the tendency of the individual to suicidal behavior on the basis of a quantitative analysis of speech production [Выявление склонности личности к суицидальному поведению на основе количественного анализа ее речевой продукции.]. Studia Humanitatis, (1).
26. *Zaporojets, K., Sterckx, L., Deleu, J., Demeester, T., & Develder, C.* (2018). Predicting Psychological Health from Childhood Essays. The UGent-IDLab CLPsych 2018 Shared Task System. In Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic (pp. 119–125).

NEWS HEADLINE GENERATION USING STEMS, LEMMAS AND GRAMMEMES

Stepanov M. A. (projectttower@gmail.com)
MIPT, Dolgoprudny, Russia

Headline generation is a task that has a good solution based on seq2seq models with an attention mechanism. However, it is still quite challenging to deal with morphologically rich languages, such as Russian, which have many word forms and therefore larger vocabularies. To deal with complex dependencies arising in such languages we propose several approaches based on using stems and grammemes. We applied these approaches to the pointer-generator network and took second place in the competition on headline generation held by the conference Dialogue-2019.

Key words: headline generation, Russian language, pointer-generator, stem, flexion, lemma, grammeme

ГЕНЕРАЦИЯ ЗАГОЛОВКОВ НОВОСТНЫХ СТАТЕЙ, ИСПОЛЬЗУЮЩАЯ СТЕМЫ, ЛЕММЫ И ГРАММЕМЫ

Степанов М. А. (projectttower@gmail.com)
МФТИ, Долгопрудный, Россия

Задача генерации заголовков имеет хорошее решение, которое базируется на использовании seq2seq моделей с механизмом внимания. Однако в случае морфологически богатых языков таким моделям приходится сталкиваться с более сложными зависимостями, которые могут проявляться в виде большого количества словоформ и их сочетаний друг с другом. Мы предлагаем несколько подходов, которые могут помочь автоматическим seq2seq генераторам заголовков учитывать зависимости таких языков, как русский. Мы также применили данные подходы к архитектуре генератора-указателя и заняли второе место на соревновании по генерации заголовков, проведённом в рамках конференции Диалог-2019.

Ключевые слова: генерация заголовков, генератор-указатель, стемминг, флексия, лемма, граммема

1. Introduction

There are two main groups of text summarization approaches: abstractive and extractive. While extractive aproaches try to find the most informative subset of the text and copy it, abstraction-based systems generate words and phrases not from the source, but from the vocabulary, using learned natural language dependencies.

Automatic headline generation is a type of the summarization task. The aim of summarization is to create a shorter version of the text (in our case, the title), which contains the main idea of the given article. Working on task of generating headings has an advantage over the traditional summarization: it is much easier to find articles with titles than with annotations, which is very convenient for systems based on machine learning methods. There is almost an infinite supply of news articles in all major languages and almost all of them have a headline.

But, despite the existence of a huge amount of data, headline generation system still should be able to deal with dependencies of natural language, and the creation of this system is a challenging task. Due to this difficulty the vast majority of past decisions use extractive methods (see [1] or [2]), but the relatively recent success of sequence-to-sequence models [3] has made the abstractive approach viable (see [4] or [5]). Now it is possible to automatically read and generate text that has the structure similar to the headings written by human.

However, the benefits that seq2seq brought were not enough to create desirable headlines: these systems have problems such as the words repeating and the inability to use out-of-vocabulary (OOV) words of a source article. To enable OOV extraction, a pointer-generator model has been developed and introduced by See et al. [6]. This model is both extractive and abstractive: it is based on seq2seq, but can copy words from text too. Additionally, the coverage mechanism and the usage of a coverage loss (penalty for repeating words) during the training phase makes this model less prone to repetition. Due to these advantages, we chose the pointer-generator network with coverage mechanism as the baseline.

Though pointer-generator network can create human-like headlines of English news, it is quite difficult for the model to achieve the same success with, for example, Russian articles. Even simple vocabulary of morphologically rich language can contain several million forms and variations. For a model it is harder to find suitable words in the space of possible variants expanded by word forms. With a larger vocabulary it takes much more memory, computing power and time to teach the network to generate desirable headings.

In this paper we propose several approaches to deal with problems of morphologically rich languages: stem+flexion encoding and grammeme embeddings. We also present results of experiments that were made with RIA corpus[1] (presented by [7]) and Lenta corpus[2] during the competition track on the headlines generation held by the conference Dialogue-2019.

[1] https://github.com/RossiyaSegodnya/ria_news_dataset

[2] https://github.com/yutkin/Lenta.Ru-News-Dataset

2. System description

2.1. Baseline model

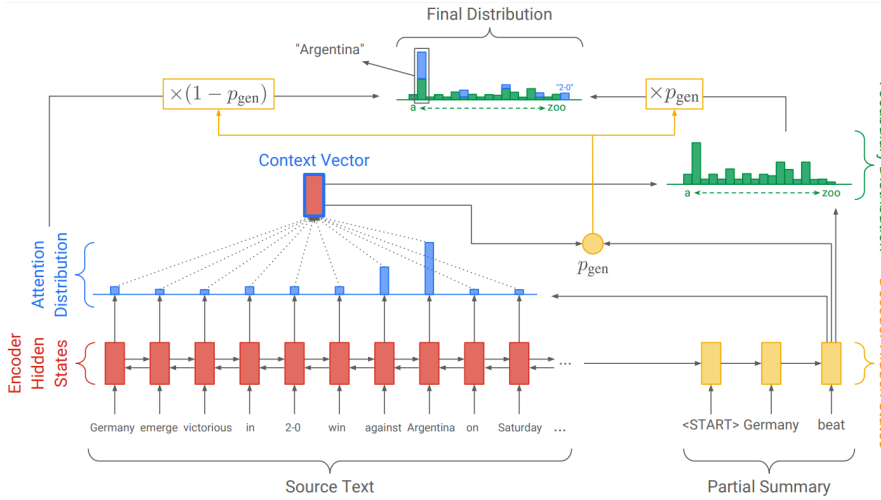


Figure 1: Pointer-generator model scheme from [6]

The pointer-generator network is based on a sequence-to-sequence model with an attention mechanism (Figure 1). It uses the encoder to make encoder hidden states h_i, which store the extracted information from the article. Article tokens w_i are fed one-by-one to the encoder's embedding layer and the single-layer bidirectional LSTM. After that the model generates words of abstract step by step, applying the decoder (unidirectional one-layer LSTM) to produce a decoder state s_t from an embedding of previously generated word y_{t-1} and so-called context vector (created by the attention mechanism) h_t^*. Then the network gets the output vocabulary distribution (that show which word is most probable as next token of the headline) from the decoder state.

The attention mechanism is a modification of the seq2seq model which helps the decoder to produce the next word indicating which words of the source article are the most important at the step t. This information is contained in the attention distribution a^t calculated by this mechanism. Next, using the attention distribution as weights in the sum of encoder states h_i, model creates context vector h_t^*—the "second ingredient" of the output vocabulary distribution.

In addition to the generation of words from the fixed vocabulary this model is able to copy tokens from the source article. It is realized by calculating the generation probability p_{gen} at each step t. Then the network use p_{gen} as a soft switch to choose between generating a word using the vocabulary distribution, or copying a word from the text using a^t, which shows the most suitable tokens for extraction. This modification makes model both extractive and abstractive and therefore more flexible for different kinds of situations.

To cope with the output repetition problem, coverage mechanism is also involved in the title generation process. This modification retains all attention distributions produced by the model at each step t, and gives an additional loss if the model use similar a^t. If pointer-generator is trained with coverage mechanism, it is more liable to extract different words from the source and use different context vectors, which makes the model less repetitive[3].

2.2. Stem+flexion encoding

In order to help model to work with larger vocabulary of morphologically rich languages, we experimented with two approaches. Both of them change the structure of input and output words to make the vocabulary sufficiently smaller with no drops in performance.

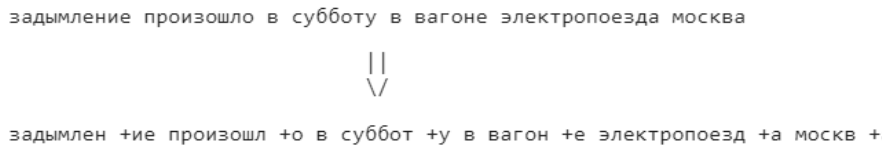


Figure 2: Example of stem+flexion encoding

The first approach is based on encoding each word as a pair of its stem and flexion (or only stem if there is no flexion). To encode n words with m forms of each word model can work with a list of n stems and a fixed number of flexions instead of a vocabulary with $n * m$ words, which makes it easier for a network to find natural language dependencies in articles. The output of model consists of stems and flexions too and it can be easily decoded into words sequence.

In our experiments we use a Porter stemmer[4] [8] for automatic encoding and a vocabulary of stems and flexions with 450 flexions[5]. Each flexion has a '+' as a prefix to distinguish them from stems (**Figure 2**) and to restore headline from the output sequence of stems and flexions.

It is important to mention that we don't make any changes in the model architecture in this part of experiments, only changes in input and output processing. But we also make attempts to use 3-layer encoder and decoder instead of single-layer to help the model to learn more sophisticated dependencies[6].

---

[3] read [6] to get more information about baseline model

[4] https://medium.com/@eigenein/стеммер-портера-для-русского-языка-d41c38b2d340

[5] https://colab.research.google.com/drive/1DEEwaFGQV6-SvoBuqalvxrSL3uLqI535

[6] https://colab.research.google.com/drive/1U6BHW2TgfnjpxnoSdJzWLf0_23mVZFqF

2.3. Grammeme embeddings

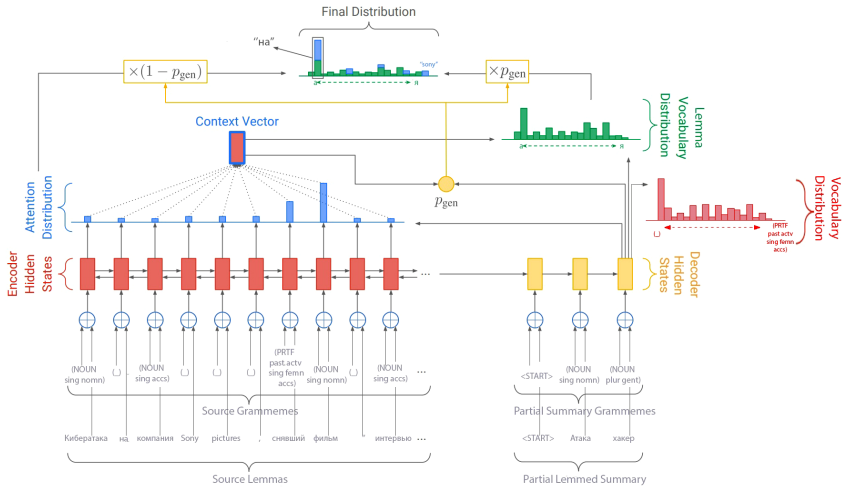


Figure 3: Pointer-generator model using grammemes

Another approach is based on the usage of lemmas and grammemes instead of words. We use a morphoparser (we choose pymorphy2[7]) to divide each word into its lemma and a string consisting of a part of speech and all values of changeable grammatical values (**Figure 3**): For example, a noun 'хакеров' is encoded to lemma 'хакер' and string '(NOUN plur gent)'. If a part of speech is not changeable (prepositions, conjunctions), then word gets string '(_)'. With this method, we created the vocabulary of lemmas and the vocabulary of strings with grammemes, which in our experiments has a size of 300.

We have changed the model architecture for these experiments: instead of the embedding layer for tokens of encoder's and decoder's input sequences we have made two independent layers for lemmas and grammeme strings. Network transforms article words to two sets of lemmas and grammemes and each of them passes through its own embedding layer. Then model concatenates two embeddings and gives the result to encoder and decoder. In addition to the vocabulary distribution (of lemmas) decoder generates distribution over the vocabulary of strings with grammemes mentioned above.

Next, in the training phase model calculates the loss. We have included additional cross entropy loss for the grammeme output sequence in order to help the network to learn how to create right word forms. If the title generator works in the test phase, it tries to create headline with morphoparser by applying grammeme strings to lemmas (if it is impossible, the model gives lemma to output)[8].

7 https://pymorphy2.readthedocs.io/en/latest/

8 https://colab.research.google.com/drive/1zIJ3Pk1oljRR8qTaZn25UkfDTeTKIL77

3. Data and training

We consider two corpora: RIA and Lenta datasets. RIA dataset was provided by Russian news agency "Rossiya Segodnya" and used in the competition track of the conference Dialogue-2019[9]. It contains 1,003,869 news articles of the time period from January 2010 to December 2014. We use this corpus as a training dataset which has an additional preprocessing such as cleaning from html-tags, lower-casing and tokenization. To speed up the learning of models, articles are also processed by the stem+flexion encoder and divided into lemmas and grammemes sequences by the morphoparser.

Lenta corpus has 739 new articles from 1999-08-30 to 2018-12-15. We chose 10,000 random articles to form the test dataset. These texts were preprocessed in the same way as the train dataset.

4. Experiments

4.1. Models

In this work there were 4 different models which trained on Ria Corpus and were tested on Lenta Corpus. Here they are: **baseline pointer-generator**, **pointer-generator using stems**, **pointer-generator using stems and 3-layer LSTM** and **pointer-generator using grammeme embeddings**.

4.2. Training

The models trained with the Adam optimizer using a scaled learning rate. All of them worked with vocabularies with 100,000 tokens and used token embeddings with the size of 128. Grammeme embeddings had the size of 32. Embedding layers were shared between encoder and decoder for all models. The size of the hidden vectors of LSTM layers was equal to 256. In addition, the length of the input sequences was limited with 600 tokens for the model with stems and 400 for other models. Reference headlines were also truncated to 20 (for the model with stems) and 12 tokens (for other models). For headline generation, beam-search size was made equal to 4.

All models trained with batches of articles with the size of 32. Baseline pointer-generator trained for 400,000 epochs, as models working with stems. Model with grammemes embeddings passed through 285,000 training epochs.

5. Results

We present our results on Lenta dataset in the Table 1. As it can be seen, all models with modifications surpassed vanilla pointer-generator on ROUGE-1, ROUGE-2 and ROUGE-L F1 scores. Model with 3-layer LSTM shows better results than the same

[9] https://vk.com/@headline_gen-announcement

model but with single-layer encoder and decoder. Both of them had the same number of training epochs, so it seems, that more complex architecture helps title generator to understand natural language dependencies arising in this dataset better.

Network using grammeme embeddings has better R-1 and R-L scores than models with stems, but it loses in R-2 scores. But this model has single-layer LSTM, and if encoder and decoder would be multi-layer, the network with grammeme embeddings could outperform models with stems in all scores and with a large margin, what makes usage of grammemes more preferable than applying stem+flexion encoding.

Table 1: ROUGE-1,2,L F1 and recall scores, on Lenta corpus

Model	R-1-f	R-1-r	R-2-f	R-2-r	R-L-f	R-L-r
Pointer-generator (baseline)	21.36	22.27	8.69	8.70	19.25	20.79
Pointer-generator with stems	23.47	23.81	10.24	10.39	21.24	22.27
Pointer-generator with stems and 3-Layer LSTM	25.16	25.82	11.32	11.63	22.78	24.13
Pointer-generator with grammeme embeddings	25.23	25.79	10.33	10.60	22.82	24.08

Using the model with stems, we took second place in headline generation contest held by Dialogue-2019. This model was evaluated on the private part of the RIA dataset and had a score of 20.29 (mean of R-1-f, R-2-f, R-3-f). Unfortunately, 3-layer stem model and the model with grammeme embedding didn't participate in competition because of lack of training time at the end of this event.

Additionally, we present headlines generated by all four models with two random texts from the dataset (Table 2).

Table 2: Samples of headlines generated by models

Original text, truncated: дамаск , 11 мая . - риа новости . президент россии дмитрий медведев считает опасным дальнейший рост напряженности на ближнем востоке . " дальнейший разогрев ситуации на ближнем востоке чреват взрывом и катастрофой " , - сказал медведев на пресс-конференции по итогам переговоров с президентом сирии башаром асадом . " с моей стороны было специально подчеркнуто , что россия будет и дальше предпринимать все от нас зависящее для того , чтобы помогать восстановлению арабо-израильского мирного процесса на основе международно-правовой базы , которая имеется ...
Original headline: медведев : " разогрев " ситуации на ближнем востоке чреват катастрофой
Headline by baseline pointer-generator: медведев считает опасным дальнейший рост напряженности на ближнем востоке
Headline by pointer-generator using stems: медведев : рост напряженности на ближнем востоке чреват взрывом
Headline by pointer-generator using stems and 3-layer LSTM: напряженность на ближнем востоке опасна , заявил медведев

Headline by pointer-generator using grammeme embeddings: медведев считает опасным рост напряжённости на ближнем востоке
Original text, truncated: москва , 5 мая - риа новости . задымление произошло в субботу в вагоне электропоезда москва - фрязино ярославского направления московской железной дороги , из-за чего пассажиров пришлось пересадить в другую электричку , сейчас движение поездов восстановлено , сообщил риа новости руководитель службы корпоративных коммуникаций мжд владимир мягков . “ сегодня в районе платформы чкаловская в электропоезде номер 6707 в пятом вагоне произошел нештатный разогрев буксы колесной пары , что дало небольшое задымление . в связи с этим электропоезд был остановлен ” , - сказал мягков ...
Original headline: в электричке в подмосковье произошло задымление вагона
Headline by baseline pointer-generator: задымление произошло в электричке на подмосковном железной дороге
Headline by pointer-generator using stems: задымление произошло в вагоне поезда москва - фрязино ярославского направления
Headline by pointer-generator using stems and 3-layer LSTM: задымление произошло в вагоне электропоезда московской железной дороги
Headline by pointer-generator using grammeme embeddings: задымление произошло в вагоне электрички мжд - фрязино

6. Conclusion

In this paper, we explore the application of two approaches to the pointer-generator network processing, such as usage of stems and grammemes, and with these described modifications model outperforms its own results on Russian news articles. The future work will focus on testing models on other datasets and experimenting with settings and subsystems of the model.

Acknowledgements

Author is thankful to Ivan Smurov for useful discussions and proofreading, organizers of competition track of Dialogue-2019 on headline generation for providing test dataset and interesting task.

References

1. *Julian Kupiec, Jan Pedersen, and Francine Chen* (1995). A trainable document summarizer. In International ACM SIGIR conference on Research and development in information retrieval
2. *Horacio Saggion and Thierry Poibeau* (2013). Automatic text summarization: Past, present and future. In Multi-source, Multilingual Information Extraction and Summarization, Springer, pages 3–21.

3. *Ilya Sutskever, Oriol Vinyals, and Quoc V Le* (2014). Sequence to sequence learning with neural networks. In Neural Information Processing Systems.
4. *Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang* (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. In Computational Natural Language Learning.
5. *Alexander M Rush, Sumit Chopra, and Jason Weston* (2015). A neural attention model for abstractive sentence summarization. In Empirical Methods in Natural Language Processing.
6. *Abigail See, Peter J. Liu, Christopher D. Manning* (2017). "Get To The Point: Summarization with Pointer-Generator Networks" arXiv:1704.04368.
7. *Daniil Gavrilov, Pavel Kalaidin, Valentin Malykh* (2019). "Self-Attentive Model for Headline Generation" arXiv:1901.07786.
8. *Martin F. Porter* (1980). An algorithm for suffix stripping.

SOME FEATURES OF THE COMPLETIVE PREFIX *DO-* IN RUSSIAN: THEORY FACES EMPIRICAL DATA[1]

Stoynova N. (stoynova@yandex.ru)
Vinogradov Russian Language Institute & Institute of Linguistics, RAS; NRU HSE; Moscow, Russia

The paper deals with some formal features of the completive prefix *do-* ('to finish, to complete'). It was claimed in previous studies, that this prefix along with some others, has a range of formal properties that differ both from formal properties of productive "superlexical" prefixes (such as the cumulative *na-*, the distributive *po-*) and "lexical" (highly integrated) ones. Two important features were mentioned among others. 1) It can attach both to the perfective stem and to the imperfective one. 2) It cannot attach to secondary imperfectives. In the paper, I verify and develop these claims on corpus data. 1) I propose the rules of choice between the perfective vs. imperfective stem and describe the pool of variation. 2) I show, that, contrary to expectations, in informal speech *do-* attaches to secondary imperfectives quite easily.

Key words: Russian, verb, prefixation, intermediate prefixes, superlexical prefixes, completive

1. Introduction

The paper deals with some formal features of the completive prefix *do-* ('to finish, to complete'). It was postulated in previous theoretical studies on Russian prefixation, that this prefix belongs to a small group of Russian prefixes, intermediate between productive "superlexical" prefixes (which are semantically transparent and occupy the external position within the stem) and "lexical" ones (which are lexicalized and attach directly to the root), cf. [Tatevosov 2008]; [2009]; [2013]. They are predicted to have a range of formal features that distinguish them from both these types.

This paper is empirically-oriented, rather than theoretical. Its aim is to consider these features on quite a large massive of empirical data (dictionary data and corpus data), to verify the theoretical predictions on this prefix and to describe its more specific properties, which do not follow from general theoretical assumptions.

The paper has the following structure. In **Section 2**, I observe briefly the existing classification of Russian prefixes with a special focus on the completive *do-*. In **Section 3**, I differentiate between the completive *do-*, which is in focus of the study,

[1] The study was funded by RFBR, project 17-29-09154 (Trends in the development of language system: a corpus-based study of synchronic variation and diachronic change in different text types).

and some other similar uses of this prefix. In **Sections 4** and **5**, I test on corpus and dictionary data two formal features of *do-*: the ability to attach both to perfective and imperfective stems (4) and the position with respect to the suffix of secondary imperfectivization (5). Section 6 contains a brief conclusion.

2. The completive *do-* within the classification of Russian prefixes

Russian verbal prefixes can be divided at least into two formal types. Prefixes of the first type are less productive, they express more concrete and less transparent meanings, and they are closer to the root within the verb stem: cf. *при-йти*, *вы-йти*, *обо-йти*, derived from *идти* 'to go'; *при-думать*, *вы-думать*, *об-думать* derived from *думать* 'to think'. Prefixes of the second type (e.g. the cumulative *на-*, the distributive *по-* and *пере-*) are very productive, they have abstract, transparent meanings, within the verb stem they can attach above prefixes of the first group, cf. *на-* in *на-при-думывать* (CUMUL-LEX.PREF-think.IPFV 'to think out a lot of things'), *по-* in *по-у-ходить* (DISTR-LEX.PREF-go 'to go one by one'). This opposition was formulated in impressionistic semantic terms of "qualifiers vs. modifiers" already in [Isatchenko 1965/2003: 222–224]. Later, this idea was developed within the modern formal approaches by [Babko-Malaya 1999]; [Ramchand 2004]; [Svenonius 2004]; [Romanova 2004]; [2006] and others (inner vs. outer prefixes, lexical vs. superlexical prefixes).

A more detailed classification was proposed by S. G. Tatevosov. In [Tatevosov 2008] he introduced one more "intermediate" type of prefixes consisting of the completive *do-*, which is in focus in the paper, and the repetitive *pere-* (and also *pod-* which was added to the list in later papers). In [Tatevosov 2009] and [Tatevosov 2013] he also divided other superlexical prefixes to several groups.

In [Tatevosov 2008] proposes a list of 6 formal features of "intermediate prefixes": some of them are shared with lexical ones, some of them are shared with superlexical ones, and some others are specific for this type.

Two of these features are relevant for this paper.

1) "Aspectual selection". Superlexical prefixes attach to the imperfective stem: *по-спрашивать* (*по-*ask.IPFV), lexical prefixes behave differently: *от-дать* *от-*give.PFV, но *за-шить* (*за-*sew.IPFV), while intermediate prefixes are predicted to combine with both types of stems without restrictions: *до-шить* (*до-*sew.IPFV) and *до-подшить* (*до-*shorten.PFV). If both options are allowed by the general model of Russian prefixation, the question arises, on which grounds the stem is chosen in practice.

2) "Position with respect to the secondary imperfective". Following [Tatevosov 2008], superlexical prefixes can attach above the secondary imperfective (*украсить* (decorate.PFV > *украш-а-ть* (decorate-IPFV) > *по-*[*у-краш-а*]*-ть* (*по-*decorate-IPFV). Intermediate prefixes, including *до-* are, in contrast, predicted to attach only below the secondary imperfective. However, in [Tatevosov 2013] notes that this feature has no connection to other formal properties of this group of prefixes and it does not follow from the general architecture of Russian prefixation, i. e. the opposite situation would not break anything in the general model. If so, it is interesting to check, whether this feature is in fact so strict or the opposite option also can be realized under some conditions.

3. The completive use among other uses of *do-*

In the paper, I discuss the formal properties of *do-* only in its completive uses, i.e. the uses with the meaning 'to finish, to complete, to carry through', as in (1). I include in this type also uses with the meaning 'to complete after a break', as in (2):

(1) До начала первого писал и, наконец, *дописал* статейку для «Смены». [RNC]
'I had been writing a paper for "Smena" till 12 o'clock, and then I finished (*do-*+write) it at last.'

(2) Потом, потом *допишу*, говорил я себе. А теперь пусть подсохнут листки. [RNC]
'I said to myself: I'll finish writing (*do-*+write) later. And now the papers must dry up.'

There are some other productive uses of *do-*, which are semantically very close to the completive one. However, I exclude them, because their formal properties differ in some respect from those of the completive *do-* and must be described separately. These are:

1) spatial uses 'to reach the (spatial) Goal', as in *дойти* (*do-*+go), *долететь* (*do-*+fly), *добежать* (*do-*+run);

2) uses with the explicitly expressed endpoint (*до* 'to' + GEN), as in: *досидеть до обеда* 'to be sitting till the lunch-time (*do-*+sit)', *дожить до старости* 'to live to an old age (*do-*+live)', *догореть до середины* 'to burn half (*do-*+burn)';

3) additive uses 'V with additional portion of object / with a new object', as in: *долить бензину* 'to pour more petrol (*do-*+pour)', *дорисовать усы* 'to draw a moustache {= to add it to the existing portrait} (*do-*+draw)', *докупить продуктов* 'to buy more food' (*do-*+buy).

The borderline between the completive *do-* and the additive *do-* is especially relevant for the further discussion. In the majority of cases it is quite clear. Usually, the object of a completive *do-*verb is definite, while the object of an additive *do-*verb is indefinite (it follows directly from their semantics). Since Russian does not express definiteness overtly, the following substitution test can be used to differentiate between these two uses. The additive *do-* allows the object modifiers, which show explicitly that a new object (or a new portion of the object) is involved: *еще* 'more', *еще один* 'one more', *новый* 'new' etc., while the completive *do-* does not. Also, only the additive *do-* can take the object in the genitive case[2] instead of the accusative case. Cf.:

(3) a. *Допив воду / *воды*, он вдруг спросил… [RNC] — completive
'Having drunk up (*do-*+drink) the water (water.ACC / *water.GEN)'
b. Лучше всего *долить* в смесь *воду* /[OK]*воды*. [RNC] — additive
'The best option is to pour (*do-*+pour) some water ('water.ACC / [OK]water.GEN') to the mixture'

[2] Ø-stem masculine nouns take the special partitive case form *-u* in this context (*чаю* 'tea.PART', *сахар-у* 'sugar.PART').

However, there are uses, in which the distinction between the completive *do-* and the additive *do-* is more subtle. A problematic class is that of verbs with the incremental object (such as 'pour'). For these verbs, a temporal phase of the event (referred to by the completive *do-*) corresponds to the degree of involvement of the object (referred to by the additive *do-*). A special case within this class is verbs with the semantic incremental object, which is referred to by the prefix, but which is not overtly expressed in the syntactic structure. Cf. *досолить* (*do-*+salt) in (4). The verb *солить* 'to pour salt to sth' (and its English equivalent 'to salt') takes food names as the direct object (*солить суп* 'to salt the soup'), and its semantic object 'salt' is not expressed. Example (4) is ambiguous, since it is unclear, whether 'salt' is definite or not and whether we deal with a new portion of salt or with the full normal portion of salt.

(4) Попробовав с ложки еще раз, Кытин засомневался, подумал, *досолил* и помешал. [RNC]
'Kytin tried (the soup) again, hesitated, salted (the soup) and stirred it.' = 'added more salt' (additive) or 'made salty enough' (closer to completive)?

4. The choice between the perfective vs. imperfective stem

As mentioned above, an important feature of the completive prefix *do-* is its ability to attach both to perfective and imperfective stems. However, the general model of Russian verbal prefixation, observed in **Section 2**, does not predict the choice of a stem for any particular verb. In this section, I formulate general empirical-based rules of the stem choice. The following data were used: first, the sample of *do*-verbs extracted from Minor Academic Dictionary (678 items); second, the corpus data of RNC (www.ruscorpora.ru) and GICR (http://www.webcorpora.ru). The first (dictionary) data source provides information on more conventionalized derivates which belong to Standard Written Russian. The second (corpus) one provides additional information on the productive derivation in the real use, particularly on occasional derivates and on those rejected by the prescriptive norm, but attested in informal speech. Such derivates are especially in focus in the study. They are checked on the data of GICR (the search on blogs: livejournal.com and vk.com).

The rules of stem choice are different for different formal types of aspectual verb pairs. I will describe them separately.

A) Aspectual pairs "unprefixed imperfective + prefixed perfective"
(*писать* — *написать* 'write')

Within this type, the completive *do-* attaches consistently to the imperfective stem:

(5) a. ^{OK}*до-писать* (*do-*+write.IPFV) — $^{???}$*до-на-писать* (*do-*+PREF-write.PFV)
b. ^{OK}*до-шить* (*do-*+sew.IPFV) — $^{???}$*до-с-шить* (*do-*+PREF-sew.PFV)

Very few counter-examples are attested in written informal texts of blogs: 1–2 uses of *до-про-читать* (*do-*+PREF-read.PFV) и *донаписать* (*do-*+PREF-write.PFV) compared to more than 10,000 uses of *дочитать* (*do-*+read.IPFV) and *дописать* (*do-*+write.IPFV) in GICR:

(6) *Допрочитал*-таки The Probability of God… [GICR]
'In the end (I)'ve finished reading (*do*-+PREF-read.PFV) "The Probability of God"….'

(7) … а сегодня *донаписал* кое-как распознавание маджонговых фишек. [GICR]
'…and today, I've somehow finished writing (*do*-+PREF-write.PFV) the code for the recognition of mahjong tokens.'

B) Aspectual pairs "unprefixed perfective + unprefixed imperfective"
(*решить — решать* 'solve, decide')
Within this type, the completive *do*- also usually attaches to the imperfective stem:

(8) Игорь сдал работу и ушел, я *дорешал* свою за двадцать минут. [GICR]
'Igor handed in his exercise and went away, I finished solving (*do*-+solve.IPFV) my task in twenty minutes.'

Uses with the perfective stem are also attested in informal speech, though they are rare (9% of uses for the verb *решить* 'solve/decide.PFV', mostly with the meaning 'to decide')[3].

(9) Ну что такого, что я не смогла *дорешить* задачу? [GICR]
'What does it matter, that I failed to solve the task completely (*do*-+solve.IPFV)?'

C) Aspectual pairs "prefixed perfective + prefixed secondary imperfective"
(*переписать — переписывать* 'rewrite')
This is the most interesting type. For aspectual pairs, in which the prefixed imperfective is derived from the prefixed perfective by means of the imperfectivizing suffix *-yva*(*~-va-~a*), a variation is attested. For one and the same aspectual pair, both the perfective stem and the imperfective one can take the completive *do*-. Cf.:

(10) – *доуложить* (*do*-+put.to.bed.PFV) and *доукладывать* (*do*-+put.to.bed-IPFV)
a. …вот *доуложу* сейчас детей и узнаю кто убийца. (www.hv-info.ru)
'Now I will finish putting children to bed and find out who is a murderer.'

– *доуложить*
b. …ты уложи Андрея, а я тогда *доукладываю* Костю.
(https://www.babyblog.ru)
'Put to bed Andrej, and I will finish putting to bed Kostya.' — *доукладывать*

D) Other cases
There are also some marginal cases that do not fit in this picture. One of them is a case of "aspectual triplets". For some verbs, the unprefixed imperfective and the prefixed secondary imperfective are synonymous or quasi-synonymous, cf. *мести* > *подмести* > *подметать* ('to sweep'). In this case, all three stems can take *do*-:

[3] This is a small heterogeneous exceptional verb class. Only *решать — решить* was consistently checked on corpus data. It is possible that the quantitative data on other verbs differ from those on *решать — решить*.

до-мести ~ *до-под-мести* ~ *до-под-мет-а-ть*. Not only two prefixed stems, but also the initial unprefixed one is involved in competition[4].

One more exceptional case is a narrow class of verbs of attachment with spatial prefixes: *при-/в-/за-крутить*; *при-/в-/за-винтить*; *при-/в-/за-вернуть* 'to screw on/in'[5]. The unprefixed imperfective stem (*крутить* etc.) is not used with the same meaning. However, *до-* can attach not only to the prefixed stem, but also to this unprefixed imperfective stem, cf.:

(11) Ну не *до-вернули* гайку, ну выпил пивка или еще чего?
[RNC] = *до-за-вернули*, the unprefixed verb **вернуть* does not exist at all
'We have not screwed on the nut completely, I've drunk a bit of beer, what else?'

Table 1 contains the quantitative data on competing derivates for 2–3 verbs of each type. These are the data on informal Internet-communication taken from GICR.

The table shows, that a real competition is in fact attested only within Type C (prefixed perfective + prefixed secondary imperfective). The majority of such competing completives (both derived from the imperfective and perfective stem) are occasional. If one of them is conventionalized and does not contradict to the prescriptive norm, then it is usually the derivate from the perfective stem (*дорассказать*). However, derivates from the perfective stem are not obviously more frequent in informal speech (cf. *довышивать* with the opposite distribution). The frequency distribution of competing variants varies a lot across particular verbs. Using such a little sample, I cannot explain which features of a verb predispose to one or another distribution.

Table 1. Competing *do*-derivates from the perfective vs. imperfective stem: GICR[6]

verb type	*do*-+PFV vs. IPFV	verb	translation	+IPFV	+PFV	% of IPFV
type A	IPFV	*читал-прочитал*	'read'	26,770	1	100%
		писал-написал	'write'	9,192	2	100%
type B	IPFV (~PFV)	*решал-решил*	'solve'	71	7	91%

[4] A more difficult case is a triplet *чесать* > *при-чесать* > *при-чес-ыва-ть* 'to brush'. It also has three competing derivates: *до-при-чесать*, *до-при-чес-ыва-ть* and *до-чесать*. The unprefixed *чесать* had been used with the meaning 'to brush' till the beginning of the XX cent. (cf. *Марьянка в одной рубахе чесала косу, собираясь спать*. [Л. Н. Толстой. Казаки (1863), RNC]). However, in modern Standard Russian it is not a neutral synonym of *причесывать*. *Мести* is also archaic compared to *подметать* 'to sweep'. At the same time, the derivates *до-мести* и *до-чесать* do not seem to be archaic.

[5] These three verbs are conventionalized and mentioned in dictionaries. Cf. also rare occasional uses of the same class, which are outside the norm, but attested in informal Internet communication: *до-вязать шнурки*, *до-стегнуть крепление*, *до-крепить унитаз*.

[6] The search on blogs: vk.com (8,720 millions) and livejournal.com (9,820 millions): the forms PST.F and PST.M. All the results were looked through manually, only completive uses were counted (see above on the distinction between different meanings of *do-*).

verb type	*do*-+PFV vs. IPFV	verb	translation	+IPFV	+PFV	% of IPFV
type C	IPFV (~PFV)	*вышила-вышивала*	'embroider'	316	56	85%
		перечитал(а)-перечитывал(а)	'read again'	22	65	25%
		переписал(а)-переписывал(а)	'write again'	3	10	23%
		рассказал(а)-рассказывал(а)	'tell'	10	202	5%

Competing completives can be used in absolutely identical contexts, (12):

(12) …потому что вензель на наволочке не успела *довышивать* :) Ну что, *довышила* и поехала! [GICR]
'… because I didn't have time to finish embroiding (*do*-+IPFV) a monogram on the pillow-case. Well, I finished it (*do*-+PFV) and went away.'

However, the following non-strict tendencies in their distribution are attested.

a) The presence of the corresponding perfective vs. imperfective base verb may predispose to the choice of stem (*укладывал, но так и не доукладывал*; *нужно уложить, вернее доуложить ребенка*).

b) There is a correlation with the frequency of a secondary imperfective derived from the completive verb, which is homonymous to the derivate from the secondary imperfective (*дорассказывать* = *do*-+tell.IPFV 'to finish telling' vs. *do*-tell+-IPFV 'to be finishing telling'), see **Section 5**.

c) The main factor is which component of the event is in focus. If the temporal semantics is in focus, then the derivate from the imperfective stem is more likely. If the argument semantics (the change/involvement of the object in the course of the event) is in focus, then the derivate from the imperfective stem is more likely. Cf.:

(13) *Контрольные допроверяла*!!! Урра!!! Еще чуть-чуть приблизилась к концу семестра. [GICR] — the imperfective stem
'I've finished checking the test. Hurrah! Now, I'm a bit closer to the end of the semester!' (the temporal semantics is more in focus[7])

(14) Доделала, *допроверила*, подправила и выслала клиенту итог двухнедельного труда — 5 свеженаписанных нарядных отчета, каждый на двух языках… [GICR] — the perfective stem
'I finished, checked and sent to the client the result of my work of the last two weeks, 5 just written accurate reports, each in two languages.' (the argument semantics is more in focus[8])

[7] The temporal interpretation is supported by the next sentence, in which the speaker refers to the time period.

[8] The speaker focuses not on the time period her work took, but rather on the result of her work, namely on the positive changes in her reports, cf. the other verbs of changing in the chain: *доделала*, *подправила*.

d) One more argument for the previous point is the fact that the additive *do-*, which is closer to uses that focus the object, than to uses that focus the temporal semantics (see Section 3 above on the semantics of additive uses), are also likely to chose the perfective stem. Cf.:

(15) a. — Ну, эти крупные планы мы *доснимем* в павильоне в Москве.
[RNC] — additive & the perfective stem
'We will make photo of these close-ups in the studio in Moscow.'
b. Завтра *доснимаю* и выложу. — completive & the imperfective stem
'Tomorrow, I will finish making photos and publish them'.

It is interesting, that the rules regulating the stem choice for *do-* are quite different from those regulating it for *pere-* (which is expected to belong to the same formal type), cf. for the rules for *pere-* [Stoynova 2014].

5. The prefix *do-* and secondary imperfectivization

One more point that requires empirical verification is the position of *do-* with respect to the suffix of secondary imperfectivization *-yva*(*~-va-~-a*).

Completive *do*-verbs, which themselves are perfective, attach the suffix *-yva* without any restrictions, cf.:

– imperfective: *мыть* (wash.IPFV) > perfective: *до-мыть* (*do*-wash) > secondary imperfective: *до-мы-ва-ть* (*do*-wash-IPFV)

(16) Разделся в прихожей и прошел на кухню — там престарелая домработница Клава *домывала* посуду после ужина. [RNC]
'(I) took off my clothes and entered the kitchen — there the old housemaid Klava was finishing washing (*do*-wash-IPFV) dishes after dinner.'

On the contrary, the attachment of the completive *do-* itself to secondary imperfectives is claimed to be forbidden. [Tatevosov 2008] considers this feature as one of the arguments to attribute this prefix to a separate intermediate formal type along with *pere-* and *pod-* (see **Section 2** above).

In fact, for *pere-* and *pod-* this restriction takes place:

– perfective: *записать* (write.down) > secondary imperfective: *запис-ыва-ть* (write.down-IPFV) > *perfective with *pere-*: **пере-*[*запис-ыва*]*-ть* (*pere*-write.down-IPFV 'to write down again')[9]

The case of *do-* is more complicated. Derivates from secondary imperfectives are not attested in the sample extracted from Minor Academic Dictionary. However, they are actively used in informal speech. Cf. the verbs *довышивать*, *доукладывать*, *до-рассказывать* and others, mentioned and exemplified in **Section 4**. Moreover, for

[9] The homonymous secondary imperfective derived from the perfective verb with *pere-* is attested: *пере-записать* 'to write again' > [*пере-запис*]*-ыва-ть* 'to be writing again'.

some verbs these variants are more frequent, than the expected derivates from the perfective stem (see **Table 1** above). Cf. one more example:

– perfective: *перечитать* 'to read again' > secondary imperfective: *перечит-ыва-ть* (read.again-IPFV) > *до-перечитывать do*-read.again-IPFV 'to finish re-reading', (17):

(17) …надеюсь *доперечитывать* и отправить с приведением любимых фрагментов. [GICR]
'I hope to finish re-reading (*do*-read.again-IPFV) (it) and to send (it) with my favorite fragments marked.'

As both derivational scenarios: *do*-V > [*do*-V]-*yva* and [V-*yva*] > *do*-[V-*yva*] are available, homophonous verbs with different structures and aspectual interpretations are imaginable. And they are in fact attested, cf. *до-рассказ-ыва-ть* (*do*-tell-IPFV) in (18) and (19):

(18) Вы не дорассказали! Зеро. *Дорассказываю*. Однажды… [RNC] — *до-рассказ-ыва-ть* is the secondary imperfective derived from *дорассказать* (*do*-tell)[10]
'— You haven't finished your story! Zero: I'm telling (*do*-tell-IPFV): Once…'

(19) *Я же ещё не дорассказывал* сказку=((risovach.ru) — *до-рассказывать* is the perfective completive derivate from *рассказ-ыва-ть* (tell-IPFV)[11]
'But I haven't finished (*do*-tell-IPFV) my tale yet'.

6. Conclusion

Thus, the empirical data involved in the study, and especially the data of informal speech which lies beyond the prescriptive norm, give the possibility to verify and enlarge the existing assumptions on the completive *do*- and on the whole system of Russian prefixes.

1) The prefix *do*- is predicted to attach both to perfective and imperfective stems. The empirical data confirm this prediction and give a possibility to formulate the rules regulating the stem choice, which do not follow from the general theoretical assumptions. The rules are complex, a large pool of variation takes place.

---

[10] The form *дорассказываю* has the progressive performative interpretation in this sentence, so this is the imperfective verb, derived from *дорассказать*. If it was the perfective derivate from *рассказывать*, it would have the future time reference, which is very unprobable in this context (cf. the following sentence with unprefixed verbs: *Рассказываю* (IPFV). *Однажды…. / ???Расскажу* (PFV). *Однажды…*).

[11] According to the general rule of interpretation of perfective vs. imperfective verbs in the not-yet context, the imperfective derivate from *дорассказать*, would have the meaning '(I) have not begun to finish telling', and not the meaning 'I (began) and have not finished telling' expected from the broader context , cf. the following contrastive pair with unprefixed verbs: *Я же еще не рассказывал сказку* ('I have not even begun to tell the tale') */ не рассказал сказку* ('I have not finished the tale yet').

2) The prefix *do-* is predicted to attach below the secondary imperfective. It is true for more conventionalized uses. However, numerous occasional uses in modern informal speech totally break this prediction.

3) The small class of "intermediate prefixes" appears to be heterogeneous: there are some features that distinguish *do-* from *pere-*, including the interaction with the secondary imperfective and the rules of stem choice.

4) So, the detailed empirically-based study does not create considerable problems for the existing theory of Russian prefixation. However, more new data involved — more detailed becomes the classification, up to one prefix classes.

References

1. *Babko-Malaya O.* (1999), Zero Morphology: A Study of Aspect, Argument Structure, and Case, Ph.D. dissertation, Rutgers University.
2. *Isatchenko A. V.* (1965/2003), Grammatical structure of Russian compared to Slovac. Morphology [Grammatičeskij stroj russkogo jazyka v sopostavlenii s slovackim. Morfologija], I-II, Moscow, JaSK.
3. *Ramchand G.* (2004), Time and the event: The semantics of Russian prefixes, Nordlyd, 32(2). Special issue on Slavic prefixes, 2004. pp. 323–366.
4. *Romanova E.* (2004), Superlexical vs. lexical prefixes, Nordlyd, 32(2). Special issue on Slavic prefixes, pp. 255–278.
5. *Romanova E.* (2006), Constructing Perfectivity in Russian, Ph.D. dissertation, University of Tromsø.
6. *Stoynova N.* (2014), Repetitive *pere-* in Russian: single vs. multiple prefixation competing, 47th Annual Meeting of the Societas Linguistica Europaea. Book of abstracts, available at: http://sle2014.eu/downloads/SLE2014BookofAbstracts_FINAL_2.pdf.
7. *Svenonius P.* (2004), Slavic prefixes inside and outside VP, Nordlyd. Special issue on Slavic prefixes, 32, pp. 205–253.
8. *Tatevosov S. G.* (2008), Intermediate prefixes in Russian, Antonenko A., Bethin C., Baylin J. (eds.), Formal approaches to Slavic linguistics, New York, Ann Arbor, MI: University of Michigan Press, pp. 423–442.
9. *Tatevosov S. G.* (2009), Multiple prefixation and anatomy of Russian verb [Množestvennaja prefiksacija i anatomija russkogo glagola], Kiseleva et al. (ed.), Corpus-based studies on Russian grammar [Korpusnyje issledovanija po russkoj grammatike], Moscow, Probel-2000.
10. *Tatevosov S. G.* (2013), Multiple prefixation and its outcome [Monožestvennaja prefiksacija i ee sledstvija], Voprosy Jazykoznanija, 3, pp. 42–89.

LANGUAGE MODELS FOR UNSUPERVISED ACQUISITION OF MEDICAL KNOWLEDGE FROM NATURAL LANGUAGE TEXTS: APPLICATION FOR DIAGNOSIS PREDICTION

Tarasov D. (dtarasov3@gmail.com),
Matveeva T., **Galiullina N.**
Meanotek, Kazan, Russia

Following recent success of neural language models in various downstream language understanding tasks, including common sense reasoning, we investigate possible utility of such models in domain specific reasoning task—proposing of preliminary diagnosis based on patient complains, presented as natural language text. We demonstrate that language model, trained on the texts collected from online medical forums posses significant accuracy in this task (73% at top 10 suggestions), when evaluated on dataset, constructed from clinical case reports, published in specialized medical journals. While preliminary, these findings indicate a possible new method that can be used to augment online symptoms checkers and clinical decision support systems.

Keywords: symptom checkers, neural language model, medical diagnosis

МОДЕЛИ ЯЗЫКА ДЛЯ ИЗВЛЕЧЕНИЯ МЕДИЦИНСКИЙ ЗНАНИЙ ИЗ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ С ЦЕЛЬЮ ПРИМЕНЕНИЯ В ЗАДАЧАХ ПРЕДСКАЗАНИЯ ДИАГНОЗА

Тарасов Д. (dtarasov3@gmail.com),
Матвеева Т., **Галиуллина Н.**
ООО «Меанотек», Казань, Россия

В связи с недавним успехом нейронных моделей языка в решении различных задач понимания естественного языка, включая рассуждения на основе здравого смысла, мы исследуем возможную полезность таких моделей в задаче рассуждений, специфичных для предметной области—предварительной медицинской диагностики на основе жалоб пациента, представленных в виде текста на естественном языке. Мы демонстрируем, что языковая модель, обученная на текстах,

собранных на медицинских форумах в Интернете, обладает значительной точностью в выполнении этой задачи (73 % из 10 лучших предположений) при оценке по набору данных, построенному на основе отчетов о клинических случаях, опубликованных в специализированных медицинских журналах. Эти результаты указывают на возможный новый метод, который можно использовать для расширения возможностей онлайн-проверки симптомов и систем поддержки принятия клинических решений.

Ключевые слова: нейронная модель языка, медицинская диагностика, системы поддержки принятия клинических решений

1. Introduction

In this paper we consider the task of proposing preliminary diagnosis based on patient complaints, presented as text in natural language (Russian). Such systems are useful for their users, because they can give them better understanding on possible causes of their symptoms, as well as provide advice on the most appropriate point of care. Currently, English-based symptom checkers have insufficient accuracy. According to recent studies [Semigran et al, 2015], most systems provide correct diagnosis first in 34% of cases, while average accuracy in top-20 (correct diagnosis within the top 20 diagnoses given) is 58%.

While new tests were published [Razzaki et al, 2018] recently for Babylon Triage and Diagnostic System, claiming near-human expert accuracy, these results are still below that of best human doctors and concerns were raised about validity of published results, due to methodological flaws [Fraser et al, 2018].

In recent years, approaches, based on deep learning gained popularity in the field of medical diagnostics. While majority of applications of deep learning are in the area of medical image processing tasks, end-to-end language processing diagnostic approaches were also proposed. For example Deep Patient [Miotto et al, 2016] system uses auto-encoders to train representation of patient medical history and then predicts probability for a given patient to develop a new disease.

However, deep learning based models require huge amounts of training data to reach good performance, and for the task, considered in this paper, such datasets are very difficult to collect. Ideally, one want a dataset of 100,000 or more patient complaints matched with clinically confirmed diagnosis. Such dataset is hard to create because of privacy issues and even if these can be solved by anonimization, hospitals rarely have full descriptions of patient conditions in their own words (as opposed to description made by physician).

It was suggested that big language models can learn useful knowledge from text in completely unsupervised manner. In particular, [Trinh et al, 2018] demonstrated that language models can obtain higher accuracy on Winograd Schema Challenge [Levesque et al, 2012] then competing methods. It is tempting to assume that such models can also learn domain specific knowledge.

In this paper we propose a method that learns diagnosis classification task using data found on online medical forums, where people discuss their health problems.

Such data is available in abundance in Russian language and in many other languages. Based on this data, this paper makes following contributions:

1. We develop term extraction model, that extracts mentions of diagnoses and symptoms from online forum postings in Russian.
2. Using this model we construct Knowledge base (KB) based on co-occurrence of diagnosis and symptoms in online forums
3. We train large language model on medical forums text collection
4. We compare co-occurrence based and new proposed language-model based approaches to diagnosis prediction. Our findings indicate that language model-based diagnosis prediction is superior to co-occurence baseline.
5. We demonstrate that diagnosis prediction from patient complaints, presented as text in natural language can have accuracy comparable to current symptom checkers based on series of multiple-choice questions

2. Related work

Symptom checkers usually use large hand-crafted knowledge bases (KB) of medical facts, than need constant revision [Blum, et al, 1991]. Semi-automatic methods for construction of such KBs were proposed [Ramnarayan et al, 2016], [Middleton et al, 2016], that are based on co-occurrence statistics in PubMed, Wikipedia, and Electronical Medical Records data. Drawbacks of these methods are requirements for complex NLP pipelines to interpret data, and limited ability of co-occurrence statistics to capture disease-symptom relationship, such as timing of symptom appearance. During their operation, most symptom checkers use multiple choice lists to collect symptoms, or allow to enter symptoms in natural language, but only one symptom in time [Kafle et al, 2018]. And even with that restriction, to achieve free-form symptoms interpretation capability, complex paraphrase generation model are employed to generate large number of paraphrases and then do look-up.

Such restrictions in data representation limit the extent to which symptom checker can leverage information that user provides about specific circumstances of the user and about disease development process. As a result, existing approaches only achieve accuracy around 59% at top 10 diagnosis suggestions [Kafle et al, 2018] and can not work with free-form descriptions of users conditions, that are typically found in online forums.

Few authors attempted to use deep learning for solving natural language based disease diagnosis problem. Current research is mostly focused on clinical question answering [Hasan et al, 2016] and medical data mining [Barnickel et al, 2009]; [Mallory et al, 2015].

3. Methods and algorithms

3.1. Datasets and data annotation

3.1.1. Online posting forums data

We used internal dataset of forum posts, collected from Russian online medical forums during September 2018. Dataset contains descriptions of medical complaints, discussions and advices from doctors on variety of medical topics. Total number of forum posts in the dataset is 30,756 (68 MB of UTF-8 encoded text).

We annotated mentions of diagnoses, symptoms and body parts in first 200 posts (60,000 words) from this dataset, and made separate training and test set (45,000 words training and 15,000 words test).

3.1.2. Diagnosis prediction test set

We collected 50 case reports published in Russian medical journals, for 50 different medical conditions randomly selected from list of diagnoses found in online forums. Each case report has final diagnosis verified by careful medical evaluation. For dataset construction, for each case report, patient complaints, described in that case report were rewritten using informal language by the person without special medical knowledge (as if written from patient perspective). In this way, we are trying to avoid systematic bias of having test set prepared by doctors (rather then lay persons for whom system is intended), which was one of important methodological issues with previous evaluation methodologies [Fraser et al, 2018]. In the same way, we avoid possible bias that can be introduced by having fictional cases that generally fit diagnosis criteria, but not based on real data, because we use data from real cases with clinically confirmed diagnosis. While larger test set is desirable, it is very labor-intensive to construct, and similar studies has used dataset of comparable size before [Semigran, 2015]; [Kafle et al, 2018], thus we consider it to be acceptable for preliminary studies.

3.2. Language model

Language modeling (LM) is one of the important tasks of natural language processing. The task involves predicting the $(n+1)$th token in a sequence given the n preceding tokens, where tokens can be words, subwords or characters. More formally, the goal of a language model is to estimate a distribution $P(x_{0:T})$ over sequences of tokens $(x_0, x_1, ..., x_T)$.

The joint distribution over long text spans can then be represented as a product of the predictive distribution over tokens conditioned on the preceding tokens:

$$P(x_{0:T}) = \prod_{t=0}^{T} P(x_t \vee x_{0:t-1}) \quad (1)$$

Neural language models [Sutskever et al, 2011] usually use recurrent neural networks (RNN) for sequence modeling. Given a sequence of vectors $\{x(t)\}$, where $t=1..T$, an RNN computes memory and output sequences:

$$h(t) = f(Wx(t) + Vh(t-1) + b) \tag{2}$$

$$y(t) = g(Uh(t) + c) \tag{3}$$

where f is a nonlinear function, such as the sigmoid or hyperbolic tangent function and g is the output function. W and V are weight matrices between the input and hidden layer, and between the hidden units. U is the output weight matrix, b and c are bias vectors connected to hidden and output units. $h(0)$ in equation (1) can be set to constant value that is chosen arbitrary or trained by backpropagation.

Recently, it was shown that by learning to predict the next character given previous characters, neural network based language models can learn internal representations that capture syntactic and semantic properties [Radford et al, 2017].

We use Long Short Term Memory (LSTM) [Hochreiter et al, 1997] based neural network. The structure of the LSTM [9] allows it to train on problems with long term dependencies. In LSTM simple activation function f from above is replaced with composite LSTM activation function. Each LSTM hidden unit is augmented with a state variable $s(t)$ The hidden layer activations correspond to the 'memory cells' scaled by the activations of the 'output gates' o and computed in following way:

$$h(t) = o(t) \times f(c(t)) \tag{4}$$

$$c(t) = d(t) \times (c(t-1) + i(t)) \times f(Wx(t) + Vh(t-1) + b) \tag{5}$$

where $\times$ denotes element-wise multiplication, $d(t)$ is dynamic activation function that scales state by "forget gate" and $i(t)$ is activation of input gate.

We train LSTM-based character level language model with 3 hidden layers, with 3,192 LSTM cells per each layer. Given that it is hard to capture all basic language structure with relatively small dataset, we pre-trained our model on subset of Russian Wikipedia, containing 2 billion characters. Model was trained by using truncated backpropogation thought time with learning rate controlled by Adam [Kingma et al, 2014] algorithm. We halted training by tracking the performance on the validation set, stopping when negligible gains were observed. Then, we use trained weights to initialize new model, that was trained on medical forums texts.

3.3. Diagnoses and symptoms extraction

We trained the sequence tagging model using mini-batch gradient descent with one sentence per mini-batch. We used simple learning rate annealing method in which we multiple the learning rate by 0.85 if test loss does not fall for 2 consecutive epochs. By performing model selection on separate development set, we found optimal number of hidden units per layer of the LSTM to be 128, and the number of LSTM layers to be 2.

We used two different sets of input features—word embeddings trained over forum texts using word2vec algorithm [Mikolov et al, 2013] and activations of last layer of LSTM language model.

3.4. Co-occurrence based diagnosis prediction

We first extracted all diagnoses from forum data using trained extraction model. We then manually assigned ICD-10 (International Statistical Classification of Diseases and Related Health Problems, revision 10) codes to each unique term extracted. We then took top 200 most frequent diagnoses and calculated co-occurrence table with symptoms, where symptoms were listed as mentioned in the text, without normalization.

To predict diagnosis for a new text, we first extract all mentioned symptoms and then find top 40 most similar entries in symptoms/diagnoses co-occurrence table, using cosine similarity between symptoms embeddings, obtained by summing embeddings of each word for a given symptom. Each diagnosis in the list was scored according to the number of matched symptoms and extend to which individual symptoms were similar. We use this method as baseline to compare against language model based method.

3.5. Language model based diagnosis prediction

Following previously proposed approaches for common-sense reasoning [Trinh et al, 2018], we concatenate full description of person's condition with diagnosis and compute joint probability of resulting text using trained language model.

3.6. Evaluation metrics

For measuring quality of term extraction models we use F-measure, computed using *Proportional Overlap*—a metric that imparts a partial correctness, proportional to the overlapping amount, for each match [Irsoy and Cardie, 2014].

For measure of diagnostic accuracy, our main outcomes were whether the system listed the correct diagnosis first or within the first 10 of potential diagnoses.This metric sometimes defined as *diagnosis recall at top N* [Middleton et al, 2016], while other authors use the term "*diagnostic accuracy at top N*" [Semigran et al, 2015], [Kafle et al, 2018]. We will use the term diagnostic accuracy here. The choice of metric is dictated by the need to compare our results to others and the fact that text descriptions alone do not provide enough information to exclude all possible causes but one, so we are interested to measure the ability of system to successfully narrow list of possible causes to a few possibilities.

4. Results and discussion

4.1. Extraction of diagnoses and symptoms mentions

After training terms extraction model, we obtained results, presented in Table 1.

Table 1. Term extraction accuracy

Input features type	F1, diagnosis	F1, symptom
Word2vec trained on forum data	0.55	0.65
Word2vec trained on Wikipedia	0.51	0.58
Activations of language model top layer (Wikipedia)	0.50	0.59
Activations of language model top layer (pre-trained on Wikipedia, fine-tuned on forum posts)	**0.57**	**0.68**

We observe that using language model contextual embeddings improves term extraction accuracy, although these improvements are relatively minor and only present when model is fine-tuned on in-domain texts. We found that term extraction models that use language model features generally have high precision (0.75 for diagnosis) compared to models that use skip-gram embeddings (0.62), which makes language model features based models more suitable for construction of co-occurence tables.

4.2. Accuracy of diagnosis prediction

Results of diagnosis prediction on test set are presented in Table 2. We found that accuracy of language-model based method is superior to that of simple co-occurrence baseline, suggesting that language model is capable of leverage additional information contained in descriptions of patients conditions, that is not present in co-occurrence statistics and skip-gram based word embeddings.

