ONLINE: ISSN 2075-7182

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной международной конференции «Диалог» (2025)

Выпуск 23
Дополнительный том

Computational Linguistics and Intellectual Technologies

Papers from the Annual International Conference “Dialogue” (2025)

Issue 23
Supplementary volume

Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог». Вып. 23, дополнительный том. 2025. С. I–1136.

Сборник включает 9 докладов международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2025», 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

23-й выпуск ежегодника «Компьютерная лингвистика и интеллектуальные технологии» содержит избранные материалы 30-й международной онлайн-конференции «Диалог». В 2025 году для публикации в дополнительном томе сборника редколлегией были отобраны 9 докладов. Работы, представленные в сборнике, отражают те направления исследований в области компьютерного моделирования и анализа естественного языка, которые по традиции представляются на Диалоге:

- **Интеллектуальный анализ документов;**
- **Лингвистический анализ текста;**
- **Глубокое обучение в компьютерной лингвистике:** методики и примеры применения в лингвистических исследованиях, содержательная интерпретация работы LLM;
- **Компьютерные лингвистические ресурсы:** включая новые датасеты и новые сценарии и типы разметки, Evaluation Benchmarks;
- **Компьютерный анализ Social Media;**
- **Корпусная лингвистика и корпусометрия:** методики создания, использования и оценки корпусов;
- **Компьютерная семантика:** аналитические и дистрибуционные модели, связь между ними;
- **Лингвистические онтологии и автоматическое извлечение знаний;**
- **Мультимодальная коммуникация:** аналитические и нейронные модели речевого акта;
- **Модели общения и диалоговые агенты;**
- **Компьютерная лексикография;**
- **Многоязычие:** language transfer, новые технологии работы с малоресурсными языками.

В соответствии с традициями «Диалога» 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, поэтому ключевым событием «Диалога» является подведение итогов технологических соревнований между разработчиками систем лингвистического анализа русскоязычных текстов — *Dialogue Evaluation*. В сборнике представлены статьи участников и организаторов 3-х соревнований:

- **CoBaLD Parsing:** Соревнование по автоматической лингвистической разметке;
- **RuOpinionNE-2024:** Соревнование по извлечению мнений из новостных текстов;
- **RuTermEval-2024:** Соревнование по выявлению терминов в научных статьях на русском языке.

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

- доклады по компьютерной лингвистике подаются 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 https://dialogue-conf.org/, 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. Все сборники размещаются на сайте конференции. С 2014 года основной том индексируются Scopus.

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

Рецензенты

Баюк Александра Михайловна
Баюк Илья Сергеевич
Беликов Владимир Иванович
Богданова-Бегларян Наталья Викторовна
Богуславский Игорь Михайлович
Бухаров Ян Михайлович
Васильев Виталий Геннадьевич
Галицкий Борис Васильевич
Добров Борис Викторович
Добровольский Дмитрий Олегович
Жарков Андрей
Ильвовский Дмитрий Алексеевич
Инькова Ольга Юрьевна
Киосе Мария Ивановна
Киселева Ксения Львовна
Клышинский Эдуард Станиславович
Клячко Елена Леонидовна
Князев Сергей Владимирович
Кобозева Ирина Михайловна
Козеренко Елена Борисовна
Коротаев Николай Алексеевич
Котельников Евгений Вячеславович
Котов Артемий Александрович
Кутузов Андрей Борисович
Лапошина Антонина Николаевна
Лепехин Михаил
Левонтина Ирина Борисовна
Лобанов Борис Мефодьевич
Лукашевич Наталья Валентиновна
Ляшевская Ольга Николаевна
Мамонтова Ангелина
Митрофанова Ольга Александровна
Мичурина Мария Александровна
Никишина Ирина Юрьевна
Пазельская Анна Германовна
Переверзева Светлана Игоревна
Петрова Мария Владимировна
Пиперски Александр Чедович
Подлесская Вера Исааковна
Рыгаев Иван Петрович
Селегей Владимир Павлович
Смирнов Иван Валентинович
Смуров Иван Михайлович
Татевосов Сергей Георгиевич
Урысон Елена Владимировна
Федорова Ольга Викторовна
Феногенова Алена Сергеевна
Хохлова Мария Владимировна
Циммерлинг Антон Владимирович
Шаврина Татьяна Олеговна
Шаров Сергей Александрович
Янко Татьяна Евгеньевна

Contents[1]

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

Russian National Corpus 2.0: corpus platform, analysis tools, neural network models of data markup (full version)

Bonch-Osmolovskaya A. A.
Vinogradov Russian Language Institute of the Russian Academy of Sciences
abonch@gmail.com

Gladilin S. A.
IITP (Kharkevich Institute), FRC CSC
gladilin@iitp.ru

Kozerenko A. D.
Vinogradov Russian Language Institute of the Russian Academy of Sciences
akozerenko@mail.ru

Lyashevskaya O. N.
HSE University
olesar@yandex.ru

Morozov D. A.
NSU
morozowdm@gmail.com

Kuznetzova Y. N.
MSU, Institute of Linguistics of the Russian Academy of Sciences
kuznetsova.yn@gmail.com

Makhova A. A.
Vinogradov Russian Language Institute of the Russian Academy of Sciences
discourse@yandex.ru

Piskounova S. V.
saruwatari.lara@gmail.com

Bujlova N. N.
Lopukhin Federal Research And Clinical Center of Physical-chemical Medicine of Federal Medical Biological Agency
bnn@rcpcm.ru

Borodina D. G.
St. Petersburg State University
daria-borodina2001@yandex.ru

Vinogradova I. I.
Prosveshchenie Publishers
irinaivinogradova@yandex.ru

Sizov V. G.
IITP (Kharkevich Institute)
victor.sizov@gmail.com

Dyachenko P. V.
IITP (Kharkevich Institute)
pavelvd@iitp.ru

Kazennikov A. O.
IITP (Kharkevich Institute)
kazennikov@gmail.com

Vlasova N. A.
A.K. Ailamazyan Institute of Program Systems of the Russian Academy of Sciences
nathalie.vlassova@gmail.com

Glazkova A. V.
University of Tyumen
a.v.glazkova@utmn.ru

Stolyarov S. S.
NSU
s.stolyarov@g.nsu.ru

Garipov T. A.
NSU
garipov154@yandex.ru

Smal I. A.
NSU
vanasmal@mail.ru

Gubar'kova Ya. N.
Yandex
karmastina-ya@yandex-team.ru

Abstract

The Russian National Corpus has existed for over 20 years and is a unique linguistic tool. However, the technical limitations of the software platform on which it was implemented significantly narrowed its development prospects. In 2020, work was launched on a comprehensive update of the RNC software platform, as a result of which the National Corpus switched to a new generation 2.0 platform. The implemented deep changes concerned both the development of functionality that meets modern approaches to corpus linguistics, and a fundamental restructuring of the platform architecture as a whole, from data preparation and indexing systems to the user interface. A separate area of development of the capabilities of the RNC was associated with the implementation of neural network models used for metadata tagging, disambiguation, word-formation markup, etc.

This article provides a detailed description of the new corpus platform as of 2024. The description includes an overview of the current technological development of corpora and corpus platforms, key parameters of changes in the architecture of the RNC platform and its user interface, descriptions of new corpus data analysis services and the specifics of their implementation, as well as a description of the experience of using neural network models for tasks related to corpus data markup.

The purpose of the article is to describe the technological layer of changes implemented in the National Corpus of the Russian Language as part of a large-scale update carried out in recent years.

Keywords: corpus linguistics; Russian National Corpus language; architecture of software platforms; markup of language data

DOI: 10.28995/2075-7182-2025-23-1001-1042

Национальный корпус русского языка 2.0: корпусная платформа, инструменты анализа, нейросетевые модели разметки данных (полная версия)

Бонч-Осмоловская А. А.
ИРЯ им. В.В. Виноградова РАН
abonch@gmail.com

Гладилин С. А.
ИППИ им. А.А. Харкевича РАН,
ФИЦ ИУ РАН
gladilin@iitp.ru

Козеренко А. Д.
ИРЯ им. В.В. Виноградова РАН
akozerenko@mail.ru

Ляшевская О. Н.
НИУ ВШЭ
olesar@yandex.ru

Морозов Д. А.
НГУ
morozowdm@gmail.com

Кузнецова Ю. Н.
МГУ, ИЯз РАН
kuznetsova.yn@gmail.com

Махова А. А.
ИРЯ им. В.В. Виноградова РАН
discourse@yandex.ru

Пискунова С. В.
saruwatari.lara@gmail.com

Буйлова Н. Н.
Федеральный научно-клинический
центр физико-химической медицины
bnn@rcpcm.ru

Бородина Д. Г.
СПбГУ
daria-borodina2001@yandex.ru

Виноградова И. И.
Издательство Просвещение
irinaivinogradova@yandex.ru

Сизов В. Г.
ИППИ им. А.А. Харкевича РАН
victor.sizov@gmail.com

Дьяченко П. В.
ИППИ им. А.А. Харкевича РАН
pavelvd@iitp.ru

Казенников А. О.
ИППИ им. А.А. Харкевича РАН
kazennikov@gmail.com

Власова Н. А.
Институт программных систем
им. А.К. Айламазяна РАН
nathalie.vlassova@gmail.com