Another surprising finding is that accuracy that we obtained from such a noisy dataset is generally high and comparable to that of much more complex systems [Kafle et al, 2018], even through our system operates directly on natural language descriptions and is not allowed to ask additional questions to the user.

We also found that model trained on Wikipedia alone does not have good performance, and using forum posts alone also leads to low accuracy, while fine-tuning Wikipedia model on forum posts leads to superior performance. This could be due to the fact that random subsample of Wikipedia contains very few medical facts (it mostly contains history, sports, and media topics) but is helpful for acquiring representations of basic natural language structure.

It is worth noting, that are goal here is not to achieve better accuracy *per se*, but to establish if unsupervised learning based on language model can obtain knowledge useful for the task of diagnosis prediction. While supervised methods may well be capable of obtaining better accuracy given proper training set, our focus here is primary on unsupervised learning.

Table 2. Accuracy of diagnosis prediction

Method	Diagnostic accuracy @ top1	Diagnostic accuracy @ top10
Co-occurrence + similarity of symptoms	22%	58%
Language model (trained on wikipedia)	1%	5%
Language model (trained on forum posts)	10%	45%
Language model (pre-trained on Wikipedia, fine-tuned on forum posts)	**33%**	**73%**

4.3. Analysis of individual cases and failure modes

In this sections we examine sample results from the system on 3 test cases and analyze possible causes of failures.

Case 1

In the first case the following description was presented to the system:

«У меня такая ситуация. Дикие боли в эпигастральной области, больше слева, от поджелудочной вниз к кишечнику. Отрыжка, метеоризм, запоры; во время приступов—расстройства желудка. Боли до еды и после(с тяжестью)»
(approximate English translation: *"I have this situation. Wild pain in the Epgaastrin area, more left, from the pancreas down to the intestines. belching, flatulence, constipation; During the attacks-indigestion. Pain before and after eating (with weight)"*)

Correct diagnosis in this case was the diagnosis of "gastritis", which coincides with the first diagnosis proposed by the system. However, it should be noted that the diagnosis of gastritis in this case is not particularly difficult. Among all the suggestions in top 10 plausible options were "colitis", "pancreatitis", "reflux disease". However, system also suggested improbable diagnoses, such as "thyroiditis".

Case 2

«першение в горле, усиление кашля с небольшим количеством мокроты сероватого цвета, повышение температуры тела до 37,8 °С, потливость. В течение примерно 25 лет беспокоит кашель, преимущественно по утрам, с небольшим количеством мокроты»
(*"Sore throat, increased cough with a small amount of sputum grayish color, increased body temperature up to 37,8°C, sweating. For about 25 years, worries cough, mostly in the mornings, with a little sputum"*)

The description for this case was compiled on the basis of an article in the Medical Journal (Internal Medicine Archive No. 2 (22) 2015-"Clinical case of tuberculosis development under the mask of exacerbation of chronic bronchitis"). As the title of the paper suggests, the reference diagnosis in this case was tuberculosis. We consider the

answer of the system in this case to be correct, because the diagnosis of tuberculosis is present in the top 10 suggestions, despite the fact that it is in the ninth place, after bronchitis, common cold, and other respiratory diseases. According to the paper, the diagnosis of tuberculosis in this case poses difficulties, and it was not established by a physician, initially despite the fact that he had the opportunity to accurately examine the patient and conduct laboratory tests, while the system relies solely on a short description.

Case 3

«Потеря сознания, продолжавшаяся 4 мин, сопровождавшаяся судорогами, которые длились около 4 с, затем утихали и потом снова появлялись еще 1–2 раза, а затем исчезали, заведением глазных яблок вверх, слюноотделением, прикусыванием языка, постприступной сонливостью, повышением температуры тела до 37,1°С. История развития настоящего заболевания. Первый приступ произошел 26 августа 2013 г. без видимой причины. Больной в это время отдыхал на море, загорал на пляже. Со слов его матери, приступ длился 1 мин: отмечалось заведение глазных яблок вверх, «потрясывание» всего тела, сначала ног, затем рук, через несколько секунд появилась пена изо рта, немного обмочился. Сам пациент не помнит приступ»
(*"The loss of consciousness lasted 4 minutes, accompanied by convulsions, which lasted about 4 s, then calmed down and then reappeared again 1–2 times, and then disappeared, the establishment of eyeballs up, salivation, the bite of the tongue, drowsiness, increase of body temperature up to 37,1°C. History of development of the present disease. The first attack occurred on August 26, 2013 for no apparent reason. The patient at this time rested on the sea, sunbathing on the beach. From the words of his mother, the attack lasted 1 minute: It was noted the establishment of eyeballs up, "shaking" the whole body, first legs, then hands, after a few seconds appeared foam from the mouth, a little wet. The patient himself does not remember the attack"*)

This description is also an adaptation of the description of the clinical case of epilepsy. Despite the fact that in this case the assumption of the diagnosis of "epilepsy" is not particularly difficult for a human doctor, the correct diagnosis was not in to 10 suggestions. Among the suggestions in this case were: "hypothermia", "arrhythmia", "mitral valve prolapse", "anxiety disorder" and "depression". It is noteworthy that this description is long and poorly adapted, as it contains the text of the description of the doctor rather than the patient. As an experiment, we introduced only a part of this description beginning with the fragment "from the words of his mother" In this case correct diagnosis was obtained on 2nd place, following allergic reaction (anaphylaxis). We hypothesize, that language model may have difficulties in processing too long texts where it is hard to select relevant symptoms, which can be mitigated in the future by using attention-based language model.

5. Conclusions

1. We found that texts, posted in online medical forums can be valuable source of data for training symptoms checkers (diagnosis prediction models), despite their noisy content.
2. Language-model based systems, trained on online medical forums posts, have considerable (73% at top-10) accuracy in detecting correct diagnosis based on user's natural language description of medical condition. This accuracy exceeds that of simple co-occurence based baseline model and possibly approaches accuracy of more complex symptom checkers, for chronic conditions, while still being inferior in diagnosing acute medical emergencies.
3. In line with previous findings, language-model features in form of activations of LSTM top layer hidden units improve medical term extraction accuracy, albeit to a small extent.
4. In summary, our findings, while being preliminary, seems to indicate that large language models can acquire significant domain-specific knowledge, possibly pointing to a completely new way for improving existing diagnostic systems.

References

1. *Barnickel, Thorsten, et al.* (2009). "Large scale application of neural network based semantic role labeling for automated relation extraction from biomedical texts." PLoS One 4.7: e6393.
2. *Blum, B. I., & Semmel, R. D.* (1991, May). Medical informatics, knowledge, and expert systems. In [1991] Computer-Based Medical Systems@ m_Proceedings of the Fourth Annual IEEE Symposium (pp. 212–218). IEEE.
3. *Fraser, H., Coiera, E., & Wong, D.* (2018). Safety of patient-facing digital symptom checkers. The Lancet, 392(10161), 2263–2264.
4. *Hasan, S. A., Zhao, S., Datla, V. V., Liu, J., Lee, K., Qadir, A., & Farri, O.* (2016). Clinical Question Answering using Key-Value Memory Networks and Knowledge Graph. In TREC.
5. *Hochreiter, S., & Schmidhuber, J.* (1997). Long short-term memory. Neural computation, 9(8), 1735–1780
6. *Irsoy, O., & Cardie, C.* (2014). Opinion mining with deep recurrent neural networks. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 720–728).
7. *Kafle, S., Pan, P., Torkamani, A., Halley, S., Powers, J., & Kardes, H.* (2018). Personalized symptom checker using medical claims. In Proceedings of the Third International Workshop on Health Recommender Systemsco located with Twelfth ACM Conference on Recommender Systems (HealthRec-Sys'18), Vancouver, BC, Canada, October 6, 2018, 5 page.
8. *Kingma, Diederik P., and Jimmy Ba* (2014). "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980.

9. *Levesque, H., Davis, E., & Morgenstern, L.* (2012). The winograd schema challenge. In Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning.
10. *Mallory, Emily K., et al.* (2015). "Large-scale extraction of gene interactions from full text literature using DeepDive." Bioinformatics: btv476.
11. *Middleton, K., Butt, M., Hammerla, N., Hamblin, S., Mehta, K., & Parsa, A.* (2016). Sorting out symptoms: design and evaluation of the babylon check automated triage system. arXiv preprint arXiv:1606.02041.
12. *Mikolov, T., Chen, K., Corrado, G., & Dean, J.* (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
13. *Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T.* (2016). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Scientific reports, 6, 26094.
14. *Radford, A., Jozefowicz, R., & Sutskever, I.* (2017). Learning to generate reviews and discovering sentiment. arXiv preprint arXiv:1704.01444.
15. *Ramnarayan, P., Kulkarni, G., Tomlinson, A., & Britto, J.* (2004). ISABEL: a novel Internet-delivered clinical decision support system. Current perspectives in healthcare computing, 245–256.
16. *Razzaki, S., Baker, A., Perov, Y., Middleton, K., Baxter, J., Mullarkey, D. & DoRosario, A.* (2018). A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis. arXiv preprint arXiv:1806.10698
17. *Semigran, Hannah L., et al.* (2015) "Evaluation of symptom checkers for self diagnosis and triage: audit study." bmj 351: h3480.
18. *Sutskever, I., Martens, J., & Hinton, G. E.* (2011). Generating text with recurrent neural networks. In Proceedings of the 28th International Conference on Machine Learning (ICML-11) (pp. 1017–1024).
19. *Trinh, T. H., & Le, Q. V.* (2018). A simple method for commonsense reasoning. arXiv preprint arXiv:1806.02847.

ASSESSING THEME ADHERENCE IN STUDENT THESIS

Tikhomirov M. M. (tikhomirov.mm@gmail.com),
Loukachevitch N. V. (louk_nat@mail.ru),
Dobrov B. V. (dobrov_bv@mail.ru)
Lomonosov Moscow State University, Moscow, Russia

In this paper we study approaches to assessing the quality of student theses in pedagogics. We consider a specific subtask in thesis scoring of estimating its adherence to the thesis's theme. The special document (theme header) comprising the theme, aim, object, tasks of the thesis is formed. The theme adherence is calculated as the similarity value between the theme header and thesis segments. For evaluation we order theses in the increased value of the calculated theme adherence and compare the ordering with expert grades using the average precision measure. The best configuration for theses ranking is based on the weighted averaged sum of word embeddings (word2vec) and keywords extracted from the theme header.

Key words: Thesis assesment, embeddings, cosine similarity, ontology

ОЦЕНКА СООТВЕТСТВИЯ ТЕМЫ И ТЕКСТА В СТУДЕНЧЕСКОЙ ВЫПУСКНОЙ РАБОТЕ

Тихомиров М. М. (tikhomirov.mm@gmail.com),
Лукашевич Н. В. (louk_nat@mail.ru),
Добров Б. В. (dobrov_bv@mail.ru)
МГУ им. М. В. Ломоносова, Москва, Россия

В данной статье изучаются подходы к оценке качества студенческих выпускных работ по педагогике. Рассматривается подзадача определения соответствия текста теме работы. Для этого формируется специальный документ (тематический заголовок), содержащий тему, цель, объект, задачи работы. Соответствие теме рассчитывается как значение сходства между тематическим заголовком и сегментами работы. Для оценивания мы упорядочиваем работы по возрастанию значений соответствия теме и сравниваем полученный порядок с оценками экспертов, используя меру средней точности.

Ключевые слова: оценивание выпускных квалификационных работ, дистрибутивное представление слов, косинусная мера близости, онтология

1. Introduction

Currently, proper assessment of a student thesis (bachelor or magister) can be a very difficult task because of availability of various informational resources in Internet, which can be plagiarized by a student. It can seem that a thesis is well-done, but in fact the share of student own work is minimal. In order to detect borrowings in the thesis texts, so-called plagiarism detection systems (antiplagiat.ru, etxt.ru) have become widespread, which allow determining the percentage of borrowings either on the basis of their own source database or by analyzing the global search engine results (Yandex, Google) [7].

However, the quality requirements to a student thesis are not limited to plagiarism restrictions. The requirements include such important characteristics of a work as the theoretical and practical significance, elements of novelty in the work, knowledge of the modern literature on the research topic, consistency in the presentation of the material, the scientific style of presentation, and others. Checking these requirements could be automated in order to provide an expert with data on different characteristics of student works. The task of automated assessment can be compared to such a known task as automatic essay scoring [5], [6], [14], when student essays to specific topics should be assessed. But also, there are significant distinctions between thesis assessment and essay scoring tasks.

One of important characteristics of a student thesis is its relatedness to the thesis theme. The proclaimed theme is usually concretized in the following terms: aim of the study, object of the study, and the tasks of the work. It is possible to gather all these information into so-called theme header. It is usually supposed that a student should develop the theme and its details in the presented work. So, there is a subtask of thesis scoring to assess its adherence to the theme header. In the essay scoring, this subtask corresponds to the prompt relatedness subtask [6], [11].

In this paper, we study approaches to determining the relatedness between the theme header and student thesis in pedagogics. To evaluate the methods, we have the collection of 40 thousand student theses, 120 student theses among them have double expert scores. The aim of the assessment is as follows: if low relatedness is detected, then the problems should be visualized to experts and some penalties to the overall score for this work should be proposed by the system. We use several means for assessing relatedness including word embeddings and a thesaurus providing knowledge about domain term relations. As a thesaurus, we use Ontology on Natural Sciences and Technologies [4], where the pedagogics domain terms have been introduced.

We consider theme adherence as one of factors needed to be calculated for comprehensive assessment of student thesis. Also thesis fragments that found irrelevant to the thesis theme are considered as good candidates for plagiarism analysis.

2. Related Works

For assessing the quality of scientific papers, Osipov et al. [10] discuss such characteristics as the presence of the necessary sections (introduction, problem statement, list of references, etc.); scientific and non-scientific vocabularies; the presence

of logical and semantic defects in the text of a scientific publication; selecting author's terms—new concepts defined by the authors of publications; highlighting the results presented in publications etc. Some authors study methods for the recognition of artificially generated scientific papers [2], [3], [8].

In the essay scoring, the most similar to our task is the task of prompt adherence that is assessing how the essay content corresponds to the announced essay topic [6], [11].

In [6] the Relatedness to Prompt feature is studied. The text of a essey fragment and the prompt (text of the essay question) must be related. If this relationship does not exist, this is perhaps evidence that the student has written an off-topic essay. The assessment was made for each sentence. The quality of the assessment was evaluated using double expert annotation for specific sentences. Most features proposed in this work are based on so-called Random indexing [13]. Random Indexing is a vector-based semantic representation system similar to Latent Semantic Analysis. In the current work, Random Indexing (RI) semantic space is trained on about 30 million words of newswire text. RI similarity to prompt for a sentence measures to what extent the sentence contains terms in the same semantic domain as compared to those found in the prompt. The SVM-classifier is trained on the calculated features and labeled data.

Persing and Ng [11] continue the study of the prompt relatedness in essay scoring using more diverse features. They try to predict the prompt relatedness for the whole essay, not for a single sentence. The predicted score ranges from one to four points at half-point intervals. 830 argumentative essays were annotated using a numerical score from one to four. Persing and Ng consider the task as a regression problem. Seven types of features were utilized in prompt-specific regressors based on linear SVM. Besides the random indexing features from the previous work, the authors used lemmatized unigram, bigram, and trigram similarity; thesis clarity keywords, which are the subdivision of the initial prompt to logical parts; LDA statistically generated topics.

3. Task, Data and Preprocessing

For experiments we usethe collection 40 thousand theses in pedagogics from various universities defended in 2017–2018 (further FullCollection). 120 theses from this collection have double scores from two experts belonging to different institutions (further AnnotatedCollection). This collection is new and the current study is the first one based on this collection.

The theses have similar structure. They include several parts: introduction, two-three chapters, sometimes recommendations, conclusion, appendices. In the introduction, a student introduces the theme of the thesis, the aim, the object, and the tasks of the work. The first chapter presents the survey of theoretical studies related to the theme of the work. The second and third chapters often describe practical experiments carried out by the student.

To have more information about the thesis' theme, we gather the above-mentioned structural elements (the theme, aim, object, and tasks of the thesis) into a specific document called theme header. The theme header conveys the main idea and direction of the thesis. It is clear that all parts of the thesis should correspond to the theme header in some extent. In this paper we assess how the first chapter of the

thesis, survey, is related to the theme header. We extract the theme header and chapters of the thesis using a specialized vocabulary and patterns. Figure 1 presents an example of a theme header.

```
ЗАГОЛОВОК = Безопасность и жизнестойкость студентов Ярославского Градостроительного колледжа
в образовательном процессе
ЦЕЛЬ = – определение безопасности и жизнестойкости студентов Ярославского Градостроительного
колледжа в образовательном процессе.
ОБЪЕКТ ИССЛЕДОВАНИЯ = – жизнестойкость студентов ЯГК и безопасность образовательного процесса.
ПРЕДМЕТ ИССЛЕДОВАНИЯ = – динамика особенностей жизнестойкости студентов Ярославского
градостроительного колледжа. Безопасность образовательного процесса.
АКТУАЛЬНОСТЬ = актуальность темы исследования заключаются в том, что в окружающем нас мире
всегда существовало и существует, много опасностей, но они недостаточно рассматривались под
углом влияния на объекты и возникающие при этом проблемы безопасности. Используя системный
подход, необходимо глубоко проанализировать и выделить объекты, на которые воздействуют
опасности, а также предложить пути решений проблем безопасности и повышения жизнестойкости.
Решение задач современного комплекса проблем безопасности может быть получено на основе общей
теории безопасности.
ЗАДАЧИ = 1. На основе теоретического анализа определить критерии и условия формирования
жизнестойкости студентов и безопасности образовательного процесса. 2. Выбрать методики,
направленные на выявление выраженности компонентов жизнестойкости и безопасности
образовательного процесса. 3.Выявить различия в уровне и содержании жизнестойкости студентов
1 и 2 курса специальностей «Автомеханик» и «Слесарь по ремонту строительных машин» в течение
2015- 2017 гг. 4. Провести анализ и сделать выводы
================================================================================================
TITLE = Safety and resilience of students of the Yaroslavl Town Planning College in the
educational process.
GOAL = – Definition of safety and resilience of students of the Yaroslavl Town Planning
College in the educational process.
OBJECT = – Student resilience of YTPC and safety of the educational process.
SUBJECT = – The dynamics of the characteristics of the student resilience of the Yaroslavl
Town Planning College. Safety of the educational process.
SIGNIFICANCE = The significance of the research topic lies in the fact that in the world
around us there always existed and there are many dangers, but they were not sufficiently
considered from the angle of influence on objects and the security problems arising from
this. Using a systematic approach, it is necessary to deeply analyze and identify objects
that are affected by hazards, as well as to offer solutions to safety problems and improve
resilience. The solution of the problems of the modern complex of security problems can be
obtained on the basis of the general theory of security.
TASKS = 1. On the basis of theoretical analysis to determine the criteria and conditions for
the formation of the student resilience and the safety of the educational process. 2. Choose
methods aimed at identifying the severity of the components of the resilience and safety of
the educational process. 3. Identify differences in the level and content of resilience of
students 1 and 2 courses of specialties "Auto Mechanic" and "Mechanic on the repair of
construction machines" during 2015-2017. 4. To analyze and draw conclusions.
```

Figure 1: Theme header for thesis “Safety and resilience of students of the Yaroslavl Town Planning College in the educational process”

The similarity between the theme header in a thesis and the prompt in essay writing can be seen. But the difference between two tasks: relatedness of a thesis to its theme header and an essay to the prompt is also significant:

- The theme elements are written by a student and can be poorly worded but the prompt in essay writing is formulated by professionals,
- There can be many essays for a given prompt. Therefore some methods can be specially tuned for a specific prompt [1]. The student thesis’s theme header is unique,
- The survey chapters are much longer than essays. In our data, they contain 250 sentences on average. Also, the theme headers are in most cases much longer than usual essay prompts,
- The survey chapters are never pervasive or argumentative in contrast to essays.

In the current study, we do not have any manual annotation of the theme relatedness of the first chapters. For evaluation, we use the overall score given to a thesis by two professional experts according to 2–5 scale, where 5 is the maximum grade in the scale. We suppose that low-scored theses should also have some problems in its surveys. As an example of "bad" segments for the same thesis, the theme header of which was presented, see Figure 2.

```
Забывая о духовности каждой мысли и каждого действия, о духовности постижения ценностей мира,
человек быстро мчится к глобальной катастрофе, вооруженный до зубов достижениями
научно-технического прогресса последнего столетия.
Пронизывающая все полезность и выгода, надежды на скорые новые прорывы в дальнейшем «обуздании»
и грабеже Природы делают хронически несвоевременным и затруднительным формирование
мировоззренческой альтернативы. Наука, как и власть, одевшись в официальные структуры, раздавая
чины и авторитеты, жестко следит за «протоколом». Сращиваясь с властью, она служит ей верой и
правдой, создавая десятки направлений различных идеализмом и материализме в, но, в конечном
итоге, исповедуя рационализм и как все общества потребления - меркантилизм. В результате Природу
растащили по кускам в дисциплинарные ниши, огородили эти ниши разного рода табу, создали
локальные языки.
================================================================================================
Forgetting about the spirituality of every thought and every action, about the spirituality of
comprehending the values of the world, a human quickly rushes to a global catastrophe, armed to
the teeth with the achievements of the scientific and technological progress of the last century
The pervasive utility and benefit, hopes for speedy new breakthroughs in further "curbing" and
robbery of Nature make chronically untimely and difficult the formation of ideological
alternatives. Science, like governance, dressed in official structures, handing out ranks and
authorities, strictly follows the "protocol". Merging with power, it serves it faithfully,
creating dozens of different directions of idealism and materialism in, but, ultimately,
professing rationalism and, like all consumer societies, mercantilism. As a result, Nature was
dragged apart in pieces into disciplinary niches, fenced in these niches of various sorts of
taboos, created local languages.
```

Figure 2: "Bad" segment for thesis "Safety and resilience of students of the Yaroslavl Town Planning College in the educational process"

We order all the theses according to the increase of the automatic scores of theme relatedness and evaluate methods of theme relatedness calculation using Average precision of 2-grade in the first positions of the created ranking. Table 1 shows the distribution of grades in 120 theses. Table 2 shows the deviation between grades of two experts. We calculate the Average precision measures according to the minimal grades of the theses. Thus, a thesis should have at least one 2 grade to be considered as a correct result in the beginning of the calculated rating.

Table 1: Distribution of marks in student theses

Mark	Number of theses with minimal mark	Number of theses with maximal mark
2	42	8
3	42	33
4	28	51
5	8	28

Table 2: Deviations between thesis marks

Points of difference	Number of works
0	49
1	51
2	16
3	4

As preprocessing, we use the procedure of segmenting the whole chapter to thematic fragments. Then we calculate similarity between specific segments and the theme header. The overall score of the theme relatedness of a chapter is based on averaging segment scores of this chapter.

4. Segmentation of Thesis Chapter to Thematic Fragments

The thematic segmentation module should break up a long sequence of the text into segments so that the sentences in one segment are thematically similar, and the boundaries of the segments signal a violation of connectivity between blocks of text. This procedure is based on the TopicTiling algorithm [12]. For splitting the text into segments, a procedure for assessing connectivity violations has been implemented, which is applied to each sentence. Based on the values of this metric, selection of separating sentences is carried out, which become the beginnings of segments. The procedure for assessing connectivity violation is as follows.

For each sentence, its "left" and "right" contexts with the length of w sentences are considered (sentences having the length of less than k words are ignored). For each sentence from the context, its vector representation is calculated using word2vec [9]. The weighted average of the word embeddings based on idf multiplier is calculated. Using cosine similarity, the similarity of all pairwise combinations of sentences between the "left" and "right" context is calculated. Based on these values, the coherence score is calculated by averaging all the similarities—*coherence_score*.

$$coherence\_score = \frac{1}{w} \sum_{v_l} \sum_{v_r} similarity(v_l, v_r) \quad (1)$$

After each sentence has its *coherence_score* calculated, in the second pass for each sentence its *depth_score* value is calculated using the following formula.

$$depth\_score(s_i) = 0.5 * (top\_left + top\_right - 2 * coherence\_score(s_i)) \quad (2)$$

Where *top_left* is the peak value of *coherence_score* to the left of the sentence, in non-descending order, *top_right* calculated by analogy.

The mean value and variance are calculated for the *depth_score* vector, on the basis of which the threshold values are chosen for the selection of candidate-sentences. If the *depth_score* of sentence is above the threshold and no separating sentences were selected in the neighborhood of several sentences, then this sentence is chosen as the separator. sentences are considered in the order of their *depth_score* values. Thus, after completing the procedure described above, the source text is broken up into segments.

5. Methods of Assessing Theme Adherence

5.1. Preprocessing

The text of the thesis was pre-processed as follows:

- The whole text was lemmatised and stop words were removed. The frequency characteristics of the words were calculated and the idf values were obtained;
- In addition, some words were removed from the text of the theme header based on a part of speech: verbs, adjectives and functional parts of speech.
- The procedure of extracting the concepts was carried out using the ontology [4]. Thus for each sentence there is a list of concepts contained in it. For concepts, idf was also counted;
- The word2vec model was trained on full collection of the theses, which contains 40 thousand documents. The parameters are: CBOW, vector length is 150, window size is 10. The training was conducted using the python gensim package;[1]

As an additional source of information about the domain, we use the ontology on natural sciences and technologies [4], which comprises terms of scientific fields (mathematics, physics, chemistry, geology, astronomy etc.) and terms of technological domains (oil and gas, power stations, cosmic technologies, aircrafts, etc.). Currently, about 6,500 terms (including term variants) were added to OENT to describe the pedagogics domain.

The main unit of the ontology is a concept, which has a unique name, the set of text entries, which express this concept in the text, and concept relations. For example, the concept *DEAF AND HARD OF HEARING EDUCATION* has the following text entries: *deaf education, deaf teaching, education of the deaf, teaching of the deaf* (translation from Russian).

5.2. Features of Theme Adherence Assessment

The chapter under analysis (further document) is presented in the form of N segments S and compared with the theme header H. Two baseline models were implemented to accomplish this task.

Baseline 1 (Tf-Idf). In this baseline model, the theme header and segments are represented as sparse vectors, where each element of the vector corresponds to a word from the collection's vocabulary. The values of vector elements are calculated as tf-idf. Based on the cosine similarity between the theme header and the S_i segment vectors, the *adherence_score* with the theme is formed.

Baseline 2 (SegWord2Vec). Each segment, including the theme header, is converted into a vector using word2vec embeddings. This operation is performed by weighted averaging of the word vectors with idf as weights, as used in the segmentation procedure. The following features were implemented and combined with the baselines:

[1] https://radimrehurek.com/gensim/index.html

NoNorm: Disabling normalization for the theme header vector. We introduce this feature, because the normalization gives lower *adherence_score*'s to larger and richer theme headers.

Concepts: Adding an additional vector for the theme header and segments, which is formed by analogy with the word vector, but based on the concepts of the ontology founded in the text of the segment/header. The concepts allow accounting for synonyms and multi-word expressions. In this case, the thematic adherence is formed as a weighted sum of the vectors similarities.

$$adherence\_score = \alpha * sim_{word} + (1 - \alpha) * sim_{conc} \quad (3)$$

Keywords: For the theme header, the most significant k_w words and k_c concepts are determined according to tf-idf. The set of keywords includes words and concepts whose tf-idf weights exceed the threshold. This threshold is calculated as 0.2 * average value to the 3 (2 for concepts) most significant (by tf-idf) words (or concepts). The weights of the keywords are multiplied by additional factors w_w and w_c, respectively. Keywords for thesis "Safety and resilience of students of the Yaroslavl Town Planning College in the educational process" are as follow:

- *words*—resilience (1.00), safety (0.47), town-planning (0.42), Yaroslavl (0.30), college (0.21), ytpc (0.17), student (0.17);
- *concepts*—urban planning (1.00), safety (0.63), college (0.57), student (0.20), system approach (0.17);

EmbedExp: There is an expansion of keywords for the theme header, by adding most similar words to them using word2vec embeddings. To do this, for each of the k_w keywords, the n_{w2v} closest words are calculated using the word2vec representation. Those words that are present in the whole thesis are added to the theme header vector. In order to calculate the weight of new words in the vector, the following formula is used:

$$weight_{wordext} = sim(word_{raw}, word_{ext}) * tf(word_{raw}) * idf(word_{ext}) \quad (4)$$

Where $word_{raw}$—the keyword on which the expansion is made, $word_{ext}$—new word. In addition, the weights of the new words are multiplied by the factor of w_{ext}. The set of expanded words for the thesis "Safety and resilience of students of the Yaroslavl Town Planning College in the educational process", includes:

- *new words*—hardiness (0.88), Maddy (0.76), tough (0.66), stories (0.62), coping (0.54), freshman (0.48), security (0.44), safe (0.41), highschool (0.14), scholar (0.13);

Regardless of the specific configuration, each segment and theme header are represented by a vector (or vectors) and *adherence_score* is calculated as cosine similarity between corresponding vectors. We calculate *adherence_score* for the whole chapter in two ways:

- *mean*: The average value of *adherence_score*'s of all segments;
- *mean_worse*: The average value of *adherence_score*'s among the worst 20% of segments;

The *mean_worse* variant corresponds to the hypothesis that a thesis is characterized by its worst fragments. Further, theses are ranked in ascending order of their *adherence_score* values.

6. Evaluation and Results

As mentioned earlier, we have 2 expert grades for each thesis. The minimum score (2) is chosen as the reference value, assuming that at least one expert could find serious problems in the theses. It was also previously shown that in the reference collection there are 42 works with the minimum grade of 2.

The evaluation methodology is proposed as follows: the algorithm for each thesis forms the values of *adherence_score* (*mean* and *mean_worse*), on the basis of which the reference collection is ranked so that the "worst" thesis was "above". For evaluation we use average precision measure.

$$average\_precision(n) = \frac{\sum_{k=1}^{n} P(k) * rel(k)}{n} \quad (5)$$

Where *P*(*k*)—precision at *k*, *rel*(*k*) is equal to 1 if *k*-th element of the list is relevant, otherwise 0.

We calculate *average_precision(25)* measure in 25th position (20% of the collection). The thesis is considered as relevant if it has grade 2 for at least one expert. In addition, the mean values of the *average_precision(n)* for positions from 1 to 25 were calculated.

The results of the evaluation of the configurations can be seen in Table 3. It also presents the result of random ordering (averaging 25000 random permutations). *NoNorm* did not give any significant improvements for any configurations. The results of average precision measures are also shown in the Figures 3 and 4.

Table 3: Evaluation results on 120 reference theses

	av_prec(25) by mean_worse	av_prec(25) by mean	mean av_prec(25) by mean_worse	mean av_prec(25) by mean
Random	0.15	0.15	0.19	0.19
Tf-Idf	0.15	0.16	0.11	0.12
Tf-Idf + Concepts	0.15	0.22	0.22	0.23
Tf-Idf + Keywords	0.23	0.21	0.16	0.29
Tf-Idf + Concepts + Keywords	0.20	0.27	0.26	0.43
Tf-Idf + Keywords + EmbedExp	0.21	0.22	0.19	0.18
Tf-Idf + Concepts + Keywords + EmbedExp	0.20	0.28	0.25	0.41
SegWord2Vec	0.28	0.22	0.37	0.18
SegWord2Vec + Concepts	0.28	0.27	0.37	0.40
SegWord2Vec + Keywords	0.30	0.19	0.47	0.36
SegWord2Vec + Concepts + Keywords	0.29	0.29	0.46	0.49

For the best configuration (accounting for both *mean* and *mean_worse*) , the optimal parameters were as follows:

- *mean*: *SegWord2Vec + Concepts + Keywords*:
  $\alpha = 0.25, k_w = 12, w_w = 1.25, k_c = 5, w_c = 1$.
- *mean_worse*: *SegWord2Vec + Concepts + Keywords*:
  $\alpha = 0.75, k_w = 12, w_w = 1.25, k_c = 3, w_c = 5$.
  Where α—the parameter which controls the participation of concepts,
  k_w—number of word keywords, w_w—multiplier of word keywords,
  k_c—number of concept keywords, w_w—multiplier of concept keywords.

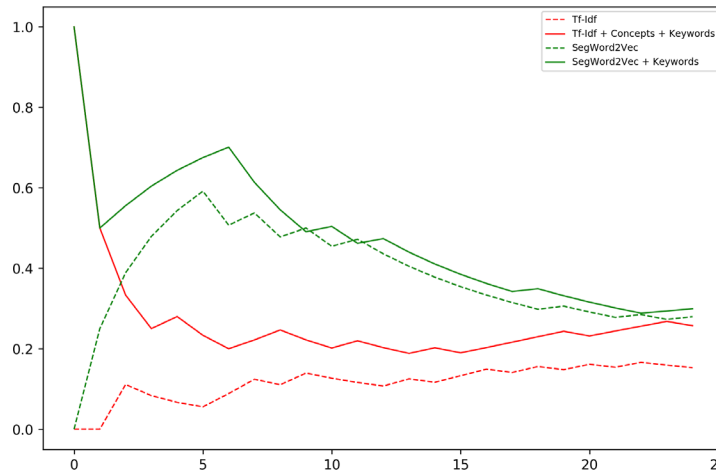


Figure 3: Average precisions graphs for base and best configurations for av_precision by mean_worse

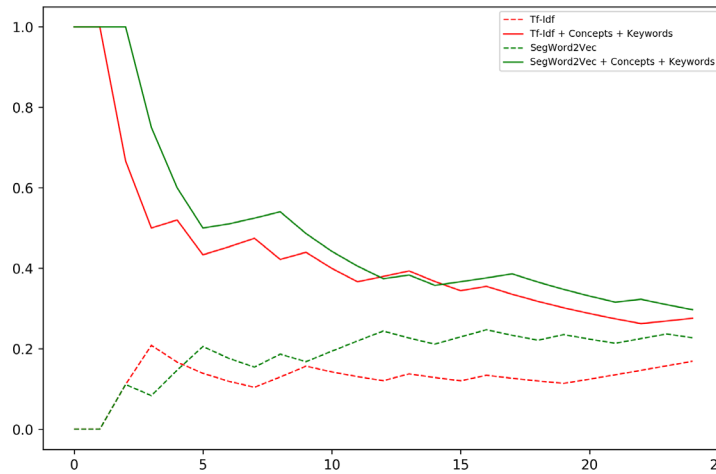


Figure 4: Average precisions graphs for base and best configurations for av_precision by mean

Based on the results, we can conclude the following:

- The configurations based on *SegWord2Vec* are better than those based on *Tf-Idf*.
- Ranking based on *mean_worse* (worst segments) at least not worse then *mean adherence_score* and better corresponds to the task of finding the worst thesis.
- Use of *Keywords* always leads to better results.
- Use of *Concepts* is very useful for *Tf-Idf* and sometimes useful for *SegWord2Vec*

7. Error analysis

For the analysis of the system, we consider two configurations: *Tf-Idf* и *SegWord-2Vec + Concepts + Keywords*.

Tf-Idf.The first 5 theses have the following scores and grades:

1) 0.0 : 4;
2) 0.001 : 3;
3) 0.001 : 2;
4) 0.001 : 3;
5) 0.001 : 4.

The scores of the worst theses are very close to each other and are practically zero. This means that the algorithm did not reveal similarities between the worst segments and theme header.

The first thesis with the grade 4 has score 0.0. Among the 13 worst segments the similarity to the theme header is zero. The thesis itself has the name “Differentiated approach in improving the physical fitness of students in 6th grade”, but in all worst segments author is talking about the biological characteristics of children and their development. The amount of specific biological information seems excessive. In addition, all the worst segments are plagiarized, and the text was deliberately distorted (every 3–4 words were simply removed from the text), which prevented the anti-plagiarism systems from detecting plagiarism.

The second thesis with grade 4 also has a low score of 0.001. Among the 10 worst segments, there are also no exact matches of words with the theme header and this is due to the fact that they also deviate from the main theme of the work. The theme of the thesis is “The implementation of a systematic approach to teaching biology in primary school” but in all the worst segments there is a serious bias in the philosophical direction, to questions of knowledge.

SegWord2Vec + Concepts + Keywords. The first 5 theses have the following scores and grades:

1) −0.099 : 2;
2) −0.098 : 3;
3) −0.076 : 2;
4) −0.056 : 2;
5) −0.037 : 2.

The spread of values here is much larger, which suggests that the model better separates different theses, among other things, it is clearly seen that the estimates are better grouped (this is also evident on the **Figure 3**).

We looked at the content of the worst theses and the text of the worst segments is poorly relevant to a given topic. But at the same time there were situations that large segments sometimes get low scores even if they contained some amount of relevant information. This is due to the fact that the averaging of word2vec vectors works worse on large texts. In addition, in comparison with *Tf-Idf*, it became more difficult to interpret, why exactly one segment is worse than the other.

As a result, we can draw the following conclusions from the error analysis:

- *Tf-Idf* badly detects links between related segments, but the text of which use different words. This leads to the fact that among the worst segments are those that do not seem to relate directly to the topic, but at the same time have some consistency with it.
- *Tf-Idf* poorly separates bad theses from each other.
- *SegWord2Vec + Concepts + Keywords* on the other hand, organizes theses well and, on average, highlights really bad segments, in which there is no useful information for thesis, but at the same time, the interpretability suffers a little.
- *SegWord2Vec + Concepts + Keywords* also sometimes puts very low weights on large segments, but at the same time in which there is some amount of relevant information.

8. Conclusion

In this paper we studied approaches to assessing the quality of student theses in pedagogics. We considered a specific subtask in thesis scoring of estimating its adherence to the thesis's theme. The special document (theme header) comprising the theme, aim, object, tasks of the thesis is formed. The theme adherence is calculated as the similarity value between the theme header and thesis segments.

For evaluation we ordered theses in the increased value of the calculated theme adherence and compared the ordering with expert grades using the average precision measure. The best configuration for theses ranking is based on the weighted averaged sum of word embeddings (word2vec) and keywords extracted from the theme header.

9. Acknowledgements

This article contains the results of the project "Developing new methods for analyzing large text data using linguistic and ontological resources, machine learning methods and neural networks", carried out as part of the Competence Center program of the National Technology Initiative "Center for Big Data Storage and Analysis", supported by the Ministry of Science and Higher Education of the Russian Federation under the Contract of Moscow State University with the Fund for Support of Projects of the National Technology Initiative No. 13/1251/2018 dated December 11, 2018.

References

1. *Attali, Y., Burstein, J.:* Automated essay scoring with e-rater v. 2. The Journal of Technology, Learning and Assessment. 4, 3, (2006).
2. *Avros, R., Volkovich, Z.:* Detection of computer-generated papers using one-class svm and cluster approaches. In: International conference on machine learning and data mining in pattern recognition. pp. 42–55 Springer (2018).
3. *Bakhteev, O. et al.:* About one method of detecting artificial and unscientific texts in an extensive collection of documents. Electronic Libraries. 20, 5, 298–304 (2017).
4. *Dobrov, B. V., Loukachevitch, N. V.:* Development of linguistic ontology on natural sciences and technology. In: LREC. pp. 1077–1082 (2006).
5. *Foltz, P. W. et al.:* Automated essay scoring: Applications to educational technology. In: EdMedia+ innovate learning. pp. 939–944 Association for the Advancement of Computing in Education (AACE) (1999).
6. *Higgins, D. et al.:* Evaluating multiple aspects of coherence in student essays. In: Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: HLT-naacl 2004. (2004).
7. *Khritankov, A. S. et al.:* Discovering text reuse in large collections of documents: A study of theses in history sciences. In: 2015 artificial intelligence and natural language and information extraction, social media and web search fruct conference (ainl-ismw fruct). pp. 26–32 IEEE (2015).
8. *Labbé, C., Labbé, D.:* Duplicate and fake publications in the scientific literature: How many scigen papers in computer science? Scientometrics. 94, 1, 379–396 (2013).
9. *Mikolov, T. et al.:* Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013).
10. *Osipov, G. et al.:* Technologies for semantic analysis of scientific publications. In: 2012 6th ieee international conference intelligent systems. pp. 058–062 IEEE (2012).
11. *Persing, I., Ng, V.:* Modeling prompt adherence in student essays. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers). pp. 1534–1543 (2014).
12. *Riedl, M., Biemann, C.:* TopicTiling: A text segmentation algorithm based on lda. In: Proceedings of acl 2012 student research workshop. pp. 37–42 Association for Computational Linguistics (2012).
13. *Sahlgren, M.:* Vector-based semantic analysis: Representing word meanings based on random labels. In: In essli workshop on semantic knowledge acquistion and categorization. Citeseer (2001).
14. *Taghipour, K., Ng, H. T.:* A neural approach to automated essay scoring. In: Proceedings of the 2016 conference on empirical methods in natural language processing. pp. 1882–1891 (2016).

ПРИТЯЖАТЕЛЬНЫЕ МЕСТОИМЕНИЯ В РУССКИХ ОБЪЕКТНЫХ ИМЕННЫХ ГРУППАХ

Тискин Д. Б. (daniel.tiskin@gmail.com)
Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

Исследование конкуренции русских лично- и возвратно-притяжательных местоимений в связанном употреблении (как в *Я*$_1$ *встретился с моими*$_1$ */ со своими*$_1$ *друзьями*) ведётся достаточно давно, однако не все аспекты этого явления были изучены квантитативными методами и получили описание в рамках той или иной теории синтаксиса и семантики. В работе исследуется поведение местоименных посессоров в прямообъектных именных группах, связанных местоимением 1 или 2 лица в позиции подлежащего. Акцент делается на связи выбора местоимения с возможностью или необходимостью коллективной интерпретации глагольной группы и отношения принадлежности.

Пользуясь данными Национального корпуса русского языка и корпуса Araneum Russicum Maximum, мы показываем, что выбор стратегии выражения посессора связан с числом субъекта (как в целом по корпусу, так и для отдельных глаголов). Проведённое анкетирование позволяет установить, что предпочтение лично-притяжательного местоимения связано с коллективным прочтением, причём в отсутствие такого прочтения при ед. ч. объекта любое выражение посессора затруднено (если лексема — вершина ИГ не является singulare tantum). Предлагается интерпретация полученных данных, основанная на том, что притяжательное местоимение имеет интерпретируемый признак числа посессора (например, *наш* обозначает коллективную принадлежность множеству посессоров), а ИГ-дополнение без посессора может реанализироваться как часть предиката.

Ключевые слова: русский язык, притяжательные местоимения, возвратность, дистрибутивность, φ-признаки, группа детерминатора

POSSESSIVE PRONOUNS IN RUSSIAN OBJECT NOUN PHRASES

Tiskin D. B. (daniel.tiskin@gmail.com)
Saint Petersburg State University, Saint Petersburg, Russia

By now, the choice between pronominal and reflexive bound possessives in Russian has been subject to prolonged investigation, but some aspects of the phenomenon have not been scrutinised by quantitative methods

or given an interpretation in terms of a formal model of syntax and semantics. The present paper deals with pronominal possessives situated in direct object noun phrases and bound by a 1st- or 2nd-person pronominal subject (including zero subjects in imperatives). We focus on the factors that have to do with the availability of the collective interpretation of the verb phrase and of the possession relation itself.

Using data from the Russian National Corpus and from the Araneum Russicum Maximum web corpus, we demonstrate a significant effect of the subject number on the choice of the possessive pronoun. To investigate the effects of collectivity, we designed a questionnaire, which showed that with plural the preference for pronominal possessives over the reflexive possessive increases as the collective interpretation becomes more salient. Moreover, both types of possessives are degraded to a certain extent within singular (but not *singulare tantum*) object NPs when the collective reading is unavailable. We suggest that the observed distribution can be explained if we assume that (a) the cardinality of the possessor is an interpretable feature of Russian possessives and is therefore able to cause interpretational conflicts, and (b) the possessive layer of the nominal domain precludes an object to be re-analysed as part of the predicate, which is required for a distributive interpretation.

Key words: Russian, possessive pronouns, reflexivity, distributivity, φ-features, DP

1. Введение

Наряду с полнозначными именными группами, личные местоимения 1–2 лица относятся к референциальным выражениям и, в соответствии с принципом C теории связывания Н. Хомского [3], не должны иметь связанных употреблений. Тем не менее, известно несколько типов исключений из этого обобщения: наряду с неприемлемыми (1) существуют (2), где приемлемость коиндексирования соотносят не собственно со связыванием, а с различием «перспективы» или способов данности одного и того же объекта [21], и (3), два значения которого интерпретируются как включающие, соответственно, свободное и связанное употребление местоимения *I* (см. обсуждение в [19]).

(1) a. **Я$_1$ увидела меня$_1$.*
b. **Mary$_1$ saw you$_1$.*

(2) *I dreamt that I was Brigitte Bardot and I$_1$ kissed me$_1$.* [9]

(3) *Only I got a question that I understood.* [7]
a. ‘Только я, x_0, обладаю свойством ‘$\lambda x.x$’у задали вопрос, который x_0 понял”
b. ‘Только я, x_0, обладаю свойством ‘$\lambda x.x$’у задали вопрос, который x понял”

Предметом рассмотрения в данной работе являются русские **притяжательные** местоимения, среди которых выделяются лично-притяжательные (*мой*, *ваш*, *его* и т. д.) и возвратно-притяжательное *свой*. В соответствии с этим делением ожидается, что лично-притяжательные будут в общих чертах повторять дистрибуцию личных и функционировать как прономинальные и референциальные

выражения, тогда как возвратно-притяжательное будет проявлять свойства анафора, чьё употребление регулируется принципом А. Однако и в этом случае известны отклонения от названной идеализации. Так, Е. В. Падучева [15: 181] отмечает вариативность в примерах типа (4), хотя и усматривает в них тонкие «различия — смысловые или какие-то ещё»:

(4) *Я не хотел становиться обузой для моих ⟨своих⟩ родителей.*

Часть ответа на вопрос о природе этих различий состоит, по-видимому, в том, что при употреблении *свой* в (4) говорящий (x_0) высказывается о нежелании иметь свойство '$\lambda x.x$ является обузой для родителей x', а в случае *мой* — свойство '$\lambda x.x$ является обузой для родителей x_0'[1]. Хотя в данном случае эти интерпретации денотативно эквивалентны, это не всегда так; в частности, в английском языке, где возвратно-притяжательного местоимения нет и лично-притяжательные местоимения всех лиц имеют связанные употребления, (5) проявляет тот же тип неоднозначности, что (3).

(5) *Only you did your homework.*

Другие возможные семантические различия между лично- и возвратно-притяжательными местоимениями, в отличие от различных значений *свой* [14] и особенностей дистрибуции притяжательных местоимений [13], [17], [5], на данный момент изучены недостаточно, хотя в книге Е. В. Падучевой [15] содержится ряд важных наблюдений. Во-первых, отмечается различие между притяжательными местоимениями 1–2 лица и 3 лица: первые могут, будучи частью объектной ИГ, иметь антецедентом подлежащее (6a), тогда как последние последовательно подчиняются принципу B и не допускают связывания в пределах клаузы (6b) (причём случаи конкуренции с возвратно-притяжательными подлежат описанию в терминах различного синтаксического размера непрозрачной области, в которой тот или иной тип местоимений должен быть связан или свободен).

(6) a. *И тут я увидел **моих** однокурсников — Зайченко с Лебедевым.*
[Сергей Довлатов. Виноград (1990)]
b. **Вася ⟨он⟩ увидел **его** однокурсников.*

Во-вторых, утверждается, что как минимум в некоторых случаях конкуренции двух типов притяжательных местоимений выбор осуществляется в зависимости от того, раздельно или совместно[2] индивиды в составе множества-посессора обладают посессумом[3]. А. Б. Андреевский [1], а позднее А. Тимберлейк [20]

---

[1] Возможно, этим следует объяснять и наблюдения относительно способности свой обозначать объект, варьирующий от ситуации к ситуации, как в *Ведь до этого я всегда дружественно работал со своим партнёром* [20]

[2] Ввиду этого противопоставления можно ожидать, что поведение посессоров в рамках объектной ИГ может быть различным при ед. ч. и мн. ч. субъекта; это отмечается ещё в [1] со ссылкой на Б. Г. Унбегауна и обсуждается нами ниже.

[3] Е. В. Падучева пользуется термином *дистрибутивность*, однако её употребление несколько расходится с принятым в литературе, посвящённым дистрибутивной предикации; см., например, [2].

и Е. В. Падучева соотносят *свой* с индивидуальным, а *наш/ваш* с коллективным обладанием.

В наиболее ранних из современных работ по теме [1], [20], [6] описание строится на анализе отдельных примеров или ограниченного корпуса. В последнее время некоторые аспекты дистрибуции и семантики притяжательных местоимений исследовались на материале больших корпусов [13], [16]. Мы продолжаем это направление исследований: задачи данной работы состоят в том, чтобы, ограничившись случаями выражения посессора в составе прямого дополнения, во-первых, оценить влияние на употребимость лично-притяжательных местоимений ранее называвшихся факторов (лица и числа глагола, а также дистрибутивной или коллективной интерпретации глагольной группы), а во-вторых, наметить контур анализа, который позволил бы объяснить наблюдаемое распределение. Первой из этих задач посвящён **раздел 2**, представляющий результаты корпусных исследований; для исследования дистрибутивности—коллективности нами было проведено анкетирование, результаты которого излагаются в **разделе 3**. Очерку анализа посвящён **раздел 4**[4].

2. Корпусное исследование

Для пилотного исследования мы отобрали 25 наиболее частотных переходных[5] глаголов русского языка по [10]. Для каждого из них в подкорпусе текстов, созданных с 1918 г., основного корпуса НКРЯ запрашивались сочетания вида «личное местоимение (в случае индикатива) + требуемая форма глагола (императив или индикатив) + притяжательное местоимение в аккузативе»[6]. Раздельно запрашивались клаузы с ед. ч. и мн. ч. субъекта. Поисковые выдачи просматривались вручную.

Результаты исследования представлены в **таблице 1**.

Таблица 1. Соотношение лично-притяжательных (ЛП) и возвратно-притяжательных (ВП) местоимений в прямообъектных ИГ в подкорпусе современных текстов НКРЯ

	императив			1 лицо			2 лицо			всего в индикативе		
мест.	SG	PL	всего	SG	PL	всего	SG	PL	всего	SG	PL	всего
ЛП	34	128	**162**	138	154	**292**	4	35	**39**	142	189	**331**
ВП	261	119	**380**	806	181	**987**	226	176	**402**	1 032	357	**1 389**
% ЛП	12	52	30	15	46	23	2	17	9	12	40	19

[4] Уже на этапе подготовки статьи нам стало известно о существовании работы [4], близкой к нашей по теме и методам, но труднодоступной ввиду того, что она написана на норвежском языке.

[5] Переходность определялась наличием соответствующей пометы хотя бы у одного вхождения данного глагола в НКРЯ.

[6] При этом мы старались исключать устойчивые выражения, оставляя, впрочем, *взять своё* и вообще субстантивированные местоимения; конструкции с творительным предикативным, как в *вы считаете свой приезд в Белград нецелесообразным*, рассматривались наряду с прочими.

Аналогичное исследование было проведено на материале веб-корпуса Araneum Russicum Maximum. Здесь поисковый запрос был ограничен случаями, где за местоимением следует существительное в винительном падеже; ручная выверка отсутствовала. Результаты представлены в **таблице 2**.

Таблица 2. Соотношение лично-притяжательных и возвратно-притяжательных местоимений в прямообъектных ИГ в корпусе Araneum Russicum Maximum

	императив			1 лицо			2 лицо			всего в индикативе		
мест.	SG	PL	всего	SG	PL	всего	SG	PL	всего	SG	PL	всего
ЛП	98	3482	**3580**	782	2929	**3711**	23	1093	**1116**	805	4022	**4827**
ВП	8932	18035	**26967**	20137	8920	**29057**	2709	13611	**16320**	22846	22531	**45377**
% ЛП	1	16	12	4	25	11	1	7	6	3	15	10

Данные обоих корпусов убедительно подтверждают различие между местоименными подлежащими ед. ч. и мн. ч. с точки зрения соотношения частот лично- и возвратно-притяжательных местоимений в составе прямого объекта: как в императиве, так и в обоих лицах в индикативе подлежащее мн. ч. в гораздо большей мере допускает лично-притяжательного посессора ($p \ll 0{,}001$ для обоих корпусов, здесь и далее критерий χ^2). Как показывает **рисунок 1**, составленный по данным Araneum, в достаточно объёмном корпусе это утверждение верно и для всех рассмотренных глаголов в отдельности.

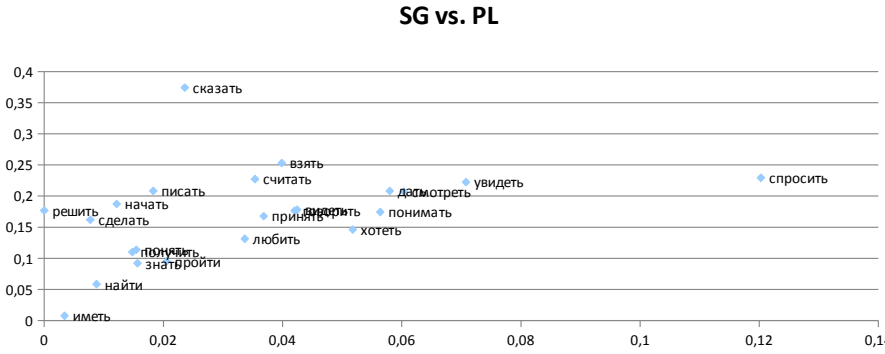


Рисунок 1. Доля лично-притяжательных местоимений для различных глаголов при множественном (по оси абсцисс) и единственном (по оси ординат) числе местоименного подлежащего в корпусе Araneum Russicum Maximum. Включены глаголы из списка, имеющие как минимум 100 вхождений в совокупности поисковых выдач

Оба корпуса позволяют утверждать, что в императиве лично-притяжательные посессоры в объектной ИГ встречаются чаще, чем при подлежащем 2 лица в индикативе ($p \ll 0{,}001$ для обоих корпусов), однако относительно сравнения императива с 1 лицом при индикативе данные корпусов расходятся.

Руководствуясь замечанием анонимного рецензента, мы проверили данные для Araneum вручную, проанализировав по 300 произвольно выбранных примеров для каждой из шести комбинаций «наклонение/лицо + число», ограничив поиск 25 глаголами из списка и отсеяв нерелевантные запросу примеры. Результаты приведены в **таблице 3**.

Таблица 3. Соотношение лично-притяжательных и возвратно-притяжательных местоимений в прямообъектных ИГ в случайной выборке из корпуса Araneum Russicum Maximum

	IMPER.SG	IMPER.PL	1SG	1PL	2SG	2PL
ЛП	4	53	18	96	5	32
ВП	290	241	268	190	293	250

И для индикатива 1 и 2 лица, и для императива различие между ед. ч. и мн. ч. подтвердилось при $p \ll 0{,}001$. Императив только во мн. ч. имеет значимо большую долю возвратно-притяжательных посессоров, чем 2 лицо индикатива. 1 лицо индикатива в обоих числах имеет значимо большую долю лично-притяжательных посессоров, чем 2 лицо и императив.

Глаголы, для которых выдача не ограничивалась единичными примерами, можно упорядочить по их склонности сочетаться с прямым объектом, в котором посессор выражен лично-притяжательным местоимением. Для корпуса Araneum результат этой операции представлен на **рисунке 2**. Можно отметить существенные различия между глагольными лексемами в отношении того, сопоставимы ли частоты лично-притяжательных посессоров у разных их форм (ср. *сказать* vs. *принять*): в перспективе анализ должен учесть эти различия, возможно объяснимые такими факторами, как эмпатия говорящего (как предлагалось в некоторых работах)[7].

[7] Заметим, что императив от *дать* и *сказать*, в отличие от *знать* или *любить*, зачастую используется как побуждение сделать нечто в интересах говорящего.

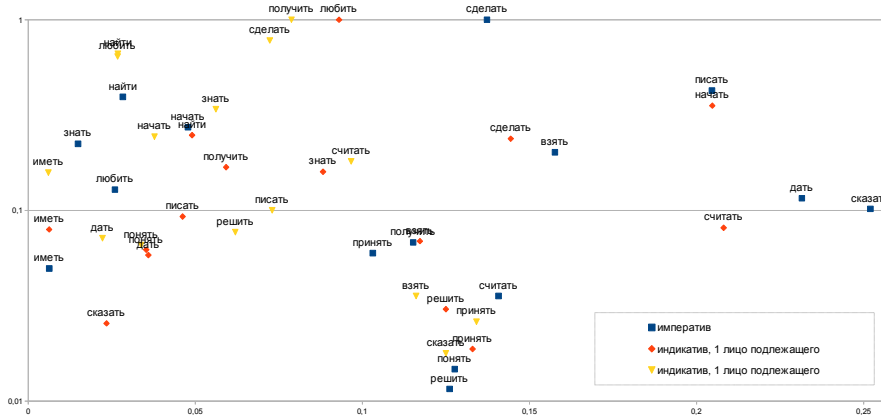


Рисунок 2. Доля лично-притяжательных местоимений (по оси абсцисс) и частотность форм глагола (по оси ординат; в долях от числа вхождений наиболее частотного из рассмотренных глаголов при данном типе подлежащего; логарифмическая шкала) для различных случаев употребления глаголов по данным Araneum Russicum Maximum. Включены глаголы из списка, имеющие как минимум 100 вхождений с каждым из типов подлежащего

3. Экспериментальное исследование

3.1. Методика

Эксперимент проводился в форме анонимного анкетирования через интерфейс Google Forms. Испытуемым предлагались в случайном порядке 20 предложений с одним пропущенным словом каждое (8 филлеров и 12 экспериментальных условий), в каждом из которых было необходимо сделать выбор из трёх вариантов (демонстрировались в случайном порядке): возможность заполнить пропуск морфологически подходящей формой лично-притяжательно местоимения; возможность заполнить пропуск морфологически подходящей формой возвратно-притяжательного местоимения; неприемлемость предложения ни в одном из этих случаев[8]. Результаты были обработаны, когда число респондентов составило 217 (173 женщины и 44 мужчины; медианный возраст — 20–29 лет).

Экспериментальные условия, распределённые по четырём вариантам анкеты методом латинского квадрата, включали переходный глагол в форме

[8] Таким образом, использованный метод является модификацией распространённой техники «forced choice», куда была добавлена возможность неприемлемости обоих вариантов, роль которой будет охарактеризована ниже.

индикатива (с подлежащим 1 или 2 лица) или императива мн. ч. и различались по следующим параметрам:

- возможность коллективной интерпретации глагольной группы, например:

(7) a. Мы скрепили *** договор рукопожатием. (обязательна)
b. Мы подняли *** багаж на немалую высоту. (возможна)
c. Мы закрыли *** голову руками. (невозможна)

- число прямого дополнения;
- является ли число прямого дополнения лексическим признаком его вершины (слова singulare tantum или plurale tantum) или выбирается свободно[9].