Глазкова А. В.
Тюменский государственный
университет
a.v.glazkova@utmn.ru

Столяров С. С.
НГУ
s.stolyarov@g.nsu.ru

Гарипов Т. А.
НГУ
garipov154@yandex.ru

Смаль И. А.
НГУ
vanasmal@mail.ru

Губарькова Я. Н.
Яндекс
karmastina-ya@yandex-team.ru

Аннотация

Национальный корпус русского языка существует уже более 20 лет и представляет собой уникальный лингвистический инструмент. Однако технические ограничения программной платформы, на которой он был реализован, существенно сужали перспективы его развития. В 2020 году были запущены работы по комплексному обновлению программной платформы НКРЯ, в результате которого Национальный корпус перешел на платформу нового поколения 2.0. Реализованные глубинные изменения касались как развития функционала, отвечающего современным подходам корпусной лингвистики, так и фундаментальной перестройки архитектуры платформы в целом, начиная от систем подготовки и индексации данных и заканчивая пользовательским интерфейсом. Отдельное направление развития возможностей НКРЯ было связано с внедрением нейросетевых моделей, использующихся для разметки метаданных, снятия омонимии, словообразовательной разметки и др.

В настоящей статье представлено подробное описание новой корпусной платформы по состоянию на 2024 г. Описание включает в себя обзор современного технологического развития корпусов и корпусных платформ, ключевые параметры изменений архитектуры платформы НКРЯ и его пользовательского интерфейса, описания новых сервисов анализа корпусных данных и специфики их реализации, а также описание опыта использования нейросетевых моделей для задач, связанных с разметкой корпусных данных.

Цель статьи заключается в описании технологического пласта изменений, реализованных в Национальном корпусе русского языка в рамках масштабного обновления, проведенного в последние годы.

Ключевые слова: корпусная лингвистика; Национальный корпус русского языка; архитектура программных платформ; разметка языковых данных

1 Введение

Национальный корпус русского языка[1] существует уже более 20 лет и представляет собой уникальный лингвистический инструмент. Однако технические ограничения программной платформы, на которой он был реализован, существенно ограничивали перспективы его развития, поэтому в 2020 году были запущены работы по комплексному обновлению программной платформы НКРЯ, в результате которого Национальный корпус перешел на платформу нового поколения 2.0. Реализованные глубинные изменения касались как развития функционала, отвечающего современным подходам корпусной лингвистики, так и фундаментальной перестройки архитектуры платформы в целом, начиная с систем подготовки и индексации данных и заканчивая пользовательским интерфейсом. Отдельное направление развития возможностей НКРЯ было связано с внедрением нейросетевых моделей, использующихся для разметки метаданных, снятия омонимии, словообразовательной разметки и др.

В настоящей статье представлено подробное описание новой корпусной платформы по состоянию на 2024 г. Описание включает в себя четыре раздела. Первый раздел представляет из себя обзор современного технологического развития корпусов и корпусных платформ, во втором разделе определены ключевые параметры изменений архитектуры платформы НКРЯ и его пользовательского интерфейса, третий раздел содержит описания новых сервисов анализа корпусных данных и специфики их реализации. Наконец, четвертый раздел посвящен описанию опыта использования нейросетевых моделей для задач, связанных с разметкой корпусных данных.

[1] https://ruscorpora.ru/

Цель статьи заключается в описании технологического пласта изменений, реализованных в Национальном корпусе русского языка в рамках масштабного обновления, проведенного в последние годы. Лингвистический аспект обновлений был подробно рассмотрен в статье (Савчук и др., 2024).

1.1 Обзор современных направлений технологического развития лингвистических корпусов

Национальный корпус русского языка был открыт для публичного доступа 29 апреля 2004 года. В этот момент объем единственного корпуса насчитывал 30 миллионов словоупотреблений. За более чем двадцать лет своего развития Национальный корпус не только заметно увеличился по объему и разнообразию данных: количество словоупотреблений, представленных в 22 корпусах НКРЯ, достигло 2 миллиардов, но и претерпел концептуальную эволюцию. Изначальная задумка «Русского Стандарта» (Сичинава 2005) состояла в подготовке представительного собрания русских текстов, снабженных морфологической разметкой и предназначенных для удобного поиска при лингвистическом исследовании. НКРЯ в его современном состоянии охватывает тысячелетнюю историю развития русского языка и «представляет как язык предшествующих эпох, так и современный, в разных социолингвистических вариантах — литературном, разговорном, просторечном, диалектном»[2]. Традиционные поисковые инструменты расширяются сервисами статистического анализа и визуализации корпусных данных, принципиально изменились внутренние возможности управления корпусом, начиная от подготовки корпусных данных и кончая гибкими настройками интерфейса. Выход Национального корпуса русского языка на текущий уровень развития потребовал технологической трансформации корпуса как электронного ресурса, происходившей в одном русле с основными процессами современной корпусной лингвистики.

Ниже будут вкратце представлены ключевые параметры развития современных корпусных технологий; это позволит очертить тот научный контекст, в котором формируется стратегия технологических изменений Национального корпуса русского языка. Такими параметрами, на наш взгляд, являются критическое увеличение объемов корпусов (от сотен тысяч словоупотреблений к миллиардам) (1.1.1), переход на стандартизованные системы лингвистической разметки (1.1.2), внедрение инструментов статистического анализа корпусных данных (1.1.3) и, как результат технологического развития, расширение областей практического применения корпусов (1.1.4).

1.1.1 Увеличение объемов корпусов

Вопрос о том, какой размер должен иметь идеальный корпус, не имеет однозначного ответа (Reppen 2021). Качественные и широко используемые лингвистами современные корпуса имеют очень разный объем. Одни из самых объемных корпусов сегодня принадлежат к типу TenTen, это целое семейство корпусов на более чем 40 языках, доступных на платформе SketchEngine[3]. По их названию понятно, что целевой размер таких корпусов — 10^{10} слов. Все корпуса этого типа собираются автоматически в Интернете по единому алгоритму, который, в частности, предполагает очистку от повторяющихся текстов и удаление нерелевантных фрагментов, и поэтому могут считаться сопоставимыми. Но вообще для корпусов, собранных автоматически, размер в 10^{10} слов не является пределом, например, почти на порядок больше размеры корпуса NOW (News on the Web)[4].

Однако существуют и корпуса принципиально иного типа, а именно репрезентативные и сбалансированные, при создании которых ставится весьма амбициозная цель подобрать коллекции текстов таким образом, чтобы корпус отражал язык в целом. Определение репрезентативности каждого корпуса — это многоэтапная работа, которая всегда ведется на разработанных специально для этого корпуса теоретических основаниях (Biber 1993). Репрезентативность невозможно измерить никакими известными метриками, поэтому создание репрезентативного корпуса представляет собой отдельную исследовательскую задачу для создателей корпуса. (McEnery, Hardie 2012). Сбалансированность предполагает задание пропорций разных текстов, входящих в корпус, с учетом разнообразных параметров, например, жанров, тем, стилей, иногда также эпохи

[2] https://ruscorpora.ru/page/corpora-about/
[3] https://www.sketchengine.eu/documentation/tenten-corpora/
[4] https://www.english-corpora.org/now/

написания (см. пример корпуса китайского языка Sinica (Chen et al., 1996), корпуса современного турецкого языка (Aksan Y. et al., 2012) или корпуса немецкого языка XX века (Geyken 2007)). Расчет и подготовка корпусных коллекций для таких корпусов — это сложная и кропотливая работа, зачастую требующая дополнительных ресурсов, например, расшифровки устных текстов, распознавания старых печатных или рукописных изданий, поэтому объемы сбалансированных корпусов заведомо существенно меньше, чем собранных автоматически.

Одним из самых больших сбалансированных корпусов является COCA (Corpus of Contemporary American English)[5] с объемом чуть больше миллиарда слов. В этом корпусе баланс понимается следующим образом: создатели корпуса выделили 8 категорий (субтитры, устный, художественная литература, журналы, газеты, научные журналы, блоги, Интернет-страницы) и равномерно распределили тексты между ними.

Также следует упомянуть один из самых известных национальных корпусов — Чешский национальный корпус, который насчитывает примерно 5 миллиардов слов. В целом, корпус является репрезентативным, но хотя бы отчасти сбалансированными можно считать только те его части, которые относятся к пятилетним коллекциям типа SYN20XX[6], а их объем на порядок меньше.

Важно подчеркнуть, что рост объемов корпусов явился ответом на изменившиеся технические возможности: хранение большого объема текстов онлайн, увеличение мощностей памяти и процессоров. Таким образом, безусловным стандартом построения современных лингвистических корпусов, претендующих на общие задачи отображения языковых явлений в синхроническом или диахроническом срезе, стали не только требования к большому объему данных, но, что принципиально, отсутствие технических ограничений для их кратного увеличения. Еще одним фактором, существенно повлиявшим на процесс создания больших корпусов, является расширение применения и улучшение качества инструментов для автоматической разметки. Не в последнюю очередь это стало возможно благодаря процессам стандартизации автоматической разметки, происходившим в последние десятилетия.

1.1.2 Стандартизованная разметка корпусов

Ключевое отличие языковых корпусов от коллекций текстовых файлов (электронных библиотек) состоит в том, что тексты в корпусе снабжены лингвистической разметкой. Как отмечается в (McEnery, Wilson 2001), разметка существенно расширяет возможности использования корпусов и спектр исследовательских вопросов, которые можно решать с их помощью. Разметки языковых корпусов представляют собой лингвистические абстракции, соответствующие базовым уровням языка, начиная с грамматики. Самая «базовая» разметка — морфологическая — обеспечивает возможность поиска по морфологическим признакам слова (например, по частям речи) и соотносит разные формы слова с общей для них «начальной формой» — леммой. Синтаксическая разметка дополняет текст информацией о связи слов в предложении, маркируя главное слово и его зависимые, а также тип синтаксической связи между главным и зависимым. Семантическая разметка корпусов соотносит лексически полнозначные слова с соответствующими семантическими категориями или классами. Дискурсивная разметка выходит за пределы одного предложения, выделяя кореферентные единицы или отношения между предложениями или их частями (элементарными дискурсивными единицами). Могут быть размечены и другие лингвистические единицы — морфемы, интонационные контуры, диалогические роли и т.д. (Newmann Cox 2021).