3.2. Результаты и обсуждение

Для каждой группы предложений нас интересуют значения двух показателей: доля случаев, в которых испытуемый счёл неприемлемыми оба варианта заполнения пропуска (%∅), и доля выборов лично-притяжательного местоимения от общего числа случаев, где пропуск **был** заполнен (%ЛП).

Полученные результаты представлены в таблице 4.

Таблица 4. Итоги анкетирования

Признаки			Реакции респондентов				
коллективная интерпретация	морфологическое число ИГ	природа числа ИГ	ВП	ЛП	ни то, ни другое	%ЛП	%∅
да	PL	—	39	168	12	81,2	5,5
да	SG	—	26	182	11	87,5	5,0
да	PL	tantum	32	167	20	83,9	9,1
да	SG	tantum	48	167	4	77,7	1,8
возможна	PL	—	84	131	4	60,9	1,8
возможна	SG	—	74	132	13	64,1	5,9
возможна	PL	tantum	83	132	4	61,4	1,8
возможна	SG	tantum	79	131	9	62,4	4,1
нет	PL	—	93	115	11	55,3	5,0
нет	SG	—	89	68	62	43,3	28,3
нет	PL	tantum	110	84	22	43,3	10,2
нет	SG	tantum	98	105	19	51,7	8,6

Анкетирование подтвердило существующие в литературе предположения о роли дистрибутивности и коллективности: при обязательности коллективной интерпретации %ЛП выше, чем при возможности, а при возможности — выше, чем при невозможности (таблица 5; $p \ll 0{,}001$ для каждой пары).

[9] В случаях, где коллективная интерпретация невозможна, в качестве «pluralia tantum» рассматривались также слова, обозначающие парные предметы: *глаза*, *щёки* и *серёжки*.

Таблица 5. Связь выбора выражения посессора с возможностью коллективной интерпретации глагольной группы по данным анкетирования

Коллективная интерпретация	Возможность вставки				Невозможность вставки	
	ВП	ЛП	всего	%ЛП	ни то, ни другое	%∅
да	145	684	829	82,5	47	5,4
возможна	320	526	846	62,2	30	3,4
нет	390	372	762	48,8	114	13,0

Статистически значимы и различия в %∅ ($p < 0{,}05$ для пары «да—возможна», $p \ll 0{,}001$ для остальных пар), но, как можно видеть, наиболее существенный вклад в эти различия вносят высокие значения %∅ для предложений с ед. ч. прямого дополнения, где невозможна коллективная интерпретация[10]; эти предложения приведены в (8).

(8) a. А давай покажем ему *** паспорт.
b. Мы закрыли *** голову руками.
c. Девочки, наденьте *** любимое платье.
d. Дети, вы опять запачкали *** рот шоколадом?

4. Анализ

Мы считаем нуждающимися в интерпретации по крайней мере следующие особенности предложений с местоименным посессором в составе прямого дополнения.

1. Существенные различия между ед. ч. и мн. ч. подлежащего (при отсутствии существенных различий, связанных с числом прямого дополнения).
2. Бóльшая предпочтительность лично-притяжательного выражения посессора при коллективных интерпретациях ГГ.
3. Затруднённость как лично-, так и возвратно-притяжательных посессоров в составе прямого дополнения в ед. ч. при дистрибутивных интерпретациях ГГ.

Что касается различий между 1 и 2 лицом, необходимо учитывать, что формы 2 лица мн. ч. не обязательно имеют референцию к множественному адресату, но могут вместо этого выражать вежливость. Эта омонимия затрудняет оценку корпусных данных, но и в случае анкетирования мы не можем исключить её влияния, хотя и использовали обращение во мн. ч. как средство отсечь нежелательную интерпретацию. В остальном исследование этого вопроса выходит за рамки нашей работы.

[10] По неизвестной нам причине для (8c) %∅ имеет сравнительно низкое значение — около 8.

Заметим, что интерпретация глагольной группы в *Мы закрыли голову руками* дистрибутивная, поскольку для каждого индивида в составе группы, называемой *мы*, верно, что он закрыл голову руками, а в *Мы закрыли головы руками* коллективная или кумулятивная, поскольку неверно, что каждый индивид закрыл рукой или руками более одной головы (если только один человек не закрывал рукой голову другого), хотя и верно, что все индивиды вместе закрыли все данные в контексте головы. В соответствии с наиболее распространённым ныне взглядом [2], дистрибутивная интерпретация предиката возникает ввиду присутствия в структуре предложения дистрибутивного оператора D, чья семантика приводится в (9):

(9) $[\![D]\!] = \lambda P\, \lambda x\, (\forall y \in \mathrm{At}(x)\ (P(y)))$

Смысл (9) в том, что D — функция, принимающая свойство P и (возможно, неатомарный) индивид x и приписывающая P всем **атомарным** индивидам в составе x.

Что касается связанных употреблений местоимений, в т. ч. в позиции посессора объектной ИГ, А. Кратцер [8] предложила считать, что семантическое связывание таких местоимений осуществляется вершиной v, в чьей сфере действия находится вся глагольная группа и чьи φ-признаки (лицо, число, род) устанавливаются путём согласования с подлежащим. Таким образом, структура предложения с посессором в составе прямого дополнения имеет вид

(10) мы 1 [t_1 [v … [D [закрыл- [нашу/свою голову] руками]] …]]

Ввиду сказанного, возможный облик прямого дополнения зависит от двух факторов: в отношении числа ИГ — от присутствия D, а в отношении φ-признаков местоименного посессора — от φ-признаков подлежащего. При этом выбор между лично-притяжательным и возвратно-притяжательным местоимением является, очевидно, лексическим (осуществляется до входа этой единицы в синтаксическую деривацию), в то время как подходящее слово из числа лично-притяжательных (например *наш* в отличие от *мой* или *ваш* при подлежащем *мы*) выбирается, с точки зрения подхода Кратцер, «после» синтаксиса, на стадии озвучивания (Spell-Out).

Принимая названные допущения, мы приходим к выводу, что в случае ед. ч. подлежащего число ИГ-дополнения зависит исключительно от обозначаемой ситуации, тогда как в случае мн. ч. подлежащего возможны варианты, связанные с D (проблема 1). При коллективной интерпретации ГГ D отсутствует, и денотат ИГ-дополнения рассматривается как единый неатомарный индивид, например мереологическая сумма всех проектов в *Мы подняли* (*наши*) *проекты на немалую высоту* в его коллективном прочтении. Такой индивид не может принадлежать кому-либо из денотата *мы* в отдельности, поскольку разные его части имеют разную принадлежность; но местоимение *наш* подходит, поскольку означает коллективную принадлежность. При связывании вершиной v конфликта признаков не возникает[11]. Относительно *свой* Кратцер предполагает, что у него нет

[11] Такое описание предполагает, что мы, в отличие от Кратцер, не считаем, что до связывания местоимение не охарактеризовано по числу посессора; см. альтернативный подход в [18].

φ-признаков, однако у нас есть основания считать, что оно в той или иной мере ориентировано на атомарные индивиды (проблема 2); ср. интроспективные отчёты некоторых респондентов нашей анкеты: «*Ваш*, — или о множественном числе или более официально»; «Например, „студенты, сдавайте вашу работу“ — если это одна работа на группу, „студенты, сдавайте свою работу“ — если у каждого отдельно»; «„Мы закрыли свою/нашу голову руками“ подразумевает, что у „нас“ на всех одна голова…»; «Поняла, что затрудняюсь в использовании слова *свой* во множественном числе». Соответственно, признак множественности посессора у *наш* мы также считаем интерпретируемым.

Наконец, затруднённость любых местоименных посессоров в составе ИГ-дополнения в ед. ч. при коллективной предикации (проблема 3) мы связываем (помимо затруднённости *наш* в силу того, что признак множественности посессора на нём конфликтует с единственным числом посессума, если последний не рассматривается как находящийся в коллективном владении) с особенностями структуры именной группы. Хотя, как показывает Е. А. Лютикова [11], [12], генитивный посессор в русском языке располагается ниже в структуре ИГ, чем детерминатор, и является адъюнктом *n*P[12], мы полагаем, что дистрибутивная интерпретация облегчается тогда, когда ИГ употребляется нереференциально и фактически реанализируется как часть предиката. Присутствие посессора[13] — само по себе (как связанное с наличием более развитой левой периферии ИГ-дополнения) или как причина референциальной интерпретации — затрудняет дистрибутивную интерпретацию. По-видимому, этому есть дополнительные подтверждения: так, именно от полноценных DP ожидается способность присоединять нерестриктивное относительное придаточное. Интроспективная оценка примеров из Google, как представляется, подтверждает это предположение: в примерах группы (11) *головы* (вне зависимости от буквальности интерпретации) понимаются как принадлежащие отдельным индивидам, а *голову* в примерах группы (12) — скорее как метафорическая часть «коллективного тела» (чем обеспечивается коллективная интерпретация).

(11) a. Наша цель — чтобы люди подняли головы, которые они опустили и боятся поднять.
b. А дальше дело было так: наши горцы, наши крестьяне вдруг подняли головы, которые им намеревались отсечь.

(12) a. Гордые тверские князья слишком рано и высоко подняли голову, которую … срубила ордынская сабля.
b. Мои сомнения тут же подняли голову, которую решительным голосом и властной рукой тут же вновь пригнул Дэн.

---

[12] П. В. Гращенков и А. Э. Гращенкова [5] после обзора существующих точек зрения помещают его в позицию спецификатора *n*P.

[13] Ср.: «*горем* — неопределённая именная группа, а *своим горем* — определённая» [14: 7].

References

1. *Andreyewsky A.* (1973), K upotrebleniyu mestoimeniya “svoy” v russkom yazyke [On the use of the pronoun *svoj* in Russian], Russian Language Journal / Русский язык, Vol. 27, no. 98, pp. 1–17. [Андреевский А. Б. (1973), К употреблению местоимения «свой» в русском языке, Russian Language Journal / Русский язык, т. 27, № 98, с. 1–17.]
2. *Champollion L.* (2017), Distributivity, collectivity and cumulativity, Companion to Semantics, Wiley.
3. *Chomsky N.* (1981), Lectures on government and binding: The Pisa lectures, Foris, Dordrecht.
4. *Fyhn S. K. A.* (2017), Eiendom på russisk, UiT Norges arktiske universitet, available at: https://munin.uit.no/handle/10037/13733.
5. *Grashenkov P., Grashenkova A.* (2006), Possessive Reflexives in Russian, Proceedings of FASL-15, Toronto, Ann Arbor.
6. *Honselaar W.* (1986), Reflections on the Russian reflexive possessive pronoun *svoj*, Dutch Studies in Russian Linguistics (= Studies in Slavic and General Linguistics), Vol. 8, pp. 235–248.
7. *Kratzer A.* (1998), More structural analogies between pronouns and tenses, Proceedings of SALT, Vol. 8, pp. 92–110.
8. *Kratzer A.* (2009), Making a pronoun: Fake indexicals as windows into the properties of pronouns, Linguistic Inquiry, Vol. 40, no. 2, pp. 187–237.
9. *Lakoff G.* (1970), Linguistics and Natural Logic, Synthese, Vol. 22, no. 1/2, pp. 151–271.
10. *Lyashevskaya O. N., Sharoff S. A.* (2009). Chastotnyy slovar’ sovremennogo russkogo yazyka [The Frequency Dictionary of Modern Russian Language]. Azbukovnik, Moscow. [Ляшевская О. Н., Шаров С. А. Частотный словарь современного русского языка. Азбуковник, Москва.]
11. *Lyutikova E. A.* (2014), Russkiy genetivnyy posessor i formal’nye modeli imennoy gruppy [Russian genitive possessor and formal models of the noun phrase], Tipologiya morfosintaksicheskikh parametrov. Materialy mezhdunarodnoy konferentsii “TMP 2014”, MSPU, Moscow, pp. 120–145. [Лютикова Е. А. (2014), Русский генитивный посессор и формальные модели именной группы, Типология морфосинтаксических параметров. Материалы международной конференции «ТМП — 2014», МПГУ, Москва, с. 120–145.]
12. *Lyutikova E. A.* (2016), Sintaksis imennoy gruppy v bezartiklevom yazyke [Noun Phrase Syntax in an Articleless Language], Lomonosov Moscow State University, Moscow. [Лютикова Е. А. (2016), Синтаксис именной группы в безартиклевом языке, МГУ им. М. В. Ломоносова, Москва.]
13. *Nedoluzhko A.* (2016), A new look at possessive reflexivization: A comparative study between Czech and Russian, Proceedings of the Workshop on Grammar and Lexicon: interactions and interfaces (GramLex), pp. 110–119.

14. *Paducheva E. V.* (1983), Vozvratnoe mestoimenie s kosvennym antetsedentom i semantika refleksivnosti [The reflexive pronoun with an indirect antecedent and the semantics of reflexivity], Semiotika i informatika, Vol. 21, VINITI, Moscow, pp. 3–33. [Падучева Е. В. (1983), Возвратное местоимение с косвенным антецедентом и семантика рефлексивности, Семиотика и информатика, т. 21, ВИНИТИ, Москва, с. 3--33.]
15. *Paducheva E. V.* (1985), Vyskazyvanie i ego sootnesennost' s deystvitel'nost'yu [The Utterance and Its Correspondence to Reality], Nauka, Moscow. [Падучева Е. В. (1985), Высказывание и его соотнесённость с действительностью, Наука, Москва.]
16. *Perevozchikova T.* (2018), Pronominal expression of possession in noun phrases in Russian, Czech, and Bulgarian, available at: https://slavicorp.ff.cuni.cz/wp-content/uploads/sites/144/2018/10/Perevozchikova_Slavicorp.pptx.
17. *Rappaport G. C.* (1986), On anaphor binding in Russian, Natural Language & Linguistic Theory, Vol. 4, no. 1, pp. 97–120.
18. *Stechow A. von* (2003), Feature Deletion under Semantic Binding: Tense, Person, and Mood under Verbal Quantifiers, Proceedings of NELS 33, University of Massachusetts, Amherst, pp. 377–403.
19. *Sudo Y.* (2012), On the semantics of phi features on pronouns, MIT, available at: https://dspace.mit.edu/handle/1721.1/77805.
20. *Timberlake A.* (1980), Reference conditions on Russian reflexivization, Language, Vol. 56, no. 4, pp. 777–796.
21. *Truswell R.* (2014), Binding theory, The Routledge Handbook of Syntax, Routledge, pp. 232–256.

CONTRAST AND COMPARISON RELATIONS IN RST FRAMEWORK: THE CASE OF RUSSIAN[1]

Toldova S. (toldova@yandex.ru)[1],
Davydova T. (tdadidik@gmail.com)[1],
Kobozeva M. (kobozeva@isa.ru)[2],
Pisarevskaya D. (dinabpr@gmail.com)[2]

[1]NRU Higher School of Economics, Moscow, Russia;
[2]FRC CSC RAS, Moscow, Russia

The Paper is devoted to a corpus study of the Contrast relation between discourse units in Russian. It is based on the data of the Ru-RSTreebank annotated within the framework of the Rhetorical Structure theory [Mann, Thompson 1988]. The research question is what cue phrases and lexical and grammatical patterns are used to express the Contrast relation as opposed to the Comparison relation. Since the simple connectives such as conjunctions *a* or *no* "but" and others are ambiguous it may be useful to single out specific cues for the Contrast relation and to find other linguistic features that can also help to differentiate Contrast and other relations, such as Comparison. The investigation of cues signalling different types of relations is an important issue for both automatic discourse mining and the theoretical researches of text coherence. We test several hypotheses presented in the reference literature on Russian against corpus data.

Keywords: discourse analysis, rhetorical structure theory, discourse connectives, corpus linguistics, corpus annotation

ОТНОШЕНИЯ КОНТРАСТА И СРАВНЕНИЯ В ТЕОРИИ РИТОРИЧЕСКОЙ СТРУКТУРЫ НА ПРИМЕРЕ РУССКОГО ЯЗЫКА

Толдова С. (toldova@yandex.ru)[1],
Давыдова Т. (tdadidik@gmail.com)[1],
Кобозева М. (kobozeva@isa.ru)[2],
Писаревская Д. (dinabpr@gmail.com)[2]

[1]НИУ ВШЭ, Москва, Россия;
[2]ФИЦ ИУ РАН, Москва, Россия

[1] This paper is partially supported by Russian Foundation for Basic Research (project No. 17-29-07033, 17-07-01477).

Статья посвящена корпусному исследованию отношения Контраста между дискурсивными единицами в русском языке. Используется материал корпуса Ru-RSTreebank, размеченного в рамках Теории риторической структуры [Mann, Thompson 1988]. Простые дискурсивные коннекторы, такие как союзы «а» или «но», неоднозначны и могут маркировать другие отношения, например, Сравнение. Поэтому цель нашего исследования — найти специфические маркеры для Контраста и дополнительные лингвистические параметры, которые помогут дифференцировать Контраст и Сравнение. Мы проверяем несколько гипотез, которые упоминаются в теоретических работах по русскому языку, на материале упомянутого корпуса. Исследование маркеров, указывающих на определенные типы отношений, является важной проблемой как для автоматического анализа дискурса, так и для теоретических исследований связности текста.

Ключевые слова: дискурсивный анализ, теория риторической структуры, дискурсивные маркеры, корпусная лингвистика, корпусная разметка

1. Introduction

It is generally assumed that discourse is not a mere chain of sentences, it is coherent [Hobbs 1985]. Text coherence presupposes that there are relations between text spans. One of the theories modeling discourse as a hierarchical structure built via rhetorical relations between text spans is the Rhetorical Structure theory (RST) [Mann, Thompson 1988]; [Taboada, Mann 2006]. The definitions of relations within RST framework are not based on the explicit linguistic features but formulated in terms of speaker's intentions and its effect on the reader. The aim of our research is to study the contrast relation (Contrast) within RST theory and to establish the repertoire of the linguistic devices (signals) that express it in Russian.

Conjunctions and other cues for Contrast are often ambiguous and unclear:

(1a) [*S tochki zreniya yazyka zdes' vs'e pravil'no,*] [***a*** *vot s tochki zreniya sootneseniya objektov real'nogo mira—net.*]
'[From the point of view of the language everything is correct here,] [**but** from the point of view of the reference to the real world entities it is not.]'

(1b) [*Funt i evro slegka ukrepilis',*] [***a*** *jena slegka upala.*]
'[The pound and the euro strengthened a little bit,] [**while** yen fell slightly.]'

The connective *a* 'while/but/and' is used both in (1a) and in (1b), however (1a) is a contrast relation while (1b) is a comparison relation according to the annotation suggested in Ru-RSTreebank (https://linghub.ru/ru-rstreebank/, a Russian corpus annotated for rhetorical relations [Pisarevskaya et al. 2017].

The present research seeks to identify linguistic features that can help to differentiate the contrast and the comparison relations. The results of this study can be helpful for the relation type recognition in the discourse parsing systems.

We use the Ru-RSTreebank as a source of data for Contrast and Comparison. As the data of RU-RSTreebank have shown, a considerable number of discourse connectives

(DC) used for Contrast relation are ambiguous as they mark other relations as well. According to the numerous studies on the adversative conjunctions and contrast constructions in Russian [Shvedova 1980]; [Uryson 2004], there is an additional set of features signalling the contrast relation. These include, among others, syntactic parallelism and additional specifiers such as particles of adverbial expressions.

The present work offers a corpus-based analysis of Contrast and Comparison in Russian News and Scientific texts. Its main Hypothesis is that the features suggested in the works on the contrast relations between clauses in Russian are also valid for detecting Contrast within RST framework. The main tasks are to establish the existence of Contrast vs. Comparison-specific connectives and to identify additional features are that are specific for these two relations.

For this purpose, we have compiled a set of corresponding discourse units from the Ru-RSTreebank and annotated them manually on different features including discourse markers (DM), additional specifiers and lexical repetitions. In our work we describe the most prominent features that we have identified.

The present paper is organized as follows. A brief description of the framework within which Contrast and Comparison determined is given in **2.1**. In 2.2. we discuss the features of conjunctions and other markers of Contrast mentioned in the literature on Russian. The overview of theoretical works on Russian serves as the basis for the hypothesis concerning signals of Contrast and Comparison that are formulated in **2.3**. **Section 3** describes the data used for the corpus research. In **Section 4** we present a survey of the signals enumerated in **2.3** that we have found in our corpus data. **Section 5** provides the conclusions.

2. Background

2.1. Contrast and Comparison rhetorical relations in RST framework

According to the RST [Mann, Thompson 1988], a text is organized as a tree whose nodes are text spans (discourse units (DU)). DUs are united into the spans of a higher level if there is a rhetorical relation between them. Relations can be multinuclear or mononuclear depending on the types of the spans, e.g.:

[*Peter went home*]$_{\text{nucleus}}$ [*because he was tired*]$_{\text{satellite}}$ vs. [*Peter went home*]$_{\text{nucleus}}$ [*while Tom stayed at work*]$_{\text{nucleus}}$.

The two relations under discussion, namely, Contrast and Comparison are multinuclear. While Contrast is the relation from the original set given in [Mann, Thompson 1988], Comparison was added later in [Carlson, Marcu 2001]. Below are the definitions of Contrast (a) and Comparison (b) described in [Carlson, Marcu 2001]:

a. “In CONTRAST, two or more nuclei come in contrast with each other along some dimension. The contrast may happen in only one or few respects, while everything else can remain the same in other respects.”
b. “In COMPARISON, two textual spans are compared along some dimension, which can be abstract. The relations can convey that some abstract entities

that pertain to the comparison relation are similar, different, greater-than, less-than, etc. In the case of a comparison relation, the spans, entities, etc. are not in contrast."

Thus, the difference between these relations is opaque (cf. "come in contrast" for Contrast—"are different" for Comparison). Moreover, many DC that are used for Contrast are ambiguous between Contrast and Comparison. For example, English *while* is a marker for both—Contrast (2) and Comparison (3) [ibid.]:

(2) [*But the staff at some of those locations will be slashed*] [***while*** *at other the workforce will be increased.*]

(3) [*Kellogg's current share is believed to be slightly under 40%*] [***while*** *General Mills' share is about 27%.*]

The indirect evidence for the fact that Contrast and Comparison are similar to some extent is the fact that for some approaches, e.g. see PDTB, [Prasad et al. 2007: 27]) Contrast is a subclass of Comparison (**Fig. 1**).

Different types of connectives are important cues that help to recognise particular relations and to differentiate them [Danlos 2018]. There are special lexicons compiled for different languages (e.g. French [Roze et al. 2012], Czech [Mirovsky et al. 2017], DiMLex for German [Scheffler, Stede 2016], DiMLex-Eng for English [Das et al. 2018], PDTB based lexicons for French [Laali, Kosseim 2017], and for Portuguese [Mendes et al. 2018].

In our work, we treat the relations annotated in the Ru-RSTreebank according to their definitions as the starting point and try to find specific connectives and other signals for Contrast and Comparison.

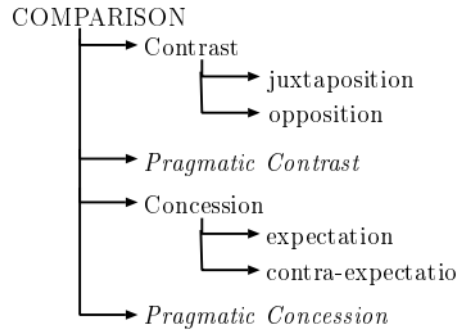


Fig. 1. A fragment of the Scheme of the sense tags for discourse relations annotation in PDTB

2.2. Studies on Contrast and Comparison relations in Russian

The majority of works on Contrast in Russian concerns a special subclass of the conjunctions used in coordination of two clauses. These are the so-called adversative conjunctions such as *a* 'while, and, but', *no* 'but' etc. (for the list of the conjunctions see, e.g. [Shvedova 1980]; [Apresyan, Pekelis 2012], [Sannikov 1989]). These conjunctions are highly polysemous (for other languages too, cf. [Spenader, Lobanova 2009]). Therefore, some other supporting linguistic features are discussed in the literature. [Apresyan, Pekelis 2012] mention discourse particles, e.g. particle *zhe* translated as 'just, but, and' in some dictionaries. [Shvedova 1980] and others mention grammatical parallelism and lexical oppositions. Besides, there are different additional lexical markers (referred to as 'discourse specifiers' further). Below we give a list of different features mentioned in the literature (**Section 2.3**).

Many studies are devoted to the fine-grained taxonomy of various senses of the conjunctions expressing Contrast [Apresyan, Pekelis 2012]; [Zaliznyak, Mikaelyan 2005]. There are also some general works concerning the semantics of particular conjunctions [Uryson 2004]; [Uryson 2012]; [Kobozeva 2011]; [Inkova-Manzotti 2001].

There have been attempts to formulate the notions of Contrast or Comparison in abstract terms. Thus, O. Inkova-Manzotti defines the relation of Contrast as a semantic configuration in which the Speaker represents two states of affairs as incompatible [ibid: 80]. In [Uryson, 2004] a mental operation of Comparison is described: it involves a human subject and at least two objects X and Y that have a common parameter Z (often implicit). Within this approach Contrast can also be treated a subtype of Comparison. In our discussion, we will use the notions of objects and parameters of comparison for both of the relations.

2.3. Summary of the Features Associated with Contrast

The primary cues for Contrast and Comparison are some of the coordinatives and other types of connectives, and some types of parentheticals. The following features are usually regarded as additional cues:

1. "**syntactic parallelism**" (syntactical similarity) of the coordinate components (DU), e.g. [Kreydlin, Paducheva 1974]; [Uryson 2004]; [Asher, Lascarides 2003];
2. the so called **discourse specifiers**—lexical items that are not generally recognized as DM but which tend to contribute to Contrast, e.g. *even, still, too* for English [Spenader, Lobanova 2009] and *zato* 'but then'*, vse zhe* 'yet', *vdrug* 'suddenly'*, dazhe* 'even' etc. [Shvedova 1980]. In RusGram [Pekelis 2018], many of these items belong to a special part of speech—correlates. [Siyjuan', Sheremet'eva 2018] distinguish discourse specifiers used for:
3. lexical **expression of negation:** particle *ne 'not'* [Syjuan', Sheremet'eva 2018], negative pronouns and some others;
4. **lexical parallelism**: synonyms and word repetitions (e.g. [Inkova-Manzotti 2001]); **antonyms, or lexical opposition**; there is some theoretical description of antonyms in Contrast, e.g. 3d type of constructions in [Siyuan', Sheremet'yeva 2018], as well as some applications of lexical oppositions

to automatic Contrast detecting [Harabagiu et al. 2006]; [Marcu, Echihabi 2002]; [Murphy et al. 2009]. On the other hand, [Spenader, Stulp 2007], [Feltracco et al. 2018] note that lexical opposition is not common in cases of Contrast.

3. Data

3.1. The corpus

The current study is based on the Ru-RSTreebank (https://linghub.ru/ru-rstreebank/) [Pisarevskaya et al., 2017]; [Toldova et al., 2018]. It consists of 179 texts (203,287 tokens in total) and represents the genres of news and popular science (79 texts) and scientific papers (100 texts). The paper uses the most recent unpublished version of the corpus, available on request. The corpus was annotated by several annotators, with the last Krippendorff's unitized alpha measurement of 81%, which is a good inter-annotator agreement level.

3.2. The data

For our research we use examples of Contrast and Comparison from Ru-RSTreebank. There are about 570 examples of Contrast and 234 examples of Comparison. The following manual was used in the annotation of discourse relations: https://rstreebank.ru/assets/docs/Manual_for_ru_RSTreebank_Annotation.pdf. Although in the manual three DC examples for Contrast are suggested (*a* 'but, and', *no* 'but', *nesmotrya na* 'despite of'), DC of this relation are much more diverse (about 50 for Contrast).

3.3. Annotation for discourse relation signals

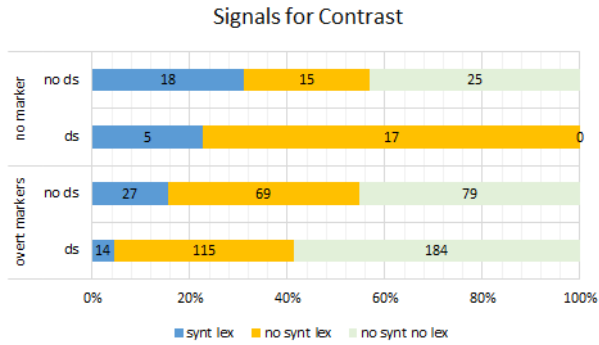
All the examples in our dataset were manually annotated for different types of linguistic signals of rhetorical relations. Firstly, the primary markers were singled out. By primary we mean those expressions that serve as overt markers for the relation between two DUs. Besides, the examples were annotated for supplementary signals listed in **2.3** and we marked different discourse specifiers (see **4.2**). Then, we marked the presence of lexical repetitions, presence of synonyms (quasi-synonyms and semantically close expressions), hypernyms and opposite lexical expressions (e.g. antonyms, conversives, for the types of opposites see also [Feltracco et al. 2018]).

4. Contrast and Comparison connectives in Ru-RSTreebank

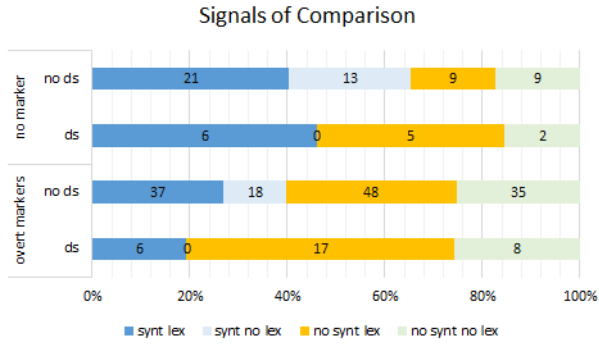
4.1. The general statistics for primary connectives and supplementary features

As it has been reported in the literature, there might be a lot of examples where the discourse relations are expressed implicitly without an overt DC (c.f. [Taboada, Das 2013]). In our set, 80 cases out of 569 (about 14%) have no overt primary markers for Contrast, while there are 65 examples out of 234 (about 28%) for Comparison. Thus, in news and scientific texts there is a tendency to express Contrast overtly. For Comparison some other devices are used.

Below, we summarise the data for different combinations of signals in Contrast vs. Comparison. We treat the syntactic parallelism separately (it was manually annotated by experts). We consider the presence of different semantically related expressions or lexical repetitions in two parts of a relation as lexical parallelism. The general statistics is given in Fig. 2 ((a) for Contrast and (b) for Comparison):



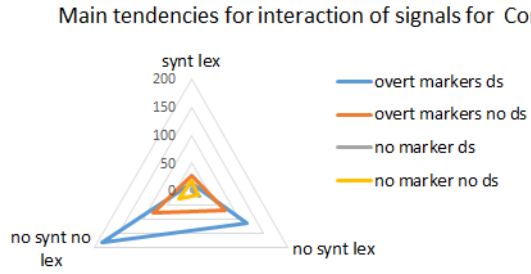
a. Contrast



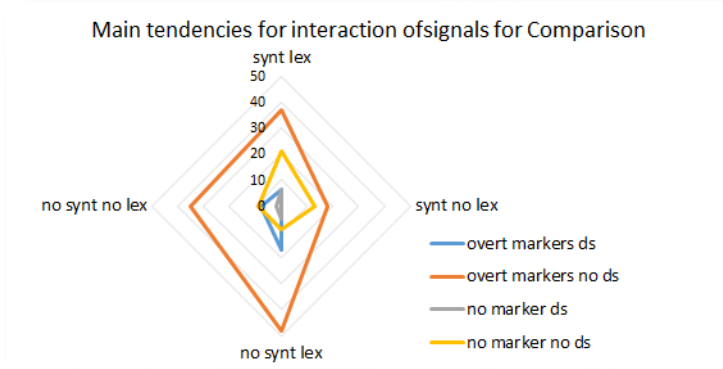
b. Comparison

Fig. 2. The distribution of different signals for Contrast and Comparison

As our data shows, syntactic parallelism is more frequent with Contrast (11%); it is also quite frequent with Comparison (26%). The main tendencies can also be seen in Fig. 3. Contrast prefers the overt markers without syntactic or lexical parallelism (38%). It is supported by a large number of discourse specifiers (cf. the bleu contour corresponding to the overt markers + discourse specifier occupies the largest area and stretches towards 'no syntactic or lexical parallelism' direction). Moreover, lexical parallelism coincides with syntactic in Contrast. On the contrary, Comparison prefers lexical parallelism, it is quite frequent irrespective of the syntactic one, c.f. the orange radar for overt markers without discourse specifiers. The latter occupies the largest area on the chart for Comparison. The discourse specifiers are rare in Comparison (the blue radar occupies one of the smallest areas).



a. Contrast



b. Comparison

Fig. 3. The main tendencies in signals for Contrast vs. Comparison distribution (a radar chart)

As for primary DMs, Table 1 provides the statistics on the most frequent DC for each relation as compared to the frequency for the opposite relation:

Table 1. DC frequency in Ru-RSTreebank

Discourse Connective	Number in Contrast	Number in Comparison
no 'but'	143 (28.98%)	2 (0.93%)
odnako 'however'	133 (27.14%)	0
a 'but, and'	49 (10%)	18 (13.02%)
ne… a 'not … but'	30 (6.12%)	0
zhe "just, but, and"	13 (2.45%)	13 (6.05%)
tem ne meneye 'nevertheless'	12 (2.45%)	0
yesli … to 'if … then'	10 (2.04%)	4 (0.94%)
v to vremya kak 'while'	9 (1.84%)	9 (3.72%)
(*i*) *v to zhe vremya* '(and) at the same time'	6 (1.22%)	3 (1.41%)
s drugoy storony 'on the other hand'	6 (1.22%)	1 (0.47%)
khotya 'although'	6 (1.22%)	0
chem … tem 'the … the'	0	6 (2.79%)
v otlichiye ot X 'unlike X'	4 (0.82%)	6 (2.79%)
analogichnyy 'similar'	0	4 (0.94%)
kak … tak i 'as … so'	0	4 (0.94%)
i 'and'	1 (0.20%)	4 (0.94%)

Other connectives for Contrast occur in less than 1% of cases, e.g. *v protivnom sluchaye* 'otherwise', *naprotiv* 'on the contrary', *togda kak* 'whereas', *v otlichije ot X* 'in contrast to', *vmesto X* 'instead X', *nesmotrya na X* 'despite of X', *vprochem* 'however' etc.

There are specific DMs for Comparison in our data set that express the similarity of two entities' properties or two states of affairs like *analogichnyj* 'analogous', *pokhozhij* 'similar to', *takoj zhe* 'similar', *kak i* 'as'. Correlative constructions (DCs consisting of two parts situated in different text spans) are among these markers (e.g. *s odnoy storony … (a) s drugoy (storony)* 'on the one hand … (a) on the other (hand)'). The majority of correlatives found in our data serve as DMs for both relations.

While Contrast has quite a distinct DM profile (the two most frequent DMs cover more than 50% examples), there is no dominant specific DM for Comparison. The most frequent conjunction covers only 13% of examples. Moreover, it is used for Contrast in the same proportion of cases. The DM diversity for Contrast is 0,1 (48 markers for 464 cases with overt DM), and it is higher for Comparison, its value is 0,5 (90 markers for 169 cases).

Another important Contrast property is that it is often marked by a sequence of connectives. One of them is a simple conjunction, the second one can be a multiword parenthetical expression or a series of particles (e.g. *no* 'but' + *tem ne meneye* 'nevertheless'):

Although there are specific DCs for each relation, there are a lot of connectives in our data that occur both with Comparison and Contrast, which means that there should be other signals that help to differentiate the two relations.

4.2. Lexical specifiers

[Shvedova 1980] and others (see also 2.2) report the usage of additional expressions specifying some aspects of the opposed states of affairs. These discourse specifiers can be helpful for differentiating the two relations. The general statistics in 4.1 shows that Contrast usually has these additional expressions (about 85% of cases for Contrast, while only 7% for Comparison). Below are some examples of discourse specifiers, grouped in different classes:

a) **temporal words** and constructions—*v to zhe vremya* 'at the same time', *vse yeshche* 'still', *uzhe* 'already'; their function is to mark the unexpected simultaneous existence of two opposed states of affairs or to emphasize the opposition of the states of affairs in different time slots.
b) markers of **epistemic modality**—*deystvitel'no* 'really', *bezuslovno* 'absolutely', *yavnyy/yavno* 'explicit'; they emphasize the reality of an unexpected state of affairs;
c) **anaphoric expressions, e.g. different types of pronouns and quantifiers:** *etot* 'this', *dannyy* 'given', *podobnyy* 'similar', *drygoj* 'other one'; one of their functions is to mark the NPs whose properties are contrasted (e.g. contrastive topics); there are many examples where they go in pairs (c.f. *odin/etot—drugoj* 'one of them/this—the other one', *nekotoryje* 'some of them'—*drugije* 'others')
d) content words such as **verbs of contradiction**—*vozrazhat'* 'to object', *protivorechit'* 'to contradict';
e) focus and topic re-activation **particles**—*imenno 'precisely', zhe* 'but, as to', *dazhe* 'even', *lish'* 'only', *tol'ko* 'only', *vse ravno* 'still'.

A special attention should be drawn to the proportion of negative particles ***ne* 'not'** and other words of negation (*e.g. nevozmozhno* 'impossible'), negative pronouns etc.

4.3. Lexical parallelism

In accordance with the definitions, relations Comparison and Contrast involve objects (X and Y) that are compared, and a parameter of comparison (Z). The implication is that some of the lexemes in the two parts of the relations are semantically related. Lexical repetitions and synonyms can specify the general grounds of comparison and the opposite notions express the difference. We treat synonymy and semantic opposition very widely here (see [Lyons 1977]). Table 2 shows the statistics on different types of lexical parallelism in our data:

Table 2. The distribution of different types of lexical parallelism in Contrast vs. Comparison

	Contrast	Comparison
Semantically opposed expressions	200 (**35%**)	42 (**18%**)
(Quasi)-synonyms, hypernyms	63 (**17%**)	94 (**36%**)
Repeated and cognate words	142 (**25%**)	62 (23%)
Total	280	108

Synonyms and repetitions are more typical for Comparison cases against (X-squared = 43.595, df = 1, p-value = 4.039e-11 (Pearson's Chi-squared test with Yates' continuity correction). They can be both the objects (X/Y) and the parameter of comparison (Z) (4), while in Contrast they are used for denoting only the objects of comparison (5):

(4) [*Naiboleye kharakternymi chertami obraza ideal'noy zhenshchiny russkoyazychnogo reklamnogo parfyumernogo diskursa* (*X*) *yavlyayutsya: privlekatel'nost', seksual'nost', zhenstvennost'* <…> (*Za*).] [*Bol'shinstvo vysheperechislennykh priznakov kharakterny i dlya nemetskoyazychnogo diskursa* (*Y*). *Predstavleniya ob ideal'noy zhenshchine takzhe svyazany s takimi ponyatiyami, kak privlekatel'nost', <…> unikal'nost'.*(*Zb*)]
'The most characteristic features of the image of an ideal woman of the Russian-language advertising perfume discourse (*X*) are: attractiveness, sexuality, femininity <…>. (*Za*) Most of the above symptoms are characteristic of the German-language discourse (*Y*). Ideas about the ideal woman are also associated with such concepts as attractiveness,<…> uniqueness.(*Zb*)'

(5) [*Vo frantsuzskom yazyke* (*X*) *prevaliruyut zvuki, obrazuyemyye v peredney chasti golosovogo apparata* (*Za*).] [*V angliyskom yazyke* (*Y*), *naprotiv, preobladayut glasnyye zvuki zadnego ryada* (*Zb*).]
'In French (X), the sounds formed in front of the voice apparatus (Za) prevail. In English (Y), on the contrary, backward vowel sounds (Zb) prevail.'

On the contrary, semantically opposed expressions can denote both X and Y or Za and Zb for Contrast, while they stand for the object of comparison only in Comparison.

Besides, Contrast can be established between the two DUs where the opposition is not between the two states of affairs, but between the implicit expectations of the first DU and the real state of affairs (stated in the second DU). For these cases no lexical parallelism is expected:

(6) [*Privivki ostayutsya naiboleye effektivnoy meroy preduprezhdeniya epidemiologicheskikh zabolevaniy.*] [*Odnako mnogiye roditeli otkazyvayutsya vaktsinirovat' svoikh detey.*]
'Vaccinations are still the most effective preventive measure for epidemiological diseases.' 'However, many parents refuse to vaccinate their children.'

In (6) the second span is in opposition not with the first one as such, but with its implication that has no explicit expression—'people need to be vaccinated to prevent diseases'. As the result, there are fewer repetitions of words in Contrast, since the second segment does not contrast directly with the first one.

5. Conclusion

In this paper, we present an analysis of different signals of the Contrast relation in Russian as compared to those used for Comparison. We examine different cues mentioned in the literature as marking the corresponding relations against our corpus

data. For this purpose, we have annotated different types of signals for all the examples. As a result, we compiled a list of Contrast and Comparison DMs based on RuRSTreebank (48 and about 90 elements respectively).

The two relations under discussion have much in common, e.g. the DC used in both relations are often ambiguous or they can be expressed implicitly. In these cases, subsidiary signals are used. These are special lexical markers (particles, adverbs etc.), syntactic and lexical parallelism. Because of these similarities, it may be difficult to draw a line between the two relations. The distinction between these two relations should be presented as a scale rather than a mere dichotomy.

However, our data confirm the role of various additional signals for detecting Contrast and differentiate it from Comparison, like discourse specifiers, markers of negation and semantically opposed expressions.

References

1. *Apresyan V. Yu., Pekelis O. Ye.* (2012), Coordinate conjunctions [Sochinitel'nyye soyuzy]. Materials for the project of Russian grammar corpus description [Materialy dlya proyekta korpusnogo opisaniya russkoy grammatiki], avaliable at: http://rusgram.ru, M., As a manuscript [Na pravakh rukopisi].
2. *Asher N., Lascarides A.* (2003), Logics of conversation. Cambridge University Press.
3. *Carlson L., Marcu D.* (2001), Discourse tagging reference manual. Technical report, Information Science Institute, Marina del Rey, CA. ISI-TR-545.
4. *Feltracco A., Magnini B., Jezek E.* (2018), Lexical Opposition in Discourse Contrast, Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10–11.
5. *Harabagiu S., Hickl A., Lacatusu F.* (2006, July), Negation, contrast and contradiction in text processing, AAAI, Vol. 6, pp. 755–762.
6. *Hobbs J. R.* (1985), On the coherence and structure of discourse. Technical Report 85–37, Center for the Study of Language and Information (CSLI), Stanford, CA.
7. *Inkova-Manzotti O.* (2001), Opposition connectives in French and Russian [= Connecteurs d'opposition en français et en russe] [Konnektory protivopostavleniya vo frantsuzskom i russkom yazykakh], M., Informèlektro.
8. *Kobozeva I. M.* (2011), Conjunctions as markers of rhetorical relations in discourse: Russian conjunction "i" [Soyuzy kak markery ritoricheskikh otnosheniy v diskurse: russkiy soyuz «i»], L'analisi linguistica e letteraria, 19, № 2, pp. 365–387.
9. *Kreydlin G. E., Paducheva E. V.* (1974), Semantics and syntactic properties of conjunction 'a' [Znacheniye i sintaksicheskiye svoystva soyuza a]. Scientific and technical information [Nauchno-tekhnicheskaya informatsiya], Seriya 2, 9, pp. 31–37.
10. *Laali M., Kosseim L.* (2017), Automatic Mapping of French Discourse Connectives to PDTB Discourse Relations, Proceedings of the SIGDIAL 2017 Conference, Saarbruecken, Germany, pp. 1–6.
11. *Lyons J.* (1977), Semantics. V. 1. Cambridge University Press.
12. *Mann W. C., Thompson S. A.* (1988), Rhetorical Structure Theory: Toward a Functional Theory of Text Organization, Text 8, 3, pp. 243–281.

13. *Marcu D., Echihabi A.* (2002), An unsupervised approach to recognizing discourse relations, Proceedings of the 40th annual meeting of the association for computational linguistics.
14. *Mendes A., del Rio I., Stede M., Dombek F.* (2018), A Lexicon of Discourse Markers for Portuguese—LDM-PT, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018), Miyazaki, Japan, pp. 4379–4384.
15. *Murphy M. L., Paradis C., Willners C., Jones S.* (2009), Discourse functions of antonymy: A cross-linguistic investigation of Swedish and English, Journal of pragmatics, 41(11), pp. 2159–2184.
16. *Pekelis O. Ye.* (2018), Correlates (that, so and others) [Korrelyaty (to, tak i dr.] Materials for the project of Russian grammar corpus description [Materialy dlya proyekta korpusnogo opisaniya russkoy grammatiki], avaliable at: http://rusgram.ru. As a manuscript [Na pravakh rukopisi], M, 2018.
17. *Pisarevskaya D., Ananyeva M., Kobozeva M., Nasedkin A., Nikiforova S., Pavlova I., Shelepov A.* (2017), Towards building a discourse-annotated corpus of Russian, Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference "Dialogue 2017", pp. 194–204.
18. *Prasad R., Miltsakaki E., Dinesh N., Lee A., Joshi A., Robaldo L., Webber, B. L.* (2007), The penn discourse treebank 2.0 annotation manual.
19. *Sannikov V. Z.* (1989), Russian coordinate construciotns: semantics, pragmatics, sintaksis [Russkiye sochinitel'nyye konstruktsii: semantika, pragmatika, sintaksis], M, "Nauka".
20. *Shvedova, N. Yu. ed.* (1980), Russian grammar [Russkaya grammatika]. V dvukh tomakh. AN SSSR Institut russkogo jazyka, M.: Nauka, 1980.
21. *Siyuan' K. H., Sheremet'yeva Ye. S.* (2018), The specificity of structures built on the basis of a combination of union and with the concretizations 'in fact', 'essentially', 'in essence' [Spetsifika konstruktsiy, stroyashchikhsya na osnove sochetaniya soyuza a s konkretizatorami posuti, posushchestvu, vsushchnosti], Scientific dialogue [Nauchnyy dialog] (1).
22. *Spenader J., Lobanova A.* (2009, January), Reliable discourse markers for contrast relations. In Proceedings of the Eighth International Conference on Computational Semantics (pp. 210–221), Association for Computational Linguistics.
23. *Spenader J., Stulp G.* (2007, January), Antonymy in contrast relations, Seventh International Workshop on Computational Semantics, Vol. 3, p. 100.
24. *Taboada M., Das D.* (2013), Annotation upon annotation: Adding signalling information to a corpus of discourse relations, Dialogue and Discourse 4(2), pp. 249–281.
25. *Taboada M., Mann W. C.* (2006), Rhetorical structure theory: Looking back and moving ahead, Discourse studies, 8(3), pp. 423–459.
26. *Toldova S., Kobozeva M., Pisarevskaya D.* (2018), Automatic mining of discourse connectives for Russian, Conference on Artificial Intelligence and Natural Language, pp. 79–87.
27. *Uryson E. V.* (2004), Some meanings of conjunction A in the light of modern semantic theory [Nekotoryye znacheniya soyuza A v svete sovremennoy semanticheskoy teorii]. Russian language in scientific coverage [Russkiy yazyk v nauchnom osveshchenii], (2), 17.

28. *Uryson E. V.* (2012), Conjunctions, connectives, and the valence theory [Soyuzy, konnektory i teoriya valentnostey]. In Computational linguistics and intellectual technologies [Komp'yuternaya lingvistika i intellektual'nyye tekhnologii], pp. 627–638.
29. *Webber B. L., Joshi A. K.* (1998), Anchoring a lexicalized tree-adjoining grammar for discourse. Discourse Relations and Discourse Markers: Proceedings of the Conference, Association for Computational Linguistics, pp. 86–92.
30. *Zaliznyak, A. A., Mikaelyan, I. L.* (2005) Russian Conjunction "a" as a language-specific word [Russkiy soyuz a kak lingvospetsifichnoye slovo], Computational Linguistics and Intellectual Technologies, Proceedings of the International Conference "Dialogue 2005", pp. 153–159.

A COMMUNICATIVE ROBOT TO LEARN ABOUT US AND THE WORLD

Vossen P. (piek.vossen@vu.nl),
Baez S. (selene.baez.santamaria@gmail.com),
Bajcetić L. (lenka.bajcetic@gmail.com),
Basić S. (suz.basic@gmail.com),
Kraaijeveld B. (bram.kraaijeveld@gmail.com)
Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

We describe a model for a robot that learns about the world and her companions through natural language communication. The model supports open-domain learning, where the robot has a drive to learn about new concepts, new friends, and new properties of friends and concept instances. The robot tries to fill gaps, resolve uncertainties and resolve conflicts. The absorbed knowledge consists of everything people tell her, the situations and objects she perceives and whatever she finds on the web. The results of her interactions and perceptions are kept in an RDF triple store to enable reasoning over her knowledge and experiences. The robot uses a theory of mind to keep track of who said what, when and where. Accumulating knowledge results in complex states to which the robot needs to respond. In this paper, we look into two specific aspects of such complex knowledge states: 1) reflecting on the status of the knowledge acquired through a new notion of thoughts and 2) defining the context during which knowledge is acquired. Thoughts form the basis for drives on which the robot communicates. We capture episodic contexts to keep instances of objects apart across different locations, which results in differentiating the acquired knowledge over specific encounters. Both aspects make the communication more dynamic and result in more initiatives by the robot.

Keywords: multimodal communication, social robots, knowledge acquisition and modeling

1. Introduction

Human-robot communication is necessary for collaboration in future societies. It is vital to build social relationships between humans and robots, to create a common ground from shared experiences and knowledge, and to build up trust. Natural language communication in multimodal environments plays a crucial role for establishing such a relationship.

Both machines and humans make errors in dealing with real-life situations. We have therefore designed a robot model that assumes that information can be wrong, has gaps and even conflicts. To deal with this, the robot needs to learn about us and the world: fill gaps and get feedback on errors and confirmation in case of uncertainty. In previous work, [17], we described a female robot model, named *Leolani*, that supports open-domain learning through communication, having a drive to learn new concepts and make new friends. The absorbed knowledge consists of everything people tell her, the situations and objects she perceives, and what she finds on the web. The results of her interactions and perceptions are kept in a triple store, enabling her to reason over her knowledge and experiences. The robot uses a theory of mind [7] to record the learning provenance (who said what, when and where).

Learning through communication results in complex knowledge states that may contain errors, false statements, conflicts or interpretations that differ across different people and situations. The functioning of the robot is at risk if the acquired information is taken as it is. It is therefore necessary that the robot knows how to reflect on the state of her brain and takes initiatives to improve this state. Furthermore, situations need to be interpreted within the unique context of an interaction. Knowledge that is accumulated within such a situation needs to be related to this context as well, e.g. my laptop is likely to be found in my office but not in other places. By differentiating these contexts, possible conflicts can be prevented and communication will be easier as there is less ambiguity and fewer conflicts.

In this position paper, we therefore describe an extension to *Leolani* that reflects on the acquired knowledge by producing so-called *thoughts*. These thoughts result in drives to improve the state of brain through communication. The robot takes initiatives to involve her human sources for that purpose. The robot model also includes a notion of context that allows us to identify different situations and the objects within it. This results in fewer conflicts and less confusion (uncertainty) and therefore more healthy states of the brain, better definitions of relevance and less need to communicate.

This paper is structured as follows: In **Section 2**, we summarize related work on social robot communication. Our data model and the way in which the robot learns through communication are described in **Section 3**. In **Section 4**, we describe the thoughts and the corresponding drives that lead to initiatives to communicate. For dealing with the world and humans, the robot needs to represent and memorize the contexts in which she encounters people. In **Section 5**), we explain how instances of contexts are created and how these result in more fine-grained and differentiated representations of situations. We conclude and discuss future work in **Section 6**.

2. Related work

Mavridis [11] gives an overview of natural language processing technologies in human-robot interaction and challenges to be tackled, including 'theory of mind', open-domain communication, varied speech acts, symbol grounding and multiple-turn dialogues. Most human-robot communication models still only handle basic communication using one or two speech acts, limited symbol grounding and single turns.

Recently, there has been an increase in chat systems that can be used for human-robot communication. Many of these models are either scripted ([14],[1]) or based on neural networks (often sequence-to-sequence (seq2seq) models), see for example: the dialogue systems built from the Ubuntu dialogue corpus [9], CoQA corpus [12], Twitter [8], the Persona-Chat dataset [18] and movie dialogues ([13] and [16]). Both types can be seen as extremes on the scales of control and fluency. Scripted conversations allow developers to control interaction, but knowledge needs to be defined manually and the conversation is limited, not robust and rarely fluent. Seq2seq models, on the other hand, are robust, fluent and respond to any input, but cannot be controlled or explained. More importantly, no explicit knowledge is derived from these conversations.

Our model is designed for open communication with the explicit result of acquiring knowledge and building a social relationship. It is designed for generic purposes defined at a low level that can support any high-level goal. This architecture provides our model with more flexibility and fluency than strictly scripted models, while the communication is more purposeful than in seq2seq models.

Another important aspect of human-robot communication is mixed-initiative interaction. Many systems leave the initiative to the human and only respond when prompted. They do not have an intrinsic drive to communicate unless they are scripted for some task, e.g. to take your order. Little work has been done on the implementation of basic drives to communicate in the systems. Our model implements low-level drives, such as the need to fill knowledge gaps and resolve conflicts and uncertainty. These drives make the communication active, lively and purposeful. We do not intend the model to fully capture human dialogue. Rather, dialogues serve to satisfy the robot's drives.

In our previous paper [17], we focused on a robot with a theory of mind [7] that acquires knowledge from people but stores the knowledge as claims from these people. In this paper we add the notions of *thought* and *context*. A thought represents a brain state that triggers drives. A context is an episodic element that explicitly gathers everything *Leolani* learns in connection with specific situations. *Thoughts* and *context* pave the way for new cognitive functionalities like relevance and permanence, as well as new intentions that exploit contextual information to drive the conversation. They also equipt the robot with new initiatives for communication and at the same time reduce conflicts, ambiguities and define relevance.

3. Data Model

3.1. Model description

Our robot model architecture is shown in **Figure 1**. We defined four layers:
1) a sensor processing layer,
2) a communication layer that responds to sensor input or inner drives,
3) a language processing layer to deal with questions and statements, and
4) a knowledge layer that queries or stores the result of communication or accesses the Web.

We utilize several ready-made modules in the sensor processing layer: WebRTC [3] for speech detection, the Inception neural network [15] for object recognition, OpenFace [2] for face recognition, and Google Cloud Speech-to-Text API [5] for speech recognition. We use the outputs of these processing modules as inputs to the other layers. Therefore, we do not address potential conflicts and ambiguities in the signal layer itself, but try to resolve them in the higher-level layers.

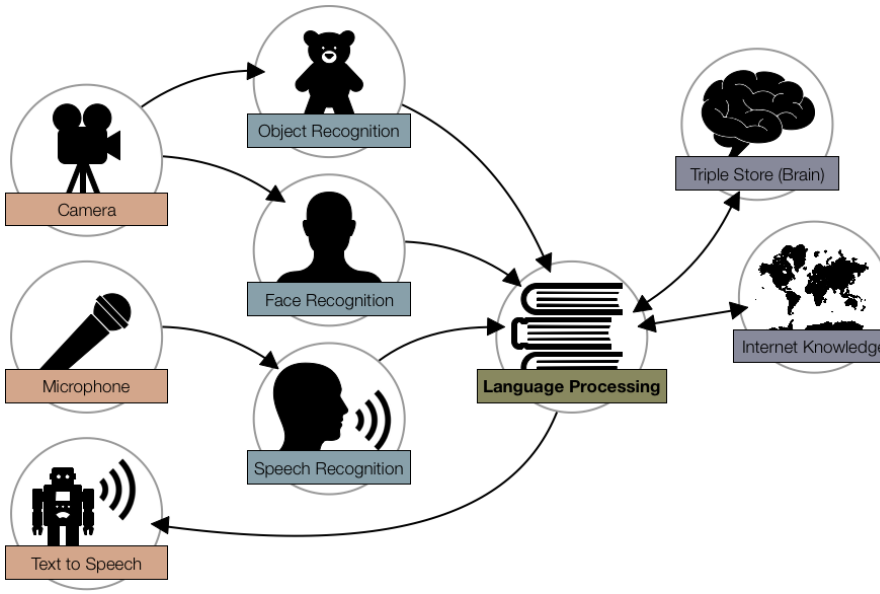


Figure 1: Global architecture of the robot model

In this paper, we focus on modeling the result of communication in an RDF triple store (called 'the brain'), which stores all interpretations of experiences. The brain forms the basis for the drives of the robot to communicate. We use the Grounded Representation and Source Perspective (GRaSP) model [4] as a basis for representing content, communication and sources. We have adapted GRaSP to deal with perception and communication by robots. Statements communicated to the robot are mapped to RDF representations, which are stored together with the source of each statement.

The model also stores the perspective of the source on a property expressed in the statement. The possible perspective values are denial/confirmation, sentiment/emotion, and certainty. Besides processing statements, the robot handles questions as SPARQL queries against the knowledge in the brain.

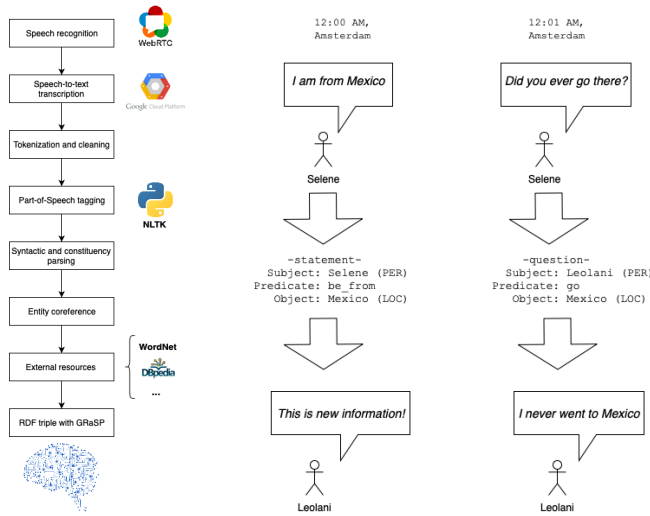


Figure 2: Natural Language Processing Pipeline

As shown in Figure 2, the NLP Pipeline consists of several external components, while some are manually implemented specifically for this task. For the sake of transparency, we resorted to rule-based parsing instead of a neural-net approach. This refers specifically to the syntactic and constituency parsing. Syntactic parsing is done with a Context-Free Grammar which captures the most typical sentence constructions in English. Since English has quite a strict word order, making such a grammar was manageable. After the CFG grammar creates a tree from the sentence, the tree is passed on to the Constituency parser, which assigns roles to the tree nodes. This is done by relying on word order, but also the POS tags and, if necessary, semantic types. The constituency parser outputs a triple, consisting of a subject, a predicate and an object, which can be stored in the brain as a claim or used to query it. Furthermore, to extract perspective information we rrsort to a simple lexicon of typical sentiment and certainty predicates, such as *like* and *think*. These lexical verbs, along with modal verbs and polarity markers, e.g. *never*, are suited for a rough estimate of the perspective expressed by the speaker.[1]

In Table 1, we show a simplified RDF representation in the brain which is the result of processing an utterance in a chat for which *Tom* is the speaker, within a specific context in Armando's office on the 24th of January 2019 during which she also perceived a chair and a person. *Tom* claimed that *Karla lived in Paris* and expressed

[1] As a next step, the model will include temporality within the perspective, using a lexicon of temporal expressions and a more advanced morphological analysis of predicate tense. Temporality indicates whether the statement is about the here and now, the past or the future (irrealis)

a perspective: he confirms the claim and he is certain and surprised. In the meantime, while *Leolani* was listening to *Tom*, she also saw a chair and recognized a person, *Gabriela* in the room where the chat took place, *Armando's office*. The event and the perceptions are all part of the same context that is anchored in time and place. The RDF representation gives further details on the source and the perspective and the entities and relations expressed in the claim.