В 1980–1990-х годах, когда, благодаря резкому увеличению электронных текстов, корпуса стали создаваться очень активно, выбор способа разметки соответствующего набора тегов был делом исключительно рабочей группы, разрабатывающей корпус. Очень характерно в этом смысле звучит один из семи принципов лингвистической разметки, сформулированных в (Leech 1993): ни одна схема разметки не может претендовать на то, чтобы быть корпусным стандартом. Существовала презумпция, что оптимальная схема разметки «вырастает» из конкретных задач, для которых собирается корпус (так называемый bottom-up подход). Однако через некоторое время стало понятно, что отсутствие единого подхода к лингвистической аннотации тормозит в целом «корпусную отрасль», мешает взаимной совместимости ресурсов и возможности их по-

[5] https://www.english-corpora.org/coca/
[6] https://wiki.korpus.cz/doku.php/en:cnk:syn:verze11

вторного использования, затрудняет применение готовых инструментов (например, готовый к использованию синтаксический парсер может опираться на иную схему разметки, чем та, которая используется в корпусе) и мешает использованию и интеграции данных лингвистических корпусов в системы, связанные с автоматическим анализом текста. Как отмечается в (Ide et al., 2017:115), «тысячи часов были потрачены на преобразование данных, представленных в одном формате, в другой, который будет работать для других целей или с другим программным обеспечением, или, что еще хуже, на воссоздание тех же ресурсов с целью решения конкретных задач». Важным фактором, повлиявшим на изменение отношения к стандартизации разметки от позиции «выбор разметки — дело рабочей группы» к использованию унифицированных единых стандартов стал переход компьютерной лингвистики от правиловых алгоритмов к методам машинного обучения и нейросетевому моделированию. Корпуса стали главным источником данных для решения всего спектра задач автоматического анализа и генерации языка, появляются открытые корпусные ресурсы, которые могут быть использованы для машинного обучения, корпуса создаются под конкретные задачи с помощью готовых инструментов и ценность их повторного использования многократно возрастает, особенно тогда, когда речь идет о коммерческих решениях. Таким образом, на сегодняшний день вопрос стандартизации лингвистической разметки является центральным и приоритетным при подготовке корпусных языковых данных. При этом по-прежнему не существует единого стандарта для лингвистического аннотирования; вариативность здесь, с одной стороны, связана с вариативностью возможных форматов разметки, а с другой стороны, с разнообразием языковых данных — лингвистических характеристик языковых структур, специфичных лингвистических паттернов отдельных языков и т.д.

Одним из самых старых стандартов текстовой разметки является Text Encoding Initiative[7] — TEI. Идея стандарта возникла в 1987 году и была направлена на формализацию документации для создания машиночитаемых текстовых объектов гуманитарного знания (литература, лингвистика, история и др.), явившись, таким образом, одним из первых стандартов, используемых для цифровизации культурного наследия. TEI — это, по сути, набор тегов и атрибутов для отображения самых разных параметров текста, которые используются в рамках синтаксиса XML (первоначально SGML). В дальнейшем развиваются близкие TEI корпусные форматы, также основанные на SGML/XML, предлагающие более последовательную и одновременно более простую схему: стандарт кодирования корпусов CES (Corpus Encoding Standard) (Ide 1998) и включающая его инициатива EAGLES (Expert Advisory Group for Language Engineering Standards) (Calzolari, McNaught, Zampolli 1996). Именно в стандарте CES была впервые сформулирована концепция standoff аннотаций — т.е. таких аннотаций, которые не вписываются непосредственно в текст, но связаны с аннотируемым словом через указатель или id. Такой формат позволяет отделять исходный текст от его разметки, а также облегчает дополнение разметки новыми уровнями. Именно этот подход является сегодня мейнстримом. Схемы и форматы, разработанные в рамках CES и EAGLES, можно рассматривать как подготовительную фазу к более масштабному процессу регулирования и стандартизации аннотации корпусов, который был запущен в начале 2000-х годов, благодаря созданию специальной группы «Управление языковыми ресурсами» (“*Languages resource management*”)[8] (Kiyong, Laurent 2010) в рамках ISO — международной организации по стандартизации. Принципиально новым решением стандарта LAF, разработанным рабочей группой, в дальнейшем реализованным в XML-формате GRAF (Graphic annotation format), стал переход от иерархической организации аннотаций к графовой модели представления данных (Ide et al., 2017). LAF основывается на двух фундаментальных принципах: во-первых, создается абстрактная модель данных, которая явным образом разграничивает структуру (физический формат) данных и их содержимое (название лейблов и категорий), во-вторых, вслед за стандартом CES утверждается принцип standoff аннотаций, хранящихся отдельно от текстов. Абстрактная модель данных LAF представляет собой нецикличный направленный граф с параметрами, связывающий с помощью «якорей» сегменты текста с узлами-лейблами. Формат GRAF, который развивает и продолжает идею LAF, предназначен для использования в качестве «стержневого формата» (pivot format), является не самостоятельным форматом, но инструментом совмещения разных форматов разметки для создания многоуровневых аннотаций, которые могут в него переводиться и, наоборот, из него извлекаться. Таким образом, был сделан важный шаг в направлении