Table 1: RDF representation representing a context taking place in a specific time and place, an utterance in a chat, the speaker, the claim made and the perspective of the speaker on the claim

Named graph: lTalk:Interactions		
lContext:context1	a	eps:Context;
	sem:hasBeginTimeStamp	lContext:2019-01-24;
	sem:hasPlace	lContext:armandosOffice;
	sem:hasEvent	lTalk:chat4;
	eps:hasDetection	lWorld:gabriela;
	eps:hasDetection	lWorld:chair1.
lTalk:chat4	a	grasp:Chat;
	sem:hasSubevent	lTalk:chat4_utterance1.
lTalk:chat4_utterance1	a	grasp:Utterance;
	sem:hasActor	lFriends:tom.
lContext:armandosOffice	a	sem:Place.
lFriends:tom	a	sem:Actor, grasp:Source.
Named graph: lTalk:Perspectives		
lTalk:chat4_utterance1 char0-25	a	gaf:Mention;
	grasp:denotes	lWorld:karla_livedIn_paris ;
	prov:wasDerivedFrom	lTalk:chat4_utterance1 ;
	prov:wasAttributedTo	lFriends:tom .
lTalk:chat4_utterance1 char0-25	ATTR1a	grasp:Attribution;
	rdf:value	grasp:CONFIRM, grasp:CERTAIN, grasp:SURPRISE;
	grasp:isAttributionFor	lTalk:chat4_utterance1_char0-25.
Named graph: lWorld:Instances		
lWorld:karla	a	n2mu:Person, gaf:Instance .
lWorld:paris	a	n2mu:Location, gaf:Instance .
lWorld:gabriela	a	n2mu:Person, gaf:Instance .
lWorld:chair1	a	n2mu:object, gaf:Instance .
Named graph: lWorld:Claims		
lWorld:karla_livedIn_paris	a	grasp:Statement, sem:Event .
Named graph: lWorld:karla_livedIn_paris		
lWorld:karla	lWorld:livedIn	lWorld:paris.

3.2. Model implementation

Following the model in Figure 1, the robot world is implemented both as a Python application, shown in Figure 3, and as an RDF representation, shown in Figure 4.

Communication modeling starts with representing the `Context`, which provides information about the situation within which conversations take place. Within

a **Context**, there are **Chat**s, which model human-robot one-to-one conversation. Within a **Chat**, **Utterance**s are spoken, both by the human and the robot. These **Utterance**s are parsed, as mentioned in Section 3.1, to obtain a subject-predicate-object RDF **Triple**. The parsed **Utterance** is sent to the brain (represented as in Table 1), which, in response, produces **Thoughts**. These **Thoughts** are the result of the inclusion of the new RDF triple and its reasoning in relation to all stored knowledge.

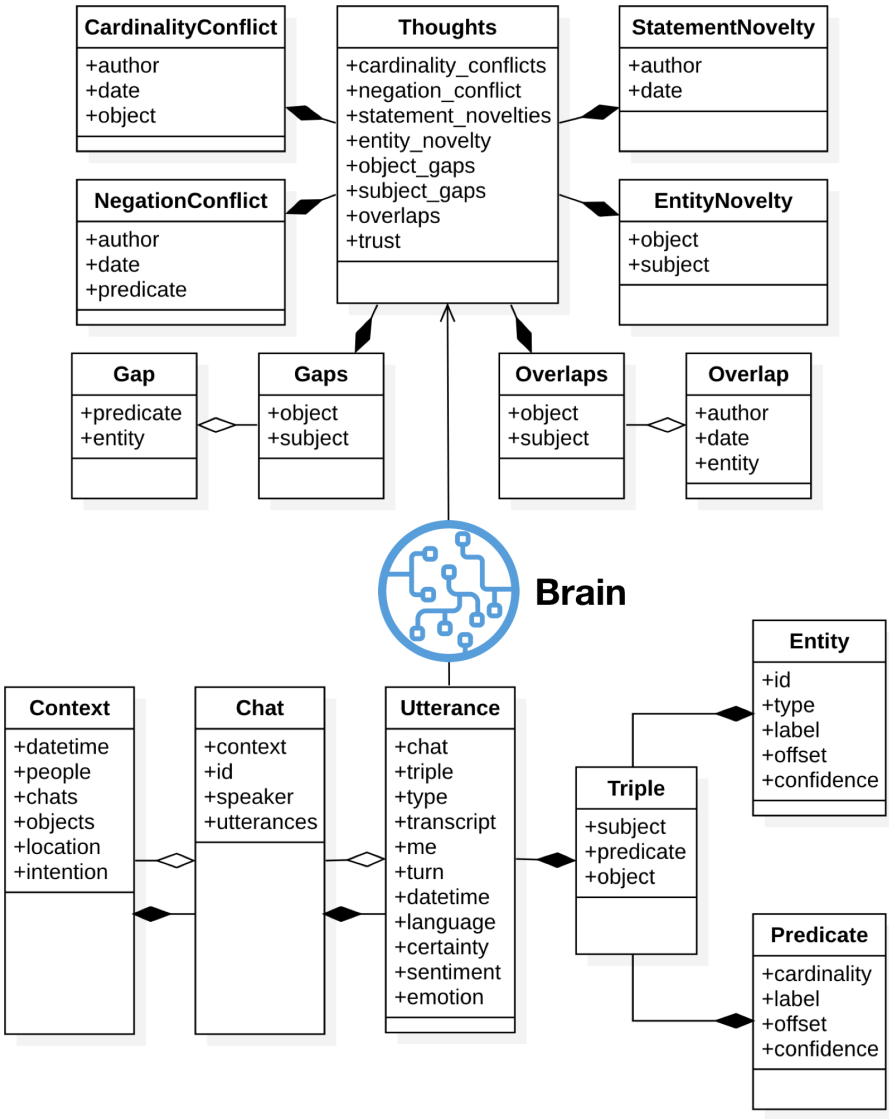


Figure 3: Data model class diagram

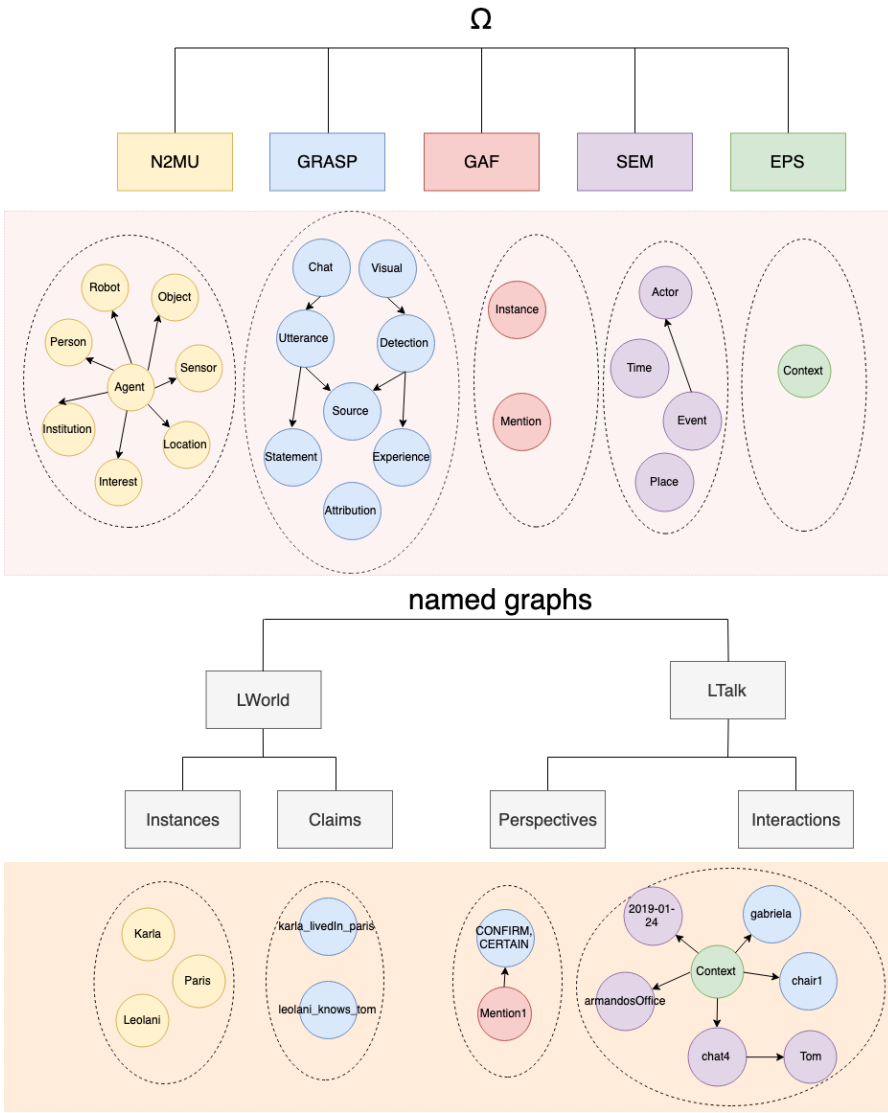


Figure 4: RDF representation

Figure 3 shows the different types of thoughts that we defined so far: *gaps*, *conflicts*, *overlap* and *novelty*. **`Gaps`** are defined by the ontologies included, and as such relate to the structure of the modelled world. **`Conflicts`**, **`Overlaps`** and **`Novelty`** are defined by the stored triples and relate to the content of the modelled world. A detailed description of what these thoughts represent is presented in Table 2. Each of these thoughts represents a state of the brain that requires a communicative action from the robot which is implemented as a drive, for example to improve this state or to inform friends. The way these **`Thoughts`** generate drives is explained in the next section.

Table 2: Types of thoughts

Cardinality Conflict	statements that cannot coexist because only one object is allowed
Negation Conflict	a previous statement is directly negated by a person
Statement Novelty	awareness that knowledge was acquired before, along with the provenance, or if it represents genuinely new information
Entity Novelty	awareness that a new entity is mentioned
Subject Gap	potential knowledge about a subject is absent and provides an opportunity to learn something new
Object Gap	potential knowledge about an object is absent and provides an opportunity to learn something new
Overlap	awareness that new statements contain shared, but not equal, information already present in the brain
Trust	a score based on how much people talked, how much the robot learned from them, and how many conflicts they generate

4. Drives

In passive robot models, people ask questions or make statements to which a robot responds. However, it may prove useful to equip a robot with drives to optimize its relation with humans and to learn from interactions. In a high-level task, e.g. finding and moving objects or showing the way, the robot can take initiative to achieve the goal. In our current model, we focus on lower-level drives that can play a role within any high-level task. Here, we specifically focus on two tasks to explain the notion of drives: 1) open-ended learning and 2) creating a personal relationship involving shared knowledge, experiences and trust. Next, we discuss some drives and thoughts related to these tasks and the corresponding communication in more detail.

4.1. Getting to know people

Knowing people is one of the robot's primary drives, as they are important sources of knowledge. The robot keeps track of her human sources through face recognition. When she meets a new person, she is triggered to learn about this person. This trigger is the result of a **`SubjectGap`**. The properties asked are predefined by the Nice2-MeetYou (n2mu) ontology, which captures social properties to start the communication, e.g. where are you from, what you like and who you know. For example, after meeting *Karla*, the triples in Table 3 inform *Leolani* that she does not know where Karla lives or what her favorite interest is.

Table 3: Sample supporting triples to infer a SubjectGap

lWorld:Karla	a	n2mu:Person .
n2mu:Person	livedIn	n2mu:City .
n2mu:Person	favorite	n2mu:Interest .

After learning about a new person, the robot queries the brain to check if other people have a similar property. An **Overlap** thought is generated if the new statement contains some shared, but not equal, information already present in the brain. For example, the triples in Table 4 show that "Karla lives in Paris" would generate an overlap with "Tom lives in Paris". The resulting **Overlap** prompts her to respond *Do you know my friend Tom who also lives in Paris?*.

Table 4: Sample supporting triples to infer an Overlap

lWorld:Karla	livedIn	lWorld:Paris .
lWorld:Tom	livedIn	lWorld:Paris .

4.2. Open-ended learning from conversation

In the above example, learning is driven by the predefined ontology. The ontology defines the properties as in a closed world, e.g. *like*, *know*, *origin*, *own*. However, we do not predefine the objects of these properties. Statements such as *I like Scrappy_Doo* or Tom likes Felix are taken seriously and the object is always stored as in instance labeled by the text coming from the speech recognition without further interpretation.

If an object is not defined in the brain by at least the type of thing it is, an **ObjectGap** thought it derived which triggers the robot to learn about it. She either asks people or consults the web. Asking people *What is Scrappy_Doo?*, she may learn it is a dog. Consulting the web what a dog is, she may learn that a *dog* is a *mammal* according to DBpedia. Asking people what a dog is, she may learn it is a *pet*. Learning about objects, can result in further thoughts such as **Overlap**, which may yield again other triggers. For example, Table 5 reflects that *Leolani* can infer that *Karla likes dogs* because she learned that *Scrappy_Doo* is a dog and *Karla likes Scrappy_Doo*. Learning that dogs are mammals may make her think that *Karla like mammals*. Knowing that cats are also mammals she can hypothesize that *Karla likes cats* and even that Karla may like *Felix*. This may make her ask Karla *Do you like Felix too?*.

Table 5: Sample supporting triples to infer a ObjectGap

lWorld:Karla	n2mu:like	lWorld:Scrappy_Doo .
dbr:Scrappy-Doo	dbo:species	dbr:Dog .
dbr:Dog	a	dbo:Mammal .
lWorld:Tom	n2mu:like	lWorld:Felix_the_cat .
lWorld:Felix_the_cat	a	n2mu:cat, dbo:Mammal .

4.3. Relevance and novelty

StatementNovelty determines if *Leolani* has acquired this knowledge before, along with the provenance information, e.g. when Karla states "I lived in Paris", *Leolani* can identify that she has heard this before from Tom. This may trigger informing Karla about this. **EntityNovelty** also signals if the statement involves a new entity, either

as the triple's subject or object. For example, "Karla visited Morocco" could lead to *Leolani* realizing she never heard about Morocco before. In general, *Leolani* comments on novelty to her friends, telling them what she learned: these are **`StatementNovelty`** thoughts.

Table 6: Sample supporting triples to infer a StatementNovelty

lTalk:chat4_utterance1_ char0-25	a	gaf:Mention;
	grasp:denotes	lWorld:karla_livedIn_paris ;
	prov:wasDerivedFrom	lTalk:chat4_utterance1 ;
	prov:wasAttributedTo	lFriends:tom .
lTalk:chat5_utterance1_ char0-16	a	gaf:Mention;
	grasp:denotes	lWorld:karla_livedIn_paris ;
	prov:wasDerivedFrom	lTalk:chat5_utterance1 ;
	prov:wasAttributedTo	lFriends:karla .

`Novelty` and **`Gap`** thoughts also yield a risk: the robot may continue talking and asking questions forever to learn more. She lacks Gricean maxims of relevance and quantity [6]. We currently limit such drives by randomly selecting responses if there are too many and mimicking relevance through recency and relatedness to the speaker. New information about the currently addressed person is considered highly relevant. Similarly, new information connecting to knowledge previously discussed with the addressee is relevant. In any case, recent information is more urgent and relevant than old information.

4.4. Uncertainties, conflicts and ambiguities

Open-ended learning also entails a risk with respect to information quality. We currently address this by capturing uncertainty scores for knowledge and perceptions, by detecting conflicts and by resolving ambiguities. Table 7 shows some of the uncertainties *Leolani* encounters.

Table 7: Types of uncertainty. * represents future work

The identity of the human participant	confidence scores of face detection confidence scores of name detection
Ambiguity in language	guessing based on immediate context
Object detection	confidence of the type mismatch with previous encounters*
Speech detection	confidence scores from the speech level of noise in the environment*
Uncertainty expressed by the human participant	classifiers that detect modal expressions classifiers that detect uncertainty from the speech itself: corrections, hesitations, volume* number of corrections, negative feedback*

The types of conflicts currently modeled are **`CardinalityConflict`**s and **`NegationConflict`**s. *Leolani* immediately addresses the source when a conflict arises and confronts other sources that provided the primary information.

A **`CardinalityConflict`** is produced whenever an author claims a statement that can not coexist with another statement as it involves a strictly one-to-one predicate. For instance, "Karla was born in France" cannot coexist with "Karla was born in Japan". A **`NegationConflict`** is returned when an author claims a direct negation of an already learned statement. For instance, "Karla lives in Paris" cannot coexist with "Karla does not live in Paris". These kinds of conflicts trigger *Leolani* to ask people for further clarification.

Table 8: Sample supporting triples to infer a CardinalityConflict

lWorld:Karla	n2mu:bornIn	lWorld:france .
lWorld:Karla	n2mu:bornIn	lWorld:japan .

Table 9: Sample supporting triples to infer a NegationConflict

lTalk:chat4_utterance1_char0-25_ATTR1 a	grasp:Attribution;
rdf:value	grasp:CONFIRM, grasp:CERTAIN, grasp:SURPRISE;
grasp:isAttributionFor	chat4_utterance1_char0-25.
lTalk:chat5_utterance1_char0-16_ATTR2 a	grasp:Attribution;
rdf:value	grasp:DENY, grasp:CERTAIN;
grasp:isAttributionFor	lTalk:chat5_utterance1_char0-16.

In our current implementation, *Leolani* only reports uncertainties and conflicts. Having a theory of mind means that conflicting information does not pose an issue. It is important that conflicting information can be stored and talked about, as this helps *Leolani* function in our conflicting and ambiguous world. In a future version, we implement more specific strategies to resolve them, e.g. consulting other (trustworthy) sources to get confirmation (e.g. DBpedia). Eventually, she could distill her own judgment based on gathered evidence.

Resolving ambiguity that is inherent to natural language is done by keeping track of the linguistic context. For instance, third person pronouns are disambiguated using the information on the last mentioned person and the information on gender. The system is equipped with a lexicon of pronouns, which contains information on the type of entity the pronoun can stand for. Cross-referencing this information with the knowledge of semantic types of previously mentioned entities allows *Leolani* to quickly guess what the pronoun might refer to. Guessing is only done when there is a high certainty level, otherwise *Leolani* will declare her confusion and ask "Which he/she do you mean?". Future plans include expanding the questions to refer to the potential guesses, like this "When you say 'she', do you mean your sister?". By relying on linguistic context and salience, we create a proactive approach to disambiguation and entity coreference, well-suited for a mixed-initiative dialogue system.

4.5. Building trust

The GRaSP model results in the accumulation of claims and the sources of those claims. Over time, the brain provides information about: 1) who shares claims with whom, 2) how many people believe or deny a claim, 3) how certain people are, both generally and individually, 4) how much emotion is expressed by whom, 5) who changes their opinion when and how often, 6) who tells things about others that are denied by the primary source, 7) who has provided most knowledge and how trustworthy that knowledge is, 8) the number of conflicts raised by a source. All this information can be used to build up trust with companions.

At the moment, **`Trust`** involves a score for people she speaks to, based on how much they have talked, how much she has learned from them, and how many conflicts they generate. Furthermore, trust can generate *thoughts* that may trigger new actions or it can be used to respond differently in case of conflicts or uncertainties in a future extension of the model. Information learned from trustworthy speakers is regarded as more likely to be correct.

5. Context awareness

One of the major problems for our robot is distinguishing between separate instances of objects of the same type. Whereas people are identified individually through face recognition, object recognition only yields types. In the first version of our model, only a single instance of each object type is represented in the brain and all knowledge is linked to this instance, i.e. all perceived chairs result in the same object instance of the type chair: all-perceptions-one-instance. The alternative is to treat each perception of an object type as a new instance of said object, but that over-generates instances, i.e. one-perception-one-instance. Failing to distinguish objects (and also people) results in unwanted errors and conflicts, as all claims made about any chair are stored as claims for the same chair. Failing to identify objects results in dispersed information over false identities and more ambiguity, making it impossible to decide which chair is being referenced. How then to define the permanence of objects and their identity, so that we achieve a natural balance for representing objects per situation and not too many?

Our current solution exploits the knowledge about locations and contexts to reason over object instances. As explained in **Section 3**, situations encountered by *Leolani* are represented as instances of a *context*. A context is anchored in time and connected to a location. All objects and people that she meets during a context are linked to this context instance together with the identified location. Identifying the location and identifying the objects mutually depend on each other and this forms the basis for making reference to situations in a context.

This is how it works. When switched on, the robot becomes aware of a new context and creates a new instance in her brain. This is shown in **Figure 5**, for *context1*, *context2* and *context3* which are created on different days during which she is switched on. Next, she scans the objects and people in her environment and relates them to this new context. People are identified through face recognition and objects

are represented as potential new object instances of a certain type based on image recognition. After this first scan, the robot tries to identify her location for which she gathers some initial information (IP, geolocation). She matches all the information of the current context with all previously modeled contexts.

context1
+ beginTimeStamp: 2019-01-23
+ ip: 192.168.1.219
+ geolocation: 52.334242, 4.866578
+ place: armandosOffice
+ events: chat4
+ detections(people): tom, gabriela
+ detections(objects): chair1, chair2, laptop1, laptop2

context2
+ beginTimeStamp: 2019-01-24
+ ip: 192.168.1.320
+ geolocation: 52.334242, 4.866578
+ place: armandosOffice
+ events: -
+ detections(people): tom, karla
+ detections(objects): chair1, chair2, laptop1, laptop2

context3
+ beginTimeStamp: 2019-05-18
+ ip: 85.113.48.148
+ geolocation: 55.753937, 37.620490
+ place: ?
+ events: -
+ detections(people): tom
+ detections(objects): chair1, potted_plant1

Figure 5: Example for context construction, and location and object identity

In **Figure 5**, the information collected for context2 is compared to context1, whereas context3 will be compared to context2 and context1. Note that only properties with so-called *endurants* as objects make sense to compare. As defined in the DOLCE ontology [10], endurants, such as objects and physical places, persist through time and place, whereas perdurants, such as events, conversations, time and situations only exist within a time and place boundary and therefore only exist at most for the duration of each instance of a context. Given the basic information on the location derived from the system, the robot thus only uses physical objects and dimensions to compare contexts for determining the potential location. If there is sufficient overlap with a previous context, *Leolani* hypothesizes that she is now in the same location. In case of uncertainty, she can ask for confirmation. If she is certain that there is no match, she assumes she is in a new location and will ask for its name. If a new location is detected and confirmed, the robot assumes all objects in this location are new instances. If a known location is recognized, she will map the physical objects of the new context to the objects of the matched location of the most recent context. If there are less objects in the new context, these objects are assumed to be absent but still exist in the brain. If there are more objects in the new context, new instances are created to match the cardinality. Object identity is thus determined in relation to location identity, where the robot tries to maximize the permanence of objects for each location across different contexts.

In **Figure 5** for example, context2 matches context1 for *Tom* and two chairs and two laptops. On the basis of the match, *Leolani* concludes she is now in *amandosOffice* and the chairs and laptops are assumed to be the same, as there is no cardinality mismatch. What is different is the presence of *Gabriela* in context1 and the presence of *Karla* in context2.[2] In contrast in the case of context3, only *Tom* and one *chair* are

[2] In the future, we plan to use properties of objects (both perceived and communicated) to help to further separate different instances, e.g. *green chair* or *my chair is close by me*.

matched while the *potted_plant* is new. Therefore, the place value remains unresolved which will trigger her to ask for the location. If that is different from previous locations, both the *chair* and the *potted_plant* will be added as new instances to the brain.

In communication, the robot treats objects in new locations as new instances unless told otherwise. For example, if somebody claims ownership of n chair within a context and location, e.g. *this is my chair*, the property *own*is assigned to that instance. In another location, a similar object can be perceived but it is considered to be a different instance. However, if the same person again claims ownership of this similar object, the robot realizes that multiple similar objects related to different locations are owned by the same person. As a weak conflict, this may trigger questions about identity: *is this the same chair?*. On the other hand, if the chair in this new location is claimed to be owned by another person, it does not result in a conflict as it was already represented as a different chair in the brain and both chairs can have different owners.

6. Conclusion

In this position paper, we described our models and implementation for a robot that can learn through communication for the purpose of building a social relationship. Our model stores knowledge as triples with the source and its perspective. It represents communication as chats and turns in which claims are made. The model allows the robot to deal with knowledge coming from different sources, handle uncertainties and conflicts, and derive trust in sources. The robot uses thoughts representing states of the brain, which trigger actions and communication as low-level drives. Finally, we have shown how the robot creates an episodic representation of a context linked to time and location, with awareness of the presence of people and objects. Awareness of contexts and locations can be used to identify object instances and model the permanence of objects. All the code of our model is available on Github[3] and project progress is reported on our website[4].

Currently, the robot has acquired knowledge regarding 296 statements through 164 conversations held with 26 distinct people. These conversations were held for testing the system and we have not evaluated the quality. In the future, we plan to carry out experiments to measure the performance of our model. Intrinsic evaluations should demonstrate the capacity to understand humans and the world, to acquire knowledge, to acquire vocabulary and expressions, and to express drives to improve the state of the brain. Extrinsic evaluations should demonstrate the user satisfaction, the quality of the relationships and any high-level task that is modeled. For evaluations, we need to create evaluation data and scenarios, define criteria and create baselines and alternative models.

[3] https://github.com/cltl/pepper

[4] http://makerobotstalk.nl/

Acknowledgements

This project was funded through the NWO-Spinoza funds awarded to Piek Vossen and by the VU University of Amsterdam. We specifically thanks Selene Kolman and Bob van Graft for their support

References

1. *Abdul-Kader, S. A., Woods, J.:* Survey on chatbot design techniques in speech conversation systems. International Journal of Advanced Computer Science and Applications. 6, 7, (2015).
2. *Amos, B. et al.:* OpenFace: A general-purpose face recognition library with mobile applications. CMU-CS-16-118, CMU School of Computer Science (2016).
3. *T. W. project:* WebRTC. In: Online publication. (2011).
4. *Fokkens, A. et al.:* Grasp: Grounded representation and source perspective. In: Proceedings of knowrsh, ranlp-2017 workshop, varna, bulgaria. (2017).
5. *Google:* Cloud speech-to-text - speech recognition. In: Online publication. (2018).
6. *Grice, H. P.:* Logic and conversation. 1975. 41–58 (1975).
7. *Leslie, A.:* Pretense and representation: The origins of "theory of mind.". Psychological review. 4, (1987).
8. *Li, J. et al.:* A persona-based neural conversation model. arXiv preprint arXiv:1603.06155. (2016).
9. *Lowe, R. et al.:* The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. arXiv preprint arXiv:1506.08909. (2015).
10. *Masolo, C. et al.:* Wonderweb deliverable d17. Computer Science Preprint Archive. 2002, 11, 74–110 (2002).
11. *Mavridis, N.:* A review of verbal and non-verbal human–robot interactive communication. Robotics and Autonomous Systems. 63, 22–35 (2015).
12. *Reddy, S. et al.:* Coqa: A conversational question answering challenge. arXiv preprint arXiv:1808.07042. (2018).
13. *Serban, I. V. et al.:* Building end-to-end dialogue systems using generative hierarchical neural network models. In: AAAI. pp. 3776–3784 (2016).
14. *Spekman, M. L. et al.:* Perceptions of healthcare robots as a function of emotion-based coping: The importance of coping appraisals and coping strategies. Computers in Human Behavior. 85, 308–318 (2018).
15. *Szegedy, C. et al.:* Going deeper with convolutions. In: Computer vision and pattern recognition (cvpr). (2015).
16. *Vinyals, O., Le, Q.:* A neural conversational model. arXiv preprint arXiv:1506.05869. (2015).
17. *Vossen, P. et al.:* Leolani: A reference machine with a theory of mind for social communication. In: Proceedings of tsd-2018, brno, https://www.tsdconference.org/tsd2018. (2018).
18. *Zhang, S. et al.:* Personalizing dialogue agents: I have a dog, do you have pets too? arXiv preprint arXiv:1801.07243. (2018).

О ПРОЕКТЕ СЛОВАРЯ «ИНТЕРТЕКСТУАЛЬНЫЙ ТЕЗАУРУС СОВРЕМЕННОГО РУССКОГО ЯЗЫКА»: КНИЖНЫЙ VS. МУЛЬТИМЕДИЙНЫЙ[1]

Вознесенская М. М. (voznes-masha@yandex.com),
Шмелева Е. Я. (eshkind@mail.ru)

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

ON A NEW DICTIONARY "INTERTEXTUAL VOCABULARY OF MODERN RUSSIAN": PAPER VS. MULTIMEDIA

Voznesenskaya M. M. (voznes-masha@yandex.com),
Shmeleva E. Ya. (eshkind@mail.ru)

Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

Russian dictionaries of idioms, winged words and quotations do not reflect "the intertextual competence" of modern Russian speakers: on the one hand, their vocabularies abound in obsolete, uncommon and even incomprehensible units; on the other hand, they are short of some well known and widely used catchwords and Internet memes. The article deals with the structure and principles for constructing a new dictionary, namely, "Intertextual Vocabulary of Modern Russian" (in paper and multimedia versions). The dictionary will be based on corpus data and include over 1000 well-known catchphrases from the 20th–21st centuries. The basic unit is a dictionary entry that will include the following parts: lexical input, meaning, source, examples, phraseological model and its transformations, comments; the last two parts are optional. The arrangement is alphabetical by the first word; however, there will be user-friendly indexes for locating all the catchphrases from the same source, same topic, etc. The multimedia version is characterized by quantitative and qualitative increase in content: in addition to text information, the dictionary will contain audio, video, photo fragments, graphics, animation, etc. referring to the relevant "multimedia" sources of intertextual units (such as movies, cartoons, paintings, songs, TV shows, etc.). Using hyperlinks, one can easily find the required information related to a given entry.

Key words: dictionary, intertextual unit, catchphrase, corpus data, multimedia

[1] Исследование проведено при поддержке гранта РФФИ № 19-012-00396 (Теоретические основы составления словаря «Интертекстуальный тезаурус современного русского языка»).

«...индивидуальный речевой опыт всякого человека формируется и развивается в непрерывном и постоянном взаимодействии с чужими индивидуальными высказываниями. Этот опыт в известной мере может быть охарактеризован как процесс освоения — более или менее творческого — чужих слов (а не слов языка). Наша речь, то есть все наши высказывания..., полна чужих слов, разной степени чужести или разной степени освоенности, разной степени осознанности и выделенности».

(М. М. Бахтин «Проблема речевых жанров»)

1. Введение

Для понимания большинства современных русских текстов недостаточно только знания языка, нужна еще и «интертекстуальная компетенция» — узнавание и понимание большого числа немаркированных и неатрибутированных цитат, квазицитат, аллюзий, реминисценций, прецедентных имен и событий и др., которые в самом общем смысле можно назвать *интертекстуальными единицами* (ср. [Фокина, Шмелева 2012]) Несмотря на существование страноведческих и фразеологических словарей, словарей крылатых слов и словарей цитат из литературы, кинематографа и др. (см., например, [Россия. Большой лингвострановедческий словарь 2007]; [Словарь-тезаурус современной русской идиоматики 2007]; [Ашукин, Ашукина 1987]; [Берков, Мокиенко, Шулежкова 2005]; [Душенко 2006]), они по целому ряду параметров не отражают интертекстуальную компетенцию современного россиянина: с одной стороны, в словники этих словарей включены устаревшие, мало употребительные и даже непонятные единицы, с другой стороны, в них не нашли отражения многие широко употребительные новые крылатые слова, цитаты и интернет-мем[2]. Поэтому актуальной лексикографической задачей является составление словаря «Интертекстуальный тезаурус современного русского языка», детально описывающего и объясняющего функционирование наиболее употребительных интертекстуальных единиц русского языка конца XX — начала XXI века. Предполагается, что словарь будет включать около 1000 единиц, отобранных на основе анализа больших массивов текстов современной литературы, СМИ и Интернета, включающего как корпусное исследование [НКРЯ], так и тотальную выборку. В настоящее время составлен предварительный словник, который в процессе работы над словарем, естественно, будет уточняться: некоторые единицы будут исключены, другие — добавлены[3] Книжный вариант словаря

[2] Из появившихся в последнее время словарей, стремящихся заполнить эту лакуну, назовем [Словарь языка Интернета.ru 2016], в котором содержатся слова и выражения, характерные для сетевого общения последних десятилетий.

[3] Так, например, одним из критерием употребительности интертекстуальной единицы является наличие трансформаций ее исходной формы (подробнее см. ниже). Полная

и его мультимедийная версия создаются параллельно. Книга будет снабжена мультимедийным вариантом, размещенным на электронном носителе, также планируется сетевое размещение мультимедийной версии.

2. Книжный вариант словаря «Интертекстуальный тезаурус современного русского языка»: общая характеристика и структура словарной статьи

Кратко охарактеризуем основные особенности устройства печатной версии словаря. Основной корпус словаря будет состоять из упорядоченных по алфавитному принципу словарных статей, снабженных буквенно-цифровым индексом (А-7, Б-5 и т. п.), что облегчит читателю поиск нужной информации. Возможность найти интертекстуальную единицу по любому из компонентов, входящих в ее состав, обеспечивается специальным указателем. Также предполагается включить в состав словаря несколько указателей, классифицирующих языковой материал по разным параметрам: Указатель источников содержит перечень источников и относящихся к ним интертекстуальных единиц, в Тематическом указателе языковой материал сгруппирован в тематические группы.

Таким образом, основной корпус словаря, с одной стороны, и указатели, с другой, демонстрируют как разные способы организации языкового материала (от интертекстуальной единицы к ее значению и источнику в словарной статье; от источника к интертекстуальной единице в Указателе источников, от значения к языковой единице в Тематическом указателе), так и разные принципы его описания (семасиологический в словарной статье, ономасиологический в Тематическом указателе).

Разрабатывается структура словарной статьи, на сегодняшний день включающая следующие зоны: лексический вход, значение, источник интертекстуальной единицы, иллюстративный материал, фразеологизованная модель, комментарии. Остановимся на каждой из зон подробнее.

1. Лексический вход — словарная форма интертекстуальной единицы, при этом начальной формой будет являться не грамматически начальная форма, а формулировка, представленная в исходном тексте. В некоторых случаях входом статьи может быть точечная цитата (имена литературных, мифологических и киноперсонажей, их производные, авторские новообразования, слова с фонетическими, морфологическими и иными особенностями, которые отсылают читателя к конкретному произведению или автору, лексемы идиолекта, вызывающие

информация о количестве и типах трансформаций будет получена только при составлении словарных досье на каждую интертекстуальную единицу (контексты употреблений в художественной литературе, языке СМИ и Интернета, разговорной речи). Кроме того, поскольку интертекстуальный тезаурус не одинаков у представителей разных поколений и социальных групп, а также может иметь региональные и индивидуальные особенности, ряд единиц, включенных авторами в предварительный словник, требует более широкого обсуждения. По результатам обсуждения, соответствующие единицы будут либо исключены из словника, либо снабжены стилистическими пометами и/или пометами *млад./стар.*, *рег.* и др,.

ассоциацию с определенной личностью и т. п.). В качестве точечных цитат могут выступать прецедентные имена и прецедентные высказывания с максимальной степенью количественной трансформации — усечения до одной лексемы.

2. Значение интертекстуальной единицы. В некоторых случаях, как правило, если речь идет о цитатах из кинофильмов или анекдотов, могут отмечаться интонационные и акцентологические особенности, жесты, которыми сопровождается произнесение фразы.

3. Указание на источник интертекстуальной единицы, время ее появления, авторство. Для переводных единиц, восходящих к текстам зарубежной литературы, зарубежным кинофильмам, афоризмам — приводится цитата на языке оригинала. В ряде случаев даются рисунки (Колобок, Чебурашка), фото картин, плакатов, кадров из кинофильмов, мультфильмов.

4. Иллюстративный материал, включающий а) цитаты — единицы, введенные во вторичный текст в исходном, неизменённом виде; б) квазицитаты, т. е. цитаты, трансформированные различными способами. В словаре предполагается отразить наиболее распространенные способы трансформации интертекстуальных единиц: количественные (усечение/расширение исходного состава: «*…И умерли в один день*»,[4] «*Не расстанусь… буду вечно…*», «*Контора «Рога и копыта» (крупного рогатого скота)*») и качественные (фонетическая трансформация — мена интонации, добавление, усечение или замена сегмента, использование омофонов: «*Зноев ковчег*», «*Вешний зов*»; лексическая замена с использованием паронимов, парономазов, омонимов, замена эпитета, замена компонента на лексему, связанную с содержанием статьи, эвфемизация: «*Пятиэтажки век недолог*», «*Вся президентская треть*», «*Когда б вы знали, из какого сюра…*», «*Человек — это звучит горько*»; морфологическая трансформация — замена числа, рода, падежа имён существительных, времени, наклонения, числа, рода глаголов и др.: «*Принцесса и нищие*», «*Каменные гости*»; синтаксическая — контаминация или агглютинация двух и более единиц и фразеологизмов, парцелляция или присоединение, перестановка компонентов интертекстуальной единицы, использование ее логико-синтаксической модели и др.: «*Все мы немножко Анны Каренины*», «*Затишье перед бурей в стакане воды*»). Частотны смешанные трансформации интертекстуальной единицы. Графически выделяются элементы исходного текста, сохраненные в составе интертекстуального фрагмента.

5. Фразеологизованная модель, по которой строятся интертекстуальные единицы, прежде всего, газетные заголовки (факультативные компоненты модели указываются в квадратных скобках). Далеко не все интертекстуальные единицы способны выступать в качестве интертекстуальной модели. В самом общем виде под фразеологизованной моделью понимаются те ядерные компоненты, которые остаются неизменными при различных вариантах трансформации языковой единицы.

---

[4] Этот пример усечения исходной цитаты Александра Грина — название рассказа Л. Улицкой. Все следующие иллюстративные примеры трансформаций интертекстуальных единиц являются заголовками из современных СМИ.

6. Комментарии, в которой содержится разнообразная дополнительная информация, в том числе релевантная экстралингвистическая информация (страноведческая, культурологическая и др.).

Приведем примеры двух словарных статей:

1. **Коня на скаку остановит, в горящую избу войдет.**
2. О сильной, решительной, смелой женщине.
3. Из ч. 1 поэмы Н. А. Некрасова «Мороз, Красный нос» (1863:

Есть женщины в русских селеньях <…>)
В игре ее конный не словит,
В беде — не сробеет, — спасет:
Коня на скаку остановит,
В горящую избу войдет!

4. Цитата: *«**Коня на скаку остановит, в горящую избу войдет!**» — увы, эти известные строки актуальны и поныне, ведь нашим женщинам и сегодня приходится решать слишком много неженских проблем…».* (Наше время); *«…о «слабых» женщинах мечтают, как правило, слабые мужчины. Ведь легче подать пальто, чем потушить **горящую избу**. Прочитать стихи, чем **останови**ть **коня на скаку**»* (Вечерняя Москва). Лексическая замена: *«Быка **на скаку остановит**»* (Комсомольская правда) — о женщине-тореадоре; *«Ворье **на бегу остановит**…»* — о женщине-милиционере (Московский комсомолец). Морфологическая трансформация: *«**В горящую избу во**шла. Рискуя жизнью, молодая женщина вытащила из огня пятерых детей, пока пожарные боролись с бездорожьем»* (Труд); *«В наш противоречивый век неравных условий, но равных возможностей уже мало кто сомневается, что предназначение Женщины не только в том, чтобы рожать детей, поддерживать тепло домашнего очага и «**остан**авливать **на скаку коня** и входить **в горящую избу**».* (Родное Подмосковье); *«В России, где женщины привыкли выносить мужчин на руках из **горящих изб** и усаживать на **останов**ленных **на скаку коней**, союзы, подобные этому, обычно осуждаются»* (Вечерний клуб).

5. (имя сущ. в Вин.пад. + на + отглагольное имя сущ. в Пр.пад.+ остановит…)

1. **«Особенности национальной охоты»**

2. О характерных для российской государственной или русской национальной действительности признаках, качествах, свойствах, отличающих явления или предметы, названные именем существительным. Обобщение, характеризация специфики какого-либо явления с точки зрения общенациональной распространённости или значимости.

3. Название фильма из трилогии А. Рогожкина (сц., реж.) «Особенности национальной охоты» (1995) «Особенности национальной рыбалки» (1998) «Особенности национальной охоты в зимний период» (2000).

4. Лексическая замена: *«**Особенности национальной** попойки»* (Аргументы и факты); *«**Особенности национальной** избирательной кампании: ФСБ и МВД сбились с ног, жена и соратники сходили с ума, Запад обвинял Кремль в немыслимых злодеяниях. А кандидат в президенты России в это время прохлаждался в Киеве»* (Комсомольская правда).

5. ([прил. во мн. ч.] + **особенности** + прил. (**национальн**ого и др.) + сущ. в Р. п.). Ядерный компонент — *особенности*.

3. Мультимедийная версия словаря «Интертекстуальный тезаурус современного русского языка»: типы информации и способы ее представления

Основные области применения мультимедийных технологий в лексикографии зависят от предназначения, целевой аудитории и описываемого материала словаря. Это могут быть как лингвистические словари (общие и специализированные, одно и многоязычные, адресованные как носителю языка, так и иностранцам, изучающим этот язык), особое место среди которых занимают мультимедийные словари жестовых языков, так и энциклопедические (см., например, [Online Multimedia Dictionary of the Polish Language]; [English Oxford Living Dictionaries]; [Korean Multimedia Dictionary]; [Kobozeva, Zakharov 2004]; [Solina, Krapež, Jaklič, Komac 2004]; [Воскресенский, Хахалин 2007], [Spreadthesign]; [Encyclopaedia Britannica]; [Универсальная энциклопедия Кирилла и Мефодия]; [Art 20C: The Thames and Hudson Multimedia Dictionary of Modern Art]; [Лингвострановедческий словарь «Россия»]).

Специфика материала словаря «Интертекстуальный тезаурус современного русского языка» — интертекстуальные единицы, отсылающие к разнообразным «мультимедийным» источникам (художественной литературе, сказкам, мифам, кинофильмам, мультфильмам, произведениям изобразительного искусства, песням, плакатам, лозунгам, телепередачам, рекламе и т. п.). Естественно, что именно мультимедийный словарь, содержащий разные типы информации (текст, графику, анимацию, аудио, фото и видео фрагменты и др.) позволяет в максимально аутентичном и эксплицитном варианте воссоздать интертекстуальный тезаурус современного носителя русского языка. В настоящее время рано говорить о детальной структуре мультимедийного варианта словаря. Остановимся лишь на тех типах информации, которые обязательно должны быть отражены в словаре, и опишем общее представление о виде интерфейса — системе поиска и представления информации.

Мультимедийная версия словаря «Интертекстуальный тезаурус современного русского языка» по сравнению с «бумажной» версией характеризуется количественным и качественным увеличением контента (типов информации). Количественное увеличение коснется, в первую очередь, контекстов употребления интертекстуальных единиц (как в стандартном, так и в трансформированном виде). Также возможно расширение комментирующей зоны словарной статьи. Качественное увеличение связано с включением в Словарь новых мультимедийных типов информации: помимо текста, это аудио, видео, фото фрагменты, графика, анимация и т. п., представляющие соответствующие «мультимедийные» источники интертекстуальных единиц.

Аналогами указателей, содержащихся в традиционном книжном варианте Словаря, в мультимедийном словаре выступают Список источников и Тематический список. Первый включает в себя перечисление источников

интертекстуальных единиц (сказки, мифы, литературные произведения, песни, кинофильмы, телепередачи, реклама, анекдоты и т. п.), снабженных краткой информацией энциклопедического характера, и сами интертекстуальные единицы, отсылающие к этим источникам. В Тематическом списке языковые единицы распределяются по тематическим группам в зависимости от выражаемых ими смыслов. Так, например, выражения *Ба, знакомые все лица! Здравствуй, племя младое, незнакомое! Откуда ты, прелестное дитя? Я пришел к тебе с приветом; Какие люди и без охраны! Здравствуй и прощай!* относятся к семантической группе ПРИВЕТСТВИЯ. Также представлен Генеральный словник, в котором все интертекстуальные единицы расположены в алфавитном порядке. Вышеперечисленным типам информации соответствуют основные элементы интерфейса мультимедийного словаря, к которым относятся Медиатека, содержащая мультимедийные ресурсы, Банк примеров, Генеральный словник, Источники и Тематический словник. Базовой единицей представления информации в мультимедийной версии Словаря, как и в «бумажном» варианте, является словарная статья, доступ к которой возможен несколькими способами: или выбором нужной языковой единицы в генеральном словнике, либо вводом в строку поиска языковой единицы или ее отдельных компонентов. В последнем случае выпадающее меню с подсказками, содержащее все выражения с заданным словом/словами, позволяет выбрать требуемую форму. Остановимся подробнее на особенностях структуры словарной статьи мультимедийной версии Словаря.

Основные зоны словарной статьи в мультимедийном Словаре те же, что и в печатном варианте. Это заглавие (лексический вход), значение, источник, примеры употребления, модель и комментарии. Помимо этого появится новая зона произношения, расположенная рядом с лексическим входом — интертекстуальной единицей и обозначенная соответствующим значком, при нажатии на который воспроизводится аудио (в ряде случаев, видео) фрагмент. Необходимо отметить, что изначально информация в основных зонах словарной статьи в мультимедийной версии отображается на экране в редуцированном (свернутом) виде. Так, в зоне Источник приводится общая характеристика источника и время его появления. Например, для литературной цитаты это название литературного произведения, его автор, время написания, в случае киноцитаты — название кинофильма, имена создателей, дата выхода на экран. Более подробная информация может получена по желанию пользователя, для чего в интерфейсе Словаря предусмотрены определенные значки. Развернутая информация может содержать текстовый фрагмент текста-источника (для литературной цитаты), отрывок из сценария и видео воспроизведение киноэпизода (для киноцитаты), текст песни и звуковое воспроизведение (для песенной цитаты), репродукцию картины (для интертекстуальной единицы, отсылающей к названию этой картины) и т. п. Иллюстративные примеры употребления так же приводятся в ограниченном объеме, для отображения большего числа примеров необходимо «нажать» на соответствующий значок.

Также предполагается заложить за разными зонами словарной статьи гиперссылки, позволяющие пользователю осуществлять быстрый переход к необходимой информации. Так, из зоны значения можно перейти в Тематический словник и получить информацию, какие еще интертекстуальные единицы выражают сходный смысл; из зоны источника попасть в список источников и увидеть другие цитаты, имеющие то же происхождение. Соответственно, возможен и обратный переход от интертекстуальной единицы в том или другом перечне к ее словарной статье. В принципе гиперссылки могут связывать любые элементы словаря, и более конкретно этот вопрос будет решаться на более позднем этапе работы при осуждении с программистами.

Целесообразность описанной выше поэтапной подачи материала обусловлена следующими обстоятельствами. Во-первых, таким образом сохраняется компактная структура словарной статьи, не перегружающая пользователя с самого начала большими объемами информации. Во-вторых, тем самым реализуется интерактивная функция словаря, когда читатель самостоятельно выбирает ту степень подробности (детализации) извлекаемой информации, которая ему необходима.

4. Особенности сетевого мультимедийного словаря «Интертекстуальный тезаурус современного русского языка»

Онлайн размещение мультимедийной версии словаря даст возможность оперативно обновлять информацию, учитывая, в том числе, обратную связь с пользователями. Так, представляется возможным расширить (увеличить) состав словника, включив в него, с одной стороны, новые интертекстуальные единицы, получившие широкое распространение в самое последнее время, и, с другой стороны, устаревающие выражения, постепенно выходящие из активного употребления. Таким образом, генеральный словник дифференцируется по временно́му признаку и будет включать устаревающий, основной и новый подкорпуса. Благодаря сетевому размещению будет производится ротация языкового материала (исключение некоторых интертекстуальных единиц из словаря, перемещение других из одного подкорпуса в другой), отражающая динамику его употребления на данном временно́м отрезке.

Кроме того, планируется привлечь пользователей к работе над словарем, дав им возможность самим пополнять словарь. Для этого будет предусмотрена специальная опция «пополнения», в которой читателям предлагается заполнить анкету, повторяющую структуру словарной статьи. Естественно, что окончательное решение о целесообразности включения новых интертекстуальных единиц в корпус словаря остается за профессиональными лексикографами. Словарная статья, добавленная пользователем, будет иметь специальную атрибуцию, указывающую, кем и когда она была создана. Все интертекстуальные единицы, добавленные пользователями, сформируют специальный «читательский» подкорпус словаря, при этом эти единицы будут включены и в генеральный словник, и в один из «временны́х» подкорпусов (основной, устаревающий или новый).

5. Заключение

«Интертекстуальный тезаурус современного русского языка» является словарем нового типа по целому ряду параметров. Словарь с самого начала создается одновременно в электронной, мультимедийной, и в традиционной, бумажной, версиях. В словаре представлен новый словник, включающий в себя ядро интертекстуальных единиц, многие из которых впервые получают лексикографическую фиксацию. Разработана новая структура словарной статьи, позволяющая дать детальное описание значений, источников и функционирования интертекстуальных единиц. Также в словаре приводится обширный иллюстративный материал, включающий как стандартные, так и трансформированные контексты употреблений описываемых языковых единиц. В мультимедийной версии словаря, благодаря современным технологиям, становится возможным количественное и качественное увеличение контента, что позволит в максимально эксплицитной форме отразить интертекстуальную компетенцию современного россиянина. Осознанная «избыточность» информации, включенной в словарь — широкий привлеченный и откомментированный материал, в том числе энциклопедического характера, помимо чисто лингвистической значимости, имеет культурологическую и страноведческую ценность. Таким образом, «Интертекстуальный тезаурус современного русского языка» (и книжный и мультимедийный варианты) характеризуется полифункциональностью, т. е. это и собственно лингвистический словарь, и культурологический и страноведческий справочник. Также словарь, особенно его мультимедийная версия, сможет служить своеобразным учебным пособием, т. к. содержит материал, который можно использовать при преподавании русского языка как российским студентам, так и иностранным учащимся. Сетевое размещение мультимедийного варианта расширит интерактивные возможности словаря, обеспечив функцию обратной связи с читателями, благодаря чему станет возможным своевременное обновление информации, коррекция допущенных неточностей, а сами читатели смогут стать соавторами словаря.

References

1. *Art 20C: The Thames and Hudson Multimedia Dictionary of Modern Art* (1998), Thames and Hudson, London.
2. *Ashukin N. S., Ashukina M. G.* (1987), Winged words: Literary Quotations. Figural Expressions [Krylatye slova: Literaturnye tsitaty. Obraznye vyrazheniya], Khudozhestvennaya literatura, Moscow.
3. *Berkov V. P., Mokienko V. M., Shulezhkova S. G.* (2005), Big dictionary of Russian winged words [Bol'shoj slovar' krylatyh slov russkogo yazyka], Russkie slovari, Astrel', AST, Moscow.
4. *Dushenko K. V.* (2006), Dictionary of Modern Quotations [Slovar' sovremennyh tsitat], Eksmo, Moskva.
5. *Encyclopaedia Britannica*, available at: https://www.britannica.com/.
6. *English Oxford Living Dictionaries*, available at: https://en.oxforddictionaries.com/.

7. *Fokina O. V., Shmeleva E. Ya.* (2012) Intertextual competence and text understanding [Intertekstual'naja kompetetsija I ponimanie teksta] Linguistic training of students at school and out of school [Lingvisticheskaja rabota so shkol'nikami na urokah raznyh predmetov i vo vneurochnom obrazovatel'nom prostranstve], MIOO, Moscow, p. 49–59.
8. *Internet.ru Language Dictionary* (2016), [Slovar' yazyka interneta.ru], edited by Krongauz M. A., AST-PRESS, Moscow.
9. *Kobozeva I. M., Zakharov L. M.* (2004), Types of information for the multimedia dictionary of Russian discourse markers, Proceedings of the 9th International Conference "Speech and Computer" SPECOM'2004, Publishing house "Anatolya", Saint-Petersburg, pp. 470–473.
10. *Korean Multimedia Dictionary*, available at: http://www.indiana.edu/~koreanrs/kordic.html.
11. *NRC* — National Russian Corpus [Natsional'nyj korpus russkogo yazyka], available at: http://www.ruscorpora.ru/.
12. *Online Multimedia Dictionary of the Polish Language*, available at: http://online-polish-dictionary.com/.
13. *Culture-Oriented Linguistic Dictionary "Russia"* [Lingvostranovedcheskij slovar' "Rossiya"], available at: https://ls.pushkininstitute.ru/lsslovar/index.php?title=Тематический_указатель:Перечень_тематик.
14. *Russia. Big Culture-Oriented Linguistic Dictionary* (2007), [Rossiya. Bol'shoj lingvostranovedcheskij slovar'], Prokhorov Yu. E. (ed.), AST-PRESS KNIGA, Moscow.
15. *Solina Franc, Krapež Slavko, Jaklič Ales, Komac Vito* (2001), Multimedia Dictionary and Synthesis of Sign Language, Design and Management of Multimedia Information Systems, Mahbubur Rahman Syed (Editor), Idea Group Publishing, Hershey PA, pp. 268–281.
16. *Spreadthesign*, available at: https://www.spreadthesign.com/ru.ru/search/.
17. *Thesaurus of modern Russian idioms* (2007), [Slovar'-tezaurus sovremennoy russkoy idiomatiki], edited by Baranov A. N., Dobrovol'skij D. O., Mir entsiklopedij Avanta+, Moscow.
18. *Universal encyclopedia of Cyril and Methodius* [Universal'naya ehntsiklopediya Kirilla i Mefodiya], available at: https://megabook.ru/.
19. *Voskresenskij A. L., Khakhalin G. K.* (2007), A Multimedia Explanatory Dictionary of Russian Sign Language [Mul'timedijnyj tolkovyj slovar' russkogo zhestovogo yazyka], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2007" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2007"] Bekasovo, pp. 115–120.

ПРОСОДИЯ ВОПРОСОВ С ЧАСТИЦЕЙ *ЛИ*[1]

Янко Т. Е. (tanya_yanko@list.ru)
Институт языкознания РАН; Государственный институт русского языка им. А. С. Пушкина, Москва, Россия

THE RUSSIAN *LI*-QUESTIONS PROSODY

Yanko T. E. (tanya_yanko@list.ru)
Institute of linguistics; Pushkin State Russian Language Institute, Moscow, Russia

The paper is aimed at the analysis of the prosody in the Russian *yes-now*-questions with particle *LI*. The three basic patterns of the Russian *LI*-questions, which are construed as semantically minimal, are singled out. (The semantically minimal sentences are considered here as such where the prosodic structure brings minimal contribution into the semantic structure of a sentence). Consequently, the prosody of the sentences composed with contrast, or discourse continuity is viewed as being derived from the prosody of the basic types.

The illocutionary force in *LI*-questions is designated not by prosody as in other Russian *yes-no*-questions but by a segmental means, namely — by *LI*. Hence, the prosody in *LI*-questions is not a cue of the illocutionary force but it forms the sentence as an autonomous prosodic unit and designates the non-illocutionary meanings: contrast and discourse continuity. The accent on the first accented word can be either rising, or falling without any reasonable difference in meaning.

In questions with particle *LI*, particle *LI* preserves its Wackernagel parameters, while the host of the clitic in the majority of cases serves as the first, or the only one, accent-bearer of the sentence. However, in the context of contrast, the first accent-bearer can be placed to the right from *LI*.

Within the discourse continuity, *LI*-questions have two accent-bearers, the first of them could be either rising, or falling, and, at the same time, either contrastive, or non-contrastive, while the second one — is always the rising one.

The prosodic patterns of *LI*-questions are exemplified here by spoken fragments taken from the Multimodal corpus of the Russian National corpus, and the minor working collection of the Russian speech recordings specifically set up for this investigation. The software program Praat was used in the process of analyzing the sounding data.

Синтаксису и порядку слов вопросительных предложений с частицей *ли* посвящена большая литература: [Halpern 1992, 1995]; [Stepanov 1998]; [Rudnickaya 2000]. См. также [Zaliznyak 2008: 30–31] о ваккернагелевских свойствах *ли* в древнерусском языке, которые в существенной степени унаследованы и русским языком. Некоторые проблемы анализа просодии вопросов

[1] Работа поддержана РНФ (проект 18-18-00462 «Коммуникативно-синтаксический интерфейс: типология и грамматика», реализуемый в Государственном институте русского языка им. А. С. Пушкина)

с *ли* также служили предметом исследования в работах [Bryzgunova 1982a: 97–122] и [Kedrova, Potapov, Omel'janova, Egorov 2002]. Из выделенных в работах основных характеристик вопросов с *ли* назовем следующие:

- частица *ли* в современном русском языке сохранила верность закону Ваккернагеля и, соответственно, располагается в предложении на втором месте после первого полноударного слова [Halpern 1992, 1995]; [Stepanov 1998]; [Rudnickaya 2000] с учетом Правила барьера ([Zaliznyak 2008: 30–31], [Zimmerling 2013: 385–395]), в известных условиях сдвигающего вторую позицию «вправо»;
- в большинстве примеров (с определенными, как показывают наши данные, исключениями, которые не были замечены ранее и которые оказались связанными с эффектом коммуникативного контраста) словоформа, предшествующая *ли*, несет на себе коммуникативно релевантный акцент [Halpern 1992, 1995]; [Stepanov 1998]; [Rudnickaya 2000];
- просодия предложений с частицей *ли* варьирует, о чем, без обсуждения деталей, говорится в работе [Kedrova, Potapov, Omel'janova, Egorov 2002].

Задача нашей работы — дать анализ просодии вопросительных предложений с *ли* в различных контекстах. Ниже анализируется просодия предложений, которые понимаются как базовые, т. е. такие, у которых просодическая структура вносит минимальный вклад в семантическую структуру предложения, а также просодия предложений, отягощенных контекстом коммуникативного контраста, дискурсивной незавершенности и композиции контраста и незавершенности. Таким образом, анализируется три функционально-семантических параметра, характеризующих вопросы с *ли* и имеющих просодию либо в качестве средства выражения, либо в качестве средства, формирующего отдельный речевой акт. Это иллокутивная сила (обязательный параметр) и два факультативных: коммуникативный контраст и дискурсивная незавершенность. Контраст и незавершенность образуют с иллокутивной силой семантические композиции, которые меняют базовую просодическую структуру вопроса.

Для анализа разработан малый рабочий корпус аннотированных звучащих предложений с *ли*. Массив содержит 164 инстанции частицы *ли* в вопросительных предложениях, звучание которых в общей сложности занимает около 60 минут. Идиоматические выражения *мало ли, что ли, чуть ли не, вряд ли, видите ли, знаете ли, видишь ли, знаешь ли*, контексты с отрицанием и вопросы с *не правда ли* и *не так ли* в работе не анализировались. Это тема для отдельного исследования. Для анализа просодической структуры звучащих примеров с частицей *ли* использовалась компьютерная система анализа речи Praat [Boersma, Weenink 2019]. Источником материала послужил мультимедийный подкорпус Национального корпуса русского языка (НКРЯ), а также записи пресс конференций и интервью на видеохостинге you-tube.ru. Звучащие версии примеров, приводимых в работе, доступны на странице сайта Института языкознания РАН http://iling-ran.ru/misc/2019_dialog_audio/. Ссылки на конкретные записи даны в тексте настоящей статьи после каждого примера.

Для разметки используется просодическая нотация, разработанная специально для этого исследования, но продолжающая традицию С. В. Кодзасова, см., например, [Kodzasov, Bonch-Osmolovskaja, Zaharov, Kobozeva, Krivnova 2005, 2006], и, одновременно, учитывающая классификацию русских интонационных конструкций Е. А. Брызгуновой [Bryzgunova 1982b: 97–122]. Это следующие обозначения.

I. \ — падение частоты тона типа ИК-1 с понижением на ударном слоге словоформы-акцентоносителя и дальнейшим понижением или ровным низким тоном на заударных слогах, если они есть [Bryzgunova 1982b: 97–122].

II. \\ — падение частоты основного тона типа ИК-2 с рельефным падением частоты на ударном слоге словоформы-акцентоносителя, причем падению частоты, как правило, предшествует небольшой подъем тона в начале ударного слога [Bryzgunova 1982b: 97–122].

III. / — подъем частоты основного тона на ударном слоге словоформы-акцентоносителя и падение на заударных слогах, если они есть. Если заударных в словоформе нет, заударное падение элиминируется (ИК-3, по [Bryzgunova 1982b: 97–122]).

IV. // — реализация ИК-3 в терминах Е. А. Брызгуновой, представляющая собой отдельную по сравнению с единицей, введенной в предыдущем пункте, просодическую фонологическую единицу. Реализуется при контрасте. При контрасте диапазон частот подъема расширяется, интенсивность звука повышается, длительность звучания слога увеличивается.

V. /- — подъем частоты на ударном слоге плюс ровные или слабо нисходящие заударные (ИК-6, по [Bryzgunova 1982b: 97–122]).

Акцентированная словоформа в примерах ниже обозначается полужирным шрифтом, показатель движения тона ставится после словоформы-акцентоносителя: ***хотим***/ (подъем частоты на ударном слоге словоформы *хотим*).

Особенностью *да-нет*-вопросов с *ли* в русском языке служит то, что иллокутивная сила в них выражена не просодически, как в других русских *да-нет*–вопросах, а сегментным способом. Это частица *ли*: *Мы* ***хотим*** *это знать?* vs. *Хотим ли мы это знать?* Таким образом, просодия в предложениях с *ли* не служит средством выражения иллокутивной силы, но оформляет предложение как отдельную фонетическую единицу и выражает другие — несобственно иллокутивные — значения: контраст и дискурсивную незавершенность.

Порядок слов в вопросительных предложениях с *ли* служит результатом передвижений составляющих в структуре, которая считается исходной [Halpern 1992, 1995]; [Stepanov 1998]; [Rudnickaya 2000]. Эта проблема изучена в указанных работах, и здесь не рассматривается.

Под линейно-просодической структурой мы понимаем последовательность релевантных движений тона в предложении. Для предложения с *ли* — это линейно-просодическая модель вида A_1/- w A_2\, где буква А обозначает словоформу, несущую коммуникативно релевантный акцент, а w — замещает ваккернагелевскую частицу *ли*. Так, модель A_1/- w A_2\ описывает структуру предложения типа ***Хотим***/- *ли мы это* ***знать***\? с подъемом типа ИК-6 на *хотим* и падением типа ИК-1 на *знать*. Пример же (1) из корпуса НКРЯ реализует

ту же лексико-синтаксическую структуру, но иную линейно-просодическую структуру с контрастом на *хотим*:

(1) *<Здесь вот начинается самое интересное — такой некоторый даже парадокс внутреннего развития лингвистики. Можем мы узнать многое, ну мы, лингвисты,> а вот* ***хотим***\\ *ли мы это* ***знать****?*
[НКРЯ] (звучащая запись доступна по ссылке: http://iling-ran.ru/misc/2019_dialog_audio/Plungian3.wav).

Контраст на *хотим* противопоставляет два понятия — желание и нежелание что-либо знать — и выражает сомнения говорящего в том, что возможность получения знания с обязательностью ведет к желанию его получить: ‘можем знать, но необязательно хотим знать’. Контраст предполагает выбор говорящим одного элемента из ограниченного и известного обоим коммуникантам множества элементов и соотнесения мнения говорящего с мнением слушающего или мнением третьего лица; о деталях значения контраста см. [Yanko 2001: 47].

В **разделе 1** ниже дается обзор параметров, релевантных для вопросов с *ли*, в **разделе 2** рассматриваются базовые линейно-просодические модели с *ли*. В **разделе 3** анализируются модели вопросов с *ли* в контексте контраста, в **разделе 4** — в контексте дискурсивной незавершенности, когда говорящий дает понять слушающему, что текущим вопросом не исчерпывается его информационная потребность и он намеревается задать еще один вопрос, или что он хочет сделать добавление к текущему вопросу в виде речевого акта другого типа. В **разделе 5** анализируются линейно-просодические модели в контексте одновременно контраста и незавершенности. В заключении обобщаются полученные результаты.

1. Параметры линейно-просодической структуры вопросов с *ЛИ*

Анализ рабочего массива говорит о том, что вопросительные предложения с *ли* характеризуются следующими параметрами: 1) какова базовая линейно-просодическая структура предложения; 2) включает ли семантическая структура предложения контраст; 3) входит ли предложение в контекст дискурсивной незавершенности; 4) расположен ли первый или единственный акцентоноситель предложения до или после частицы *ли* (иначе — имеется ли в предложении «сдвиг» акцента).

Параметры линейно-просодической структуры вопросов с *ли* объединены в **Таблице**. В заголовках столбцов приведены типы базовых линейно-просодических структур предложений с *ли*. Заголовки строк содержат имена параметров, которые модифицируют базовые структуры. Это контраст, незавершенность и композиция контраста и незавершенности. На пересечении строк и столбцов — расположено по одному из примеров рабочего массива, которые характеризуют соответствующее сочетание значений параметров. Кроме примера, каждая клетка **Таблицы** содержит схему вида A_1/- w A_2\, которая формализует линейно-просодическую модель, соответствующую примеру.

Таблица. Параметры линейно-просодической структуры вопросов с частицей *ли*

Модифицирующие значения		Линейно-просодические последовательности		
	Базовые последовательности	A_1/- w A_2\ ***Есть***/- *ли масса у* ***нейтрино***\? 1	A_1/ w A_2\ ***Предложили***/ *ли вы работу бы* ***Познеру***\? 2	A\ w ***Жива***\ *ли?* 3
Контраст	без сдвига акцента	A_1// w A_2\ ***Весь***// *ли будет капитал возвращаться*\? *<Или не весь?>* 4		A\\ w ***А башня***\\ *ли это Шухова?* 5
	со сдвигом акцента	w A// *Вопрос ли это* ***веры***// *<или научной проекции, научной новой технологии?>* 6		w A\\ *Будет ли усилена* ***российская***\\ *группировка в Калининградской области?* 7
Дискурсивная незавершенность		A_1/- w A_2/ ***Воспринимаете***/- *ли вы это как определенный* ***сигнал***/, *<и как вы считаете, его собственные дипломаты воспримут этот сигнал?>* 8	A_1/ w A_2/ ***Будете***/ *ли вы по-прежнему так* ***смелы***/, *<что придете в суд/…>* 9	A_1\ w A_2/ ***Была***\ *ли явная* ***поддержка***/ *какого-либо течения, <и в каких трудах отражалось их отношение к российским мыслителям?>* 10
Контраст в композиции с незавершенностью	без сдвига акцента	A_1// w A_2/ ***Будет***// *ли Русский* ***марш***/ *<и кто будет его вести…>* 11		A_1\\ w A_2/ Хороши\\ *ли* эти произведения по шкале, так сказать, литературной ценности/ *<мы находим столь противоположные точки зрения/,* что они наводят грустные мысли…> 12
	со сдвигом акцента	w A_1// A_2/ *А есть ли у него* ***вообще***// *представления о том, что его права чего-то* ***стоят***/ *<ваще он чего-то стоит…>* 13		w A_1\\ A_2/ *<Во Франции Жанна Д'Арк была, которая воодушевила все практически войско на то, чтобы встать и продолжить войну,> а были ли в* ***Англии***\\ *какие-то* ***люди***, *<которые тоже вот так помогали войску?>* 14

Строки, соответствующие контрастным вопросам с *ли*, имеют дополнительное подразделение, отражающее возможный перенос акцента на акцентоноситель, «правее» *ли*. Примеров каждого класса, зафиксированного клеткой

таблицы, в рабочем массиве не менее пяти. Только подборка 5–10 примеров, которые не нарушают грамматической корректности, осмысленности и не кажутся автору контринтуитивными, рассматривалась как возможность выделить определенный класс. Выделенные классы описывались содержательно и формально. Неожиданно для нас в рабочем массиве фактически не оказалось примеров с единственным подъемом перед частицей *ли* типа A/ w, если не считать фрагмента стихотворения А. С. Пушкина «Певец»: *Вздохнули/ ль вы?* (чтение М. Лангермана, (звучащая запись доступна по ссылке: http://iling-ran.ru/misc/2019_dialog_audio/pevets.wav). С точки зрения языковой интуиции примеры типа A/ w представляются вполне естественными. В дальнейшем изложении класс A/ w отсутствует. Этот вопрос требует дополнительного анализа. И наконец, отметим следующее: наш материал говорит о том, что между вопросами с *ли* и косвенными вопросами с *ли* принципиальных просодических различий нет. В дальнейшем мы исходим из этого положения.

Перейдем к обсуждению значений параметров просодии.

2. Базовые последовательности

Выделяются три типа линейно-просодических последовательностей вопросов с *ли*. Модель A_1/- w A_2\ — это наиболее распространенная модель, см. Таблицу в предыдущем разделе, где клетка 1 содержит пример наиболее частотной базовой структуры A_1/- w A_2\, а клетки 2 и 3 — менее частотные, но также нейтральные структуры с *ли*. Попутно отметим, что та же модель — A_1/- w A_2\ — реализуется и в повествовательных предложениях (без *ли*) базового (или нейтрального, семантически «минимального») типа (*Пришла/- весна*\; *Дедушка/- сердится*\), а также в вопросах с вопросительным словом: *Который/- час*\?. Совпадение объясняется тем, что в вопросах с *ли* и с *который* (а также с другими вопросительными словами) иллокутивная сила выражается сегментно, то есть просодия в этих случаях играет лишь роль, формирующую предложение как отдельную единицу, а в повествовательном предложении эта же просодия по умолчанию (в силу отсутствия сегментных показателей иллокутивной силы) маркирует речевой акт сообщения (и, одновременно, формирует сообщение как отдельную фонетическую единицу). Назовем модель A_1/- w A_2\ Первой моделью. Эту модель реализует пример (2):

(2) ***Есть****/- ли масса у* ***нейтрино***\? [НКРЯ] (звучащая запись доступна по ссылке: http://iling-ran.ru/misc/2019_dialog_audio/nejtrino.wav).

В примере (2) на первом акцентоносителе реализуется подъем в небольшом диапазоне частот, затем следует практически ровный тон на центральной части предложения вплоть до падения на ударном слоге словоформы *нейтрино* (клетка 1 **Таблицы**).

В примере (3) реализована модель A_1/ w A_2\ с более крутым подъемом, чем в Первой модели, на ударном слоге первого акцентоносителя словоформе *предложили* плюс падение на заударных слогах. Завершается предложение падением на втором акцентоносителе *Познеру* (клетка 2).

(3) ***Предложили****/ ли вы работу бы* ***Познеру****\?*
[НКРЯ] (звучащая запись доступна по ссылке:
http://iling-ran.ru/misc/2019_dialog_audio/Pozner.wav).

В предложении также имеются своего рода риторический акцент на *вы* и небольшой подъем на первой фазе ударного слога словоформы *Познеру*. Это указывает на четкость произнесения, присущего человеку, который умеет следить за своей речью и старается говорить ясно. Риторические акценты не меняют в данном случае исходной линейно-просодической модели предложения. Назовем модель A_1/ w A_2\ Второй моделью.

Линейно-просодическая модель примера (4) (клетка 3) сформирована единственным акцентоносителем — словоформой *жива*. Модель реализуется с падением: A\ w.

(4) ***Жива****\ ли*? [НКРЯ] (звучащая запись доступна по ссылке:
http://iling-ran.ru/misc/2019_dialog_audio/ZhivaLI.wav).

Назовем модель A\ w Третьей моделью.

Примеры (1)–(4) позволяют сделать достаточно парадоксальный вывод о том, что начало вопросов с *ли* может быть сформировано как по восходящей, так и по нисходящей модели движения частоты основного тона, ср. *Слыхали/ ль вы?* vs. *Слыхали\ ль вы?*.

3. Вопросы с *ЛИ* и контраст

Наша гипотеза состоит в том, что соединение предложений Первой и Второй модели с контрастом дает модель A_1// w A_2\ с подъемом в больших диапазонах частот на ударном слоге первого акцентоносителя (плюс падение на заударных) и падением — на втором акцентоносителе (клетка 4):

(5) ***Весь****// ли будет капитал возвращаться\? <Или не весь?>*
(звучащая запись доступна по ссылке:
http://iling-ran.ru/misc/2019_dialog_audio/VesjLI.wav).

Третья модель в контексте контраста дает падение (типа ИК-2) на единственном акцентоносителе предложения A\\ w (клетка 5):

(6) ***А башня****\\ ли это Шухова?* [НКРЯ] (звучащая запись доступна по ссылке:
http://iling-ran.ru/misc/2019_dialog_audio/bashnyaShuxova.wav).

Далее. В контексте контраста первый акцентоноситель модели может находиться «справа» от *ли*. В примере (7) (клетка 6) реализуется модель w A_1// с подъемом частоты основного тона на первом акцентоносителе, а в примере (8) (клетка 7) — модель w A\\ — с падением на единственном акцентоносителе:

(7) *Вопрос ли это* ***веры****// <или научной проекции\, научной новой технологии?>* [НКРЯ] (звучащая запись доступна по ссылке:
http://iling-ran.ru/misc/2019_dialog_audio/voprosVery.wav).

(8) *Будет ли усилена* ***российская***\\ *группировка в Калининградской области?* (звучащая запись доступна по ссылке: http://iling-ran.ru/misc/2019_dialog_audio/Kaliningrad.wav).

Основное обобщение, которое служит итогом данного раздела состоит в том, контекст контраста способен смещать коммуникативно релевантный акцент с первой словоформы в предложении в позицию после *ли*.

4. Вопросы с *ЛИ* и дискурсивная незавершенность

В контексте дискурсивной незавершенности вопрос с *ли* имеет восходящее движение частоты основного тона на втором акцентоносителе. Первый акцентоноситель в Первой, Второй и Третьей моделях сохраняет исходное движение тона. Обратимся к примерам.

Модель, производная от Первой модели, в контексте незавершенности сохраняет первый акцент типа ИК-6. В контексте незавершенности возникает модель A_1/- w A_2/ (пример (9), клетка 8).

Производная от Второй модели сохраняет первый акцент типа ИК-3 (модель A_1/ w A_2/, пример (10), клетка 9), производная от Третьей модели — нисходящий акцент типа и ИК-1 (модель A_1\ w A_2/, пример (11), клетка 10):

(9) ***Воспринимаете****/- ли вы это как определенный* ***сигнал****/, <и как вы считаете, его собственные дипломаты воспримут этот сигнал?>* [НКРЯ] (звучащая запись доступна по ссылке: http://iling-ran.ru/misc/2019_dialog_audio/signal.wav).

(10) ***Будете****/ ли вы по-прежнему так* ***смелы****/, <что придете в суд/ …>* (звучащая запись доступна по ссылке: http://iling-ran.ru/misc/2019_dialog_audio/svidetelj.wav).

(11) ***Была***\ *ли явная* ***поддержка****/ какого-либо течения, <и в каких трудах отражалось их отношение к российским мыслителям?>* [НКРЯ] (звучащая запись доступна по ссылке: http://iling-ran.ru/misc/2019_dialog_audio/podderzhka.wav).

Итак, в контексте дискурсивной незавершенности вопрос с *ли* имеет два акцентоносителя, второй из которых следует восходящей модели типа ИК-3 по Е. А. Брызгуновой с подъемом на ударном слоге и падением на заударных (если они есть). Восходящее движение частоты основного тона и выражает значение незавершенности.

5. Вопросы с *ЛИ* в композиции с контрастом и дискурсивной незавершенностью

Композиция контраста и дискурсивной незавершенности в применении к вопросам с *ли* дает ожидаемые классы. Это контрастный подъем на первом акцентоносителе и менее рельефный подъем на втором акцентоносителе

(модель A_1// w A_2/, пример (12), клетка 11) и контрастное падение типа ИК-2 на первом акцентоносителе и подъем на втором акцентоносителе (модель A_1\\ w A_2/, пример (13), клетка 12).

(12) ***Будет***// *ли Русский* ***марш***/ *<и кто будет его вести…>*
[НКРЯ] (звучащая запись доступна по ссылке: http://iling-ran.ru/misc/2019_dialog_audio/RusskijMarch.wav).

(13) Хороши\\ *ли* эти произведения по шкале, так сказать, литературной ценности/ <мы находим столь противоположные точки зрения/, что они наводят грустные мысли …> [НКРЯ] (звучащая запись доступна по ссылке: http://iling-ran.ru/misc/2019_dialog_audio//Zaliznyak.wav).

И, как и в примерах (7) и (8), линейно-просодические последовательности могут быть реализованы с расположением первого акцентоносителя в позиции после частицы *ли*: пример (14) с подъемом (модель w A_1// A_2/, клетка 13) и пример (15) — с падением (модель w A_1\\ A_2 /, клетка 14).

(14) *А есть ли у него* ***вообще***// *представления о том, что его права чего* ***стоят***/ *<ваще он чего-то стоит…>*[НКРЯ] (звучащая запись доступна по ссылке: http://iling-ran.ru/misc/2019_dialog_audio/Petranovskaja.wav).

(15) *<Во Франции Жанна Д'Арк была, которая воодушевила все практически войско на то, чтобы встать и продолжить войну,> а были ли в* ***Англии***\\ *какие-то* ***люди***/, *<которые тоже вот так помогали войску?>*
[НКРЯ] (звучащая запись доступна по ссылке: http://iling-ran.ru/misc/2019_dialog_audio/JeanneDArc.wav).

Контрастная интерпретация примера (14) сформирована лексемой *вообще*, склонной к контрастному употреблению. Контраст здесь основан на противопоставлении *вообще* vs. *в частности*, ср.: *А он* ***вообще*** *ничего не знает* (≈ 'не только этого, а совсем ничего'); *Грамотный психолог* ***вообще*** *никогда не станет давать советов* [НКРЯ]; *Рыбок мужчины* ***вообще*** *не замечают* [НКРЯ].

Заключение

Предложен анализ просодии вопросов с частицей *ли*. Выделено три типа вопросов с *ли*, которые понимаются как исходные, т. е. такие, у которых просодическая структура вносит минимальный вклад в семантическую структуру предложения. Просодия предложений в контексте коммуникативного контраста, дискурсивной незавершенности и композиции контраста и незавершенности рассматривается как производная от просодии исходных предложений.

В вопросах с *ли* частица *ли* сохраняет свои ваккернагелевские свойства, а словоформа — хозяин клитики служит в большинстве предложений носителем первого (или единственного) коммуникативно релевантного акцента. Однако в контексте контраста носитель коммуникативно релевантного акцента может находиться и «правее» *ли*.

Иллокутивная сила в вопросах с *ли* выражена не просодически, как в других русских *да-нет*–вопросах, а сегментным способом: с помощью частицы *ли*. Таким образом, просодия в предложениях с *ли* не служит средством выражения иллокутивной силы, но оформляет предложение как отдельную фонетическую единицу и выражает другие — несобственно иллокутивные — значения: контраст и дискурсивную незавершенность. При этом, в вопросе с *ли* акцент на первой акцентированной словоформе может быть и восходящим, и нисходящим без заметного различия в значении.

В контексте дискурсивной незавершенности вопрос с *ли* имеет два акцентоносителя, первый из которых может быть как восходящим, так и нисходящим, в частности, несущим контрастный подъем или падение, а второй — следует восходящей модели типа ИК-3 по Е. А. Брызгуновой.

References

1. *Bryzgunova E. A.* (1982a) Expressing the unknowns in question by lexical, contextual, and intonational means [Sredstva vyrazhenija neizvestnogo v voprose (vzaimodejstvie leksiki, konteksta i intonacii] // Russian Grammar [Russkaya grammatika]. Vol. 2, Nauka, Moscow, pp. 397–402.
2. *Bryzgunova E. A.* (1982b) Intonation [Intonatsiya], Russian Grammar [Russkaya grammatika]. Vol. 1, Nauka, Moscow, pp. 98–118.
3. *Boersma P., Weenink D.* (2019). Praat: Doing phonetics by computer. Version 6.0.47. Online: http://www.praat.org/.
4. *Halpern A. L.* (1992) Topics in the Placement and Morphology of Clitics. PhD Dissertation, Stanford University.
5. *Halpern A. L.* (1995) On the Placement and Morphology of Clitics, CSLI Publications, Stanford.
6. *Kedrova G. E., Potapov V. V., Omel'janova E. B., Egorov A. M.* (2002) Sentences with li particle [Predlozhenija s chasticej li] // Russian Phonetcs [Russkaja fonetika], available at: http://fonetica.philol.msu.ru/intonac/m321.htm.
7. *Kodzasov S. V., Bonch-Osmolovskaja A. A., Zaharov L. M., Kobozeva I. M., Krivnova O. F.* (2005) Data Base 'Intonation of Russian Dialogue: Interrogative Phrases' [Baza dannyh «Intonacija russkogo dialoga»: voprositel'nye repliki] // Proceedings of the International Conference "Dialog 2005". P. 245–247.
8. *Kodzasov S. V., Arhipov A. V., Bonch-Osmolovskaja A. A., Zaharov L. M., Krivnova O. F.* (2006) Data Base 'Intonation of Russian Dialogue: Commanding Propositions' [Baza dannyh «Intonacija russkogo dialoga»: pobuditel'nye repliki] // Proceedings of the International Conference "Dialog 2006". P. 236–242.
9. *Rudnickaya E.* (2000). The derivation of yes–no li questions in Russian: syntax and/or phonology? // Formal approaches to Slavic linguistics: the Philadelphia meeting, 1999, ed. Tracy h. King and Irina Sekerina, 347–362. Ann Arbor, Mich: Michigan Slavic publications.
10. *Stepanov A.* (1998) On wh-fronting in Russian // P. Tamanji and K. Kusumoto (eds.), Proceedings of the North Eastern Linguistic Society (28), UMASS, Amherst. 453–467.

11. *Yanko T.* (2001) Communicative strategies of the Russian speech [Kommunikativnye strategii russkoj rechi]. Moscow: Jazyki slavjanskoj kul'tury.
12. *Zaliznjak A. A.* (2008) The Old-Russian enclitics [Drevnerusskie enklitiki]. M.: "Jazyki slavjanskih kul'tur".
13. *Zimmerling A. V.* (2013) Word order systems in Slavic languages from a typological perspective [Sistemy porjadka slov slavjanskih jazykov v tipologicheskom aspekte]. Moscow: Jazyki slavjanskoj kul'tury, 2013, 544 p.

РУССКОЕ *ЧТО-ТО* КАК ДИСКУРСИВНОЕ СЛОВО[1]

Зализняк Анна А. (anna.zalizniak@gmail.com)
Институт языкознания РАН, Москва; Институт проблем информатики ФИЦ ИУ РАН, Москва

Падучева Е. В. (elena.paducheva@yandex.ru)
Институт проблем информатики ФИЦ ИУ РАН, Москва; Государственный институт русского языка имени А. С. Пушкина, Москва

В докладе демонстрируется, что в русском языке имеется дискурсивное слово *что-то*, которое может выражать определенный спектр установок говорящего по отношению к некоторому (наблюдаемому им) обстоятельству, отклоняющемуся от нормы. А именно, дискурсивное *что-то* может маркировать: желание говорящего обратить внимание слушающего на сообщаемый факт, не интересуясь специально его причиной (ср. *Что-то я на склоне лет стал сентиментален*), желание говорящего выразить осуждение (ср. *Что-то она слишком вырядилась сегодня*) или просто сообщить о чем-то негативном (ср. *Что-то сегодня пасмурно*, но ??*Что-то сегодня светит солнце*); выразить свою тревогу или подозрение (ср. *Что-то в детской слишком тихо*); желание ослабить категоричность негативного или потенциально обидного для собеседника высказывания, в частности — смягчить резкость отказа (ср. — *Давай чай пить! — Что-то не хочется*) и др. Показано, что выделяемое в словарях значение *что-то* ‘непонятно почему’ возникает лишь в определенных контекстных условиях. Выявлены условия возникновения этого значения и его место в цепи семантической деривации, исходной точкой которой является значение неопределенного объекта. Исследование проведено на материале Национального корпуса русского языка, в том числе его параллельных подкорпусов.

Ключевые слова: дискурсивные слова, русский язык, неопределенные местоимения, говорящий, причина, семантическая деривация, параллельный корпус, перевод

[1] Работа выполнена при поддержке Российского научного фонда, проект РНФ 18-18-00462, реализуемый в Государственном институте русского языка им. А. С. Пушкина (разделы 1–2) и при частичной поддержке РФФИ, проект № 17-29-09124 (раздел 3).

RUSSIAN CHTO-TO AS A DISCOURSE MARKER

Zalizniak Anna A. (anna.zalizniak@gmail.com)
Institute of Linguistics of the RAS; Research Centre of Computer Science and Control of the RAS

Paducheva E. V. (elena.paducheva@yandex.ru)
Federal Research Centre of Computer Science and Control of the RAS; Pushkin State Russian Language Institute

The paper demonstrates that the Russian indefinite pronoun *chto-to* ('something') can function as a discourse marker, which expresses a range of attitudes of the speaker with respect to a situation he/she considers deviating from the norm. Namely, using the discursive *chto-to* the speaker may draw the listener's attention to the reported fact, not being interested in its cause (cf. *Chto-to ja na sklone let stal sentimentalen* '*Chto-to* in my declining years, I became sentimental'); he/she may express reprobation (cf. *Chto-to ona slishkom vyrjadilas' segodnja* '*Chto-to* she is too dressed up today') or simply report something negative (cf. *Chto-to segodnja pasmurno* '*Chto-to* it is cloudy today' and ??*Chto-to segodnja svetit solnce* '*Chto-to* the sun is shining today'; *chto-to* may also express anxiety or suspicion (cf. *Chto-to v detskoj slishkom tixo* '*Chto-to* it is too quiet in the nursery'), the desire to soften the effect of a negative or potentially offensive for the interlocutor utterance, in particular, to soften the sharpness of the refusal (cf. *Davaj chaj pit'! — Chto-to ne xochetsja* 'Let's have tea! — *Chto-to* I don't want it'), and other attitudes. It is demonstrated that the meaning 'unclear why' attributed to *chto-to* by dictionaries arises only in certain contexts. The conditions for the emergence of the discursive meaning of *chto-to* are identified and its place in the semantic derivation chain is revealed. The research is based on Russian National Corpus.

Key words: Russian language, indefinite pronouns, speaker, cause, semantic derivation, parallel corpora, translation

1. Вводные замечания

В своем основном значении *что-то* — это неопределенное местоимение, относящееся к объекту, который говорящий не в состоянии идентифицировать, т. е. '(мне) неизвестно что' [Падучева 1985: 210–211], ср. *Там что-то случилось*; *Он что-то натворил*. Назовем это значение **актантным**.

Слово *что-то* может быть также наречием. В словаре МАС указано два наречных значения. Первое обозначено как «в некоторой степени, несколько, вроде» и иллюстрируется примерами:

(1) Ее находят *что-то* странной (Пушкин. Евгений Онегин);

(2) [Лизе] становилось *что-то* неприятно (Толстой. Два гусара).

Оба эти примера демонстрируют устаревшее употребление. В качестве подзначения указано «приблизительно, примерно», иллюстрируемое предложением (3), которое, как кажется, не вписывается в предлагаемую экспликацию, но при этом, наоборот, полностью соответствует современной норме.

(3) Однажды при дворе она проиграла герцогу Орлеанскому *что-то* очень много. (Пушкин. Пиковая дама)

В Словаре языка Пушкина этот пример входит в группу, обозначенную «как-то, как будто, несколько, в некоторой степени». Если исходить из предположения, что в данной точке норма не изменилась[2], то *что-то* выражает здесь смесь удивления с некоторым неодобрением (об этом значении пойдет речь ниже).

Что касается значения «приблизительно, примерно», то в современном языке оно представлено лишь в синтаксических группах с количественным числительным; при этом значение приблизительности должно быть продублировано либо конструкцией с постпозицией числительного (ср. *Мы проработали там что-то месяца полтора*), либо словом *около* (*Он вернулся что-то около одиннадцати часов*); нельзя сказать **что-то полтора месяца*, **что-то в одиннадцать часов*. Тем самым реально на долю *что-то* приходится лишь акцент на уже обозначенной приблизительности.

Назовем это значение слова *что-то* **количественным** (= значение количественной неопределенности).

Второе наречное значение для *что-то* в МАС отмечено как «почему-то, неясно почему» и иллюстрируется примерами:

(4) — Максим Максимыч, не хотите ли чаю? — закричал я ему в окно. — Благодарствуйте; *что-то* не хочется. (Лермонтов. Герой нашего времени)

(5) *Что-то* уж очень строга Сашенька! Все приказывает! (М. Горький. Мать)

Приведем аналогичные примеры из НКРЯ:

(6) — Я позавтракал дома, и вдруг *что-то* голова разболелась, захотелось проехаться, я проезжал мимо… совсем нечаянно, не думал… [И. И. Панаев. Раздел имения (1850–1860)]

(7) — Не знаю, *что-то* беспокойно на душе, — ответила тетя Рая. [Маша Трауб. Замочная скважина (2012)]

(8) Я кончал институт, а кафедра не принимала мою дипломную работу. *Что-то* она им там не понравилась. Она даже в какой-то мере их испугала. [Фазиль Искандер. Должники (1968)]

[2] Подчеркнем, что всего лишь гипотеза, обсуждение которой выходит за рамки настоящей статьи.

(9) — Ты закусывай, закусывай, — сказал Шура. — *Что-то* ты, Серёга, воодушевился сверх меры… [Андрей Волос. Недвижимость (2000) // «Новый Мир», 2001]

Как легко убедиться, для всех примеров (4)–(9) экспликация «почему-то, неясно почему» является весьма неточной (в особенности — для (5), (9)). В частности, замена *что-то* на *почему-то* не всегда возможна (ср. ??*Почему-то уж очень строга Сашенька!*); в любом случае такая замена ведет к существенному искажению смысла, одновременно сужая значение высказывания и увеличивая уровень эксплицитности выражения идеи неясной причины.

2. Семантика и дискурсивные функции слова *что-то*

Значение слова *что-то*, представленное примерами (4)–(9), будем называть **дискурсивным** — и далее пойдет речь только о нем. Как мы попытаемся показать, слово *что-то* может выражать целый спектр коммуникативных и прагматических установок — как связанных, так и не связанных с идеей неясной причины. Существенно при этом, что все эти установки выражаются при помощи *что-то* в неявной форме.

В книге [Николаева 1985: 55] отмечается, что высказывания типа *Что-то я устала*; *Что-то сон одолевает* описывают «несоответствие, отклонение от нормы, от позитивного сценария событий»; это ситуация, «генезис которой неясен и сама она еще как бы не определилась точно. Чаще всего это касается человека, его душевного и физического состояния». Мы полагаем, что значение «неясной причины» (мы будем называть его «квазипричинным») возникает в результате конвенционализации следующей импликатуры: если говорящий наблюдает некоторое странное положение вещей, то естественно возникает желание найти его причину, которая объяснила бы наблюдаемое отклонение от нормы. В этом случае компонент ‘непонятно почему’ становится частью значения. Но это происходит не всегда.

В целом можно сказать, что дискурсивное *что-то* чаще всего употребляется в высказываниях от 1-го л., описывающих внутреннее (ментальное, эмоциональное, перцептивное, физиологическое) состояние говорящего, так или иначе отклоняющееся от «нормы» — и тем самым воспринимаемое как негативное; часто оно содержит отрицание, ср: *что-то не понимаю, не помню, не по себе, не хочется, не нравится, не спится; устал, разболелась голова* и т. п., но не ??*что-то понимаю, помню, хорошо себя чувствую* и т. п.[3]

О тесной связи дискурсивного *что-то* с описанием внутреннего состояния свидетельствуют следующие факты.

Поисковый запрос в НКРЯ «*что-то*[Nom] + *я* + *не* + V,praes» дает 113 примеров. Это: *что-то я не понимаю* (14 примеров плюс еще 6 с другими глаголами

[3] Ср. [Сахно 1983] о связи показателей неопределенности с отрицанием, [Арутюнова 1998: 821–823] о «негативности» как благоприятных контекстах для употребления неопределенных наречий.

с тем же значением: *не улавливаю, не врубаюсь, не въезжаю, не просекаю*, т. е. всего с глаголами, означающими ‘не понимаю’, т. е. всего 20 примеров); *что-то мне не нравится* (17), *что-то я не помню* (16); кроме того, встречаются: *что-то мне не верится, не кажется, не хочется, не спится*; *что-то не вижу, не знаю, не верю, не могу, не чувствую, не доверяю*. Таким образом оказывается, что 100 % примеров, содержащих последовательность «*что-то*[Nom] + *я* + *не* + V,praes», включают глаголы внутреннего состояния; глаголы других классов в этой конструкции не встречаются.

Из 458 примеров на поисковый запрос «*что-то*[Nom] + *не* + V,praes,sg,3p» около половины составляют примеры с дискурсивным *что-то*, из них 73 (т. е. около 30 %) — *что-то не хочется*.

Еще более показательным представляется результат поискового запроса «*что-то*[Nom] + *не* + V,fut,1p». Он дает 156 примеров; все они (за исключением трех, представляющих собой «шум») содержат форму praesens perfectivum и описывают некоторое актуальное отрицательное неконтролируемое ментальное состояние[4], а именно: *не припомню* (71 пример), *не пойму* (50), *не разберу* (13), а также: *не упомню, не вспомню, не врублюсь, не соображу, не уловлю, не соберусь с мыслями*.

Тем самым, корпусные данные позволяют выделить в русском языке конструкцию «*что-то*[Nom] +(*я*) + *не* + V,praes», обозначающую негативное внутреннее состояние и конструкцию «*что-то* + *не* + V,fut,1p» с еще более узким значением нечеткого или неполного ментального образа.

Констатация некоторого отклоняющегося от нормы положения вещей естественным образом вызывает ощущение странности, непонимания. Поэтому для дискурсивного *что-то* особенно характерен контекст глагола со значением ‘не понимать’, в котором *что-то* как бы дублирует его значение, ср.:

(10) — Гениально-то оно, конечно, гениально, — сказал Сергей Борисович, — но ***что-то*** *я не совсем понимаю*, кого это вы собираетесь ревизовать? [В. П. Катаев. Алмазный мой венец (1975–1977)]

(11) Ген, а каких хороших людей она имела ввиду, ***что-то*** *я не въехала*? [Марина Зосимкина. Ты проснешься. Книга первая (2015)]

— Ну хорошо, — сказал напоследок я. — ***Что-то*** *я ничего не пойму*. Короче, скоро увидимся. [Андрей Волос. Недвижимость (2000) // «Новый Мир», 2001]

Показательно, что контекст ‘не понимать’ «стимулирует» появление слова *что-то* в переводе на русский язык, ср. примеры (48), (49) ниже.

С другой стороны, для дискурсивного *что-то* характерен контекст предикатов, обозначающих неуверенное умозаключение (сделанное, в том числе, на основе обработки перцептивных данных): *похоже*, *кажется* ср.:

[4] Ср. [Зализняк 2015b] о связи формы praesens perfectivum с отрицанием и семантикой внутреннего состояния.

(12) Мне позарез надо встретиться с товарищами, разработать план действий, выяснить, как жили и работали без нас молодые, ***что-то*** *похоже*, что они решили идти по легкой дорожке, хотят свернуть движение на экономическую борьбу. [З. И. Воскресенская. Сердце матери (1963–1965)]

(13) — ***Что-то*** ваше лицо *кажется* мне знакомым, — сказал он, пристально разглядывая миллиардера. [Елена Туева. Свет в Windows'е (2000) // «Карьера», 2000.02.01]

Ср. также примеры (42), (43) ниже, где глаголы *to seem* и *to look* выполняют роль непосредственного «стимула» при переводе с английского.

2.1. Квазипричинное значение *что-то*

Идея «неясной причины» обнаруживает себя, прежде всего, в том случае, когда за высказыванием с *что-то* следует обсуждение причины описываемого положения дел, ср.:

(14) — *Что-то* плохо мне. Видно, давление поднялось на нервной почве. [Даниил Корецкий. Менты не ангелы, но… (2011)]

(15) Я *что-то* еле-еле потом наверх выползаю… Это почему? [Юлия Лавряшина. Улитка в тарелке (2011)]

(16) — *Что-то*, я смотрю, от жены ты бегаешь, чего вы там не поделили? [А. А. Фадеев. Разгром (1925–1926)]

Показательно, что в примере (17) переводчик с английского языка даже добавил слово *что-то* — в функции текстового коннектора, проясняющего логическую структуру фразы (слово *что-то* в русском переводе маркирует тот факт, что вторая часть является объяснением причины странного обстоятельства, отмеченного в первой):

(17) "*I don't remember this cliff*," said Jack, crestfallen, "so this must be the bit of the coast I missed". [William Golding. Lord of the Flies (1954)]
Что-то *я этого утеса не помню*, — сказал Джек, он заметно увял, — значит, я пропустил это место. [Уильям Голдинг. Повелитель мух (Е. Суриц, 1985)]

Для данного типа употребления *что-то* характерно наличие вводного глагола *смотреть*, эксплицирующего тот факт, что умозаключение было сделано на основе перцептивных данных, ср.(16), а также (18), (19):

(18) — А товарищ гвардии полковник ***что-то***, *я смотрю*, заскучал… [Г. Н. Владимов. Генерал и его армия (1994)]

(19) ***Что-то***, *я смотрю*, ты про начальников-то легко говорить стал… смотри, худа не было бы! [Максим Горький. Мещане (1901)]

У такого квазипричинного *что-то* имеется соотносительное *то-то*, указывающее на факт внезапного обнаружения говорящим причины обратившего ранее на себя внимание странного обстоятельства[5]; реплика с *то-то* возникает как реакция на высказывание (собеседника или самого говорящего), содержащее указание на эту причину, ср. примеры (20)–(25):

(20) А к вашему ребенку, по всей видимости, он относится с особым неравнодушием. *То-то* он так запереживал, когда вы перестали водить сына в бассейн! [Станислав Акимов. Чужие письма: журнал «Малые народы» (2004) // «Хулиган», 2004.07.15]

(21) Это и боцман наш учуял, Страшной, *то-то* он ему и врезал. [Георгий Владимов. Три минуты молчания (1969)]

Для этого значения особенно характерна конструкция *то-то я смотрю*[6], которая может использоваться как прием комического, ср. примеры (24), (25):

(22) — У нас света не было. — А-а. Тогда понятно. *То-то я смотрю* в подъезде темно. — [Андрей Геласимов. Фокс Малдер похож на свинью (2001)]

(23) — […] Госдолг Японии достиг рекорда — почти 12 триллионов долларов. — *То-то я смотрю*, два десятка лет, как ни одной новой японской машины не видел. [коллективный. Форум: Скоро Японию ждёт дефолт? (2012)]

(24) — Вы знаете, Рабинович умер! — *То-то я смотрю* его вчера хоронят. (анекдот)

(25) Мама, правда, что люди от обезьяны произошли? — Правда. — *То-то я смотрю*: обезьян так мало стало. [К. И. Чуковский. От двух до пяти (1933–1965)]

Особо следует отметить класс употреблений *что-то*, реализующий следующую риторическую фигуру. При помощи *что-то* говорящий подает описываемое положение вещей как странное, не имеющее понятной причины; между тем, причина эта на самом деле ему ясна (или по крайней мере он имеет предположение, представляющееся ему весьма вероятным), и именно свое знание или подозрение об этой причине говорящий хочет сообщить слушающему, ср.:

(26) [муж спрашивает жену] — Где ты сегодня была? — На выставке Тинторетто. — Я тоже там сегодня был, но *что-то* тебя там не видел[7].

---

[5] В МАС про это значение слова *то-то* говорится: «Употребляется для выражения внезапно возникшей догадки в значении: так вот почему»: *Значит старик-то теперь один; то-то он и повадился ко мне ходить*. (Островский. Не все коту масленица). Следует уточнить, что речь здесь идет о догадке, касающейся объяснения некоторого факта, вызывавшего ранее его недоумение.

[6] Сочетание *то-то я смотрю* составляет 18,8% (67 из 356 примеров на поисковый запрос «*то-то я*»); плюс «то-то я |1| гляжу» (38) и «то-то я |1|вижу» (27), т.е. в сумме 36,8%; другие наиболее частотные глаголы в этом контексте: *замечаю*, *думаю*, *чувствую*, *наблюдаю*, *примечаю*.

[7] Такое же «разоблачительное» риторическое значение есть у неопределенного наречия *почему-то*, ср. *…но почему-то я там тебя не видел*.

Говорящий, очевидно, хочет таким образом передать смысл ‘Ты не была на выставке’ и таким образом уличить жену во лжи: тот факт, что она не была на выставке, является наиболее вероятной (для говорящего) причиной того *якобы странного* обстоятельства, что он ее там не видел. При этом данный смысл не выражается явно, а имплицируется. Эта риторическая фигура «выведения на чистую воду» порождается именно словом *что-то*: то же высказывание без *что-то* вышеупомянутой импликации не содержит. Этот тип квазипричинного значения *что-то* будем называть **риторическим.** Риторическое *что-то не вижу* (*не видел, не встречал* и т. п.) — это косвенный способ передать смысл ‘этого нет’, ср:

(27) — Два или более погибших в результате грубого нарушения. От четырёх до десяти лет. В зависимости от смягчающих. Только я пока смягчающих *что-то не вижу*. [Андрей Волос. Недвижимость (2000) // «Новый Мир», 2001]

(28) А я *что-то не видел* мужей, бегающих по магазинам в поисках особой бесцветной помады или вышивающих на подушках петухов. [В. Орлов. Как я стал подводным охотником (Исповедь, содержащая ряд практических советов) // «Спортсмен-подводник», 1963]

Обратим внимание на тот факт, что смысл ‘этого нет’ может быть эксплицирован при переводе, например, на английский язык, ср.:

(29) — *Что-то я ни разу в жизни не встречал* стоголовых драконов! — недоверчиво усмехнулся Знайка. [Н. Н. Носов. Приключения Незнайки и его друзей (1953–1954)]
“Pooh, pooh!” scoffed Doono. “*As if there was such a thing as* a hundred-headed dragon!” [Nikolay Nosov. The Adventures of Dunno and his Friends (Margaret Wettlin, 1980)]

Наблюдение некоторого отклоняющегося от нормы положения вещей может вызывать тревогу, обусловленную предположением о каком-то негативном обстоятельстве, послужившем его причиной (т. е. подозрением); это подозрение может быть высказано в форме вопроса, ср.:

(30) Тут ее мобильник запел голосом Юлии Савичевой про белогривые лошадки, значит, Викуся рвется к общению. *Что-то рано*. Или она в школу опять не пошла? [Марина Зосимкина. Ты проснешься. Книга первая (2015)] рано, и это странно (может быть, плохо), что рано.

(31) — Где ж этот рыженький, что с вами плавал, сердитый такой? *Что-то* я не вижу его. Он часом, не утоп ли? [Георгий Владимов. Три минуты молчания (1969)] странно, подозрительно

Отметим, что само слово *подозрительно* часто появляется в контексте *что-то*, ср.:

(32) — «Иду, — говорю, — вчера вечером по набережной, а впереди мужичок с этой вот сумкой крадется… И ***что-то*** показался он мне *подозрительным*… [Евгений Лукин. В стране заходящего солнца (1999)]

(33) Я листаю исписанные страницы. Которых ***что-то*** *подозрительно* много.
[Мариам Петросян. Дом, в котором… (2009)]

2.2. Прочие дискурсивные значения *что-то*

До сих пор речь шла о тех контекстах употребления дискурсивного *что-то*, где оно так или иначе соотносится с идеей неясной причины описываемого положения вещей и попыткой ее прояснения. Однако имеется обширный класс контекстов, не содержащих отсылки к причине.

Так, высказывание *Что-то вылет нашего самолета задерживается* может выражать как беспокойство относительно причины этого факта (какая-то неполадка в самолете), так и беспокойство относительно его последствий, независимо от его причины, ср. возможное продолжение: *Как бы нам не опоздать на пересадку.*

Идея отклонения от нормы, выражаемая дискурсивным *что-то*, обнаруживает себя в характерной сочетаемости дискурсивного *что-то* с наречиями *очень*, *слишком*, *больно*, *совсем* часто сопровождаемыми частицей *уж*, ср.:

(34) — Пардон, дорогая, — перебила я. — Это ***что-то*** *уж слишком* изысканно. Мой слабый ум не выдерживает… [Вера Белоусова. По субботам не стреляю (2000)]

(35) ***Что-то*** *уж* цифра *больно* нескромная. Три вагона тротила…
[коллективный. Форум: Фукусима — природа взрывов (2011)]

(36) — Мам, ты ***что-то*** *совсем* замоталась, — начала Маринка.
[Маша Трауб. Замочная скважина (2012)]

Во всех примерах (34)–(36) речь идет не о поиске причины, здесь говорящий просто **привлекает внимание** к описываемой ситуации при помощи подачи факта как отклоняющегося от нормы; ср. также:

(37) … ***Что-то*** я на склоне лет стал сентиментален…
[В. П. Катаев. Алмазный мой венец (1975–1977)].

Дискурсивное *что-то* может также выполнять функцию смягчения резкости отказа (ср. пример (4) выше), а также **уменьшения категоричности** утверждения — переводя высказывание из утверждения в предположение, т. е. выступает в функции «ограничителя» («hedge», ср. Lakoff 1972), ср.:

(38) ***Что-то*** они хотят, я вижу, нас голодом заморить.
[Ю. О. Домбровский. Факультет ненужных вещей, часть 5 (1978)]

Поскольку неопределенность тесно связана с неконтролируемостью (ср., в частности [Арутюнова 1998: 820–823]), дискурсивное *что-то* может использоваться для снятия с себя ответственности путем представления своих действий как не полностью контролируемых[8], ср.:

[8] В этой своей функции дискурсивное *что-то* входит в богатый арсенал средств русского языка, описанный в [Зализняк, Левонтина 1996].

(39) Но, живя в этой стране вот уже пятый десяток, присматриваясь к нашей жизни, ***что-то*** *я потерял* из виду этого ортодокса. [Владимир Войнович. Иванькиада, или рассказ о вселении писателя Войновича в новую квартиру (1976)] речь не идет о причине

(40) Ладно, давай выпьем, а то я ***что-то*** совсем тебя *заболтал*…
[Даниил Корецкий. Менты не ангелы, но… (2011)]

В (40) неконтролируемость, вносимая словом *что-то*, служит основанием для своего рода **извинения**.

Отклоняющееся о нормы положение дел может вызывать **осуждение**, ср.:

(41) ***Что-то*** Люба сегодня очень уж раздухарилась, колоколит и колоколит, накрывая на стол, прыгает, галдит, того и гляди чего-нибудь на ней из туго её облегающей одежды от резвости лопнет!..
[Виктор Астафьев. Обертон (1995–1996)]

Во всех этих случаях значение *что-то* не связано с идеей неясной причины. Тем самым, «неясность причины» — это одно из производных значений дискурсивного *что-то*, наряду с другими.

2.3. Данные параллельных корпусов

В подтверждение сказанного приведем некоторые результаты анализа параллельных текстов из английского подкорпуса НКРЯ, содержащих дискурсивное *что-то*.

А именно, дискурсивное *что-то* в русском переводе может иметь в английском оригинале следующие «стимулы»[9]:

1. Глаголы *to seem* и *to look*, выражающие идею некоторой неопределенности в описании ситуации:

(42) "*Nothing* ***seems*** *to happen*," said Zeb, doubtfully. [L. Frank Baum. Dorothy and the Wizard in Oz (1908)]
— ***Что-то*** *ничего не происходит*, — с сомнением сказал Зеб.
[Л. Фрэнк Баум. Дороти и Волшебник в Стране Оз (Т. Д. Венедиктова, 1992)]

(43) *You* ***look*** *as ruddy as your native vine*, but we are not getting any younger, as the amerlocks say, and that pretty messenger of mine must have been waylaid by some younger and more fortunate suitor.' [Vladimir Nabokov. Ada, or Ardor (1968)]
Что-то *ты стал багров, точь-в-точь вино твоей родины*, впрочем, все мы, как говорят американкашки, не очень-то молодеем, вот и мою

[9] Термин «стимул перевода» в значении «фрагмент иноязычного текста, реакцией на который оказывается появление в русском переводе интересующей нас языковой единицы», введен в [Loiseau et al. 2013]; ср. также [Сичинава 2014], [Зализняк 2016]. Эквивалент исследуемой единицы русского языка в переводе на иностранный язык называется «моделью перевода».

прелестную посланницу перехватил дорогою какой-то ухажер поудачливей и посвежее. [Владимир Набоков. Ада, или Радости страсти (С. Ильин, 1996)].

2. Слова *a bit* и *a little*:

(44) *I’m feeling* ***a bit*** *off.* Rémy and I need to pop up to the Isles for my treatments. [Dan Brown. The Da Vinci Code (2003)]
Что-то я расхворался, и нам с Реми надо на острова, чтобы я мог подлечиться… [Дэн Браун. Код Да Винчи (Н. Рейн, 2004)]

(45) *You took to it* ***a little*** *too readily* if you ask me, he told himself. [Ernest Hemingway. For Whom The Bell Tolls (1940)]
Ты *что-то уж очень* охотно взялся за это, если хочешь знать. [Эрнест Хемингуэй. По ком звонит колокол (Н. Волжина, Е. Калашникова, 1968)] ‘слишком’ — стимул для *что-то*.

3. Слово *sort of*:

(46) I pulled the peak of my hunting hat around to the front all of a sudden, for a change. *I was getting* ***sort of*** *nervous, all of a sudden*. I’m quite a nervous guy. [J. D. Salinger. The Catcher in the Rye (1951)]

(47) Я вдруг перевернул свою красную шапку по-другому, козырьком вперед. *Что-то* я начинал нервничать. Нервы у меня вообще ни к черту. [Дж. Д. Сэлинджер. Над пропастью во ржи (Р. Райт-Ковалёва, 1965)]

При этом обращает на себя внимание высокая доля случаев, когда русское *что-то* остается без перевода и, в особенности, когда оно имеет «нулевой» стимул перевода, т. е. в переводе на русский язык появляется как бы «из ничего»[10]. Однако анализ таких случаев позволяет вывить те признаки контекста, которые «стимулируют» появление слова *что-то* в русском переводе; мы будем называть их «стимулирующими контекстами», или «**контекстными стимулами**»:

1. Контекст глагола со значением ‘не понимать’:

(48) “At this gathering,” Teabing said, “many aspects of Christianity were debated and voted upon — the date of Easter, the role of the bishops, the administration of sacraments, and, of course, the divinity of Jesus.” “*I don’t follow.* His divinity?” [Dan Brown. The Da Vinci Code (2003)]

[10] Нулевые «модели» и «стимулы» перевода являются одним из наиболее характерных признаков лингвоспецифичности; о понятии степени лингвоспецифичности и количественных методах ее определения см. [Зализняк 2015а], [Инькова 2017]. В Каталоге дискурсивных слов, созданном в рамках проекта «Контрастивное корпусное исследование дискурсивных слов русского языка» (см. http://a179.frccsc.ru/PublicLingvoProjects/main.aspx) нулевая модель перевода на французский язык у дискурсивного *что-то* встречается в 69 % примеров

На этом собрании обсуждались, принимались и отвергались многие аспекты христианства — дата Пасхи, роль епископов, церковные таинства и, разумеется, божественность самого Иисуса Христа. — ***Что-то*** *я не совсем понимаю*, — с недоумением нахмурилась Софи. — Божественность Иисуса?.. [Дэн Браун. Код Да Винчи (Н. Рейн, 2004)]

(49) "My name is Rémy." Silas was amazed. "*I don't understand*. If you work for the Teacher, why did Langdon bring the keystone to your home?"
[Dan Brown. The Da Vinci Code (2003)]
Кстати, я Реми. Сайлас удивился еще больше: — ***Что-то*** *я не пойму…* Если вы работаете на Учителя, как мог Лэнгдон принести краеугольный камень к вам в дом? [Дэн Браун. Код Да Винчи (Н. Рейн, 2004)]

2. Контекст обозначения негативного внутреннего состояния:

(50) If it doesn't work the first time, we keep doing it until it does. "*I'm scared green*," Hunton said. "As a matter of fact, so am I." [Stephen King. The Mangler (1972)]
Если не сработает в первый раз, будем повторять еще. ***Что-то*** *мне страшно*, — сказал Хантон. — Это естественно. Мне тоже.
[Стивен Кинг. Давилка (В. Эрлихман, 1992)]

(51) "I don't know," said Carrie. "*I feel real bad*". She hung about the stove, suffered a chattering chill, and went to bed sick, the next morning she was thoroughly feverish. [Theodore Dreiser. Sister Carrie (1900)]
— Я сама не знаю, — ответила Керри. — *Мне* ***что-то*** *плохо*. Она не отходила от печки, и зубы у нее стучали от озноба, совсем больная, легла она в постель, а на следующее утро у нее оказался жар.
[Теодор Драйзер. Сестра Керри (М. Волосов, 1927)]

3. Контекст обозначения неконтролируемой ситуации:

(52) I said (having gushed for at least an hour) *I'm talking too much*.
[John Fowles. The Collector (1963)]
Я сказала (после того, как меня несло чуть ли не целый час), ***что-то*** *я разговорилась*. [Джон Фаулз. Коллекционер (И. Бессмертная, 1991)]

(53) I had the cold, I knew it wasn't much. *You talk too much*, I said. You forget who's boss. [John Fowles. The Collector (1963)]
Она же от меня заразилась, а у меня был всего-навсего насморк. ***Что-то*** *вы много болтаете*, говорю ей. Забыли, кто тут хозяин. [Джон Фаулз. Коллекционер (И. Бессмертная, 1991)] нет стимула для *что-то*.

4. Контекст слова *a little*[11]:

(54) *‘I am a little deaf in my left ear,’* Mr. Wonka said.
[Roald Dahl. Charlie and the Chocolate Factory (1964)]
— ***Что-то*** *я стал глуховат на левое ухо*, — сказал мистер Вонка.
[Роальд Даль. Чарли и шоколадная фабрика (М. Барон, Е. Барон, 1991)]

(55) You look, she said shakily. *I’m feeling a little weak*. Scarlett tore off the rag and with trembling hands opened the leather folds. [Margaret Mitchell. Gone with the Wind, Part 1 (1936)]
— Ты погляди, — проговорила она дрогнувшим голосом. — *А у меня* ***что-то*** *немного закружилась голова*. Скарлетт сорвала тряпку и дрожащими руками раскрыла кожаный бумажник.
[Маргарет Митчелл. Унесённые ветром, ч. 1 (Т. Озерская, 1982)]

3. Генезис дискурсивного *что-то*

Проведенный анализ позволяет следующим образом реконструировать процесс семантической деривации, обеспечившей возникновение дискурсивных значений слова *что-то*.

Первый этап семантической деривации состоит в распространении признака неопределенности с объекта на всю пропозицию. Если мне что-то не нравится в некоторой ситуации, то это может означать, что мне (в некоторой степени) не нравится ситуация в целом[12]. Сама по себе последовательность *что-то не нравится* омонимична. Выбор актантного или дискурсивного прочтения *что-то* может быть задан синтаксической структурой предложения: так, в (56) и (57) *что-то* однозначно является неопределенным местоимением, выполняющим роль подлежащего, а в (58) — выпадающим из синтаксической структуры предложения дискурсивным словом.

(56) Если мне ***что-то*** не нравится, я выхожу на улицу и протестую.
[Светлана Алексиевич. Время second-hand // «Дружба народов», 2013]

(57) — Мне ***что-то*** не нравится, а что именно — не понимаю.
[Борис Васильев. Дом, который построил Дед (1990–2000)]

(58) ***Что-то*** не нравится мне вся эта история.
[Александра Маринина. Последний рассвет (2013)]

В примере (59) *что-то* также является дискурсивным словом, но это вытекает только из смысла предложения в целом:

---

[11] В примере (54) англ. *a little* передается при помощи прилагательного с суффиксом *-оват-*, в примере (55) — словом *немного*; одновременно англ. *a little* выступает в роли контекстного стимула для появления слова *что-то*.

[12] Из 60 примеров на поисковый запрос «*мне что-то не нравится*» в НКРЯ в 44 примерах *что-то* выступает в функции неопределенного местоимения в позиции подлежащего, 16-ти является дискурсивным словом.