[7] https://tei-c.org/

[8] https://www.iso.org/committee/297592/x/catalogue/

совместимости форматов. Разработанная абстрактная графовая модель, лежащая в основе лингвистических аннотаций корпусов, изоморфна многим современным форматам, в том числе и такому общему формату, как формат Linked Open Data (открытых связанных данных) RDF/OWL.

Одновременно с институциональной разработкой форматов лингвистической разметки корпусов происходил процесс развития форматов «де факто», т.е. таких, которые являлись результатом развития индустрии компьютерной лингвистики. Так, оказалось, что формализм синтаксических зависимостей, представляющий предложение как систему связанных элементов «вершина — зависимое», является наиболее удобным для автоматического синтаксического парсинга. Рост популярности этого подхода выразился в распространении трибанков (корпусов синтаксически размеченных предложений). Использование трибанков для обучения парсеров естественным образом привело к необходимости стандартизации синтаксической разметки. Важным этапом для формирования общего формата представления синтаксических отношений стало соревнование (shared task) по зависимостному парсингу, проведенное на конференции CONLL в 2006 году и включавшее в себя задачу установления синтаксических зависимостей для текстов на 13 языках (Buchholz, Marsi 2006). Формат представления данных CONLL-X, предложенный на конференции, стал де факто стандартом. В этом формате каждый токен представляет собой строку из 10 колонок, которые заполняются лингвистической информацией, включая отношения между вершиной и его зависимым. Этот формат задает модель данных, но не устанавливает никаких ограничений собственно на те условные обозначения (лейблы), которыми заполняются колонки, а также на те критерии, по которым устанавливаются синтаксические отношения. В результате даже следующие CONLL-X формату трибанки на разных языках чрезвычайно затруднительно сравнивать. Ответом на эту проблему стала инициатива «универсальных зависимостей» Universal Dependencies (Nivre et al., 2016), (De Marneffe et al., 2021), созданная в 2014 году группой ученых под руководством Иоакима Нивра. Universal Dependencies (UD) предлагает готовый универсальный набор из трех лейблов — часть речи, морфологические параметры и название синтаксического отношения, а также единые правила, определяющие то, как устанавливаются синтаксические зависимости. К моменту запуска инициативы были представлены трибанки на 10 языках, выполненные в этом формате, а в настоящий момент количество представленных трибанков превышает 200, включая данные на древних языках, таких как латынь, древнегреческий, церковнославянский. Безусловному распространению принципов UD способствовало развитие парсера UDPipe, обученного на существующих в коллекции UD трибанках а также приспособленного для обучения на любой коллекции в формате CONLL-X, обогащенного лейблами универсальных зависимостей, этот формат получил название CONLL-U (Straka, Hajic, Straková 2016). На сегодняшний день этот формат наиболее широко используется для подготовки корпусных данных для задач NLP, а также для корпусной лингвистики в целом.

В России распространение CONLL-U было во многом стимулировано соревнованием морфосинтаксических парсеров GRAMEVAL, прошедшим в 2020 году (Lyashevskaya et al., 2020).

Автоматическая разметка корпусов текста является ключевым решением для обработки больших объемов данных, поскольку ручная разметка невозможна в разумные сроки. Принципиально важным является не только приписывание потенциально возможных тегов словоформе, но выбор единственно правильных тегов и соответствующей леммы, т.е. снятие морфологической омонимии. Эта задача решается с помощью современных методов, основанных на глубинном обучении. Модель Rubic, дополнившая традиционно используемый в НКРЯ алгоритм MyStem, которая была применена для разметки корпусов НКРЯ, будет подробно описана в 4.1. Важным результатом применения модели Rubic стало внедрение синтаксической разметки и автоматическое снятие омонимии для больших объемов данных Основного и Газетных корпусов (более 1,24 миллиарда словоупотреблений). Этот результат позволил не только существенно расширить возможности поисковых запросов, но разработать сервисы комплексного анализа выдачи, отвечающие современным требованиям корпусных инструментов и методов лингвистических исследований.