(59) ***Что-то*** *не нравится* ему там, никак с начальством сладиться не может. [Ю. О. Домбровский. Хранитель древностей, часть 1 (1964)]

Приведем еще один ряд примеров, иллюстрирующий механизм возникновения дискурсивного значения путем переноса признака неопределенности, касающейся объекта, на неопределенность ситуации в целом. Так, *что-то не получается* может содержать как актантное (пример (60)), так и дискурсивное (пример (61)) значение *что-то*:

(60) Когда ***что-то*** *не получается* — разбросает и уйдет. [коллективный. Форум: Помогите разобраться со случаем (2011–2012)] (актантное)

(61) Хотел было сменить тему, но ***что-то*** *не получается*. [Олег Зайончковский. Счастье возможно: роман нашего времени (2008)] (дискурсивное)

Однако в некоторых случаях даже контекст не позволяет сделать однозначный выбор — как в примерах (62), (63) и в особенности (64):

(62) Тем более что с женским полом у тебя сегодня ***что-то*** *не клеится*. [Ашот Аршакян. Шведский дебют Ивана Денисовича // «Сибирские огни», 2012] (скорее дискурсивное)

(63) Ну вот… так вот… Хм… ***что-то*** *не складывается* у нас, Маша. [Виктор Ремизов. Воля вольная // «Новый мир», 2013] (скорее дискурсивное)

(64) Однажды я спросила Коонен: «Алиса Георгиевна, у вас бывает так, что вы играете спектакль, а ***что-то*** *не получается*? Чувствуете, что пустая и ничего не можете сделать?» [Лидия Смирнова. Моя любовь (1997)] (неоднозначно: *что-то* может быть и подлежащим и дискурсивным словом)

Таким образом, интерпретация *что-то* как выполняющего актантную или дискурсивную функцию может быть детерминирована контекстом, но неопределенность может и сохраняться. Это и есть та точка совмещения, которая позволяет реконструировать семантический переход[13].

***

Итак, первая ступень семантической деривации слова *что-то* состоит в переносе признака неопределенности с объекта на всю пропозицию, сопровождаемом превращением слова *что-то* из члена предложения в дискурсивное слово; неопределенность при этом переинтерпретируется как некое отклонение от нормы. Возможен также следующий шаг, на котором отклонение от нормы, вызывающее ощущение дискомфорта, провоцирует поиск причины этого отклонения. Само по себе высказывание типа *Что-то у меня голова разболелась* отсылки к идее неясной причины не имеет, оно наводится

[13] Имеется в виду принцип семантической реконструкции, предложенный в [Бенвенист 1974].

последующим комментарием (говорящего или его собеседника) типа *С чего бы это?* или *Наверное давление подскочило* и т. п.; в другом дискурсивном контексте причинное значение не появляется, ср. возможную реакцию собеседника: *Пойди прогуляйся, и голова пройдет*.

Таким образом, отмечаемое словарями значение неясной причины у слова *что-то* является производным от значения отклонения от нормы и реализуется лишь в определенном круге контекстов. Слово *что-то* имеет еще ряд других дискурсивных значений, возникших на основе значения отклонения от нормы, а также одно — производное от значения неясной причины (оно было названо риторическим).

Обратим также внимание на тот факт, что причинное значение, выделяемое словарями у вопросительного местоимения *что*[14], тоже не является собственно причинным. Вопрос *Что ты плачешь?* может быть задан, в том числе, в случае, когда говорящему прекрасно известна причина слез. Такое *что* выполняет косвенную иллокутивную функцию: *Что ты не отвечаешь?* — это не вопрос, а побуждение одновременно с выражением нетерпения; *Что ты на меня уставился?* — побуждение прекратить действие одновременно с выражением его отрицательной оценки, и т. д. И наоборот, если преподавателя интересует причина опоздания студента, он не может спросить об этом в форме *Что ты опоздал?*

Литература

1. *Арутюнова Н. Д.* (1998), Язык и мир человека. М., 1998.
2. *Бенвенист Э.* (1974), Семантические проблемы реконструкции // Бенвенист Э. Общая лингвистика. М.: Прогресс, 1974.
3. *Зализняк Анна А.* (2015a), Лингвоспецифичные единицы русского языка в свете контрастивного корпусного анализа // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции Диалог’2015. М., 2015. С. 651–662.
4. *Зализняк Анна А.* (2015b) Презенс совершенного вида в современном русском языке // Dekonstruktion und Konstruktion. Zwischen Sprach- und Literaturwissenschaft. Festschrift für Ulrich Schweier zum 60. Geburtstag. Kubon&Sagner. München-Berlin-Leipzig-Wien, 2015. (Wiener Slawistischer Almanach. Sonderband 86). S. 293–316.
5. *Зализняк Анна А.* (2016), База данных межъязыковых эквиваленций как инструмент лингвистического анализа. // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции Диалог’2016. М., 2016. С. 763–775.
6. *Зализняк Анна А., Левонтина И. Б.* (1996), Отражение «национального характера» в лексике русского языка (размышления по поводу книги: Anna Wierzbicka. Semantics, Culture, and Cognition. Universal Human Concepts

[14] Ср. 4-е значение в МАС «Почему?»

in Culture-Specific Configurations. N.Y., Oxford, Oxford Univ. Press, 1992) // Russian Linguistics, vol. 20, 1996, pp. 237–264.
7. *Инькова О. Ю.* (2017), Принципы определения степени лингвоспецифичности коннекторов. // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции Диалог'2017. М., 2017.
8. *Николаева Т. М.* (1985), Функции частиц в высказывании. На материале славянских языков. М., 1985.
9. *Падучева Е. В.* (1985), Высказывание и его соотнесенность с действительностью. М.: Наука, 1985.
10. *Сахно С. Л.* (1983), Приблизительное именование в естественном языке. // Вопросы языкознания, №6, 1983. С. 29–36.
11. *Сичинава Д. В.* (2014), Использование параллельного корпуса для количественного изучения лингвоспецифичной лексики // Язык, литература, культура: Актуальные проблемы изучения и преподавания. Вып. 10. М.,МАКС ПРЕСС, с. 37–44.
12. *Lakoff G.* (1972), Hedges: A study in meaning criteria and the logic of fuzzy concepts // CLS, v. 8, 1972, 183–228.
13. *Loiseau S., Sitchinava D. V., Zalizniak Anna A., Zatsman I. M.* (2013), Information technologies for creating the database of equivalent verbal forms in the Russian-French multivariant parallel corpus // Информатика и ее применения, 2013. Том 7, вып. 2. С. 100–109.

КОРПУСНАЯ ГРАММАТИКА КОЛИЧЕСТВЕННЫХ ГРУПП В РУССКОМ ЯЗЫКЕ[1]

Циммерлинг А. В. (fagraey64@hotmail.com)
Государственный институт русского языка им. А. С. Пушкина; Московский педагогический государственный университет; Институт языкознания РАН, Москва, Россия

THE CORPUS GRAMMAR OF RUSSIAN QPS

Zimmerling A. V. (fagraey64@hotmail.com)
Pushkin State Russian Language Institute; Moscow Pedagogical State University; Institute of Linguistics, Russian Academy of Science, Moscow, Russia

The paper is addressed the corpus grammar of Russian quantifier phrases (QPs), with focus on two issues: (i) subject-predicate agreement patterns in sentences with a QP in the position of a grammatical subject, (b) the choice of the agreeing/non-agreeing form of the adjective in QPs with an embedded NP with the head noun in the feminine gender. QPs license both the plural and the singular form of the predicate. I argue that the singular form optionally shown on the predicate instantiates non-canonic agreement controlled by the QP and does not pattern with the so called default agreement in 3Sg.N. The analysis is based on the complete statistics of all Russian cardinal numerals used in the RNC in QPs of the type 'два человека/ пять человек' in the Russian National Corpora. I show the correlations between plural/singular agreement forms, word order (QP—V ~ V—QP) and communicative status of QP. The choice of the agreeing preposed NP-level adjective as in *dve interesnye knigi* does not constrain the form of the predicate agreement, while agreeing DP-level elements as in *eti dve knigi* blocks the singular form on the predicate. Russian subject QPs are non-canonic arguments, since in the two thirds of the corpus data they lack the status of a theme.

Keywords: numerals, agreement, case, number, phrases, QP, NP, DP, information structure, communicative-syntax interface, Russian language, corpus grammar

[1] Работа выполнена при поддержке проекта РНФ 18-18-00462 «Коммуникативно-синтаксический интерфейс: типология и грамматика», реализуемого в Государственном институте русского языка имени А. С. Пушкина. Я благодарю анонимных рецензентов, а также Е. А. Лютикову и А. А. Герасимову за высказанные замечания и дискуссию.

1. Понятие количественной группы и синтаксис русских числительных

Синтаксис количественных групп (далее — КГ[2]), возглавляемых количественным числительным (*два, три..., пять... десять, сто...*), кванторным наречием (*много, мало, немного, несколько*) или выражениями с дистрибутивной и аппроксимативной семантикой (*по два <человека>, <человека> два, около двух <человек>* и т. п.), является одним из самых сложных фрагментов русской грамматики. В настоящей статье на материале НКРЯ изучаются два параметра — 1) свойства русских КГ как контролеров согласования сказуемого в роде и числе — *пришло (ед. ч. ср. р.) два человека ~ пришли (мн. ч.) два человека*; 2) Выбор согласуемой и несогласуемой формы прилагательного при существительных ж. р. в контексте числительных *две, три, четыре — две красивые (им. п. мн. ч.) женщины (им. п. мн. ч.) ~ две красивых (род. п. мн. ч.) женщины (им. п. мн. ч.)*. Собирательные числительные (*двое*, *трое*, *четверо*, *пятеро* и т. п.) тоже могут формировать КГ, но их соотношение с количественными числительными заслуживает отдельного разбора. В настоящей статье их материал не обсуждается.

1.1. Внутренний и внешний синтаксис количественных групп

Описательные работы констатируют, что русский язык сохраняет варьирование в обеих областях, ср. [Шведова 1982: 242–243] для первого параметра, [ibid:, 78] — для второго. Модели, где предлагался анализ в терминах КГ — NumP в нотации [Franks 1995: 93–180], или QP, в нотации [Pereltsvaig 2006], не нейтральны теоретически. Кроме того, приводимые в данных работах тестовые предложения должны быть уточнены. Утверждение о теоретической ненейтральности не является формой полемики с описаниями языка, однако мы полагаем, что внешняя проверка утверждений о русской грамматике, а также их валидация методами корпусной лингвистики возможна в той мере, в какой эти утверждения, в том числе, анализ в терминах КГ, переводимы на метаязык других теорий.

Не нейтральны также описания, предложенные А. А. Шахматовым и В. В. Виноградовым. Первый из них, вопреки наблюдаемым фактам, утверждал, что в современном русском языке форма род. п. не может зависеть от числительного [Шахматов 1925: 14], а второй полагал, что допустимость сочетаний типа *три огненные глаза* (А. П. Чехов) «доказывает, что такое сочетание числительного с существительным понимается как неразложимое грамматическое целое» [Виноградов 1972: 238]. Непривычная ныне терминология этих авторов предвосхищает анализ в терминах синтаксических групп, но выделение КГ как выражений особого типа не значит, что к ним неприменимы понятия согласования и управления. Внутренний синтаксис сочетаний количественного числительного с зависимыми словами изучался в [Мельчук 1980]; [Yanko 2005]; [Mikaelian 2013] в терминах грамматики зависимостей. Анализ в терминах грамматики

[2] Здесь и далее используется русскоязычная нотация, соответствующая нотации, принятой в англоязычной генеративной традиции. Сокращение КГ соответствует символам QP и NumP в синтаксическом минимализме.

универсальной фразовой структуры предложен в [Лютикова 2017: 304], где вслед за [Franks 1995: 95] принимается постулат о том, что именная группа (далее — ИГ) в безартиклевых языках имеет иерархическую структуру. В КГ типа [$_{QP}$ *пять* [$_{NP}$ *новых студентов*]] вложена ИГ т. н. малой структуры (small nominal, в нотации [Pereltsvaig 2006]), ср. [$_{NP}$ *новые студенты*]. В то же время, КГ вложена в т. н. ИГ полной структуры (иначе — группу детерминатора, DP): [$_{DP}$ Эти [$_{QP}$ *пять* [$_{NP}$ *новых студентов*]]]. В схематичном виде иерархию вершин D, Q, N и возглавляемых ими групп[3] обобщает гипотеза (i):

(i) [$_{DP}$ D^0 [$_{QP}$ Q^0 [$_{NP}$ N^0...]]].

Эмпирическое обоснование гипотезы (i) состоит в том, что в позиции подлежащего русские КГ допускают как согласование как с мн. ч., так и постановку глагола в ед. ч. ср. р., в то время как ИГ полной структуры, в составе которых числительному предшествуют согласуемые детерминативы, допускают только согласование во мн. ч.

(1) а. В последний момент **пришли**^{PL} [$_{QP}$ два студента].
б. В последний момент **пришло**$^{SG.N}$ [$_{QP}$ два студента].

(2) а. В последний момент **пришли**^{PL} [$_{DP}$ **эти**^{PL} [$_{QP}$ два студента]]^{PL}.
б. *В последний момент **пришло**$^{SG.N}$ [$_{DP}$ **эти**^{PL} [$_{QP}$ два студента]]^{PL}.

Стоящее за (1а)–(1б) и (2а)–(2б) обобщение не ново, ср. «если при количественном слове есть согласуемое определяющее слово (*все, эти, остальные* и под.) или другое конкретизирующее определение, то глагол всегда имеет форму мн. ч.» [Шведова 1982: 243]. Оно не проверялось статистически, хотя корпуса текстов дают возможность это сделать.

Столь же давнюю историю обсуждения имеют модификации ИГ малой структуры в контексте количественных слов. Основное внимание уделялось трем параметрам: 1) постановке ИГ в род. п. мн. ч. при числительных, начиная с '5'; — *пришли*$_{PL}$ [$_{NP}$ *новые студенты*]$^{NOM.PL}$ ⇒ *пришли*$_{PL}$/*пришло*$_{SG.N}$ [$_{QP}$ ***пять*** [$_{NP}$ *новых студентов*]$^{GEN.PL}$]; 2) постановке сущ. м. р. в контексте числительных *два, три, четыре*, составных числительных с компонентами *два, три, четыре*, а также слов *оба* и *полтора* в т. н. счетную форму, окончание которой омонимично окончанию род. п. ед. ч. [Зализняк 1967: 47–48]: < *пришло*$_{SG.N}$/*пришли*$_{PL}$ > [$_{QP}$ ***два***^{M} [$_{NP}$ *новых*$^{GEN.PL}$ ***студента***$^{GEN.SG.M}$]^{GEN}]; 3) вариантной реализации согласования/управления при сущ. ж. р. в контексте числительных *два, три, четыре*, составных числительных с компонентами *два, три, четыре* и слов *оба* и *полтора* — < *пришло*$_{SG.N}$/*пришли*$_{PL}$ > [$_{QP}$ ***две***^{F}[$_{NP}$ *красивых*$^{GEN.PL}$ ***женщины***$^{NOM.PL.F}$]] ~ [$_{QP}$ ***две***^{F}[$_{NP}$ *красивые*$^{NOM.PL}$ ***женщины***$^{NOM.PL.F}$]$^{NOM.PL}$].

Таким образом, гипотеза (i) о том, что КГ — промежуточное звено в иерархии двух типов ИГ, резюмирует на метаязыке грамматик универсальной фразовой структуры накопленные за последнее столетие наблюдения над количественными

[3] Для дальнейшего изложения не имеет значения, интерпретируются ли элементы D и Q как собственно лексические категории или как т. н. функциональные категории, т. е. как позиции операторных слов, которые могут замещаться или не замещаться некоторыми выражениями синтаксиса, имеющими ненулевую форму.

словами. В работах Е. А. Лютиковой и А. А. Герасимовой утверждается, что внешний и внутренний синтаксис КГ связаны между собой следующим образом. Русские КГ лишены падежа [Лютикова 2017: 306], в силу чего сказуемое может стоять в несогласуемой форме Зл.ед.ч. ср.р. [Лютикова 2018: 46], а согласование существительного с прилагательным в контексте КГ исчезает [Герасимова, Лютикова 2019]. Предикатное согласование выражений типа *два студента* с глаголом во мн. ч. связано с реализацией абстрактного признака [+ D] 'квантифицируемость', 'синтаксический им. п.', который может вноситься детерминативом (*эти два студента*), согласуемым определением (*два студента, пришедшие позже*), либо не выражаться внешне [ibid.]. Данная гипотеза частично поддается проверке, поскольку можно оценить, в какой мере она предсказывает дистрибуцию согласовательных моделей, и в какой мере делаемые на ее основе предсказания взаимосвязаны.

1.2. Категория числа и контроль предикатного согласования

Ключевым для анализа КГ является тезис о том, что у количественных числительных нет категории числа [Шведова 1970: 334]; [Виноградов 1972: 237]. То, что в предложениях (1а) и (2а) сказуемое стоит во мн. ч., подтверждает, что на уровне КГ реализуется признак [+ множественность]. Семантически этот признак вносится именем (*студенты* ⇒ *два студента*, *женщины* ⇒ *две женщины*), но существительное (N^0) нельзя считать вершиной КГ хотя бы потому, что в контексте количественного слова (Q^0) происходит обязательная или факультативная модификация зависимых адъективных элементов:

(3) а. Пришли **новые** студенты^0 ⇒ Пришли [$_{QP}$ два^0 [$_{NP}$ **новых** студента^0]].
б. Пришли интересные женщины ⇒ Пришли [$_{QP}$ две^0 [$_{NP}$ **интересных** женщины^0]] ~ пришли [две^0 [$_{NP}$ **интересные** женщины^0]].

Примеры типа (1б), где глагол стоит в Зл. ср.р. ед.ч., могут описываться либо как согласование с подлежащным контролером [Шведова 1982: 243], либо как дефолтное согласование, реализующееся в отсутствие лексически выраженного контролера [Лютикова 2018: 46, 148]. В терминах традиционной грамматики, где не принимается тезис о том, что абстрактный признак 'согласование' является характеристикой структуры предложения, дефолтное согласование равнозначно отсутствию согласования.

Неясно, есть ли связь между внутренним и внешним синтаксисом КГ, т. е. между их поведением в позиции подлежащего и наличием согласуемых определений уровня малой ИГ. В [Герасимова, Лютикова 2019] утверждается, что согласование прилагательного с существительным ж. р. в контекстах типа *две интересные женщины* носит реликтовый характер и что в позиции подлежащего такие КГ с согласуемым прилагательным почти не встречаются. Этот тезис проверяется ниже в **разделе 4**.

1.3. Статус счетной формы как особого падежа

Признание счетной формы особым падежом возможно благодаря небольшому числу словоформ (*два шарá, шагá, рядá, следá, часá, шагá, шарá, две гримерные, мастерские, столовые, примерочные, уборные*), которые не омонимичны

формам род. п. ед. ч. [Зализняк 1967: 47]. Данное решение принято в модуле лексико-грамматического поиска НКРЯ. В модели А. А. Зализняка падеж считается признаком словоформ. Падежное значение соответствует закрытому ряду алломорфов показателей проверяемого падежа и т. н. однопадежному ряду, который определяется как перечень всех сегментов, выражающих словоформы с общей граммемой некоторого падежа [Зализняк 1967: 40][4].

В современной лингвистической типологии падеж считается признаком синтаксических групп [Arkadiev 2016]. Такой подход не исключает гипотезу о счетной форме как особом падеже, но делает описание громоздким. Приходится допустить, что счетная форма у прилагательных ж. р. реализуется в виде алломорфов *-ые/-ых*, в зависимости от наличия или отсутствия согласования с существительным. В записи (4a-b) тэг 'Count' использован и как признак всей ИГ малой структуры, и как признак ее элементов — существительных и прилагательных.

(4) a. $[_{QP}$ *две*^{0} $[_{NP}$ *интересных*$^{GEN.PL/COUNT}$ *женщины/книги*$^{NOM.PL/COUNT}]^{COUNT}]$.
b. $[_{QP}$ *две*^{0} $[_{NP}$ *интересные*$^{NOM.PL/COUNT}$ *женщины/книги*$^{NOM.PL/COUNT}]^{COUNT}]$.

Итак, при анализе в терминах КГ гипотеза о том, что падеж ИГ меняется в контексте количественного слова на счетную форму, не имеет преимуществ по сравнению с анализом, согласно которому существительные и прилагательные получают в этом контексте род. п. (при сущ. всех трех родов), либо сохраняют исходный падеж (при сущ. ж.р.)[5].

1.4. Гипотеза о малых числительных

В [Зализняк 1967: 48] высказана мысль, что особенности сочетаний со словами *два*, *три*, *четыре*, *оба*, *полтора* объясняются действующей в языках мира тенденцией снабжать числительные малого количества (paucal) дополнительными грамматическими признаками. В [Madariaga & Igartua 2017; 2018] гипотеза о том, что числительные *два*, *три*, *четыре* имеют в русском языке особый паукальный синтаксис, принимается без оговорок. Для употреблений в позиции прямого дополнения, ср. *увидел* ***двух*** *женщин*, но *увидел двадцать* ***две*** *женщины/* ****двадцать двух*** *женщин*, такое решение оправдано. Для подлежащных

[4] Почти у всех указанных А. А. Зализняком словоформ есть варианты с накорневым ударением, делающие их омонимичными род. п. ед. ч.: *три воздушных ша́ра, четыре широких ша́га, три первых ря́да, два долгих ча́са*.

[5] Анонимный рецензент замечает, что принятое в минимализме, ср. [Chomsky 1995]; [Bailyn 2012] различение т. н. с-(интаксического) падежа (абстрактного признака ИГ) и м-(орфологического) падежа (маркировка семантических ролей средствами флективной морфологии) дает дополнительные возможности для описания русской падежной системы. В общей форме это верно, поскольку один и тот же м-падеж может соответствовать разным периферийным и грамматическим употреблениям русских с-падежей, ср. посессивный род. п., род. п. меры, род. п. качества и род. п. отрицания. Для оценки статуса счетной формы это различение не важно, поскольку счетная форма (*два трудных шара́*), как и род. п. II (*выпить чаю*) и предл. п. II (*в дыму*) не являются с-падежами и не приписываются ИГ в целом. Ключевой вопрос — является ли счетная форма м-падежом, или же это вариант род. п. [Кустова 2011].

употреблений КГ требуется проверка. Подтверждением можно считать такое распределение корпусных данных, при котором простые числительные *два, три, четыре* имеют иные дистрибутивные признаки в плане контроля согласования сказуемого и выбора согласуемого определения, по сравнению с составными числительными типа *двадцать два*, *сто тридцать четыре* и т. п.

1.5. Предложно-падежные сочетания с числительными

В [Виноградов 1972: 244–245]; [Мельчук 1980]; [Yanko 2005]; [Mikaelian 2013] обсуждается возможность употребления форм *два, две, оба, обе, три, четыре* после предлога в контексте *в* ***обе*** *руки*, *через* ***два*** *человека в очереди*. Исторически данная форма является вин. п., но после становления категории одушевленности и появления новых форм вин. п. *двух, трех, четырех, обоих, обеих*, она реинтерпретируется как им. п. И. А. Мельчук и принимающие его анализ русисты толкуют форму числительного в сочетаниях типа ***два*** *человека* как неодушевленный алломорф вин. п. Т. Е. Янко стремится доказать, что в конструкциях типа *на два Х-а/пять Х-в* признак одушевленности нерелевантен, поскольку выражения *человек, тенор, Чацкий* в контексте *Хлеб кончился* ***за два человека до меня***. *Театру нужно* ***два героических тенора***. *Было подготовлено* ***два Чацких*** выступают не в качестве обозначений живых существ, а в функции единицы измерения [Yanko 2005].

2. Количественные и сочиненные группы как контролеры согласования

2.1. Подлежащная vs неподлежащная форма КГ

Выбор формы ср. р.ед. ч. в предложениях типа *пришло два студента* может объясняться трояко:

А) Русские КГ являются контролерами подлежащного согласования в Зл. ед. ч. ср. р., те же КГ в большинстве контекстов допускают форму Зл. мн. ч. [Шведова 1982: 243].

Б) Подлежащный контроль согласования со стороны выражений типа *два студента* есть лишь в предложениях с глаголом в Зл. мн. ч. В предложениях с глаголом в Зл. ед. ч. ср. р. реализуется дефолтное согласование, эта форма не контролируется подлежащим [Лютикова 2018: 46, 148].

В) Русские предложения типа *пришло два студента* — безличные конструкции, где отсутствие согласования связано со сниженной топикальностью КГ [Malchukov, Ogawa 2011].

Ответ В) опровергается примерами типа (5а), где пассивизация КГ сочетается с формой ср. р. ед. ч.

(5) а. **Было**$_{SG.N}$ **аттестовано**$_{SG.N}$ два студента.
б. Были$_{PL}$ аттестованы$_{PL}$ два студента.

В литературном русском языке безличного пассива с дополнением в вин. п. (т. н. ленивый пассив) нет. Поэтому идея, что форма *два* в ***два*** *студента* в (5а) является неодушевленным вариантом вин. п. [Mikaelian 2013], не спасает. В данной позиции возможен только им. п. и исключен вин. п.: *****двух студентов*** *(вин. п.) было аттестовано*, ср. нормативное *двух студентов* (вин. п.) ***аттестовали***, в действительном залоге [Zimmerling 2013]. Сниженная топикальность КГ в (5а) — не основание для того, чтобы считать такую КГ неподлежащной. К тому же, (5а) и (5б) едва ли различаются по степени топикальности.

2.2. Сочиненные группы

Выбор между ответами А) и Б) зависит от принимаемых допущений. Примем, что согласование глагола во мн. ч. — свойство именной вершины (N), а постановка глагола в ед. ч. ср. р. — свойство количественного оператора (Q). В таком случае, двойственность согласования во мн. ч./ед. ч.ср. р. связана с тем, что КГ является гибридным контролером Q/N. Такое решение формализует ответ А), который условно можно назвать моделью Н. Ю. Шведовой. Ответ Б) — модель Е. А. Лютиковой — связывает согласование глагола во мн. ч. со статусом ИГ полной структуры (свойством +D), а форму глагола в 3 л. ед. ч. ср. р. — со свойством Q (= −D). Проверка утверждений А) и Б) связана с синтаксисом выражений вида КГ & ИГ, которые предположительно образуют в позиции подлежащего группу особого типа — сочиненную группу (СочГ). К сожалению, понятие СочГ не нейтрально. При одной из трактовок, СочГ постулируется независимо от формы согласования, а в предложениях типа *пришел*$_{\text{SG.M}}$ *адвентист*$_{\text{SG.M}}$ и *адвентистка*$_{\text{SG.F}}$ усматривается согласование с ближайшим конъюнктом [Willer Gold et alii 2018], при этом используется постулат о синтаксической однородности конъюнктов. Подобная трактовка СочГ совместима с моделью Е. А. Лютиковой, но не может считаться ее независимым подтверждением. Поскольку по (i) КГ встроена в ИГ полной структуры, а ИГ малой структуры встроена в КГ, единственный способ описать неоднородные СочГ вида x_{α} & y_{β}, где α и β — разные категории, сохраняя постулат об однородности конъюнктов, состоит в приписывании им статуса общей вышестоящей категории γ, т.е. статуса ИГ полной структуры:

(ii) $[_{\text{CoP}}\ [_{\text{DP}}\ [_{\text{QP}}$ два $[_{\text{NP}}$ леща] & $[_{\text{DP}}\ [_{\text{QP}}\ [_{\text{NP}}$ щука]]].

При альтернативном подходе СочГ постулируется, если форма согласования целого не совпадает с формой согласования, предписываемой конъюнктами (например, при конъюнкции ИГ в м. р. и ИГ в ж. р. выбирается форма ср. р. или форма мн. ч., или форма дв. ч. и т. п.). Согласование с ближайшим конъюнктом признается несовместимым со статусом СочГ.

(iii) СочГ требует формы предикатного согласования, отличной от формы согласования, предписываемой ее конъюнктами.

При принятии (iii), ответы А) и Б) могут быть проверены в позициях, где возможна особая форма согласования. Независимым подтверждением Б) были бы примеры, где согласование во мн. ч. блокируется в СочГ вида КГ1 & КГ2, где ни один из конъюнктов не является ИГ. Однако таких предложений нет.

(6) a. [$_{CoP}$ [$_{QP1}$ **Два участника конференции**] **и** [$_{QP2}$ **три девушки**]] **пришли**$_{PL}$/пришло$_{SG.N}$ на выставку в шортах.
b. [$_{CoP}$ [$_{QP1}$ **Два офицера**] **и** [$_{QP2}$ **двадцать солдат**]] **были**$_{PL}$ **убиты**$_{PL}$/было$_{SG.N}$ убито$_{SG.N}$ осколками.

Вариативность согласования в числе — общее свойство всех КГ и СочГ вида КГ1 & КГ2, независимо от залога. В СочГ вида ИГ1 & ИГ2 согласование в ед. ч. блокируется, см. (7). При подлежащих вида КГ & ИГ согласование в ед. ч. невозможно или сильно затруднено, см. (8а)–(8б). Допустимость примера (9) может объясняться тем, что в нем реализуется структура с эллипсисом глагола. При порядке VS и согласовании со смежной с глаголом ИГ в предложениях типа рус. ***Пришел***$_{SG.M}$ ***Вася***$_{SG.M}$ *и две девушки*$_{PL}$ наличие СочГ не доказано и противоречит условию (iii).

(7) *__Вася и Света пришло__ на выставку.

(8) а. *__Вася и две девушки пришло__ на выставку.
б. *__Две девушки и Вася пришло__ на выставку.[6]

(9) На выставку пришло две девушки и ~~пришел~~ Вася.

Таб. 1. Согласование КГ и СочГ с предикатом в ед. ч. ср. р. и мн. ч.

	КГ	СочГ		
		КГ1 & КГ2	КГ & ИГ	ИГ1 & ИГ2
SG.N	+	+	?	*
PL	+	+	+	+

Таким образом, выбор формы ед. ч. реализует согласование с гибридным контролером Q/N, а не дефолтное согласование. Те же свойства гибридного контролера Q/N имеют те СочГ, где по крайней мере два коньюнкта имеют статус КГ.

(10) На этом стуле **были**$_{PL}$ **казнены**$_{PL}$ [$_{CoP}$ [$_{QP1}$ двести мужчин] и [$_{QP2}$три женщины]], между тем стул выглядел совсем как новый.
[Илья Ильф, Евгений Петров. Одноэтажная Америка (1936)]

(11) С обеих шхун **высажено**$_{SG.N}$ [$_{CoP}$ [$_{QP1}$ сто двадцать мужчин], [$_{QP2}$ триста женщин] и [$_{QP3}$ около сотни подростков]].
[Роберт Штильмарк. Наследник из Калькутты (1950–1951)].

В дальнейшем изложении примеры с сочиненными группами не учитываются.

[6] В русских пассивных предложениях с порядком SV и смежной с глаголом ИГ согласование с ближайшим конъюнктом исключено: **Две девушки*$_{PL}$ *и* ***Вася***$_{SG.M}$ ***был***$_{SG.M}$ *отчислен*.

3. Количественные группы в позиции подлежащего в НКРЯ

3.1. Методика анализа

Подлежащные КГ вида ‘n Х-в’ при большинстве стимулов дают в НКРЯ слишком малую выгрузку для того, чтобы проверять модели предикатного согласования при разном порядке слов при одном и том же существительном в контексте одного и того же числительного и получить значимое распределение. Это касается и обозначений профессий — *солдат, офицер, студент, спортсмен, врач* и т. п. Для получения репрезентативной статистики КГ как контролеров предикатного согласования были просмотрены все комбинации слова *человек* с простыми числительными от 2 до 900 и составными числительными с последним компонентом от 2 до 900. Выгрузка проверялась аннотатором вручную. Зачитывались употребления КГ в позиции единственного подлежащего в предложениях с проверяемой согласовательной формой сказуемого. Отсеивались следующие примеры:

1) предложения, где КГ не стоит в позиции подлежащего или не является единственным контролером согласования, в том числе, все случаи употребления КГ в составе СочГ;
2) все употребления ИГ полной структуры, содержащие согласованные определения, ср. *эти два человека*, *два человека, возмущенные произволом властей* и т. п.;
3) все употребления числительных после предлога, ср. *по два человека*, *через два человека* и т. п.;
4) все случаи аппроксимативной инверсии, ср. *человека два*. При таких критериях большинство примеров выборки оказалось связано с точным обозначением количества.

В то же время, не исключались примеры с аппроксимативным наречием, не меняющим порядок числительного и существительного и падеж числительного, ср. *примерно пятнадцать человек*, *почти триста человек, два-три человека* и т. п. Собирательные числительные не учитывались. Не учитывались слова *оба*, *обе* и дроби.

Релевантные примеры делились на случаи контроля сказуемого в ед. ч. и мн. ч. Предложения в действительном и страдательном залоге подсчитывались отдельно. Для всех подтипов отдельно подсчитывалась препозиция подлежащной КГ (VS) и ее постпозиция (SV). Тем самым, каждая строка таблицы, т. е. комбинация слова *человек* с простым или составным числительным, дает сведения о 8 потенциально возможных случаях употребления КГ в позиции контролера предикатного согласования. Статистика каждой комбинации учитывает простые и составные числительные. Так, например, в строке ‘7 человек’ учтены как употребления типа *пришло/пришли семь человек*, так и употребления *пришло/пришли сто двадцать семь человек*.

3.2. Дистрибутивные свойства конструкции ‘n X-ов’ в НКРЯ

3.2.1. Конструкция ‘n человек’

Наибольшее число релевантных примеров получено для следующих простых и составных числительных с последним компонентом от 2 до 900:

Таб. 2. Соотношение количественных числительных и числа подлежащных употреблений КГ ‘n человек’ в основном корпусе НКРЯ

Цифра	2	3	5	6	10	4	7	20	100	9	12	40	30	15	50	200
Примеры	1362	1166	1081	858	591	742	523	349	332	289	266	213	204	184	183	160

Без сравнения с другими корпусами неясно, в какой степени такое распределение подлежащных КГ связано с частотой употребления числительных, а в какой — со спецификой текстов, вошедших в корпус.

Таб. 3. Конструкция ‘n человек’ с количественными числительными с последним компонентом от 2 до 10 в позиции подлежащего в основном корпусе НКРЯ

Всего примеров в корпусе		В позиции подлежащего в позиции с проверяемой согласовательной формой сказуемого							
		SG				PL			
		Active		Passive		Active		Passive	
		VS	SV	VS	SV	VS	SV	VS	SV
2	1362	93	1	8	0	389	392	14	10
3	1166	126	3	7	0	281	249	4	9
4	742	101	1	23	0	150	104	2	8
5	1081	184	9	28	7	104	130	15	2
6	858	167	11	3	7	110	70	1	5
7	523	113	0	12	8	53	36	6	7
8	598	111	10	17	2	57	35	6	6
9	289	77	4	10	4	21	27	5	6
10	591	84	11	10	2	61	66	1	1
		1086	50	118	30	1226	1109	54	54
		95,6%	4,4%	79,8%	20,2%	52,5%	47,5%	50%	50%

Есть корреляция между ед. ч. и порядком VS: на подтип SV_{SG} приходится лишь 4,4% случаев. Предложения VS_{PL} ~ SV_{PL}. представлены сопоставимой выгрузкой. Число примеров с глаголом в страдательном залоге невелико, но вероятность реализации порядка SV_{SG} в страдательном залоге возрастает до 20,2%. Суммарная статистика конструкции, с учетом более редких комбинаций, дает сходную картину.

Таб. 4. Конструкция ‘n человек’ с количественными числительными с последним компонентом от 2 до 900 в позиции подлежащего в основном корпусе НКРЯ

Всего примеров в корпусе		В позиции подлежащего в позиции с проверяемой согласовательной формой сказуемого							
		SG				PL			
		Active		Passive		Active		Passive	
		VS	SV	VS	SV	VS	SV	VS	SV
2–900	4816	1512	107	168	46	1404	1438	65	77
		93,4%	5,6%	78,5%	21,5%	49,4%	50,6%	45,8%	54,2%

При порядке VS обе модели согласования находятся в отношении свободной конкуренции, в то время как при порядке SV явно предпочитается форма мн. ч. Тем не менее, 9% от общей выгрузки предложений SV, приходящихся на подтип SV_{SG} (153 примера из 1668) подтверждают, что конкуренция есть и в препозиции глаголу. Такое распределение не соответствует предсказанию [Malchukov, Ogawa 2011] о том, что отсутствие согласования во мн. ч. в предложениях типа *пришло два человека* связано со снижением топикальности: постпозиция КГ статистически более значима, чем форма согласования. НКРЯ не дает оснований полагать, что предложения вида $V_{PL}S$ более топикальны, чем предложения была $V_{SG}S$.

3.2.2. Конструкции ‘n X-в’ с гендерными существительными

Для контроля мы проверили две гендерные конструкции — ‘*n мужчин* ’ и ‘*n женщин*’, которые фиксируют большинство комбинаций с количественными числительными. В общей сложности на эти конструкции пришлось 698 релевантных примеров. Соотношение препозиции и постпозиции подлежащего сходно с данными по конструкции ‘n человек’, но доля примеров с глаголом в ед. ч. при гендерных существительных — 5,7% от общей выгрузки, существенно ниже соответствующего показателя (38%) для конструкции ‘n человек’.

Таб. 5. Конструкция ‘n мужчин’ и ‘n женщин’ с числительными с последним компонентом от 2 до 900 в основном корпусе НКРЯ

Всего примеров в корпусе		В позиции подлежащего в позиции с проверяемой согласовательной формой сказуемого							
		SG				PL			
		Active		Passive		Active		Passive	
		VS	SV	VS	SV	VS	SV	VS	SV
2–900	698	31	6	3	0	363	274	13	8
		83,8%	16,2%	100%	0%	57%	43%	61,9%	38,1%

Уменьшение доли согласования в ед. ч. для слов *мужчина* и *женщина* объяснимо на основе двух гипотез: 1) гендерные обозначения в русском языке ненейтральны,

в том числе в контексте ‘*n X-ов*’; 2) предикатное согласование во мн. ч. подчеркивает, что имя исчисляемой сущности не только выступает в роли единицы измерения, но и указывает на наличие у нее некоторого индивидуального ‘профиля.[7]

3.2.3. Паукальность и составные числительные

Материал не подтвердил гипотезу о том, что различие простых и составных числительных влияет на форму сказуемого. ‘222 человека’ согласуется так же, как ‘2 человека’, КГ с простыми числительными от 2 до 10 служат образцом для обозначений больших чисел. Последние допускают обе числовые формы предиката. Словами из корпусного примера:

(12) А то, что за каждым из нас, коммунистов, стоит$_{SG}$ [$_{QP}$ девятьсот миллионов плюс миллиард **шестьсот человек**], вы забыли! [Г. Е. Николаева. Битва в пути (1959)].

3.3. Лингвистическая интерпретация

3.3.1. ИГ полной структуры и числительные

Материал подтвердил прогноз о том, что наличие согласованных определений блокирует согласование сказуемого в ед. ч. Примеров с согласованным препозитивным определением перед КГ и сказуемом в ед. ч. обнаружено не было. На 4816 примеров употреблений конструкции ‘n человек’ пришелся 1 случай с согласованным постпозитивным определением при сказуемом в ед. ч.:

(13) И на весь этот люд **было**$_{SG.N}$ только **три человека смущенные**$^{NOM.PL}$ **и озабоченные**$^{NOM.PL}$: преосвященный, настоятель и молодой Каменский. [Е. А. Салиас. На Москве (1880)].

Также встретилось три примера контроля предикативного атрибута (вторичного предиката) в тв. п. мн. ч. при основном сказуемом в ед. ч. Все они приводятся ниже:

(14) Начался выбор его из всего дворянства. **Кандидатами**$^{INSTR.PL}$ **представлено**$_{SG.N}$ **пять человек**. [И. М. Долгоруков. Повесть о рождении моем, происхождении и всей моей жизни, писанная мной самим и начатая в Москве, 1788-го года в августе месяце, на 25-ом году моей жизни / Часть 4 / 1799–1806 (1788–1822)]

(15) В течение первой недели из 35-ти человек казаков **лежало**$_{SG.N}$ **тифозными**$^{INSTR.PL}$ **восемь человек**, о других амбулаторных больных я уже не говорю. [В. В. Корсаков. Пекинские события. Личные воспоминания участника об осаде в Пекине. Май-август 1900 года (1901)].

---

[7] Статистическая проверка этих гипотез возможна после того, как на корпусе будет определено некоторое среднее значение (коэффициент V_{PL}/V_{SG}) для нейтральных единиц измерения в контексте ‘*n X-ов*’. В этом случае превышение порогового значения V_{PL}/V_{SG} для конкретной единицы, например, для слов *учитель*, *девушка, самолет*, будет свидетельствовать о наличии у них ‘индивидуального профиля’.

(16) **Оказалось**$_{SG.N}$ на месте катастрофы **пять человек убитыми**$^{INSTR.PL}$ и несколько раненых. [Народная воля. Социально-революционное обозрение. №6 // «Народная воля», 1881]

Для проверки гипотезы о том, что согласуемые элементы в составе подлежащной КГ блокируют согласование сказуемого во ед. ч., эти примеры нерелевантны, поскольку предикативный тв. п. относится к группе сказуемого, а не подлежащего и может рассматриваться как признак, приписываемый в группе особого типа, проекции PredP [Matushansky 2008]; [Bailyn 2012]; [Циммерлинг 2018]. В то же время, они подтверждают, что КГ вида 'n X-в' реализуют семантический признак [+ множественность], который может проявляться в числовой форме вторичного предиката даже при постановке основного глагола в ед. ч.

3.3.2. Линейный порядок и топикальность КГ

Отличительной чертой коммуникативно-синтаксического интерфейса русского языка является непрозрачность линейного порядка: один и тот порядок обычно допускает разные интерпретации. К-(оммуникативный) статус элемента (обычные и контрастные тема и рема, эмфаза, незавершенность) маркируется интонационно, и лишь во вторую очередь — порядком слов [Янко 2008: 27–43]. Просодическая маркировка *к*-статусов, которую мог иметь в виду автор письменного текста, не всегда расшифровывается однозначно. Тем не менее, вопрос о топикальности КГ при порядках SV и VS поддается проверке.

При порядке SV подлежащее не может быть акцентно выраженной темой ни в расчлененных предложениях типа {*На ↗вечеринку*} (тема) {*пришло два↘ человека*} (*рема*)[8], ни в нерасчлененных предложениях типа *пришло два↘ человека* [ibid., 59]. Обсуждаемый в [Циммерлинг 2016] вопрос о том, порождаются ли все русские предложения с порядком VS, ср. *пришла ↘ весна* и *пришла ↘ бабушка*, одинаково, не важен для статуса подлежащной КГ в постпозиции глаголу — такая КГ темой быть не может, независимо от того, стоит ли глагол во мн. ч. или в ед. ч. Отсюда вытекает, что КГ — неканонический подлежащный актант, поскольку почти в двух третях случаев (3171 из 4816 для 'n человек' — 65,9%, 405 из 698 для двух гендерных конструкций — 58%, среднее значение — 64,8% для всех трех конструкций) она встретилась в составе ремы или нерасчлененного предложения.

Порядок SV обычно предполагает возможность интерпретации подлежащной КГ как темы. Но порядок SV может также быть производным и порождаться сдвигом акцентоносителя влево, т. н. механизм Left Focus Movement: V — ↘X ⇒ ↘↘X [Циммерлинг 2016: 89]. В этом случае препозитивная КГ будет инвертированной ремой. Такая интерпретация вероятна в контексте (17), где исходная *к*-структура восстанавливается в виде *во мне боролись два ↘человека,* и в контексте (18), где исходная *к*-структура имеет вид *в ↗машине было два ↘человека*.

[8] Здесь и далее фигурные скобки обозначают границы коммуникативных составляющих, помета '↗', записанная слева от словоформы, указывает на акцентоноситель темы в повествовательных предложениях, а помета '↘' — на акцентоноситель ремы в том же типе предложений.

(17) Точно **два ↘↘человека** боролись$_{PL}$ во мне — один веселый, легкий, который старался припомнить и живо представить себе все самое хорошее в жизни, и другой — мрачный и мстительный, не забывающий обид, томящийся от невозможности отплатить за унижение. [Вениамин Каверин. Два капитана (1938–1944)]

(18) Я напрягся немного, потому что в машине **два ↘↘человека** было. [Александра Маринина. Шестерки умирают первыми (1995)]

Общее число примеров подтипа SV_{SG} невелико — 159 из 5514 (2,9 % от общей выгрузки трех конструкций), из них — 113 в действительном залоге (2 %). Тем не менее, топикальность препозитивной КГ при согласовании в ед. ч. во многих контекстах диагностируется надежно, см. (19), где предполагается, что с группой X-в произошло p, в то время как с Y-ами случилось q.

(19) {**Двадцать четыре ↗ человека**} {потонуло$_{SG.N}$ буквально в версте от ↘берега}, не умея добраться не то, что до справедливого устройства, а до песчаного пляжа. [Вячеслав Рыбаков. Гравилет «Цесаревич» (1993)]

Коммуникативная разметка корпуса является трудоемкой задачей, но с учетом того, что часть примеров с порядками SV_{SG} и SV_{PL} может порождаться инверсией ремы, мы вправе заключить, что доля топикальных употреблений подлежащных КГ ниже общего процента предложений SV в выгрузке (35, 2 %).

3.3.3. Конкуренция двух моделей согласования

Материал подтвердил, что обе модели согласования конкурируют при одних и тех предикатах, с поправкой на коммуникативную перспективу. Как указано выше, при порядке VS обе модели имеют сопоставимую частотность: 1845 примеров VS_{PL} против 1714 примеров VS_{SG}. Однако конкуренция есть и в зоне, где преобладает мн. ч. Так, в публикации 1989 г. автор А. Терехов проставил форму мн. ч., а в публикации 1991 г. заменил ее в том же контексте на ед. ч.:

(20) а. На свадьбе **шестьсот человек гуляли**$_{PL}$ семь дней. [Александр Терехов. Секрет // Библиотека «Огонек», 1989]

б. На свадьбе **шестьсот человек гуляло**$_{SG.N}$ семь дней. [Александр Терехов. Клоун // «Огонек». № 6, 1991]

4. Внутренний синтаксис КГ

4.1. Согласование прилагательных с существительными м.р и ср. р.

Принято считать, что при существительных м. р. и ср. р. согласование с прилагательным в составе ИГ малой структуры, вложенной в КГ, носит реликтовый характер. Проверка этого утверждения в НКРЯ затруднена из-за того, что модуль поиска по триграммам не дает контекстов и дат и не показывает снятой омонимии, выдаваемые цифры о 1700 примерах недостоверны. После проверки коллокаций мы получили следующие данные:

Таб. 6. КГ с согласуемыми прилагательными при существительных м. р. и ср. р., по Основному корпусу НКРЯ

	Беспредложный им. и вин. п.		Предложные употребления	
	Всего	После 1917 г.	Всего	После 1917 г.
ДВА + [ПРИЛ. + СУЩ] в м./ср. р. им. п. мн. ч.	259	30	37	8
ТРИ + [ПРИЛ. + СУЩ] в м./ср. р. им. п. мн. ч.	27	5	7	3
ЧЕТЫРЕ + [ПРИЛ. + СУЩ] в м./ср. р. им. п. мн. ч.	14	0	1	0
ИТОГО:	300	35	45	11

4.2. Согласование прилагательных с существительными ж. р.

После простых числительных *два, три, четыре* и составных числительных с последними компонентами *два, три, четыре* прилагательное либо согласуется с существительным в падеже — *две интересные*$^{\text{NOM-ACC.PL}}$ *книги*$^{\text{NOM-ACC.PL}}$, либо стоит в род. п. мн. ч. — *две интересных*$^{\text{GEN.PL}}$ *книги*$^{\text{NOM-ACC.PL}}$. В [Голуб 1997] утверждается, что в предложно-падежных конструкциях предпочитается род. п. Эта точка зрения не подтвердилась — конструкция с согласуемым прилагательным встречается в НКРЯ более чем в 2 раза чаще[9].

Таб. 7. Предложные конструкции с прилагательными и существительными ж. р. в контексте числительных ***две, три, четыре***, по Основному корпусу НКРЯ

	Им. п.	Род. п.
Предлог + *две* + прил. + сущ. ж. р.	542	31
Предлог + *три* + прил. + сущ. ж. р.	144	241
Предлог + *четыре* + прил. + сущ. ж. р.	41	76
	727	346

В работе [Герасимова, Лютикова 2019] со ссылкой на исследование М. В. Шкапы утверждается, что согласование прил. с сущ. ж.р. практически невозможно, «если количественная конструкция находится в позиции не контролирующего согласование подлежащего» и что в НКРЯ всего 5 исключений. Это утверждение не соответствует актуальному состоянию НКРЯ. Были проверены предложения с порядком $V_{SG}S$, где глагол в форме ед. ч. ср. р. непосредственно предшествует подлежащной КГ с прилагательным и существительным ж. р. Из 120 обнаруженных примеров употребления прилагательных в составе подлежащной КГ с числительными *две*, *три*, *четыре*, в 53 использована модель с согласованием

[9] Анонимный рецензент справедливо отмечает, что целесообразно провести отдельные подсчеты для конструкции меры *на две интересн-ых/-ые книги больше* и для прочих предложных конструкций, но это задача отдельного исследования.

(44,2%), 32 из которых относятся к периоду 1986–2015 гг. На тот же период приходится 33 примера с управляемой формой прилагательного в род. п. мн. ч. Во всех примерах, кроме одного[10], прилагательное предшествует существительному, см. (21) и (22).

(21) Допустим, в популяции бесполых микробов **возникло**$_{SG.N}$ [$_{QP}$ две^{F} [$_{NP}$ **полезные** мутации]$^{NOM.PL}$]. [Александр Марков, Елена Наймарк. Эволюция. Классические идеи в свете новых открытий (2014)]

(22) — А то, что на сегодняшний день в Доме опять **образовалось**$_{SG.N}$ [$_{QP}$ две^{F} [$_{NP}$ **враждующие** группировки] $^{NOM.PL}$], вас не настораживает? [Мариам Петросян. Дом, в котором… (2009)]

Общая статистика двух конструкций приводится в **таб. 8**.

Таб. 8. Согласуемая и управляемая формы прилагательных в предложениях с подлежащной КГ при глаголе в ср. р. ед. ч., по Основному корпусу НКРЯ

	‘2’, ‘3’, ‘4’ + им. п.	‘2’, ‘3’, ‘4’ + род. п.	Всего
Действительный залог	36	60	96
Страдательный залог	17	7	24
	53	67	120
	44,2%	55,8%	100%

Тем самым, прогноз о том, что конкуренция двух моделей внутреннего синтаксиса КГ при существительных в ж. р. устранена в пользу род. п., не подтвердился. Внутренний синтаксис КГ (выбор согласуемой/несогласуемой формы прилагательного при вложенном существительном ж. р.) не ограничивает ее внешний синтаксис, т. е. выбор модели предикатного согласования и возможность использования в предложно-падежной конструкции, и наоборот.

4.3. Непротиворечивость модели и верификация

Мы можем теперь оценить выдвинутую в [Лютикова 2017] формальную модель КГ в плане непротиворечивости и точности предсказаний.

[10] В примере-исключении Ф. И. Буслаев (1887) пересказывает или цитирует былину: *вместо гривы было прибито две лисицы* ***бурнастые****, вместо хвоста повешено два медведя* ***белых, заморских***. [Ф. И. Буслаев. Русский богатырский эпос (1887)]

Таб. 9. Апробация модели о беспадежной КГ в русском языке

	Прогноз	Проверка
1	Внешне выраженный D-элемент (детерминатив или согласованное определение) во мн. ч. блокирует согласование глагола в ед. ч.	ДА
2	КГ без признака [+ D] лишена падежа, поэтому согласованные определения внутри КГ устраняются.	НЕТ
3	Согласование КГ, не содержащей детерминатива или согласованных определений, с глаголом во мн. ч. связано с реализацией синтаксически не выраженного абстрактного признака [+ D].	Непроверяемо
4	Постановка глагола в ср. р. ед. ч. реализует дефолтное согласование.	Зависит от соглашений о синтаксисе СочГ

Можно заключить, что модель КГ по [Лютикова 2017] непротиворечива, но прогноз 2. должен быть уточнен. Прогноз 3. непроверяем. Проверка прогноза 4. связана с дополнительными соглашениями о статусе СочГ. Универсалистский подход, основанный на постулате (ii) о синтаксической однородности конъюнктов СочГ, совместим с прогнозом 4, но не является его самостоятельным подтверждением. Операционалистский подход к определению СочГ, основанный на постулате (iii), не подтверждает прогноз 4. Поэтому есть условия для применения обоснованной в настоящей статье альтернативной модели, эксплицирующей предположение Н. Ю. Шведовой о том, что КГ является лексическим контролером предикатного согласования в ср. р. ед. ч.

5. Выводы

Выбор согласуемой vs несогласуемой формы прилагательного в ИГ малой структуры при существительном ж. р., ср. *две интересные книги ~ две интересных книги* не ограничивает тип предикатного согласования — *было издано/были изданы две интересные/интересных книги*, а также возможность использования КГ в конструкции *на две интересные/интересных книги больше*. Тем самым, внутренний и внешний синтаксис КГ не ограничивают друг друга при современном состоянии русской грамматики. Подлежащные КГ являются неканоническими аргументами в плане тема-рематического членения: ок. 65% их употреблений приходится на позицию ремы или на коммуникативно-нерасчлененные высказывания без темы. Утверждение, что постановка глагола в ед. ч. при подлежащной КГ отражает сниженную топикальность, ложно. Вариативность предикатного согласования в числе — диагностическое свойство русских КГ и СочГ вида КГ1 & КГ2. Постановку сказуемого в ср. р.ед..ч. уместно трактовать как идиосинкратическое согласование, контролируемое КГ, а не как отсутствие согласования.

References

1. *Arkadiev 2016* — Arkadiev, P. Case. Article for WSK Linguistic Typology (De Gruyter). Ms, 2016. Available at: https://www.academia.edu/27206077/Case
2. *Bailyn 2012* — Bailyn J. The Syntax of Russian. Cambridge: CUP, 2012.
3. *Chomsky 1995* — Chomsky, N. (1995) The Minimalist Program. Cambridge, Mass: MIT Press.
4. *Franks 1995* — Franks, S. L. Parameters of Slavic Morphosyntax. Oxford: OUP.
5. *Madariaga & Igartua 2017* — Madariaga, N. & I. Igartua. Idiosyncratic (Dis)agreement Patterns: The Structure and Diachrony of Russian Paucal Subjects. Scando-Slavica 63(2): 99–132.
6. *Madariaga & Igartua 2018* — Madariaga, N. & I. Igartua. The interplay of semantic and formal factors in Russian morphosyntax: Animate paucal constructions in direct object function // Russian linguistics 42: 27–55.
7. *Malchukov & Ogawa 2011* — Malchukov, A. & A. Ogawa. Towards a typology of impersonal sentences/ A. Malchukov, A. Siewierska (eds.). Impersonal constructions. A cross-linguistic study. [Studies in Language Companion Series 124]. Amsterdam: John Benjamins, 2011. P. 19–56.
8. *Matushansky 2008* — Matushansky O. A Case Study of Predication. Marušič, F. and R. Žaucer (eds.), Studies in Formal Slavic Linguistics. Contributions from Formal Description of Slavic Languages 6.5. Frankfurt am Main: Peter Lang. P. 213–239.
9. *Mikaelian 2013* — Mikaelian, I. Cardinal numeral constructions and the category of animacy in Russian. Russian linguistics 37 (1).
10. *Pereltsvaig 2006* — Pereltsvaig, A. Small Nominals. Natural Language and Linguistic Theory 24: 433–500.
11. *Willer Gold et alii 2018* — Willer Gold J., Arsenijević B., Batinić M., Becker M., Čordalija N., Kresić M., Lekoe N., Marušić F. L., Milićev T., Milićević N., Mitić I., Peti-Stantić A., Stanković B., Šuligoj T., Tušek J. and Nevins A. When linearity prevails over hierarchy in syntax. Proceedings of the National Academy of Sciences of the USA, Vol. 115, № 3, 495–500.
12. *Yanko 2005* — Yanko, T. Russian Numerals with Nouns Denoting Human Beings // General Linguistics, Vol. 43, No. 1–4.
13. *Zimmerling 2013* — Zimmerling, A. Transitive impersonals in Slavic and Germanic. Zero Subjects and Thematic Relations. Computational linguistics and intellectual technologies, Proceeding of the international conference “Dialogue 2013”, vol. 12 (19).
14. *Виноградов 1972* — Vinogradov, V. V. The Russian language. Grammatical theory of word. [Russkij Yazyk. Grammatičeskoe učenie o slove]. Moscow: Vysšaja škola.
15. *Герасимова, Лютикова 2019* — Gerasimova, A. A., Lyutikova, E. A. Case variation in Russian paucal constructions [Var’irovanie padežnogo oformlenija prilagatel’nogo v konstrukcijax s malymi čislitel’nymi]. The VI International Congress “The Russian Language. Historical trends and the present time” [VI meždunarodnyj kongress «Russkij yazyk. Istoričeskie sud’by i sovremennostj». Tezisy dokladov. Moscow: Lomonosov Moscow State University.

16. *Голуб 1997* — Golub, I. B. Russian stylistics [Stilistika russkogo yazyka]. Moscow: Iris-press.
17. *Зализняк 1967* — Zalizniak, A. A. Russian inflexion [Russkoe imennoe slovoizmenenie]. Moscow: Nauka.
18. *Кустова 2011* — Kustova G. I. Case. [Padež]. Russian Corpus Grammar [Русская корпусная грамматика] http://rusgram.ru/%D0%9F%D0%B0%D0%B4%D0%B5%D0%B6.
19. *Лютикова 2017* — Lyutikova E. A. Formal models of case. [Formal'nye modeli padeža]. Moscow: LRC.
20. *Лютикова 2018* — Lyutikova, E. A. The syntax of the nominal phrase in articles languages [Синтаксис именной группы в безартиклевом языке]. Moscow: LRC.
21. *Мельчук 1980* — Mel'čuk, I. A. On the case of the numeral expression in Russian constructions of the type bol'še na dva mal'čika or pot roe bol'nyx [O padeže čislovogo vyraženija v russkix slovosočetanijax tipa (bol'še) na dva mal'čika ili pot roe bol'nyx // Russian Linguistics. 1980. Vol. 5. No. 1: 55–74.
22. *Циммерлинг 2016* — Zimmerling, A. V. The linear-accent grammar and Russian thetic sentences [Linejno-akcentnaja grammatika i kommunikativno-nerasčlenennye predloženija v russkom yazyke] // Zimmerling A. V., Lyutikova E. A. (Eds). Clause architecture in the Parametric models: syntax, information structure, word order. [Zimmerling A. V., Lyutikova E. A. (red.). Arxitektura klauzy v parametričeskix modeljax: sintaksis, informacionnaja structura, porjadok slov]. Moscow: LRC: 76–103.
23. *Циммерлинг 2018* — Zimmerling A. V. Two Dialects of Russian Grammar: Corpus Data and a Model. [Dva dialekta russkoj grammatiki: korpusnye dannye i model'. Computational linguistics and intellectual technologies. Issue 17 (24). Proceedings of the international conference "Dialogue 2018" [Компьютерная лингвистика и интеллектуальные технологии. Вып. 17 (24)]. P. 818–830.
24. *Шахматов 1925* — Šaxmatov, A. A. Russian syntax [Sintaksis russkogo yazyka. Leningrad.
25. *Шведова 1970* — Švedova, N. Yu. (ed.). The Grammar of the present-day contemporary Russian [Grammatika sovremennogo russkogo literaturnogo yazyka. Moscow: Nauka.
26. *Шведова 1982* — Švedova, N. Yu. (ed.). Russian grammar [Russkaja grammatika. Vol 2. Moscow: Nauka.
27. *Янко 2008* — Yanko T. The intonation strategies of the Russian speech [Intonačionnye strategii russkoj reči]. Moscow: LRC.

THE ROLE OF ORIENTED GESTURES DURING ROBOT'S COMMUNICATION TO A HUMAN[1]

Zinina A. (zinina_aa@nrcki.ru),
Arinkin N. (arinkin_na@nrcki.ru),
Zaydelman L. (zaydelman_ly@nrcki.ru),
Kotov A. (kotov_aa@nrcki.ru)

Kurhcatov Institute, Moscow, Russia;
Russian State University for the Humanities, Moscow, Russia

The role of oriented gestures is crucial while solving spatial problems. We analyze the influence of a robot, using oriented gestures, on a human. In an experimental situation robot F-2 was helping a human to solve a "tangram" puzzle. Robot was indicating in speech, which game element to take and where to place it. In a half of the tasks the robot was using oriented communicative actions (hand gestures, head movements and gaze) to indicate the required game element, and then—the game position to place it in. In the other half of tasks, the robot was using non-oriented gestures. We show, that the use of oriented gestures increases the attractiveness of a robot to human and rises the general satisfaction of the interaction with the robot.

Keywords: Multimodal communication, oriented gestures, robot-to-human interaction

Several sciences—linguistics, psychology and social robotics—cooperate in order to explore the capabilities of a robot to maintain natural communication with humans. This area has a wide research potential—in particular, the search for "natural" and easy-to-use user interfaces is of fundamental importance [Breazeal, Scassellati, 2002]; [Beuter, Spexard et al., 2008]; [Klamer et al., 2011]. At the same time, this behavior of a robot should be as close as possible to the communicative behavior of a human, therefore—should be complex and diverse.

The interaction between a robot and a human can be modelled and studied in three modules: multimodal, cognitive and emotional [Lee et al., 2005]. Cognitive interaction is the ability of a robot to understand user's commands, emotional interaction is necessary to maintain positive relations between a robot and a user, and the multimodal module represents the means of interaction that are most convenient and familiar to humans. Many studies have shown that non-verbal communication plays an important role in coordinating actions when a robot and a person work together. In [Breazeal, 2003] it is experimentally proved that on the one hand, people send

[1] The research is supported by the Russian Science Foundation (project No 19-18-00547).

a robot to perform a physical task using speech and gesture. On the other hand, the non-verbal behavior of the robot has a positive effect on the success of solving the problem during human-computer interaction. In [Cabibihan, et al, 2009] it is shown that the pointing gestures accompanying the speech facilitate the understanding of spatial information in a videoconference. The authors in an experimental study proved that the use of pointing gestures increases the speed and accuracy of the task. Researchers [Häring, et al, 2012] studied the influence of gaze and pointing gestures of a humanoid robot on human performance in solving abstract puzzles of varying complexity. The authors confirmed that the directional gaze of the robot usually improves the interaction. At the same time, the authors have shown that additional pointing gestures are often necessary to make the interaction between the robot and a user more effective.

In [Salem et al., 2012] it is shown that participants evaluate the robot more positively when its nonverbal behavior (hand gestures) is reproduced along with speech, even if speech and gestures are semantically incongruent. The interaction between the robot and a user was evaluated in three experimental conditions: (1) unimodal—speech statements only, (2) congruent multimodal—semantically matching speech and gestures and (3) incongruent multimodal—semantically non-matching speech utterances and gestures. The authors have revealed an interesting effect: in the third condition, the robot was evaluated as more lively, active, friendly, sociable and cheerful compared to the robot in the second condition. That is, the robot was perceived more positively when the gesture did not correspond to the statement. The researchers suggested that the communicative behavior of the robot is positively evaluated by the user when it is potentially less predictable, and the robot is "imperfect".

In our work, we decided to test the effect of robot's oriented gestures on participants in an experimental study. We assumed that participants would prefer the robot that helps them and also uses pointing communicative actions. The study was conducted using F-2 robot, an experimental platform for studying the interaction between humans and robots. On the one hand, F-2 robot can construct a semantic text representation using the syntax parser [Kotov et al, 2015], [2017], [2018]. On the other hand, the robot selects communicative responses to the constructed meaning and reproduces the gestures and expressional patterns using the behavior management system [Kotov et al, 2019]; [Zinina et al., 2018].

1. Research procedure

In this experiment the robot helped participants to complete a Tangram puzzle. This puzzle is a well-known experimental media for studies of natural human communication [Clark, Wilkes-Gibbs, 1986], development of linguistic resources [Shore et al, 2018], [Gnjatović, Rösner, 2019] as well as for the design of robot communicative strategies [Kirschner et al., 2016]. The puzzle consists of 7 elements of different color, shape and size (two big triangles: red and blue, two small triangles: yellow and dark-blue, a middle green triangle, an orange parallelogram and a purple square—see **Fig. 1**). The task of a participant was to arrange the elements within a given contour on a white sheet. It was allowed to turn the tangram elements upside down.

Fig. 1. Tangram puzzle

During the experiment a participant was to complete: *Parallelogram*, *Whale*, *Triangle* and *Ship* figures (**Fig. 2**). The order of tasks presentation was random. The robot was located on the table in front of the participant and used speech instructions, gestures and gaze, instructing the participant to put a certain element on a certain place. Before each task the game elements were placed in front of the participant on the left and right sides of the playing field. Two paired elements (large triangles; small triangles) were always placed on different sides of the playing field. In its speech instructions the robot has always been referring to an element by its shape and size (not by color). Thus, an ambiguous reference in speech had been appearing when the robot mentioned one of the paired elements, such as *Take a big triangle!*

The behavioral scripts of the robot were organized as sequences of BML (Behavior Markup Language) packets: one sequence per task. The experiment was organized in the paradigm of the Wizard of Oz [Kelley, 1984] in which the moves by the player were evaluated as successful or not by a remotely located human operator. The robot was controlled through a Python script that has been sending BML packets to the robot. Special groups of BMLs were developed for a successful move by the player and for an unsuccessful move: wrong position or wrong game element. If the user's actions were correct, the operator gave the robot a command to praise the user and move on. If the participant was mistaken, the robot, according to the operator's command, informed the participant of the error and repeated the previous instruction. If the user has solved the entire figure before the end of BML protocol, the operator gave a command to the robot to praise the participant for successful work. The whole experiment was recorded from two viewpoints: the experimental situation from the side with the view on a player (Fig. 3) and the top view of the playing field (**Fig. 4**).

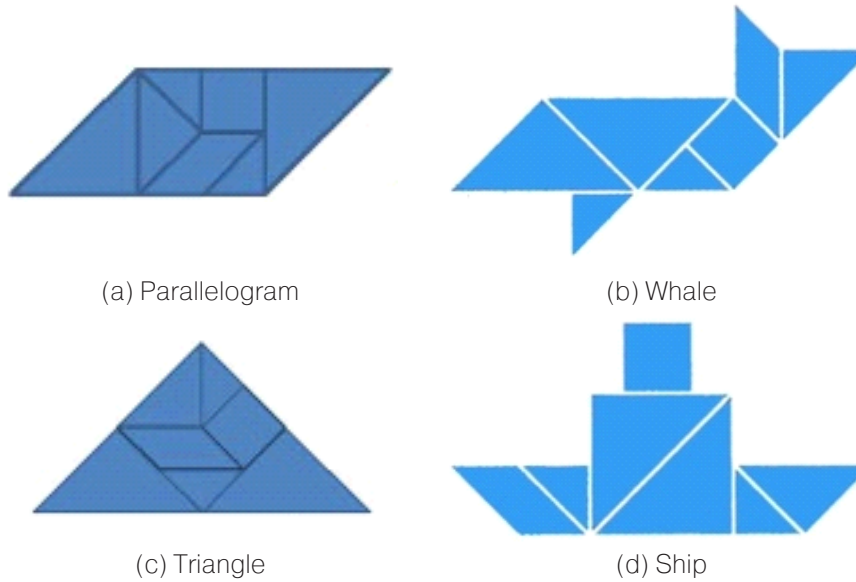


Fig. 2. Tangram figures for experimental tasks

Fig. 3. Experimental setup: side view

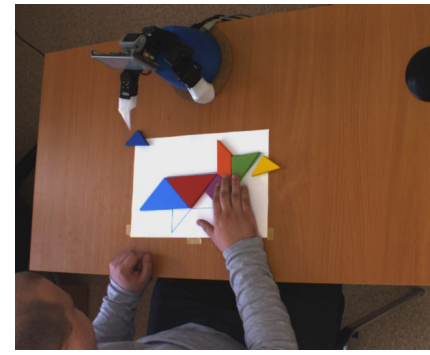


Fig. 4. Experimental setup: top view

The experiment involved 31 participants (12 female, 19 male), at a mean age 27.4. Among respondents 29 (90,7%) held tertiary education qualifications, 2 (9.3%) were students.

2. Experimental conditions

Two experimental conditions had to examine the role of pointing gestures/gaze within the game: in these conditions the robot helped a participant in different ways.

Condition 1: The robot accompanied its instructions by oriented communicative actions: it used pointing hand gestures, head and eye movements (Fig. 5). The robot

performed the oriented gestures while instructing the participant, which element to take and where to place it. For example, the robot pronounced: *Take this* (pointed and looked at the element) *little triangle*. In this experimental condition each participant completed *Parallelogram* and *Whale* figures.

Condition 2: The robot did not use pointing gestures, its speech instructions were accompanied by non-oriented movements of hands, head and eyes (Fig. 6). The movements of the robot were selected in such a way as to exclude any directions to the game elements or sides of the playing field. For example, the head tilts were performed along a straight vertical path, eyes could only move up and down, and hand gestures were strictly symmetrical, performed by both hands. In this experimental condition the participant completed *Triangle* and *Ship* figures.

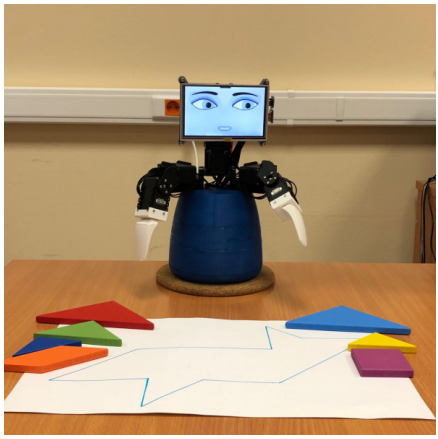


Fig. 5. Condition 1: Pointing gestures

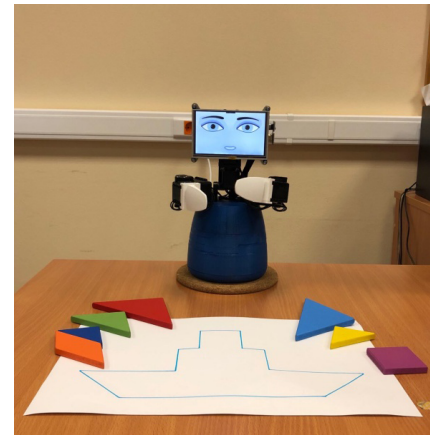


Fig. 6. Condition 2: Symmetric gestures

Tasks with different experimental conditions were presented in random order. After the experiment participants had to fill out a questionnaire: describe the difference in robot's behavior, choose the preferred condition and rate the robot in two experimental conditions on five-point semantic differential scales. In addition, the experiment recorded objective indicators—the speed of solving each task and the number of participants' errors.

3. Results

The experiment took 8.5 minutes on average. *Whale* took the longest time to complete (1 minute 44 seconds) and *Ship* took the shortest time (1 minute 14 seconds) (Fig. 7). *Whale* was also the most complex figure—the participants made the most mistakes while completing this figure. The simplest figure for the participants was *Triangle* (Fig. 8).

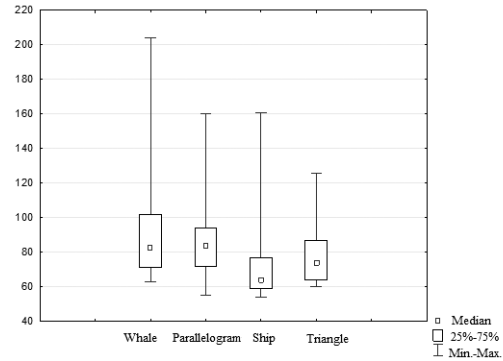


Fig. 7. Tasks solving time (s)

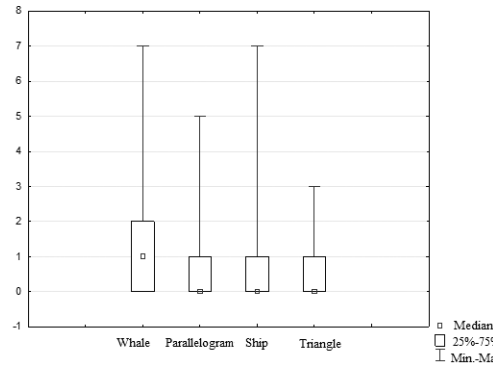


Fig. 8. The number of mistakes

64.5% of the participants preferred the robot pointing to the tangram elements and their locations; 32.3% of respondents equally liked the robot in two experimental conditions. The participants noted that they were waiting for clues from the robot when the conditions were changed. As soon as the robot stopped using pointing gestures, the participants chose the elements longer and more often doubted their positiosn inside the contour. Only one participant (3.2% of the entire sample) preferred the robot that did not use oriented gestures in its instructions. This could be explained, as the participant was mistaken four times while completing *Whale* figure (a task with pointing gestures). Because of this, the robot repeated the instruction four times and the participant, as he has indicated in the self-report, "felt foolish."

The difference between the experimental conditions was not obvious for participants: only half of the players (15 people; 48.4% of the total group) noticed the difference between robot's performance with and without pointing gestures. According to the results of the experiment, most participants (20 people, 64.5% of the total group) significantly preferred the robot with oriented gestures (**Condition 1**) (chi-squared 18.1, $p < 0.01$). Therefore, the assessment can be implicit because the participants didn't always distinguish two experimental conditions, but much more

often preferred the robot that indicates the necessary element and its location in the contour with the help of head, eye and hand movements.

At the same time, a nonverbal instruction was significant even for the participants, who did not notice the difference. We have evaluated the participants' choice for a Tangram paired element to analyze the implicit perception of robot's pointing gestures. When two paired elements are not yet used and are placed on the two sides of the playing field, the robot may refer to such an element ambiguously as *big triangle* or *small triangle*. When the robot used a pointing gesture with this reference (on left or right side), the participants followed this nonverbal indication in 91.1% of cases and took the element the robot was pointing at. This can indicate the substantial influence of oriented gestures on the user's behavior, even if the user did not reflect this influence in the self-report.

In the study, we did not identify a link between the preferred condition and whether the participants noticed the difference between the conditions (chi-squared 2.7, $p > 0.05$). According to the collected data, 16 respondents (51.6% of the whole group) stated that they did not notice the difference between the robot that used pointing gestures and the robot that accompanied its instructions with non-oriented movements. These participants in 43.7% of the cases equally evaluated the attractiveness of the robot. However, even these participants followed the robot's gestural instructions in 78.5% of cases. Therefore, one could speak of the implicit influence of the robot oriented gestures on a user behavior, even if the user did not clearly realize this influence.

Moreover, there are cases when participants for several seconds wait for the robot's pointing gesture to resolve the ambiguity (for example, when the participant selects one of the small or large triangles). In addition, we can observe "reverse" in participants' gestures. Reverse is a reciprocating motion when choosing a certain figure. For example, the robot instructs the user: *Take the little triangle*. After that the user brings one hand to one triangle, and another—to the second triangle. The participant can look like "frozen" while he waits for the pointing gesture of the robot.

11 people (73.3%) out of those who noticed the difference in the robot's behavior (15 people, 48.4% of the whole sample) prefer robots using pointing gestures (**Condition 1**). This distribution does not correspond to the normal (chi-squared 9.9, $p < 0.01$), therefore, these parameters are correlated. There are also those participants who equally evaluate robots in different conditions. It can be assumed that these participants preferred verbal instructions, when evaluating robots. The differences between the conditions obtained with the standard scales of the semantic differential is not revealed.

4. Conclusion

We conducted a study that provides further insights into the question of robot-to-human interaction. As we have shown, the robot's oriented gestures in solving a spatial problem are important to give a positive impression on a user. Results showed that participants significantly preferred when the robot used oriented gestures rather than it did not use pointing gestures. This effect is observed even if the participants did not explicitly notice the difference between pointing and non-pointing behavior.

The obtained results open up perspective for a further research on the interaction between robot and humans. Within future studies we plan to evaluate the contribution of expressive means of the robot to its attractiveness for a user, the influence of the shape designation method on the efficiency of solving the problem, etc.

Moreover, the developed system becomes helps to test communicative strategies (for example, using positive or negative feedback) and styles in a dialogue between robot a user as well as is to evaluate the effectiveness of such strategies in different communication situations and in different socio-demographic groups.

References

1. *Beuter N., Spexard T., Lutkebohle I., Peltason J., Kummert F.* (2008) Where is this?—gesture based multimodal interaction with an anthropomorphic robot, 8th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2008), pp. 585–591.
2. *Breazeal C.* (2003) Emotive qualities in lip-synchronized robot speech, Advanced Robotics, V. 17(2), pp. 97–113.
3. *Breazeal C., Scassellati B.* (2002) Robots that imitate humans, Trends in Cognitive Science, V. 6(11), pp. 481–487.
4. *Brugman H., Russel A.* (2004) Annotating Multimedia Multi-modal resources with ELAN, Proceedings of the 4th International Conference on Language Resources and Language Evaluation (LREC 2004), pp. 2065–2068.
5. *Cabibihan J. J., So W. C., Nazar M., Ge S. S.* (2009) Pointing Gestures for a Robot Mediated Communication Interface. In: Xie M., Xiong Y., Xiong C., Liu H., Hu Z. (eds) Intelligent Robotics and Applications. ICIRA 2009. Lecture Notes in Computer Science, V. 5928, pp. 67–77.
6. *Clark H. H., Wilkes-Gibbs D.* (1986) Referring as a collaborative process, Cognition, pp. 1–39.
7. *Gnjatovic, M., Rosner, D.* (2010). Inducing genuine emotions in simulated speech-based human-machine interaction: The nimitek corpus. IEEE Transactions on Affective Computing, 1(2), 132–144.
8. *Häring M., Eichberg J., André E.* (2012) Studies on Grounding with Gaze and Pointing Gestures in Human-Robot-Interaction, Social Robotics. ICSR 2012. Lecture Notes in Computer Science, V. 7621, pp. 378–387.
9. *Kelley J. F.* (1984) An iterative design methodology for user-friendly natural language office information applications, ACM Transactions on Office Information Systems, March 1984, V. 2(1), pp. 26–41.
10. *Kirschner, D., Velik, R., Yahyanejad, S., Brandstötter, M., Hofbaur, M.* (2016). YuMi, come and play with Me! A collaborative robot for piecing together a tangram puzzle. In International Conference on Interactive Collaborative Robotics, pp. 243–251.
11. *Klamer T., Allouch S. Ben, Heylen D.* (2011) Adventures of Harvey: Use, acceptance of and relationship building with a social robot in a domestic environment, Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, V. 59 LNICST, pp. 74–82.

12. *Kotov A., Arinkin N., Filatov A., Zaidelman L., Zinina A.* (2017) Semantic Comprehension System for F-2 Emotional Robot, BICA 2017: Biologically Inspired Cognitive Architectures, pp. 126–132.
13. *Kotov A., Zinina A., Filatov A.* (2015) Semantic Parser for Sentiment Analysis and the Emotional Computer Agents, Proceeding of the AINL-ISMW FRUCT Conference, pp. 167–170.
14. *Kotov A. A., Arinkin N. A., Zaydelman L. Y., Zinina A. A.* (2019) Linguistic Approaches to Robotics: From Text Analysis to the Synthesis of Behavior, Language, Music and Computing. LMAC 2017. Communications in Computer and Information Science, V. 943, pp. 207–214
15. *Lee K. W., Kim H.-R., Yoon W. C., Yoon Y.-S., Kwon D.-S.* (2005) Designing a human-robot interaction framework for home service robot, ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, DOI: 10.1109/ROMAN.2005.1513793
16. *Salem, M., Kopp, S., Wachsmuth,* (2012) Generation and Evaluation of Communicative Robot Gesture, International Journal of Social Robotics, V. 4(2), pp. 201–2017
17. *Shore, T., Androulakaki, T., Skantze, G.* (2018). KTH Tangrams: A Dataset for Research on Alignment and Conceptual Pacts in Task-Oriented Dialogue. In Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pp. 768–775.
18. *Zinina A. A., Arinkin N. A., Zaydelman L. Y., Kotov A. A.* (2018) Development of communicative behavior model for f-2 robot basing on «REC» multimodal corpora [Razrabotka modeli kommunikativnogo povedeniya robota f-2 na osnove mul'timodal'nogo korpusa «REC»], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2018" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2018"], V. 17 (24), pp. 831–844.

CROSS-LANGUAGE TEXT ALIGNMENT FOR PLAGIARISM DETECTION BASED ON CONTEXTUAL AND CONTEXT-FREE MODELS

Zubarev D. V.[1], **Sochenkov I. V.**[2,1]

[1]Federal Research Center 'Computer Science and Control' of Russian Academy of Sciences, Moscow, Russia

[2]Skolkovo Institute of Science and Technology, Moscow, Russia

In this paper, we present a dataset for cross-language (Russian-English) text alignment subtask of plagiarism detection. We compare different models for detecting translated plagiarism. One is based on different textual similarity scores, which exploit word embeddings. Another model extends the previous one with the features obtained via neural machine translation. The last model is built on top of pre-trained language representation (Bert) via fine-tuning for our task. The Bert model shows great performance and outperforms other models. However, it requires much more computation resources than simpler models. Therefore, it seems reasonable to use both context-free models and contextual models together in modern plagiarism detection systems.

Keywords: cross-language text alignment, cross-language plagiarism detection, word embeddings, neural machine translation, Bert

ОБНАРУЖЕНИЕ ФРАГМЕНТОВ В ЗАДАЧЕ КРОСС-ЯЗЫКОВОГО ПОИСКА ЗАИМСТВОВАНИЙ С ИСПОЛЬЗОВАНИЕМ ЭМБЕДДИНГОВ СЛОВ

Зубарев Д. В.[3], **Соченков И. В.**[4,3]

[3]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, Москва, Россия

[4]Сколковский институт науки и технологий, Москва, Россия

1. Introduction

Plagiarism is a serious and known problem in education and research, especially in developing countries like Russia. The availability of the huge amount of texts on the Web and free machine translation services makes it easier to create an "original" study.

Systems for detecting plagiarism are very common now, and they are capable of detecting monolingual plagiarism even with obfuscations.

However, detecting translated plagiarism is a very challenging task, and there are no such tools on the Russian market of plagiarism detection systems. However, recent studies showed that cross-language text reuse is common in research and education [Bahteev et al., 2018]. Therefore, it is important for the state-of-the-art system to detect cross-language plagiarism too.

Commonly plagiarism detection is divided into two stages: source retrieval and text alignment.

- On the source retrieval stage for a given suspicious document, we need to find all sources of probable text reuse in a large collection of texts. For this task, a source is a whole text, without details of what parts of this document were plagiarized. Typically we get a large set of documents (around 1,000 or more) as a result of this stage. Those documents are called "candidates".
- On the text alignment stage: we compare suspicious document to each candidate to detect all reused fragments, and identify its boundaries.

For a review of the state-of-the-art monolingual methods of plagiarism detection see [Stamatatos et al., 2015]. The same stages are valid for cross-language plagiarism detection too. In this work, we study only the second task. It means that we operate with pairs of a suspicious document and a candidate. To simplify this task further, we consider only pairs of sentences, as if we split the suspicious document and a source into sentences and combine all of them (Cartesian product of sentences in the source and target documents). Then our task is to identify the pairs of sentences that are translated from English to Russian.

2. Related work

The overview of different approaches for solving this task is presented in [Potthast et al., 2011]. Also, there made an evaluation and a detailed comparison of some featured methods. In [Ferrero et al., 2017] is described a method for cross-language similarity detection based on the distributed representation of words (word embeddings). Experiments were conducted on English-French corpus. In [Kutuzov et al., 2016] is described a training of word embeddings on monolingual comparable corpora and learning the optimal linear transformation of vectors from one language to another (there were used Russian and Ukrainian academic texts). Also there were discussed usage of those embeddings in source retrieval and text alignment subtasks.

SemEval 2017 [Cer et al., 2017] was focused on multilingual and cross-lingual semantic textual similarity (STS) of sentence pairs. Unfortunately, the Russian language was not presented in any task. Most participants used common machine translating systems to transform the task of cross-lingual STS to monolingual task.

However, the reverse translation of the suspicious document is not enough (and the translation of potential source as well), since machine translation generates multiple variants, and unscrupulous authors modify reused fragments.

Another task similar to plagiarism detection is the cross-lingual language understanding (XLU) [Conneau et al., 2018] that is derived from Multi-Genre Natural Language Inference (MNLI) Corpus. Only the development and test sets of MNLI were translated, and that corpus is mostly dedicated for the zero-shot models.

3. Cross-language sentence similarity scores

In this section, we describe how we measure the similarity between two sentences. For this, we trained cross-language word embeddings based on a large parallel corpus. We used these embeddings to estimate two different similarity scores: one is based on sentence embeddings, and other is calculated after the substitution of all words with the most similar ones in the other language. We trained the neural machine translation system to obtain an additional similarity score by comparing various N-grams of sentences. Finally, we used these similarity scores to construct a dataset for cross-language text alignment task.

3.1. Preprocessing

On a preprocessing stage, we split each sentence into tokens and lemmatize tokens of texts using methods described in [Osipov et al., 2013]. In addition, we removed words with non-important part of speech: conjunction, pronoun, preposition, etc., and common stop-words (be, являться). The result of preprocessing is the source to learn embeddings.

3.2. Cross-language word embeddings

We train cross-language word embeddings for a Russian-English pair on parallel corpora available on the Opus site[1], namely:

- OpenSubtitles2018
- News Commentary
- TED Talks 2013
- MultiUN
- ParaCrawl
- Wikipedia

We extended this dataset with sentences from the Yandex Parallel corpus[2] [Antonova et al., 2011]. We used texts of various genres since we wanted to have a large vocabulary. The final goal of training word embeddings was to be able to find most common translations of Russian word/phrases to English. Generally, we were

[1] Opus: the open parallel corpus, http://opus.nlpl.eu/

[2] Англо-русский параллельный корпус: https://translate.yandex.ru/corpus?lang=en

interested in multiple translations from different genres since we do not know in advance what text will be checked (e.g., for word “chick” there were translations девчонка and цыпленок). All parallel sentences were preprocessed. After that, all pairs that had a difference in the size of more than five words were filtered out.

We extended our model for cross-language word embeddings adding phrases representing some concepts/terms. Thus, we used parallel concepts from Wikidata[3], which consist of multiple words, as a single phrase (e.g., military_law) when training embeddings. Thus we can learn the similarity between words and phrases (sustainable word combinations) that are represented differently in two languages: such as “зубная_щетка” and toothbrush. We excluded concepts that are too rare for our corpus (<10 occurrences) and those concepts that had an irrelevant category (film/song/books titles) for our purposes.

Finally, we assembled a corpus of more than 44 million sentences (for each language). The dictionary size was around 507,000 words/phrases.

We applied the method proposed in [Vulić et al., 2015], designed for learning bilingual word embeddings from non-parallel document-aligned data, but it can be used for learning on parallel corpora too. According to the method, two comparable texts in different languages are combined into one pseudo text. In our case, we interleaved two parallel sentences, e.g. if we were given two sentences: “Мама мыла раму” and “Mother washed the frame”, the result of their merging is: “мама mother мыла washed раму the frame”. Since we removed auxiliary words from sentences, we assumed that corresponding Russian and English words were in the same context window. It would not be the case if there is a different words order, and it can be somehow fixed with larger context window, but in our experiments and in [Upadhyay et al., 2016] the context window of 5 words was enough. After that, the word2vec skip-gram model [Mikolov et al., 2013] is used on the resulting corpus of merged texts. We used gensim word2vec implementation with those parameters: dimensionality of embeddings was 300, a window size of 5 words, the minimal corpus frequency of 10, negative sampling with 10 samples, no down-sampling, 15 iterations over the corpus.

3.3. Sentence embeddings

Sentence embedding in our approach is defined by averaging embeddings of its words and phrases.

$$Emb(s) = \frac{1}{|s|}\sum_{w \in s} Emb(w)$$

We tried various approaches to obtain sentence embeddings from the word embeddings [Rücklé et al., 2018], but their performance was slightly worse than an average of word vectors.

We chose cosine similarity as a sentences similarity score.

[3] Wikidata: https://www.wikidata.org/wiki/Q321

3.4. Words substitution

The natural approach to measure the cross-language similarity between sentences is to substitute Russian words from the sentence with the N_s most similar English words. The similarity is determined based on cosine similarity between embeddings of words/phrases. On the first step, we tried to replace the whole phrases in the sentence if they were found in the dictionary. On the second step, we replaced single words with their most similar English analogues. We precalculated top N_s of similar words for each Russian word/phrase in our dictionary to increase computational performance.

Let S_r and S_e denote the Russian and English sentences respectively. For the simplicity, we will consider the case when we map words from Russian sentence S_r to English words.

$Si(w_1, w_2)$ is a cosine similarity score between two words w_1 and w_2, $Top(w, l)$ is a function that for word w returns N_s most similar words of language l.

We then define for each English word w_e a set of Russian words that have w_e in their top of similar words. $(w_e) = \{w_r \mid w_r \in S_r \wedge w_e \in Top(w_r, \text{'en'})\}$. Then non-normalised similarity score is calculated by the following formula.

$$NSubst(S_r, S_e) = \sum_{w \in S_e} \frac{\sum_{w_r \in M(w)} Sim(w, w_r) \cdot (\upsilon(w) + \upsilon(w_r))}{|M(w)|}$$

where (w) is the number of words in a phrase if w is a phrase and 1 otherwise.

A normalized variant is presented below:

$$N_r = \sum_{w \in S} \upsilon(w) \quad N_e = \sum_{w \in S} \upsilon(w)$$

$$Subst(S_r, S_e) = \frac{NSubst(S_r, S_e)}{N_r + N_e}$$

3.5. Neural machine translation (NMT)

We used OpenNMT-py[4] library to train a machine translation (MT) system as an additional criterion to estimate the pairwise similarity between the sentences. For this purpose, we employed a subset of the parallel corpus, which was used for learning embeddings. We adjusted corpus in the following ways:

- sentences from OpenSubtitles2018 were removed since they are generally too short after stop-words removing;
- the maximum difference in length between sentences should have been no more than 2;
- some sentences from ParaCrawl and MultiUN were dropped to reduce the time of training.

Thus we got about 7 millions of parallel sentences.

[4] An open source neural machine translation system: http://opennmt.net/. Pytorch implementation https://github.com/OpenNMT/OpenNMT-py v0.5

We did not use pre-trained embeddings for training MT system. We used default architecture RNN encoder-decoder with attention and mostly default settings provided by OpenNMT, with Russian and English dictionary sizes: 95k and 80k respectively, with 64 batch size and with 600k train steps.

We used the trained model for translating Russian sentences and measuring Jaccard similarity between sentence pairs without any further preprocessing.

$$NMT(S_{rt}, S_e) = \frac{|S_{rt} \cap S_e|}{|S_{rt} \cup S_e|},$$

where S_{rt} is a translated Russian sentence, and S_e is an English sentence. We measured similarity on 1-grams (NMT) and on 2-grams (NMT2).

4. Dataset

To address our cross-language text alignment task we created a dataset for Russian-English plagiarism detection. 16k sentence pairs were taken from Yandex parallel corpus (those sentences were not used for learning word embeddings), and 4k sentences were manually written by students. Students should have searched sources in English using common Web search engines. After that, they must have translated them. They were allowed to use common translation tools like Google Translate or Yandex Translate, but the adjustment of the translated text was required to produce correct Russian text. Some percent of sentences should have been translated without any automatic tools. Those pairs are positive examples of plagiarized pairs of sentences.

Modern methods for training language models and learning word embeddings need negative sampling to get meaningful results. To obtain negative samples, we compared each Russian sentence to all English sentences (except one that is a translation) using various sentence similarity scores (described in the previous sections). The most similar sentences were selected for each Russian sentence as negative examples. We controlled that for each Russian sentence there should be distinct English translations, because the same English translation may be the most similar one by various scores. However, the same English translation could have been selected for different Russian sentences. The rationale behind this approach is that random negative sampling would not have any reasonable effect due to involving quite different sentences into the training set. However, the desired behavior of the model is to separate plagiarized sentences from others, which contain some similar lexis but have different semantics. To achieve this, we generated two corpora with different amount of negative samples per each positive sentence pair (containing source and plagiarized sentences):

- Negative-1: One negative example was selected randomly from the most similar sentences. According to [Belyy et al., 2018] one negative sample is enough for the text alignment task. We used this dataset for training and tuning models.
- Negative-4: 4 negative examples were selected (one most similar sentence for each used similarity score). By its nature, this task requires a comparison of many pairs of sentences. Moreover, most of these pairs are negative examples. Therefore, we used this dataset for testing purposes, to check how models handle a larger amount of negative examples.

The obtained corpora characteristics are presented in the table (sizes are in sentences).

Table 1. Corpora characteristics

	training set size	hold-out set size	test set size
Negative-1	28,320	4,000	7,998
Negative-4	65,962	9,266	18,613

These corpora are freely available[5].

5. Models

We used previously described sentence similarity scores as simple baselines. If a score for some pair was greater than some threshold, this pair was considered to be a case of plagiarism. The process of tuning thresholds is described in the next section. As a more complex model, we used a classifier with similarity scores as features. In addition, we tried to use some pre-trained language representations for this task.

5.1. Similarity scores

We tuned thresholds and parameters for each score independently. We selected parameters using grid search to maximize F1 on the hold-out Negative-1 dataset. The obtained values of parameters are presented in the table.

Table 2. Thresholds and parameters values

	T_{cos}	T_{subst}	N_s	T_{nmt}	T_{nmt2}
Negative-1	0.7	0.25	3	0.2	0.06

5.2. Logistic Regression Classifier

We trained two Logistic Regression (LR) classifiers (with L2 regularization and C=1.0) that used the previously described sentence similarity scores as features. The first classifier (LR-1) used all features, whereas the second (LR-2) used only sentence embeddings similarity score and words substitution similarity score as features.

[5] http://nlp.isa.ru/ru-en-text-align-corp

5.3. Bert

We fine-tuned Bert [Devlin et al., 2018] Multilingual model[6] for our classification task. We considered a simple linear layer for the sentence pair classification on top of the pooled output of Bert. Pooling is done by simply taking the hidden state corresponding to the first token ('CLS' in this case) in the input sequence. The same pooling was used to train Bert originally to perform Next-Sentence-Prediction task. The input sequence consists of two sentences separated with the special token 'SEP'. The task is the same as paraphrase detection MRPC[7]. The main difference that the multilingual model is used and the two sentences are in a different language. Bert was trained on Negative-1 training data with the following parameters: `max_seq_length - 128`, `train_batch_size - 32`.

5.4. Laser sentence embeddings

Another approach to get sentence embeddings is called Language-Agnostic SEntence Representations (LASER). LASER [Artetxe et al., 2018] provides a BiLSTM encoder, which was trained on 93 languages. The encoder was coupled with an auxiliary decoder and trained on publicly available parallel corpora.

We obtained sentence embeddings from the encoder via max pooling of the last layer outputs. We apply cosine similarity on corresponding sentence embeddings of each sentence pair to determine whether this is a plagiarism case. The threshold was tuned on Negative-1 hold-out set to maximize F1 score. The value of the tuned threshold was 0.72.

6. Evaluation Results

In this section, we present evaluation results, obtained on test sets of two corpora negative-1 and negative-4, for each score independently and for the classifiers.

The results are in the **table 3**.

Bert outperforms the classifiers on both datasets. Bert's performance drops while testing on the larger corpus, but this decrease was lesser than for the classifiers.

The classifiers outperform all standalone similarity measures. The best F1 score has the classifier trained on all features (LR-1). Its performance is decreased on the 0.09 when moving to the larger data set (Negative-4), whereas the performance of the best result among similarity scores (NMT) dropped on 0.13 when comparing F1 scores. It is clear that the higher number of negative examples is the lesser precision, although recall stays the same.

---

[6] BERT-Base, Multilingual Cased: https://storage.googleapis.com/bert_models/2018_11_23/multi_cased_L-12_H-768_A-12.zip.

[7] Microsoft Research Paraphrase Corpus: https://www.microsoft.com/en-us/download/details.aspx?id=52398.

Table 3. Results on the test set

	Negative-1			Negative-4		
	Precision	Recall	F1	Precision	Recall	F1
Sentence embeddings	0.75	0.77	0.76	0.45	0.77	0.57
Words substitution	0.84	0.76	0.80	0.6	0.76	0.66
NMT	0.85	0.8	0.82	0.61	0.8	0.69
NMT2	0.83	0.64	0.72	0.54	0.64	0.58
LR-2 (2 features)	0.87	0.8	0.83	0.64	0.8	0.71
LR-1 (all features)	0.91	0.8	0.85	0.73	0.8	0.76
Laser	0.9	0.89	0.89	0.7	0.89	0.78
Bert	**0.96**	**0.93**	**0.95**	**0.88**	**0.93**	**0.9**

It is debatable what score should be used for evaluation of the PD system. There are multiple scenarios of usage for such systems, which requires different qualities of the system. Precision is more important when searching for literal plagiarism. The PD system may be used as a source of evidence during some legal procedures. Few plagiarized fragments are enough in this case, and it is also important to minimize manual checking of not plagiarized fragments. High recall is vital for researchers that scrupulously study origins of some piece of work or for checking students' essays. One may argue that it is easy to obtain high precision and recall scores independently: for the former, it is enough to find one literal plagiarism case, for the latter one can return all fragments marked as plagiarism. We think that the balance of precision and recall is quite important for this task and the PD system. It is required to find all plagiarism cases, and in the same time not to clutter the results with many false positives, which can make the inspection of a final report quite challenging for a human expert. We think that the real world PD system should optimize F1 score if there are no specific requirements by default, and should be tunable for various specific use-cases, which require either higher precision or higher recall.

Also, the PD system should be quite efficient to be able to cope with a large amount of checks in a small amount of time (typical use-case scenario during exam session). The computation times of all models are presented in the table.

Table 4. Computation time of models in seconds

	Negative-1	Negative-4
Sentence embeddings	2.89	4.02
Words substitution	2.63	3.30
NMT on GPU	34.15	34.31
NMT on CPU	240.13	240.29
LR-2 (2 features)	5.53	7.34
LR-1 (all features)	39.68	41.65
LASER	7.63	11.04
Bert	91.95	197.45

Bert is quite slow: it requires about 90 seconds to classify all pairs in Negative-1 test set, using one GPU GTX 1080. Bert is more than two times slower than the classifier and orders of magnitude slower than the simple classifier (LR-2). The LASER embeddings show good balance between F1 and computational cost. Considering that we did not pre-learn English embeddings, its computation time may be reduced further.

As a side note, the LR-1 classifier trained on Negative-1 and Negative-4 showed roughly the same performance, when were tuned on Negative-4 hold-out set.

Table 5. Results on test set

	Recall	Precision	F1
LR-1 trained on Negative-1	0.75	0.81	0.779
LR-1 trained on Negative-4	0.75	0.82	0.782

To achieve these results, we tuned classifier's margin b on the hold-out Negative-4 set $S = (x, y)$ [Belyy et al., 2018].

$$b = argmax_b F1(y, [cls(x) - b > 0])$$

Results of tuning:

Table 6. Classifier's margin values after tuning

	b
LR-1 trained on Negative-1	0.6
LR-1 trained on Negative-4	0.33

The tuning of the classifier on the Negative-1 hold-out set yielded $b == 0.5$, which is not a surprise since the set was balanced.

7. Conclusion

In this paper, we presented a dataset for cross-language (Russian-English) text alignment task as an alternative to existing datasets. We compared different models for detecting translated plagiarism. One is based on various textual similarity scores that exploit word embeddings and neural machine translation. Another model is built on top of pre-trained language representation via fine-tuning for our task. The Bert model showed great performance and outperformed our custom model. However, in the production usage, it is common to process a hundred millions of sentence pairs only for one suspicious document. It is not practical to employ Bert model for such a computationally expensive task. It is reasonable to filter out many negative pairs with a more efficient method. As such a method, it is possible to use our classifier with reduced feature space: only with sentence embeddings and word substitution measures. This classifier, tuned to maximize recall, can significantly decrease the load on the more complex processing downstream. Also, it seems promising to employ

cross-language sentence embeddings (LASER) for the preprocessing step. Since the embeddings for the source and suspicious sentences could be built only once. After that, it is possible to use efficient nearest-neighbor search algorithms [Johnson et al., 2017] to find similar vectors.

Extending vocabulary is another important issue, which should be considered for any real cross-language plagiarism detection system. Most available parallel corpora contain common lexis. However, plagiarism detection should also work for scientific papers, patents, etc. containing a lot of special lexis and terms. One of the possible solutions is to create parallel corpora from comparable corpora [Zweigenbaum et al., 2018] using the system for translated plagiarism detection and extend vocabulary with new parallel data. It requires additional study.

This work was dedicated to text alignment subtask of plagiarism detection task. The source retrieval subtask is the first and crucial step in plagiarism detection. We plan to address this problem in future research, using cross-language word embeddings described in this paper.

8. Acknowledgments

The reported study was funded by RFBR according to the research projects № 18-37-20017 & № 18-29-03187.

References

1. *Antonova, A., & Misyurev, A.* (2011, June). Building a web-based parallel corpus and filtering out machine-translated text. In Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web (pp. 136–144). Association for Computational Linguistics.
2. *Artetxe, M., & Schwenk, H.* (2018). Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. arXiv preprint arXiv:1812.10464.
3. *Bahteev, O. Y.* (2018). Cross-language plagiarism detection in large collections of scientific documents: http://www.machinelearning.ru/wiki/images/f/ff/Sample_final.pdf.
4. *Belyy A. V., Dubova M. A.* (2018). Framework for Russian plagiarism detection using sentence embedding similarity and negative sampling. In Proceedings of the International Conference "Dialogue 2018" (pp. 96–110).
5. *Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L.* (2017). Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. arXiv preprint arXiv:1708.00055.
6. *Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S. R., Schwenk, H., & Stoyanov, V.* (2018). Xnli: Evaluating cross-lingual sentence representations. arXiv preprint arXiv:1809.05053.
7. *Devlin, J., Chang, M. W., Lee, K., & Toutanova, K.* (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

8. *Ferrero, J., Agnes, F., Besacier, L., & Schwab, D.* (2017). Using Word Embedding for Cross-Language Plagiarism Detection. arXiv preprint arXiv:1702.03082.
9. *Johnson, J., Douze, M., & Jégou, H.* (2017). Billion-scale similarity search with GPUs. arXiv preprint arXiv:1702.08734.
10. *Kutuzov et al.* (2016). Clustering Comparable Corpora of Russian and Ukrainian Academic Texts: Word Embeddings and Semantic Fingerprints. Proceedings of the Ninth Workshop on Building and Using Comparable Corpora (LREC 2016, Portorož, Slovenia, May 23, 2016). pp. 3–10
11. *Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J.* (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111–3119).
12. *Osipov G., Smirnov I., Tikhomirov I., Shelmanov A.* (2013), Relational-situational method for intelligent search and analysis of scientific publications, Proceedings of the Integrating IR Technologies for Professional Search Workshop, pp. 57–64.
13. *Potthast, M., Barrón-Cedeño, A., Stein, B., & Rosso, P.* (2011). Cross-language plagiarism detection. Language Resources and Evaluation, 45(1), 45–62.
14. *Rücklé, A., Eger, S., Peyrard, M., & Gurevych, I.* (2018). Concatenated $ p $-mean Word Embeddings as Universal Cross-Lingual Sentence Representations. arXiv preprint arXiv:1803.01400.
15. *Stamatatos, E., Potthast, M., Rangel, F., Rosso, P., & Stein, B.* (2015, September). Overview of the pan/clef 2015 evaluation lab. In International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 518–538). Springer, Cham.
16. *Upadhyay, Shyam, et al.* (2016). Cross-lingual models of word embeddings: An empirical comparison. arXiv preprint arXiv:1604.00425.
17. *Vulić, I., & Moens, M. F.* (2015). Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (Vol. 2, pp. 719–725).
18. *Zweigenbaum, P., Sharoff, S., & Rapp, R.* (2018, May). Overview of the third BUCC shared task: Spotting parallel sentences in comparable corpora. In Proceedings of 11th Workshop on Building and Using Comparable Corpora (pp. 39–42).

Abstracts

PRAGMATICS IN THE INTERPRETATION OF SCOPE IN WRITTEN RUSSIAN TEXTS

Apresjan V. Ju. (valentina.apresjan@gmail.com, vapresyan@hse.ru), National Research University Higher School of Economics, Vinogradov Russian Language Institute of the Russian Academy of Sciences

The paper is a corpus study of pragmatic factors involved in disambiguating sentences with negation and universal quantifier in written Russian and English, such as *Ja ne pozval vseh svoih dal'nih rodstvennikov,* 'I haven't invited all of my distant relatives.' Ambiguity results from differences in scope. If negation scopes over the quantifier, we get partial negation: 'I have invited some, but not all of my distant relatives.' If negation scopes over the verb, we get total negation: 'I haven't invited any of my distant relatives.' Our study is based on Russian and English data extracted from a variety of corpora.

We demonstrate that despite syntactic differences, Russian and English rely on similar mechanisms of disambiguation via pragmatic reasoning. We show that quantifier 'all' has different interpretations with verb vs. quantifier negation: emphatic in the former case and quantificational in the latter. Contextual markers for each reading are consistent with this difference. V-negation occurs with demonstrative pronouns, negatively connoted nouns and temporal modifiers, which add emphasis (*I don't want to talk to all these idiots*; *I haven't eaten all day*), while Q-negation occurs in the context of quantitative verbs that consolidate the interpretation of quantity (*I haven't listed all the options*).

Certain pragmatically plausible readings are lexicalized in patterns, similar in the two languages and reflecting common background knowledge; e. g. *ne spat' vsju noch'* and *not to sleep all night* both mean 'not to sleep at all during the night'.

In both languages, Q-negation is more frequent than V-negation because of its semantic and pragmatic non-markedness. Q-negation is the default interpretation option which is changed to V-negation in the presence of V-negation markers. Due to syntax, in English its share is much higher than in Russian. Finally, we show that language speakers are able to infer intended scope readings in written language.

SEMANTIC TYPES OF IMPLICATURES AND THEIR CONTEXTUAL TRIGGERS (BASED ON THE CORPUS OF NEWS HEADLINES)

Apresyan V. Ju. (valentina.apresjan@gmail.com, vapresyan@hse.ru), National Research University "Higher School of Economics", Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia, **Orlov A. V.** (alexander.orlov98@gmail.com), National Research University "Higher School of Economics", Moscow, Russia

LEARNING MULTI-PARTY DISCOURSE STRUCTURE USING WEAK SUPERVISION

Badene S. (sonia.badene@irit.fr), **Thompson K.** (catherine.thompson@irit.fr), **Lorré J-P.** (jplorre@linagora.com), **Asher N.** (asher@irit.fr)

Discourse structures provide a way to extract deep semantic information from text, e.g., about relations conveying causal and temporal information and topical organization, which can be gainfully employed in NLP tasks such as summarization, document classification, sentiment analysis. But the task of automatically learning discourse structures is difficult: the relations that make up the structures are very sparse relative to the number of possible semantic connections that could be made between any two segments within a text; furthermore, the existence of a relation between two segments depends not only on "local" features of the segments, but also on "global" contextual information, including which relations have already been instantiated in the text and where. It is natural to try to leverage the power of deep learning methods to learn the complex

representations discourse structures require. However, deep learning methods demand a large amount of labeled data, which becomes prohibitively expensive in the case of expertly-annotated discourse corpora. One recent advance in the resolution of this "training data bottleneck", data programming, allows for the implementation of expert knowledge in weak supervision system for data labeling. In this article, we present the results of our application of the data programming paradigm to the problem of discourse structure learning for multi-party dialogues.

DISCURSIVE WORDS IN CORPUS DIMENSION: *ODNIM SLOVOM* IN THE WORKS OF DOSTOEVSKY AND HIS CONTEMPORARIES

Baranov A. N. (baranov_anatoly@hotmail.com), **Dobrovol'skij D. O.** (dobrovolskij@gmail.com), Russian Language Institute of the RAS, Moscow, Russia

The starting point of the present paper is the hypothesis that discursive words characterize the individual style of the author. The subject of the study is the fixed expression *odnim slovom* 'in one word' in the works of Dostoevsky, Tolstoy, Saltykov-Shchedrin, Turgenev and Goncharov. The analysis of representative corpora yields the conclusion that Dostoevsky and Saltykov-Shchedrin differ from other contemporaries both in the frequency of using this expression in the discursive function and in the variety of its semantics.

Particularly interesting in this respect is Dostoevsky, whose prose presents all discursive functions of the expression *odnim slovom*:

- interpretation (including interpretation proper, conclusion, and clarification/explanation),
- introducing a new idea,
- regulatory uses of *odnim slovom*, such as interruption of discourse, marking difficulties in choosing a nomination,
- marking the change of nomination, as well as
- the introduction of someone else's speech.

As for the non-discursive uses of the expression *odnim slovom*, they are distributed more or less evenly among the authors under consideration.

LANGUAGE MODEL EMBEDDINGS IMPROVE SENTIMENT ANALYSIS IN RUSSIAN

Baymurzina D. R. (dilyara.rimovna@gmail.com), **Kuznetsov D. P.** (kuznetsov.den.p@gmail.com), **Burtsev M. S.** (burtsev.m@gmail.com), Neural Networks and Deep Learning Lab, Moscow Institute of Physics and Technology, Moscow, Russia

Sentiment analysis is one of the most popular natural language processing tasks. In this paper we introduce pre-trained Russian language models which are used to extract embeddings (ELMo) to improve accuracy for classification of short conversational texts. The first language model was trained on Russian Twitter dataset containing 102 million sentences, while two others were trained on 57.5 million sentences of Russian News and 23.9 million sentences of Russian Wikipedia articles. Although classifiers trained on top of language models perform better than in the case of utilizing of fastText embeddings of the same language style, we show that domain of language model also has a significant impact on accuracy. This paper establishes state-of-the-art results for RuSentiment dataset improving weighted F1-score from 72.8 to 78.5. All our models are available online as well as the source code which allows everyone to apply them or fine-tune on domain-specific data.

BERT FINETUNING AND GRAPH MODELING FOR GAPPING RESOLUTION

Belkin I. (ilya.belkin-trade@yandex.ru), Moscow Institute of Physics and Technology, Moscow, Russia

This paper reports our participation in the Automatic Gapping Resolution for Russian shared task (AGRR-2019) within Dialogue Evaluation 2019. Our team took the first place among other nine teams in all subtasks which includes gapping presence-absence classification, gap resolution and full annotation.

The phenomenon of gapping is well theoretically studied. However, the problem of automatic gapping resolution is new and there is no baseline for it. We found it possible to bring this

task into sentence classification and token tagging problems and solve them using recent advances in Natural Language Processing and deep learning. Training large language models with millions of parameters on small data became possible with the development of transfer learning methods. Using pretrained models for computer vision problems is straightforward and since BERT language model was realized it became possible to benefit from transfer learning in NLP. Our solution is heavily based on BERT, but we found that parsing gapping constructions, which are very structured, benefit from special postprocessing which includes modeling a gapping in the form of a directed graph. Our solution may be considered as the first public baseline for the task of automatic gapping resolution which is based on NLP modern practices.

PRAGMATIC MARKERS ANNOTATION IN RUSSIAN SPEECH CORPUS: RESEARCH PROBLEM, APPROACHES AND RESULTS

Bogdanova-Beglarian N. V. (n.bogdanova@spbu.ru), **Blinova O. V.** (o.blinova@spbu.ru), **Martynenko G. Ya.** (g.martynenko@spbu.ru), **Sherstinova T. Yu.** (t.sherstinova@spbu.ru), **Zaides K. D.** (kristina.zaides@student.spbu.ru), **Popova T. I.** (tipopova13@gmail.com), Philological Faculty of St. Petersburg State University, St. Petersburg, Russia

The article describes the experience of pragmatic markers (PM) annotation in two Russian speech corpora: "One Speaker's Day" (ORD; dialogues) and "Balanced Annotated Textotec" (SAT; monologues). To prepare an optimal PM annotation scheme, 4 pilot annotations were conducted on samples from ORD and SAT. It made it possible to form the final list of PM: 450 units, representing variants of 53 basic structural types. Processing the results of the pilot annotation allowed to obtain preliminary data on frequency of individual pragmatic markers and their types, as well as on the dependence of PM usage on sex and the level of speech competence of the speaker. As a result of statistical data processing, frequency lists of both PMs and their functions were obtained. The most commonly used in the dialogue are the PM *вот*, which is usually used as a «boundary marker» (G), and the PM *там*, which is usually used as a hesitative and/or rhythm-forming marker. In the monologue, the upper zone of the frequency list of the PMs is also full of boundary markers (G), marking the beginning/end of the monologue or serving as navigators in the text (*вот/ну вот, значит, так*). The most frequent types of PMs in dialogue are: X (hesitative markers), M (meta-communicative marker), GX (boundary/hesitative marker), K (xeno-indicator marker that introduces someone's speech), RX (rhythm-forming/hesitative marker). In the list of the most frequent types of PMs in monologue speech, the markers of the type GX (boundary/hesitative marker) and X (hesitative marker) are in the lead. The analysis of the frequency lists of PMs showed that we can talk about statistically significant differences in the use of PMs in dialogue and monologue.

KNOWLEDGE-BASED APPROACH TO WINOGRAD SCHEMA CHALLENGE

Boguslavsky I. M. (bogus@iitp.ru)[1,2], **Frolova T. I.** (tfrolova@gmail.com)[1], **Iomdin L. L.** (iomdin@gmail.com)[1], **Lazursky A. V.** (lazursky@mail.ru)[1], **Rygaev I. P.** (irygaev@gmail.com)[1], **Timoshenko S. P.** (nyrestein@gmail.com)[1]; [1]A. A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia; [2]Universidad Politécnica de Madrid, Madrid, Spain

We propose a method to resolve anaphoric pronouns in the framework of Winograd Schema Challenge (WSC) by means of SemETAP—a knowledge-based semantic analyzer. WSC is a modern version of the famous Turing test. Its objective is to check a machine's ability to exhibit intelligent behavior indistinguishable from that of a human. In contrast to other approaches to WSC, which are based on machine learning, our method uses explicit knowledge. An important advantage of this approach is that it gives an opportunity to provide an explanation of the result understandable for humans. SemETAP interprets the text using both linguistic and extralinguistic (background) knowledge. The former is stored in the grammar and the dictionary of the ETAP-4 system, and the latter is provided by the SemETAP ontology, inference rules and the repository of individuals. We show how this knowledge is used for resolving WSC. At the moment, the performance of the algorithm is not high—54%. This is due to the incompleteness of the background knowledge supplied to the system. It is shown, however, that if the background knowledge is complete and accurate enough, the WSC test is resolved well and it is easily understandable why the system arrived at a particular conclusion.

COMPARING MODELS OF MORPHEME ANALYSIS FOR RUSSIAN WORDS BASED ON MACHINE LEARNING

Bolshakova E. I. (eibolshakova@gmail.com), Moscow State Lomonosov University; National Research University Higher School of Economics, Moscow, Russia,
Sapin A. S. (alesapin@gmail.com), Moscow State Lomonosov University, Moscow, Russia

The paper reports on the experimental comparison of several machine learning models proposed in recent years for automatic morpheme segmentation of Russian words, including conditional random fields (CRF), sequence-to-sequence neural network (Seq2seq), convolutional neural network (CNN) model, as well as a new model we have developed with the aid of gradient boosted decision trees (GBDT). For more complete research, in our experiments we have also evaluated the semi-supervised method of Morfessor. All the morpheme analysis models being compared are briefly described in the paper, some of them perform only segmentation of words into morphs, the other produce segmentation with classification of resulted morphs. Since for Russian language linguistics rules for splitting words into morphs (and also the classification of some morphs) may differ, the experiments were performed for two data sets differing in labeling, which are obtained respectively from CrossLexica's dictionary and Tikhonov's dictionary. The experimental evaluation has shown that two best models of morpheme segmentation with classification, namely GBDT and CNN models have comparable quality, giving about 86–94% of word-level accuracy.

MULTILINGUAL PARALLEL CORPORA AS A SOURCE FOR QUANTITATIVE CROSS-LINGUISTIC GRAMMAR RESEARCH (THE CASE OF VOICE CONSTRUCTIONS)

Bonch-Osmolovskaya A. A. (abonch@gmail.com), **Nesterenko L. V.** (lnesterenko@hse.ru), National Research University "Higher School of Economics", Moscow, Russia

Multilingual parallel corpora make possible the application of quantitative methods in cross-linguistic research. Due to the lack of appropriate resources, this has not become a widespread technique among linguists, but the studies based on this idea tend to emerge. In our work, we focus on the application of logistic regression for the research of passive voice constructions with an overtly expressed agent. The study is conducted on the data extracted from a multilingual parallel corpus that was created for this purpose. The issue we find noteworthy about voice alternation is the motivation for choosing active instead of passive, i.e. when a person would say 'This essay was written by Mary' instead of 'Mary wrote this essay'. Relying on theoretical studies, we selected a bunch of features claimed to be important for this kind of choice and used them for training logistic regression models. As a result, based on the model coefficients we can detect which features appear to be passive triggers.