1.1.3 Инструменты статистического анализа

Наличие встроенных статистических инструментов существенно упрощает лингвисту работу с корпусными данными, потому что, во-первых, позволяет отказаться от статистических подсчетов

вручную, а во-вторых, помогает увидеть важные статистические особенности и распределения языковых явлений.

Большинство крупных корпусов мирового уровня уже немыслимы без статистических инструментов. В целом, в корпусах существует всего три основных типа выдачи: конкорданс (чаще всего в формате KWIC), коллокации и частотные списки (Stefanowitsch 2020). Конкорданс — исторически первый и наиболее просто реализуемый способ выдачи, без которого корпус не может считаться корпусом. Под коллокациями в корпусной лингвистике понимаются привычные и повторяющиеся сочетания слов (Firth 1957). Все статистические метрики, применяемые для извлечения коллокаций, идейно основаны на том, что в текстах естественного языка слова встречаются друг с другом не случайно: их сочетаемость ограничена, по крайней мере, грамматикой и семантикой, поэтому она подчиняется определенным статистическим распределениям, а значит, коллокации могут извлекаться автоматически (Evert 2008). Коллокации позволяют исследователю получить в концентрированном виде информацию о сочетаемости того или иного слова, найти слова, которые связаны с ним более тесными связями. Некоторые современные корпуса ориентированы именно на поиск коллокаций, наиболее яркий пример — корпуса, работающие на платформе Sketch Engine[9]. Главная функция, реализованная на этой платформе, а именно построение скетчей слова, основана на обнаружении коллокаций с учетом синтаксических связей. Еще один корпус, нацеленный в том числе на поиск коллокаций, это COCOCO[10] (Kopotev et al., 2015). Встроенные инструменты для поиска коллокаций также существуют, например, в таких корпусах, как Британский национальный корпус[11], Чешский национальный корпус[12], Корпуса Университета Лидса[13].

Частотные списки могут иметь разнообразные применения, в частности, для составления пособий для изучения некоторого языка как иностранного, но если говорить о чисто исследовательских целях, то частотные списки особенно полезны для выявления отличий корпусов друг от друга или корпуса и его подкорпуса. Помимо лемм, упорядоченных по частоте встречаемости, частотный список также может включать грамматические формы слов некоторого языка. Впервые такая задача в неавтоматическом режиме была решена еще в 1982 году на данных Брауновского корпуса (Francis, Kučera 1982). Было разработано несколько приложений для построения частотных списков, среди известных следует упомянуть, по крайней мере, WordSmith Tools[14] и MonoConc[15] (оба приложения коммерческие). Однако открытых корпусов со встроенными инструментами для построения частотных списков (в том числе для пользовательских подкорпусов) мало. Впервые эта возможность была реализована на платформе Sketch Engine. В Британском национальном корпусе функционал построения частотных списков существует только в очень ограниченном виде.

Корпусом с наиболее разнообразными статистическими инструментами можно без сомнения назвать Чешский национальный корпус. Помимо коллокаций, корпус позволяет получить данные о частотных распределениях лемм и любых сочетаний грамматических признаков. Еще одним чрезвычайно наглядным инструментом является Word at a Glance (Machálek 2020a). Он позволяет увидеть в одном окне самую разнообразную информацию о слове: его частотность в жанровых подкорпусах, сочетаемость в рамках коллокаций, похожие слова, изменение частотности во времени, а также несколько примеров с заданным словом из корпуса.

Развитие инфраструктуры для многомерных корпусных исследований — унификация разметки, развитие статистических инструментов и корпусных приложений, наконец, значимое увеличение объемов самих корпусов способствовали тому, что корпусные методы стали применяться не только для изучения лингвистических явлений, но для решения более широких прикладных лингвистических задач, таких как преподавание, переводоведение, автоматический анализ языка, а также в разных областях социальных наук.

[9] https://www.sketchengine.eu/
[10] https://cococo.cosyco.ru/about.html
[11] https://www.english-corpora.org/bnc/
[12] https://www.korpus.cz/
[13] http://corpus.leeds.ac.uk/internet.html
[14] https://www.lexically.net/wordsmith/
[15] https://monoconc.com/

1.1.4 Расширение аудитории корпусных исследований

Первоначально корпуса создавались для лингвистических исследований и сбора естественных примеров для описания разнообразных языковых явлений. Однако по мере развития корпусной инфраструктуры и методов корпусного анализа, корпуса стали применяться и для других задач. В первую очередь корпуса стали использоваться в преподавании языка как иностранного, начало этой методике было положено еще несколько десятилетий назад (Wray 2013). Можно сказать, что существует отдельный подход data-driven learning, который подразумевает, что изучающий язык использует корпус как конкорданс для обнаружения характерного для некоторого слова языкового поведения (Johns, King 1991), (Boulton 2011). На основе корпусов создаются учебные материалы, упражнения и словари (McCarthy 2008). Параллельные корпуса находят наиболее широкое применение в переводоведении и обучении переводчиков (Doval, Sánchez Nieto 2019), (Beeby, Rodríguez, Sánchez-Gijón 2009), (Zanettin 2013).