REFERENTIAL CHOICE IN MULTIMODAL COMMUNICATION

Budennaya E. V. (jane.sdrv@gmail.com), Institute of Linguistics, RAS, Moscow, Russia

This article deals with an application of referential markup to a large multimodal resource "Russian Pear Chats and Stories", annotated for vocal, oculomotor, manual and cephalic channels. Despite a large number of works on referential choice, it has never been investigated within the framework of multimodal communication. For this purpose, a special annotation scheme in the ELAN environment is proposed, allowing one to annotate different types of referential units and to conduct a simultaneous tracking of referential expressions (full NPs, pronouns, demonstratives, zeroes, etc) with accompanying verbal and non-verbal units. The analysis of three recordings (overall duration equals to 141 minute), where the new referential annotation was introduced in addition to the existing multimodal markup, reveals a range of understudied peculiarities of the referential choice. It was found that the role of the Commentator in the conversation entails a significantly larger amount of constructions with a zero subject pronoun, compared to the monologue discourse of the Narrator and the Reteller. The analysis of referential expressions and accompanying pointing gestures complied with more general data previously obtained on the English material and showed that nouns are significantly more often accompanied by a pointing stroke than personal pronouns, while demonstratives occupy an intermediate position between nouns and personal pronouns as units potentially accompanied by a gesture.

APPLYING AN AUTOMATIC FTD CLASSIFIER TO THE ANNOTATION OF THE GICR CORPUS

Bulygin M. V. (bulyginmv1996@gmail.com), Radboud University, Nijmegen, Netherlands,
Sharoff S. A. (s.sharoff@leeds.ac.uk), Russian State University of Humanities, Moscow, Russia; Leeds University, Leeds, UK

This paper addresses the task of automatic genre classification for Russian within the Functional Text Dimensions (FTD) framework. Our aim in this study was to build the optimate FTD classification model to annotate web texts from the GICR corpus. For training data, we used an extended GICR dataset. We used the Support Vector Machine method with linear kernel for classification and converted training data to lower case to increase accuracy. During our research we experimented with several classification parameters, such as types of features, C-value and feature filtering to determine the best option for the classification model of the GICR dataset. The resulting model was able to achieve satisfactory classification accuracy and was used for GICR annotation. We also looked at the most significant features for each FTD in our best performing model and compared them to the most frequent words in which these features occur. Finally, we applied our model to segments of the GICR and looked at the FTD components in these segments.

A SIMPLE FINGERPRINT APPROACH TO EXTRACTING THE GLOBAL PROSODIC PROPERTIES FROM FIELD DATA

Chechuro I. Yu. (ilyachechuro@gmail.com)[1], **Lyashevskaya O. N.** (olesar@yandex.ru)[1, 2]; [1]National Research University Higher School of Economics, Moscow, Russia; [2]Vinogradov Institute of the Russian Language RAS, Moscow, Russia

The paper reports a method to create a speaker's prosodic fingerprint based on the global characteristics of the pitch movement. Prosodic fingerprint is the distribution of f0 in the low, middle, and high ranges and the distribution of pitch movements from one range into other [Šimko et al. 2017]. This fully automated method can be used to classify the records and to provide the reference level for more sophisticated analysis of the pitch movement and intonation strategies. We evaluate the method by applying it to the spontaneous Russian spoken data recorded in different regions. We model the correlation between the fingerprint and sociolinguistic features such as age, gender, and region. The results of this analysis allow to formulate several sociolinguistic hypotheses that can further be tested with a more detailed analytic technique.

CLASSIFICATION MODELS FOR RST DISCOURSE PARSING OF TEXTS IN RUSSIAN

Chistova E. V. (chistova@isa.ru)FRC CSC RAS, Moscow, Russia;
RUDN University, Moscow, Russia, **Shelmanov A. O.** (shelmanov@isa.ru), Skoltech, Moscow, Russia, FRC CSC RAS, Moscow, Russia, **Kobozeva M. V.** (kobozeva@isa.ru), **Pisarevskaya D. B.** (dinabpr@gmail.com), **Smirnov I. V.** (ivs@isa.ru), FRC CSC RAS, Moscow, Russia, **Toldova S. Yu.** (toldova@yandex.ru), NRU Higher School of Economics, Moscow, Russia

The paper considers the task of automatic discourse parsing of texts in Russian. Discourse parsing is a well-known approach to capturing text semantics across boundaries of single sentences. Discourse annotation was found to be useful for various tasks including summarization, sentiment analysis, question-answering. Recently, the release of manually annotated Ru-RSTreebank corpus unlocked the possibility of leveraging supervised machine learning techniques for creating such parsers for Russian language. The corpus provides the discourse annotation in a widely adopted formalisation—Rhetorical Structure Theory. In this work, we develop feature sets for rhetorical relation classification in Russian-language texts, investigate importance of various types of features, and report results of the first experimental evaluation of machine learning models trained on Ru-RSTreebank corpus. We consider various machine learning methods including gradient boosting, neural network, and ensembling of several models by soft voting.

SIMULATION OF BACKGROUND KNOWLEDGE AND BRIDGING IN RUSSIAN

Dikonov V. G. (dikonov@iitp.ru), IITP RAS, Moscow, Russia

This paper introduces a knowledge-based semantic approach towards bridging annotation of Russian texts. Our method simulates human background knowledge by using compact domain descriptions based on an extended version of SUMO ontology and lexical-semantic data from the "Universal Dictionary of Concepts". Our approach supports a wide and extensible range of bridging relations. The tagger that implements it can build complex bridges with multiple arcs, supports making assumptions and can be adapted to annotate other languages supported by the underlying dictionary of concepts.

AN APPROACH TO CUSTOMIZATION OF PRE-TRAINED NEURAL NETWORK LANGUAGE MODEL TO SPECIFIC DOMAIN

Dudarin P. V. (p.dudarin@ulstu.ru), **Tronin V. G.** (v.tronin@ulstu.ru),
Svyatov K. V. (k.svyatov@ulstu.ru), Ulyanovsk State Technical University, Ulyanovsk, Russia

Nowadays the majority of tasks in NLP field are solved by means of neural network language models. These models already have shown state-of-the-art results in classification, translation, named entity recognition and so on. Pre-trained models are accessible in the internet, but the real life problem's domain could differ from the origin domain which the network was learned. In this paper an approach to vocabulary expansion for neural network language model by means of hierarchical clustering is presented.This technique allows to adopt pre-trained language model to a different domain. In the experimental part the proposed approach is demonstrated on specific domain of textual artifacts of software development process. This field is actively studied this days due the expensiveness of the process and its impact on the modern world and society.

GAPPING PARSING USING PRETRAINED EMBEDDINGS, ATTENTION MECHANISM AND NCRF

Emelyanov A. A. (login-const@mail.ru), **Artemova E. L.** (echernyak@hse.ru),
Moscow Institute of Physics and Technology; Nationa Research University Higher School of Economics, Moscow, Russia

The article is devoted to the problem of automatic gapping resolution for the Russian language. We use BERT Language Model as embeddings with bidirectional recurrent network, attention, and NCRF on the top. Unlike other models these are using BERT, we apply BERT only as embedder without any fine-tuning. As a result, our implementation took second place in the AGRR-2019 competition.

TRACING CULTURAL DIACHRONIC SEMANTIC SHIFTS IN RUSSIAN USING WORD EMBEDDINGS: TEST SETS AND BASELINES

Fomin V. (wadimiusz@gmail.com), **Bakshandaeva D.** (dbakshandaeva@gmail.com),
Rodina Ju. (julia.rodina97@gmail.com), National Research University Higher School of Economics, Moscow, Russia, **Kutuzov A.** (andreku@ifi.uio.no), University of Oslo, Oslo, Norway

The paper introduces manually annotated test sets for the task of tracing diachronic (temporal) semantic shifts in Russian. The two test sets are complementary in that the first one covers comparatively strong semantic changes occurring to nouns and adjectives from pre-Soviet to Soviet times, while the second one covers comparatively subtle socially and culturally determined shifts occurring in years from 2000 to 2014. Additionally, the second test set offers more granular classification of shifts degree, but is limited to only adjectives.

The introduction of the test sets allowed us to evaluate several well-established algorithms of semantic shifts detection (posing this as a classification problem), most of which have never been tested on Russian material. All of these algorithms use distributional word embedding models trained on the corresponding in-domain corpora. The resulting scores provide solid comparison baselines for future studies tackling similar tasks. We publish the datasets, code and the trained models in order to facilitate further research in automatically detecting temporal semantic shifts for Russian words, with time periods of different granularities.

IMPORTANCE OF COPYING MECHANISM FOR NEWS HEADLINE GENERATION

Gusev I. O. (ilya.gusev@phystech.edu), MIPT, Moscow, Russia

News headline generation is an essential problem of text summarization because it is constrained, well-defined, and is still hard to solve. Models with a limited vocabulary can not solve it well, as new named entities can appear regularly in the news and these entities often should be in the headline. News articles in morphologically rich languages such as Russian require model modifications due to a large number of possible word forms. This study aims to validate that models with a possibility of copying words from the original article performs better than models without such an option. The proposed model achieves a mean ROUGE score of 23 on the provided test dataset, which is 8 points greater than the result of a similar model without a copying mechanism. Moreover, the resulting model performs better than any known model on the new dataset of Russian news.

ANNOTATION OF PARALLEL TEXTS: THE CONCEPT OF DIVERGENT TRANSLATION

Inkova Olga Yu. (Olga.Inkova@unige.ch), Institute of Informatics Problems, FRC CSC RAS, Moscow, Russia; University of Geneva, Geneva, Switzerland

The annotation of parallel corpora, as well as building of supracorpora databases, challenges linguists with the question of how to define a functional equivalent of the linguistic units that serve as an object of a given study. The paper discusses the concept of divergent translation and whether it is theoretically important for the analysis of logical-semantic relations (LSR). It is shown that relations between states of things can be expressed not only by connectives but also by lexical means (referred to as "alternative lexicalizations" in the works of the Penn Discourse Treebank group) and grammatical tools (syntactic constructions and morphological forms), and by marks of punctuation. While the two latter ways are mentioned in grammars, they are usually not taken into account when the alternative ways of tagging LSR are described, nor are they annotated in corpora or databases. The supracorpora database of connectives, built on the basis of the French and Italian parallel subcorpora of the Russian National Corpus, introduces new functional capabilities. It stores a representative array of annotations tagged as "divergent translation" (more than 1,250, i. e. 7.7 per cent of the total number), which allows users to collect various statistical data. With these data, one could establish: (1) which LSR tend to be expressed by alternative means and how often they occur compared to connectives, (2) what these alternative means are, (3) which divergent translations may be used to render a given marker of LSR and how often each of them is used, (4) which alternative markers of LSR are specifically employed to convey one or another relation and which of them are able to express several LSR. The conclusive part of the paper suggests that, for the analysis of divergent equivalents, it is central that one and the same alternative means is used by different translators when translating one and the same textual fragment into one and the same language as well as into several languages, which speaks for its productivity. The further development of multi-language and polyvariant parallel corpora and databases would let us find out to what extent the means conveying LSR differ in various languages.

AN ANAPHORA RESOLUTION SYSTEM FOR RUSSIAN BASED ON ETAP-4 LINGUISTIC PROCESSOR

Inshakova E. S. (e.s.inshakova@gmail.com), Laboratory of Computational Linguistics, A. A. Kharkevich Institute for Information Transmission Problems, Moscow, Russia

The paper presents a rule-based system of automated anaphora resolution for Russian. The system is based on the resources of ETAP-4 linguistic processor: the Russian combinatorial dictionary (RCD), the ETAP parser, and the ontology OntoEtap. In this paper, I describe the ordered algorithms for resolution of different pronouns and provide the results of their evaluation.

ADDING TO THE TREASURY OF RUSSIAN MICROSYNTACTIC CURIOSITIES: TWO ANTONYMIC SYNTACTIC IDIOMS WITH COMPARATIVES

Iomdin L. L. (iomdin@iitp.ru), Institute for Information Transmission Problems (Kharkevich Institute), Russian Academy of Sciences, Moscow, Russia

The paper continues a series of research studies into the microsyntax of Russian. Two constructions that are sufficiently close to each other in syntactic structure and semantics are considered in detail: these are linguistic units of the type *kak možno lučše* ≈ 'in the best way possible' and *kak nelzja lučše* '≈ 'it can never be better'. In both constructions, the first two elements are determined lexically while the third one is fixed grammatically since it can be instantiated by (almost) any comparative form. It is demonstrated that the two units possess substantial semantic differences; in particular, the former unit is oriented prospectively (cf. *sygraj kak možno lučše* 'play as well as you possibly can' but hardly ?*sygral kak možno lučše* ≈ 'he has played as well as he possibly could') while the latter unit is, rather, oriented respectively (cf. *vse složilos' kak nel'zja lučše* ≈ 'everything turned out in a way that could never be better' but hardly ?*Reši etu zadaču kak nel'zja lučše, čtoby sdat' ekzamen* ≈ 'solve this problem in a way that could never be better, to pass the exam'. The material under consideration is also used to discuss certain general subtleties of the Russian comparative.

THE CORPUS OF CONTACT-INFLUENCED RUSSIAN OF NORTHERN SIBERIA AND THE RUSSIAN FAR EAST

Khomchenkova I. A. (irina.khomchenkova@yandex.ru), Lomonosov Moscow State University; Vinogradov Russian Language Institute & Institute of Linguistics, RAS; Moscow, Russia, **Pleshak P. S.** (polinapleshak@yandex.ru), Lomonosov Moscow State University; Institute of Linguistics, RAS; Moscow, Russia, **Stoynova N. M.** (stoynova@yandex.ru), Vinogradov Russian Language Institute & Institute of Linguistics, RAS; NRU HSE; Moscow, Russia

The paper presents a spoken corpus of contact-influenced Russian, which consists of oral spontaneous Russian speech of bilingual speakers of indigenous languages of Northern Siberia and the Russian Far East (Samoyedic, Tungusic, Chukotko-Kamchatkan). The texts included in the corpus were transcribed in ELAN in Standard Russian orthography and provided with a special system of manual annotation of contact-induced features developed for the corpus. The paper focuses mainly on this system of annotation, which is relevant in a wider context of annotating any kind of speech with "deviations" from the standard language variety (bilinguals', learners', dialectal speech etc.). The annotation tags are grouped in several separate levels: contact-induced morphological, syntactic, phonetic, lexical features etc. The exact meanings for the annotation tags were proposed on empirical grounds. Transcribed and annotated texts gain morphological annotation and search implementation based on the Tsakorpus platform. The aim of the project is to provide a useful resource for linguistic studies on language contact.

UNIFIED MULTICHANNEL ANNOTATION: A TOOL FOR ANALYSING NATURAL COMMUNICATION

Kibrik A. A. (aakibrik@gmail.com), Institute of Linguistics RAS, Lomonosov Moscow State University, Moscow, Russian Federation, **Korotaev N. A.** (n_korotaev@hotmail.com), RSUH, Institute of Linguistics RAS, Moscow, Russian Federation, **Fedorova O. V.** (olga.fedorova@msu.ru), Lomonosov Moscow State University, Institute of Linguistics RAS, Moscow, Russian Federation, **Evdokimova A. A.** (arochka@gmail.com), Institute of Linguistics RAS, Moscow, Russian Federation

This paper contributes to the research field of multichannel discourse analysis. Multichannel discourse analysis explores numerous channels involved in natural communication, such as verbal structure, prosody, manual gesticulation, head movements, eye gaze, torso postures, etc., and treats them as parts of an integrated process. For the purposes of investigating the way participants interact with one another and the way different communication channel correlate, we introduce the notion of an integrated multichannel annotation created with ELAN software. In particular, we consider three topics: (1) temporal alignment between participants' speech and manual gesticulation; (2) distribution of participants' visual attention as they watch their interlocutors talking and gesticulating manually; (3) interrelationship between participants' torso postures and head movements.

EVOLUTION OF DIALECTAL UNSTRESSED VOWELS SYSTEM IN MOSCOW: 4 GENERATIONS

Knyazev S. V. (svknia@gmail.com)[1,2], **Malykhina P. A.** (malyhinapolina@rambler.ru)[2];
[1]National Research University Higher School of Economics;
[2]Moscow State Lomonosov University, Москва, Россия

The paper deals with evolution of one part of dialectal phonetic system (neutralization of non-high unstressed vowels' in different allophones as a function of stressed vowel's length or/and quality) over the course of three generations of speakers from one family, moved from a village to Moscow, Russian capital city. We discuss some methods of phonetic analysis that could be utilized in order to present sound changes observed and argue that the result obtained from a large data volume could be not so informative as compared to those, achieved from thorough analysis of every token. Our results show that the phonetic system starts to change immediately after the resettlement of a family: in the first generation of a family moved. The second and third generation displays yet more dramatic changes with only few markers of previous dialectal peculiarities remaining; along with this, the qualitative dissimilation survives somewhat longer than the quantitative one.

INTROSPECTIVE AND PERCEPTUAL LABELING OF PROSODIC PHRASING (A COMPARATIVE ANALYSIS ON THE MATERIAL OF R. I. AVANESOV TEXTS COLLECTION)

Krivnova O. F. (okrivnova@mail.ru), **Smirnova O. S.** (kisaolga@mail.ru),
Lomonosov Moscow State University, Moscow, Russia

This paper discusses the problems and results of a comparative analysis of two fundamentally different types of prosodic phrasing labeling realized for some literary Russian texts. The introduction examines the theoretical basis of the study and formulates specific tasks, the solution of which was necessary for comparative analysis and the achievement of the final goal of the study. The first section of the paper describes the experimental material, methods of research and the basic principles of experimental data processing. In the second, central section of the work, a detailed description of the parameters of comparative analysis of introspective labeling and perceptual one is given. The following parameters were taken into account in the comparative analysis: the general distribution of frequency of occurrence of text spaces with different indexes of word boundary strength; their contextual distribution with respective frequency data; relationship of prosodic breaks' strength with pauses. This section also contains many illustrations that demonstrate the main results of the comparative analysis of the target prosodic labeling of the experimental text material. Section 3 analyzes the relationship between the prosodic breaks' strength and pauses' duration in both types of labeling analyzed. In conclusion results of the study are summarized and promising areas for further research on the relevant topics are noted.

ADAPTATION OF DEEP BIDIRECTIONAL MULTILINGUAL TRANSFORMERS FOR RUSSIAN LANGUAGE

Kuratov Yu., **Arkhipov M.,** Neural Networks and Deep Learning Lab, Moscow Institute of Physics and Technology

CONCEPTUALIZATION OF NON-FULLY CONTROLLED SITUATIONS: VERBS AND PRONOUNS

Kustova G. I. (galinak03@gmail.com), Vinogradov Russian Language Institute of the Russian Academy of Sciences; Moscow, Russia

The paper introduces the opposition "level of the situation" vs. "level of the story". Within this opposition, features of the verbs denoting non-fully controlled situations are considered (*to succeed* vs. *to happen*): government (infinitive vs. clause), combinability with negation and propositional pronouns. Propositional pronouns *tak* ('so') and *eto* ('it') and the matrix verbs which they are combined with, imply a different conceptualization of the antecedent situation: *My proigrali. Tak poluchilos'* ('We lost. So it turned out') vs. *My hoteli pobedit', i nam eto udalos'* ('We wanted to win, and we succeeded'). *Tak* is semantically related to the mode of action and in other meanings implies a variable factor or aspect.

LEXICAL ANALYSIS OF THE RUSSIAN LANGUAGE TEXTBOOKS FOR PRIMARY SCHOOL: CORPUS STUDY

Laposhina A. N. (antonina.laposhina@gmail.com),
Veselovskaya T. S. (TSVeselovskaya@pushkin.institute),
Lebedeva M. U. (m.u.lebedeva@gmail.com), **Kupreshchenko O. F.** (ofkupr@gmail.com),
Pushkin State Russian Language Institute (Moscow, Russia)

Annotation: This paper presents the first results of a comparative corpus-based research of the modern Russian language textbooks for primary school children. Volume and diversity statistics of textbooks' vocabulary, the results of the vocabulary's analysis included in frequency and thematic groups are given.

SENTENCE LEVEL REPRESENTATION AND LANGUAGE MODELS IN THE TASK OF COREFERENCE RESOLUTION FOR RUSSIAN

Le T. A. (anhlt@vimaru.edu.vn)[1,2], **Petrov M. A.** (maksimallist@gmail.com)[1], **Kuratov Y. M.** (yurakuratov@gmail.com)[1], **Burtsev M. S.** (burtcev.ms@mipt.ru)[1]; [1]Neural Networks and Deep Learning Lab—Moscow Institute of Physics and Technology, Moscow, Russia;
[2]Faculty of Information Technology—Vietnam Maritime University, Hai Phong, Viet Nam

Coreference Resolution (CR) is one of the most difficult tasks in the field of Natural Language Processing due to the lack of deeply and comprehensively understanding the semantic meaning of the mention in not only the sentence-level context but also the entire document-level context. To the best of our knowledge, the previous proposed models often address the coreference resolution task in two steps: 1) detect all possible mention candidates, 2) score and cluster them into chains. We instead propose a new approach which reforms the coreference resolution task to the task of learning sentence-level coreferential relations. Additionally, by leveraging the power of state-of-the-art language representation models such as BERT, ELMo, it was possible to achieve cutting edge results on Russian datasets.

RELAXING COOCCURRENCE RESTRICTIONS: THE DISTRIBUTION OF THE RUSSIAN PARTICLE *-KA*

Levontina I. B. (irina.levontina@mail.ru), RLI RAS, Moscow, Russia

The development of corpus linguistics quite often makes it necessary to revisit the items studied and comprehensively described in the "pre-corpus" epoch. As a result we obtain a more voluminous or even radically different picture of their functioning. This is especially true of linguistic units with bizarre compatibility, in a complex way motivated by their semantics, such as the Russian particle *-ka*. It is a study of a large array of linguistic data that makes it possible to notice relatively rare, but regularly arising types of combinations that reveal the semantic potential of this particle. In the present work, we used the Russian National Corpus, as well as Yandex search, which allowed us to assess if this or that type of combination is relevant for nowaday live speech. The study of corpus data not only contributes to our understanding of the properties of linguistic units — in this case, the distribution of a particle, but also makes it possible to observe the linguistic mechanisms involved in relaxing cooccurrence restrictions. Thus, the analysis of the corpus material allowed us to find two fairly common, but very nontrivial types of combinations of *-ка* with non-imperative expressions: *лучше-ка* and *знаешь-ка/знаете-ка*. As we show, their occurrence is due to the effect of completely different linguistic mechanisms.

DRIVING US CRAZY WITH YOUR INFINITIVES! THE RISE OF A NEW CAUSATIVE CONSTRUCTION IN RUSSIAN

Levontina I. B. (irina.levontina@mail.ru), RLI RAS, Moscow, Russia,
Polinsky M. S. (mpolinsk@gmail.com), National Research University Higher School of Economics, Moscow, Russia; Professor: Faculty of Humanities / School of Linguistics

Russian has an impressive set of psych-verbs with the general meaning of causing extreme irritation and exhausting one's patience, which we will henceforth refer to as EXASPERATE-verbs: *достать*; *задолбать, заколебать, замучить, бесить*, etc. With these predicates, the experiencer is in the accusative, and the non-salient, inanimate or abstract causer of irritation can be

expressed by a noun phrase in the nominative, or by an infinitival clause, e. g., *Меня достало это выражение/разбирать эти выражения*. In addition, these verbs participate in another causative construction, with a salient, agentive causer expressed by a noun phrase in the nominative case, and the manner in which irritation is brought about expressed by the instrumental phrase, with or without a preposition: *Ты меня достал* (*с*) этими *выражениями*. In modern spoken Russian, we also find a new agentive causative construction (NACC): *Ты меня достал ныть!* 'You drive me up the wall by your whining.' The NACC is colloquial and is largely used by younger speakers. Among the verbs that participate in the NACC are vulgar lexical items, which further adds to its colloquial nature. (The use of vulgar expressions to vent frustration is attested cross-linguistically, so Russian is not exceptional in that regard.) We provide a detailed analysis of the syntax of the NACC and argue that it instantiates obligatory adjunct control by the subject. We hypothesize that the rise of the NACC is driven by the analogy with the existing constructions with EXASPERATE-verbs in standard Russian, and we address several other factors that contribute to the development of the new construction.

AUTOMATIC VOCABULARY POSITIONING IN A THESAURUS

Likhonosov A. (andrew.likhonosov@abbyy.com), **Indenbom E.** (Eugene_I@abbyy.com), **Yudina M.** (maria_yu@abbyy.com), ABBYY, Moscow, Russia

Thesauri are one of the most widely used resources in natural language processing. At the same time, many of them are built manually, which takes a lot of time and, due to human errors, can affect their quality and completeness. We propose a procedure for automatic positioning of vocabulary in the ABBYY Compreno thesaurus using large monolingual corpora, a regular bilingual dictionary and a subset of already positioned words.

ANALYSIS OF PROSODIC FEATURES OF THE EMOTIONAL INTONATION USING "INTONTRAINER" SYSTEM (ON THE EXAMPLE OF RUSSIAN PHRASES)

Lobanov B. M. (Lobanov@newman.bas-net.by), **Zhitko V. A.** (zhitko.vladimir@gmail.com), United Institute of Informatics Problems NAS Belarus, Minsk, Belarus

The main results of the update of the IntonTrainer system for the purposes of analyzing and studying the prosodic signs of emotional intonation are described. A distinctive functional feature of the updated system is the creation of an expanded set of prosodic signs of emotional intonation. The paper presents preliminary assessments of their effectiveness using the created experimental database of emotional phrases of Russian speech.

A REUSABLE TAGSET FOR THE MORPHOLOGICALLY RICH LANGUAGE IN CHANGE: A CASE OF MIDDLE RUSSIAN

Lyashevskaya O. N. (olesar@yandex.ru), National Research University Higher School of Economics; Vinogradov Institute of the Russian Language RAS, Moscow, Russia

The paper discusses the standardization efforts to create a morphological standard for the Middle Russian corpus, which is part of the historical collection of the Russian National Corpus (RNC). To meet the needs of different categories of corpus researchers as well as NLP developers, we consider two styles of the morphological annotation (RNC schema and Universal Dependencies schema). A number of specifications of the feature list proposed to facilitate data reusability, linking and conversion.

POSTPOSITIONAL CONSTRUCTIONS IN TATAR: METHODOLOGIES FOR MEASURING INTRALINGUAL VARIATION

Lyutikova E. A., MSU, MPSU, Pushkin State Russian Language Institute,
Gerasimova A. A., MSU, MPSU, Pushkin State Russian Language Institute

The paper addresses the issue of intralingual variation in Tatar postpositional phrases. The nominal in Tatar postpositional phrases demonstrates differential case marking: the choice between genitive and unmarked case form is determined by the morphosyntactic class of the

nominal. With postpositions derived from nouns with locative or abstract semantics variation in case assignment is accompanied by presence/absence of the ezafe marker on the postposition. In this paper we use corpus-based and experimental methods to investigate the distribution of grammatical variants and estimate the current status of the variation. We argue that the existing grammatical descriptions do not capture the current state of affairs.

We show that pronouns and nouns do not form a homogeneous class with respect to case marking in the postpositional phrase. The genitive case marking is common for 1st/2nd person personal pronouns and 3rd person singular personal pronoun. All other pronouns and nouns are primarily used in an unmarked form, an observation supported by both corpus and experimental data.

We argue that the grammaticalization of denominal postpositions is not complete. In both corpus and experimental studies, we observe a wide range of features that unite postpositional phrases with nominal embedding ezafe constructions. First, genitive case marking for the complement is acceptable for non-personal pronouns and nouns. Second, the absence of the ezafe marker is acceptable only with 1st / 2nd person personal pronouns and partially with 1st / 2nd person reflexive pronouns. Third, the case marking of the nominal and the choice of the ezafe marker for the postposition are interrelated. When the complement is genitive, speakers prefer the agreeing form of the postposition. When the complement is unmarked, the postposition shows no agreement with the possessor. This contrast reflects the opposition between ezafe-3 and ezafe-2 constructions, respectively.

Interestingly, the denominal postpositions demonstrate different degrees of grammaticalization. For instance, the postposition *turında* 'about' is mostly used with a possessive affix that shows no agreement. We suppose that the form with the non-agreeing ezafe affix is reanalyzed by the speakers as uninflected.

Another crucial observation concerns the reflexive pronoun *üz*. In both experiments 1st / 2nd person reflexive pronouns show syntactic behavior similar to the one of personal pronouns, while 3rd person singular reflexive pronoun patterns with interrogative pronouns.

As the result of the study, we compare different methodologies for investigation of the intralingual variation. We suggest that the combination of different sources of data, both corpus-based and experimental, provides the fuller description for cases of intralingual variation than a single method. The experimental methods that we used differ in sensitivity to various aspects of language phenomena: the elicited production is better in distinguishing deviation from the grammatical pattern; the acceptability judgements show to what extent a grammatical innovation is used. Remarkably, the comparison of the different sources of data allows us to determine the direction of language change and estimate the current status of the variation.

DERIVATIVE MEANINGS OF THE RUSSIAN INDEFINITE ADVERB *KAK-TO*: A CORPUS-BASED STUDY

Mikaelian I. L. (irina-mikaelian@yandex.ru), Pennsylvania State University, USA,
Zalizniak Anna A. (anna.zalizniak@gmail.com), Institute of Linguistics of the RAS; Institute of Informatics Problems of the FRC CSC RAS, Moscow, Russia

The paper analyzes derivative meanings of the Russian indefinite adverb *kak-to*, which are insufficiently described in the existing grammars and dictionaries. Besides its primary meaning of indefinite manner, cf. *grabitel' kak-to pronik v dom* 'the buglar somehow got into the house', *kak-to* has two derivative meanings. 1) It can refer to an indefinite moment in time, cf. *on kak-to mne rasskazal etu istoriju* 'he told me this story once'; 2) it can function as a discursive marker of 'general indefiniteness,' which has two varieties: a) *kak-to* can point to an underspecified aspect of a situation—'in some respect/in some mesure/kind of' (*ona kak-to stranno posmotrela na menja, on kak-to smutilsja, on kak-to po-brastki obnjal menja* 'she gave me an odd glance, he felt somewhat confused, he hugged me in a kind of brotherly way'); b) it can accentuate the idea of uncontrollability of a situation ('it happened so'): *ja kak-to upustil iz vidu* 'I somehow overlooked'. Using data form the RNC, we have identified contexts correlating with each of the meanings of *kak-to*. We have also demonstrated that its use as a discursive marker is much more frequent than its occurrences as an adverb of manner proper. We used data from Russian-English and English-Russian parallel subcorpora to demonstrate that in many instances, translators from Russian leave the discursive *kak-to* without a translation, and, vice-versa, translators into Russian frequently insert *kak-to* without a specific stimulus for it in the original English text. We conclude that usage of *kak-to* is regulated by a highly language specific discursive strategy in Russian.

AN ATTENTION-BASED APPROACH TO AUTOMATIC GAPPING RESOLUTION FOR RUSSIAN

Movsesyan A. A. (derise@iitp.ru), Institute for Information Transmission Problems (Kharkevich Institute), Russian Academy of Sciences, Moscow, Russia

Gapping is a type of ellipsis in which a finite verb is elided in a coordinate structure. Reconstruction of the elided material is essential for different NLP tasks. However, from a practical point of view, the problem did not receive considerable attention for Russian language because of lack of training data and rarity of the phenomenon itself. This paper is one of the first works of deep learning-based automatic gapping resolution in Russian as a part of AGRR-2019 competition. We used a recurrent neural network-based approach to determine presence/absence of gapping in a sentence and for the full annotation we applied a Universal Transformer neural network that combines self-attention mechanism with recurrence in depth. Also using pretrained fastText word embeddings, we achieved 85% standard F-measure on test set for binary classification task and 62% symbol-wise F-measure for full annotation task. We assume that fixed word embedding like fastText does not contain enough syntactic information to properly match remnants in sentences with gapping. Also we show that our model generalize better if punctuation marks were ignored during training and evaluation.

THE WORD *ÈTO* IN A *WH*-QUESTION: ON THE DIFFERENCES BETWEEN A PRONOUN AND A PARTICLE

Pekelis O. E. (opekelis@gmail.com), Russian State University for the Humanities, Moscow, Russia

The paper examines the grammatical and semantic features of the word *èto* when it precedes or follows a *wh*-word (cf. *Gde èto ty byl?*). In this context, *èto* is usually considered to be a particle, with the only—and not clear-cut—exception being a question with the *wh*-words *kto* and *čto*. However, the data presented below suggest that as many as four different types of *èto* used in an interrogative context have to be distinguished. It is demonstrated that these types differ in their meaning, their syntactic distribution, and their position within the "pronoun-particle" continuum.

TENSE AND LAX BODY PARTS IN THE RUSSIAN DEICTIC GESTURES: THE CASE OF INDEX FINGER POINTING

Pereverzeva S. I. (P_Sveta@hotmail.com), National Research University Higher School of Economics, Moscow, Russia

The article regards the way in which the deictic gestures with the active index finger are executed in Russian body language and focuses on the role of the tension of the index finger (slightly curved vs. extended). Using the data retrieved from the Russian Multimedia Corpus, we discover the dependency between the tension of the index finger and the tension of the arm, which is engaged in executing the deictic gestures. We also reveal correlations between the tension of the index finger and (a) the primary / secondary reference to the pointed object, (b) the closest and the farthest distance between the speaker and the pointed object. We examine the difference in meaning and usage of the deictic gestures with the slightly curved vs. extended index finger. We argue that the choice between these types of pointing may be influenced both by physical and pragmatic factors.

AN ANATOMY OF A LIE: DISCOURSE PATTERNS IN ULTIMATE DECEPTION DATASET

Pisarevskaya D. (dinabpr@gmail.com), Institute for Systems Analysis FRC CSC RAS, Moscow, Russia, **Galitsky B.** (boris.galitsky@oracle.com), Higher School of Economics, Moscow, Russia and Oracle Corp, Redwood Shores CA, USA

We propose a hypothesis that a deception in text should be visible from its discourse structure. The problem of deception detection is then formulated as classification of a discourse tree of this text, according to the Rhetorical Structure Theory. This discourse tree (DT) is extended by the speech acts expressions attached as the labels for the edges. We employ what we call an ultimate deception dataset: a set of customer complaints for English, that includes descriptions

of problems customers experienced with certain businesses. It contains about 2,400 complaints about banks and provides clear ground truth, based on available factual knowledge in the financial domain. The complaints are written by non-professional writers. We conduct experiments to explore correlation between implicit cues of the rhetorical structure of texts and how truthful/deceptive are these texts. The results show that a deception in text can be detected reliably enough to assure industrial applications. Automated detection of text with misrepresentations such as fake reviews is an important task for online reputation management.

PROSODY AND GRAMMAR OF CLAUSAL AND VP COORDINATION: THE RUSSIAN CONJUNCTION *I (AND)* VIEWED THROUGH THE PRISM OF PROSODICALLY ANNOTATED CORPUS DATA

Podlesskaya V. I. (vi_podlesskaya@il-rggu.ru), Russian State University for the Humanities, Moscow, Russia

The paper focuses on Russian constructions with clauses or VPs combined by means of the conjunction *I* 'and'. Prosodically, the construction may come up in two forms: (a) integrated, i.e.—as a single illocution with the first clause pronounced with a rising pitch that projects discourse continuation, and (b) disintegrated, i.e. as two separate illocutions with the first clause pronounced with a falling pitch that projects no continuation. Basing on the data from the Prosodically Annotated Corpus of Spoken Russian, prosody and grammar of coordinate constructions with the conjunction I 'and' were analyzed qualitatively and quantitatively. The results show that coordinated clauses and VPs are more frequent than coordinated NPs and other types of groups; in spoken narratives, coordinated clauses are more frequent than VPs, while in written narratives, coordinated VPs are more frequent than clauses; coordinated clauses and VPs more often come up as prosodically integrated than as prosodically disintegrated; the rate of integrated constructions is higher in coordinated VPs than in coordinated clauses.

A CORPUS STUDY OF SELF-REPAIRS IN RUSSIAN MONOLOGUES AND DIALOGUES

Podlesskaya V. I. (vi_podlesskaya@il-rggu.ru), **Korotaev N. A.** (n_korotaev@hotmail.com), **Mazurina S. I.** (svet.mazurina95@gmail.com), RSUH, Moscow, Russia

Self-initiated and other-initiated self-repairs (N=632) were investigated in a subcorpus (1 h 14 min) extracted from the multichannel corpus "Russian Pear Chats and Stories". The subcorpus consists of three communication sessions where participants retell and discuss the "Pear stories" film, hence each session contains both monologue and dialogue discourse parts. The overall rates of self-repairs and the distribution of their particular types were compared in monologues and dialogues. The results show that while, overall, speakers tend to repair more often in conversational than in retelling parts, particular types of repairs are distributed differently, e. g. (a) repetitions and restarts have higher rates in conversational parts, while corrections appear more often in retellings; (b) in retellings, reparandum and reparans appear more often within the same discourse unit, while in conversational parts, they tend to appear in separate discourse units.

MEASURE CLUSTERING APPROACH TO MWE EXTRACTION

Rossyaykin P. O. (petrrossyaykin@gmail.com), **Loukachevitch N. V.** (louk_nat@mail.ru), Lomonosov Moscow State University, Moscow, Russia

In this paper we present an unsupervised and resource-independent approach to the well-known task of discovery of multiword expressions (MWE) in text corpora. We experimented on extracting Russian nominal phrases (Adj-N and N-N.Gen) relevant for lexical resources (thesauri, WordNet, etc.). Our approach is based on the assumption that idiosyncrasy of MWEs can be due to different properties (morphosyntactic, semantic, pragmatic and statistical), and thus, different types of measures (statistical, context, distributional) are efficient at extracting different MWEs. We propose new context measures as well as an unsupervised method of combining measures in which we cluster vectors of ranks assigned by individual measures. The proposed method accounts for different properties of MWEs and allows surpassing both individual measures and their simple sum/product.

WORD VECTOR MODELS AS AN OBJECT OF LINGUISTIC RESEARCH

Shavrina T. O. (rybolos@gmail.com), NRU HSE, Moscow, Russia; Sberbank, Moscow, Russia

This article launches a series of studies in which popular vector word2vec models are considered not as an element of the architecture of an NLP application, but as an independent object of linguistic research. The linguist's view on the surrogate of contexts on the corpus, as which vector models can be considered, makes it possible to reveal new information about the distribution of individual semantic groups of vocabulary and new knowledge about the corpus from which these models are derived. In particular, it is shown that such layers of English and Russian vocabulary, such as the names of professions, nationalities, toponyms, personal qualities, time periods, have the greatest independence from changing the model and retain their position relative to their neighbour words—that is, they have the most stable contexts regardless of the corpus; it is shown that the vocabulary from the Swadesh list is statistically more resistant to changing the model than the frequency vocabulary is; it is shown which word2vec models for the Russian language preserve best the ontological structures in vocabulary.

CHURCH SLAVONIC TEXT IN THE RUSSIAN SCRIPT: CAN ONE USE ANY FORMAL PROCEDURE TO GET IT?

Shmelev A. D. (shmelev.alexei@gmail.com), Moscow Pedagogical State University, Moscow, Russia; Vinogradov Institute of Russian Language, Russian Academy of Sciences, Moscow, Russia; St Tikhon's Orthodox University

The paper discusses the problem of rendering Church Slavonic text in the modern Russian script, which is a common practice at present. The relevant procedure would include the following stages: spelling out words with titla, replacing the letter-based denotation of numerical values with Arabic numerals, replacing characters that are absent from the Russian alphabet with characters with the same phonetic value, removing breathings, replacing different accent marks with a unified stress accent. Certain semantic and grammatical information will be lost in the resulting text while the sound will be kept. In other words, the resulting text may be regarded as a practical transcription of the original text. At the next point, the procedure should aim at replacing the original punctuation with the common Russian punctuation (within certain limits) and at the capitalization of certain words (the latter task might require a system of determining co-reference links). The need for a system of automatic punctuation (when the input is a written text) and a system of automatic resolution of referential ambiguity poses challenges to computational linguistics.

AGRR-2019: AUTOMATIC GAPPING RESOLUTION FOR RUSSIAN

Smurov I. M., **Ponomareva M.,** ABBYY, Moscow, Russia, **Shavrina T. O.,** NRU HSE, Sberbank, Moscow, Russia, **Droganova K.,** Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic

The 2019 Shared Task on Automatic Gapping Resolution for Russian (AGRR-2019) aims to tackle non-trivial linguistic phenomenon, gapping, that occurs in coordinated structures and elides a repeated predicate, typically from the second clause.

In this paper we define the task and evaluation metrics, provide detailed information on data preparation, annotation schemes and methodology, analyze the results and describe different approaches of the participating solutions.

PHRASE-BASED ATTENTIONAL TRANSFORMER FOR HEADLINE GENERATION

Sokolov A. M. (sokolov.andrej.m@gmail.com), SPBU, Saint-Petersburg, Russia

Nowadays the task of selecting key information from large amount of text data is becoming more and more relevant. This article proposes a model of deep neural network with phrase-based attentional mechanism used for automatic generation of news headlines. The proposed architecture achieves a new state-of-the-art on the RIA news dataset.

FILLING THE GAPS WITH RULES AND NETWORKS

Sorokin A. A. (alexey.sorokin@list.ru), Moscow Institute of Physics and Technology, Neural Networks and Deep Learning Lab, Dolgoprudny, Russia; Moscow State University, Faculty of Mathematics and Mechanics, Moscow, Russia

In this paper we describe rule-based and neural approaches to gapping resolution task for Russian language. Our study was conducted on the material of AGRR-2019 Shared Task. We demonstrate that neural model definitively outperforms the rule-based one even when only 2000 annotated sentences are available. The rule-based model took the 6th place in AGRR-2019 competition (2nd in terms of precision), while the neural one was better than the second-ranked system.

MORPHOLOGICAL PARSING OF LOW-RESOURCE LANGUAGES

Sorokin A. A. (alexey.sorokin@list.ru), Moscow Institute of Physics and Technology, Neural Networks and Deep Learning Lab, Dolgoprudny, Russia; Moscow State University, Faculty of Mathematics and Mechanics, Moscow, Russia

It this paper we study morphological parsing and lemmatization on the material of Evenk and Selkup language. We compare basic neural models with their extensions that attempt to utilize additional linguistic information from the training data. We show that the augmented model does not improve over the baseline even decreasing performance for the task of lemmatization. We hypothesize that to be helpful additional information should be extracted from external resources, if available, not the corpus itself.

PREDICTING DEPRESSION FROM ESSAYS IN RUSSIAN

Stankevich M. A. (stankevich@isa.ru)Artificial Intelligence Research Institute, FRC CSC RAS, Moscow, Russia, **Smirnov I. V.** (ivs@isa.ru), Artificial Intelligence Research Institute, FRC CSC RAS; Peoples' Friendship University of Russia (RUDN University), Moscow, Russia, **Kuznetsova Y. M.** (kuzjum@yandex.ru), Artificial Intelligence Research Institute, FRC CSC RAS, Moscow, Russia, **Kiselnikova N. V.** (nv.pirao@gmail.com), Psychological Institute of Russian Academy of Education, Moscow, Russia, **Enikolopov S. N.** (enikolopov@mail.ru), Department of Medical Psychology, Mental Health Research Centre, Moscow, Russia

The study is focused on the detection of depression by processing and classification of short essays written by 316 volunteers. The set of 93 essays was provided by two different teams of psychologists who asked patients with clinically confirmed depression to write short essays on the neutral topic. The other 223 essays on the same topic were written by volunteers who completed questionnaires, which are designed to reveal depression status and did not demonstrate any signs of mental illnesses. The study describes psycholinguistic and classic text features which were calculated by utilizing natural language processing tools and were used to perform on the classification task. The machine learning classification models achieved up to 73% of f1-score for the task of revealing essays written by people with depression.

NEWS HEADLINE GENERATION USING STEMS, LEMMAS AND GRAMMEMES

Stepanov M. A. (projectttower@gmail.com), MIPT, Dolgoprudny, Russia

Headline generation is a task that has a good solution based on seq2seq models with an attention mechanism. However, it is still quite challenging to deal with morphologically rich languages, such as Russian, which have many word forms and therefore larger vocabularies. To deal with complex dependencies arising in such languages we propose several approaches based on using stems and grammemes. We applied these approaches to the pointer-generator network and took second place in the competition on headline generation held by the conference Dialogue-2019.

SOME FEATURES OF THE COMPLETIVE PREFIX *DO-* IN RUSSIAN: THEORY FACES EMPIRICAL DATA

Stoynova N. (stoynova@yandex.ru), Vinogradov Russian Language Institute & Institute of Linguistics, RAS; NRU HSE; Moscow, Russia

The paper deals with some formal features of the completive prefix *do-* ('to finish, to complete'). It was claimed in previous studies, that this prefix along with some others, has a range of formal properties that differ both from formal properties of productive "superlexical" prefixes (such as the cumulative *na-*, the distributive *po-*) and "lexical" (highly integrated) ones. Two important features were mentioned among others. 1) It can attach both to the perfective stem and to the imperfective one. 2) It cannot attach to secondary imperfectives. In the paper, I verify and develop these claims on corpus data. 1) I propose the rules of choice between the perfective vs. imperfective stem and describe the pool of variation. 2) I show, that, contrary to expectations, in informal speech *do-* attaches to secondary imperfectives quite easily.

LANGUAGE MODELS FOR UNSUPERVISED ACQUISITION OF MEDICAL KNOWLEDGE FROM NATURAL LANGUAGE TEXTS: APPLICATION FOR DIAGNOSIS PREDICTION

Tarasov D. (dtarasov3@gmail.com), **Matveeva T.**, **Galiullina N.**, Meanotek, Kazan, Russia

Following recent success of neural language models in various downstream language understanding tasks, including common sense reasoning, we investigate possible utility of such models in domain specific reasoning task—proposing of preliminary diagnosis based on patient complains, presented as natural language text. We demonstrate that language model, trained on the texts collected from online medical forums posses significant accuracy in this task (73% at top 10 suggestions), when evaluated on dataset, constructed from clinical case reports, published in specialized medical journals. While preliminary, these findings indicate a possible new method that can be used to augment online symptoms checkers and clinical decision support systems.

ASSESSING THEME ADHERENCE IN STUDENT THESIS

Tikhomirov M. M. (tikhomirov.mm@gmail.com), **Loukachevitch N. V.** (louk_nat@mail.ru), **Dobrov B. V.** (dobrov_bv@mail.ru), Lomonosov Moscow State University, Moscow, Russia

In this paper we study approaches to assessing the quality of student theses in pedagogics. We consider a specific subtask in thesis scoring of estimating its adherence to the thesis's theme. The special document (theme header) comprising the theme, aim, object, tasks of the thesis is formed. The theme adherence is calculated as the similarity value between the theme header and thesis segments. For evaluation we order theses in the increased value of the calculated theme adherence and compare the ordering with expert grades using the average precision measure. The best configuration for theses ranking is based on the weighted averaged sum of word embeddings (word2vec) and keywords extracted from the theme header.

POSSESSIVE PRONOUNS IN RUSSIAN OBJECT NOUN PHRASES

Tiskin D. B. (daniel.tiskin@gmail.com), Saint Petersburg State University, Saint Petersburg, Russia

By now, the choice between pronominal and reflexive bound possessives in Russian has been subject to prolonged investigation, but some aspects of the phenomenon have not been scrutinised by quantitative methods or given an interpretation in terms of a formal model of syntax and semantics. The present paper deals with pronominal possessives situated in direct object noun phrases and bound by a 1st- or 2nd-person pronominal subject (including zero subjects in imperatives). We focus on the factors that have to do with the availability of the collective interpretation of the verb phrase and of the possession relation itself.

Using data from the Russian National Corpus and from the Araneum Russicum Maximum web corpus, we demonstrate a significant effect of the subject number on the choice of the possessive pronoun. To investigate the effects of collectivity, we designed a questionnaire, which showed that with plural the preference for pronominal possessives over the reflexive possessive increases as the collective interpretation becomes more salient. Moreover, both types of possessives are degraded to a certain extent within singular (but not *singulare tantum*) object NPs when the collective reading

is unavailable. We suggest that the observed distribution can be explained if we assume that (a) the cardinality of the possessor is an interpretable feature of Russian possessives and is therefore able to cause interpretational conflicts, and (b) the possessive layer of the nominal domain precludes an object to be re-analysed as part of the predicate, which is required for a distributive interpretation.

CONTRAST AND COMPARISON RELATIONS IN RST FRAMEWORK: THE CASE OF RUSSIAN

Toldova S. (toldova@yandex.ru)[1], **Davydova T.** (tdadidik@gmail.com)[1], **Kobozeva M.** (kobozeva@isa.ru)[2], **Pisarevskaya D.** (dinabpr@gmail.com)[2]; [1]NRU Higher School of Economics, Moscow, Russia; [2]FRC CSC RAS, Moscow, Russia

The Paper is devoted to a corpus study of the Contrast relation between discourse units in Russian. It is based on the data of the Ru-RSTreebank annotated within the framework of the Rhetorical Structure theory [Mann, Thompson 1988]. The research question is what cue phrases and lexical and grammatical patterns are used to express the Contrast relation as opposed to the Comparison relation. Since the simple connectives such as conjunctions *a* or *no* "but" and others are ambiguous it may be useful to single out specific cues for the Contrast relation and to find other linguistic features that can also help to differentiate Contrast and other relations, such as Comparison. The investigation of cues signalling different types of relations is an important issue for both automatic discourse mining and the theoretical researches of text coherence. We test several hypotheses presented in the reference literature on Russian against corpus data.

A COMMUNICATIVE ROBOT TO LEARN ABOUT US AND THE WORLD

Vossen P. (piek.vossen@vu.nl), **Baez S.** (selene.baez.santamaria@gmail.com), **Bajcetić L.** (lenka.bajcetic@gmail.com), **Basić S.** (suz.basic@gmail.com), **Kraaijeveld B.** (bram.kraaijeveld@gmail.com), Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

We describe a model for a robot that learns about the world and her companions through natural language communication. The model supports open-domain learning, where the robot has a drive to learn about new concepts, new friends, and new properties of friends and concept instances. The robot tries to fill gaps, resolve uncertainties and resolve conflicts. The absorbed knowledge consists of everything people tell her, the situations and objects she perceives and whatever she finds on the web. The results of her interactions and perceptions are kept in an RDF triple store to enable reasoning over her knowledge and experiences. The robot uses a theory of mind to keep track of who said what, when and where. Accumulating knowledge results in complex states to which the robot needs to respond. In this paper, we look into two specific aspects of such complex knowledge states: 1) reflecting on the status of the knowledge acquired through a new notion of thoughts and 2) defining the context during which knowledge is acquired. Thoughts form the basis for drives on which the robot communicates. We capture episodic contexts to keep instances of objects apart across different locations, which results in differentiating the acquired knowledge over specific encounters. Both aspects make the communication more dynamic and result in more initiatives by the robot.

ON A NEW DICTIONARY "INTERTEXTUAL VOCABULARY OF MODERN RUSSIAN": PAPER VS. MULTIMEDIA

Voznesenskaya M. M. (voznes-masha@yandex.com), **Shmeleva E. Ya.** (eshkind@mail.ru), Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

Russian dictionaries of idioms, winged words and quotations do not reflect "the intertextual competence" of modern Russian speakers: on the one hand, their vocabularies abound in obsolete, uncommon and even incomprehensible units; on the other hand, they are short of some well known and widely used catchwords and Internet memes. The article deals with the structure and principles for constructing a new dictionary, namely, "Intertextual Vocabulary of Modern Russian" (in paper and multimedia versions). The dictionary will be based on corpus data and include over 1000 well-known catchphrases from the 20th–21st centuries. The basic unit is a dictionary entry that will include the following parts: lexical input, meaning, source, examples,

phraseological model and its transformations, comments; the last two parts are optional. The arrangement is alphabetical by the first word; however, there will be user-friendly indexes for locating all the catchphrases from the same source, same topic, etc. The multimedia version is characterized by quantitative and qualitative increase in content: in addition to text information, the dictionary will contain audio, video, photo fragments, graphics, animation, etc. referring to the relevant "multimedia" sources of intertextual units (such as movies, cartoons, paintings, songs, TV shows, etc.). Using hyperlinks, one can easily find the required information related to a given entry.

THE RUSSIAN *LI*-QUESTIONS PROSODY

Yanko T. E. (tanya_yanko@list.ru), Institute of linguistics; Pushkin State Russian Language Institute, Moscow, Russia

The paper is aimed at the analysis of the prosody in the Russian *yes-now*-questions with particle *LI*. The three basic patterns of the Russian *LI*-questions, which are construed as semantically minimal, are singled out. (The semantically minimal sentences are considered here as such where the prosodic structure brings minimal contribution into the semantic structure of a sentence). Consequently, the prosody of the sentences composed with contrast, or discourse continuity is viewed as being derived from the prosody of the basic types.

The illocutionary force in *LI*-questions is designated not by prosody as in other Russian *yes-no*-questions but by a segmental means, namely — by *LI*. Hence, the prosody in *LI*-questions is not a cue of the illocutionary force but it forms the sentence as an autonomous prosodic unit and designates the non-illocutionary meanings: contrast and discourse continuity. The accent on the first accented word can be either rising, or falling without any reasonable difference in meaning.

In questions with particle *LI*, particle *LI* preserves its Wackernagel parameters, while the host of the clitic in the majority of cases serves as the first, or the only one, accent-bearer of the sentence. However, in the context of contrast, the first accent-bearer can be placed to the right from *LI*.

Within the discourse continuity, *LI*-questions have two accent-bearers, the first of them could be either rising, or falling, and, at the same time, either contrastive, or non-contrastive, while the second one — is always the rising one.

The prosodic patterns of *LI*-questions are exemplified here by spoken fragments taken from the Multimodal corpus of the Russian National corpus, and the minor working collection of the Russian speech recordings specifically set up for this investigation. The software program Praat was used in the process of analyzing the sounding data.

RUSSIAN *CHTO-TO* AS A DISCOURSE MARKER

Zalizniak Anna A. (anna.zalizniak@gmail.com), Institute of Linguistics of the RAS; Research Centre of Computer Science and Control of the RAS,
Paducheva E. V. (elena.paducheva@yandex.ru), Federal Research Centre of Computer Science and Control of the RAS; Pushkin State Russian Language Institute

The paper demonstrates that the Russian indefinite pronoun *chto-to* ('something') can function as a discourse marker, which expresses a range of attitudes of the speaker with respect to a situation he/she considers deviating from the norm. Namely, using the discursive *chto-to* the speaker may draw the listener's attention to the reported fact, not being interested in its cause (cf. *Chto-to ja na sklone let stal sentimentalen* '*Chto-to* in my declining years, I became sentimental'); he/she may express reprobation (cf. *Chto-to ona slishkom vyrjadilas' segodnja* '*Chto-to* she is too dressed up today') or simply report something negative (cf. *Chto-to segodnja pasmurno* '*Chto-to* it is cloudy today' and ??*Chto-to segodnja svetit solnce* '*Chto-to* the sun is shining today'; *chto-to* may also express anxiety or suspicion (cf. *Chto-to v detskoj slishkom tixo* '*Chto-to* it is too quiet in the nursery'), the desire to soften the effect of a negative or potentially offensive for the interlocutor utterance, in particular, to soften the sharpness of the refusal (cf. *Davaj chaj pit'! — Chto-to ne xochetsja* 'Let's have tea! — *Chto-to* I don't want it'), and other attitudes. It is demonstrated that the meaning 'unclear why' attributed to *chto-to* by dictionaries arises only in certain contexts. The conditions for the emergence of the discursive meaning of *chto-to* are identified and its place in the semantic derivation chain is revealed. The research is based on Russian National Corpus.

THE CORPUS GRAMMAR OF RUSSIAN QPS

Zimmerling A. V. (fagraey64@hotmail.com), Pushkin State Russian Language Institute; Moscow Pedagogical State University; Institute of Linguistics, Russian Academy of Science, Moscow, Russia

The paper is addressed the corpus grammar of Russian quantifier phrases (QPs), with focus on two issues: (i) subject-predicate agreement patterns in sentences with a QP in the position of a grammatical subject, (b) the choice of the agreeing/non-agreeing form of the adjective in QPs with an embedded NP with the head noun in the feminine gender. QPs license both the plural and the singular form of the predicate. I argue that the singular form optionally shown on the predicate instantiates non-canonic agreement controlled by the QP and does not pattern with the so called default agreement in 3Sg.N. The analysis is based on the complete statistics of all Russian cardinal numerals used in the RNC in QPs of the type 'два человека/ пять человек' in the Russian National Corpora. I show the correlations between plural/singular agreement forms, word order (QP—V ~ V—QP) and communicative status of QP. The choice of the agreeing preposed NP-level adjective as in *dve interesnye knigi* does not constrain the form of the predicate agreement, while agreeing DP-level elements as in *eti dve knigi* blocks the singular form on the predicate. Russian subject QPs are non-canonic arguments, since in the two thirds of the corpus data they lack the status of a theme.

THE ROLE OF ORIENTED GESTURES DURING ROBOTS COMMUNICATION TO A HUMAN

Zinina A. (zinina_aa@nrcki.ru), **Arinkin N.** (arinkin_na@nrcki.ru), **Zaydelman L.** (zaydelman_ly@nrcki.ru), **Kotov A.** (kotov_aa@nrcki.ru), Kurhcatov Institute, Moscow, Russia; Russian State University for the Humanities, Moscow, Russia

The role of oriented gestures is crucial while solving spatial problems. We analyze the influence of a robot, using oriented gestures, on a human. In an experimental situation robot F-2 was helping a human to solve a "tangram" puzzle. Robot was indicating in speech, which game element to take and where to place it. In a half of the tasks the robot was using oriented communicative actions (hand gestures, head movements and gaze) to indicate the required game element, and then—the game position to place it in. In the other half of tasks, the robot was using non-oriented gestures. We show, that the use of oriented gestures increases the attractiveness of a robot to human and rises the general satisfaction of the interaction with the robot.

CROSS-LANGUAGE TEXT ALIGNMENT FOR PLAGIARISM DETECTION BASED ON CONTEXTUAL AND CONTEXT-FREE MODELS

Zubarev D. V.[1], **Sochenkov I. V.**[2,1]; [1]Federal Research Center 'Computer Science and Control' of Russian Academy of Sciences, Moscow, Russia; [2]Skolkovo Institute of Science and Technology, Moscow, Russia

In this paper, we present a dataset for cross-language (Russian-English) text alignment subtask of plagiarism detection. We compare different models for detecting translated plagiarism. One is based on different textual similarity scores, which exploit word embeddings. Another model extends the previous one with the features obtained via neural machine translation. The last model is built on top of pre-trained language representation (Bert) via fine-tuning for our task. The Bert model shows great performance and outperforms other models. However, it requires much more computation resources than simpler models. Therefore, it seems reasonable to use both context-free models and contextual models together in modern plagiarism detection systems.

Авторский указатель

Author Index

Научное издание

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной
международной конференции «Диалог»

Выпуск 18 (25). 2019

Ответственный за выпуск **А. В. Ульянова**
Вёрстка **К. А. Климентовский**

Издательский центр «Российский
государственный гуманитарный университет»
125993, Москва, Миусская пл., д. 6
Тел.: +7 499 973 42 06