Безусловно, корпуса имеют огромное значение для решения разных задач, связанных с обработкой естественного языка, в том числе и как обучающие коллекции для машинного обучения. Среди классических задач можно назвать обучение чат-ботов (Shawar, Atwell 2005), разрешение омонимии (Roll, Correia, Berger-Tal 2018), определение тональности (Yang, Lin, Chen 2007), (Schrauwen 2010), классификации текстов (Curtotti, Mccreath 2010) и многоге другое.

Корпусные методы широко применяются в судебной лингвистике, например, для установления авторства текста или для целей семантического анализа и определения значения слова (Coulthard 1994), (Heffer 2005), (Coulthard, Johnson, Wright 2017), (Баранов 2023).

Использование корпусного подхода в социальных науках определяет специфику и тематику собираемых корпусов (Wiedemann 2013). Корпуса используются для решения таких задач, как обнаружение «языка ненависти» (Poletto et al., 2021) или для анализа сообщений из соцсетей при разного рода катастрофах (Imran, Mitra, Castillo 2016). При этом узкая тематическая направленность задачи не обязательно влияет на размер корпуса. Так, в период с января 2020 года до декабря 2022 года был собран мультиязыковой корпус сообщений прессы, посвященный коронавирусу, который достиг объема 1,2 млрд словоупотреблений (Davies 2021). В целом, можно сказать, что на сегодняшний день мы видим примеры использования корпусных методов практически в любой области знаний, так или иначе связанной с естественным языком, от библиотечного дела (Bowker 2018) до экспериментальной психологии (Chatrand 2022).

1.2 НКРЯ 2.0 в свете основных тенденций развития современных корпусов

Обзор современного состояния развития корпусной лингвистики позволил выдвинуть принципиальные требования к новой платформе по сравнению со старой. Ниже обобщены ключевые результаты развития платформы 2.0.

Объемы корпусов

Старая платформа: К 2020 году общий объем всех корпусов НКРЯ составляет около 1 миллиарда словоупотреблений. При этом грамматическая омонимия снята лишь менее чем в 1% от всех словоупотреблений (только вручную снятая омонимия в Основном, Обучающем и Устном корпусах, а также снятая при помощи процессора Этап-3 и проверенная вручную омонимия в СинТагРусе). Дальнейшее расширение корпуса затруднено в связи с архитектурными ограничениями платформы.

Новая платформа: К 2024 году общий объем всех корпусов НКРЯ составляет около 2 миллиардов словоупотреблений, грамматическая омонимия снята примерно в 65% от всех словоупотреблений (добавилась автоматически снятая омонимия в Основном корпусе и обоих корпусах СМИ). Новая платформа позволяет увеличить объем данных НКРЯ до 100 миллиардов словоупотреблений.

Разметка данных

Старая платформа: Корпусные данные снабжаются морфологической разметкой, осуществленной с помощью программы MyStem (Зобнин, Носырев 2015), представляющей собой контаминацию недетерминированного конечного автомата и наивного байесовского классификатора. Алгоритм позволяет строить гипотетические разборы для слов, которых нет в грамматическом словаре Зализняка, и ранжирует леммы по вероятности в случае омонимии. Используется собственный тип разметки НКРЯ.

Новая платформа: Сохраняется морфологическая разметка алгоритмом MyStem, к ней добавлена разметка данных с помощью нейросетевой модели Rubic. Модель размечает не только морфологические, но и синтаксические характеристики словоформ, а также снимает омонимию не только по леммам, но и по словоизменительным признакам. Используется разметка формата CONLL-U, разработаны принципы взаимной трансформации разметок НКРЯ и CONLL-U. Подробнее о том, как работает модель Rubic, будет изложено в разделе 4.1.

Развитие инструментов корпусного анализа

Старая платформа: Основным инструментом корпусного анализа является выдача по поисковому запросу в формате конкорданса или KWIC (key word in context). Пользователь имеет возможности сортировки результатов по дате создания текста и другим релевантным параметрам, а также по правому/левому контексту в формате KWIC. Обобщенная информация об изменениях частотностей слова представлена в виде диахронического графика, который может быть построен только по конкретной словоформе. Пользователь имеет доступ к n-граммам, предпосчитанным по словоформам.

Новая платформа: Инструменты корпусного анализа существенно расширены как на уровне запроса, так и на уровне представления выдачи. На уровне поисковых запросов появился поиск по коллокациям, на уровне выдачи — анализ частотности запроса, допускающий разные способы сортировки и представления данных (например, анализ частотности не только встретившихся словоформ, но и обобщение результатов на уровне грамматических параметров). Появилась возможность получить предпосчитанную информацию о слове в корпусе в целом — «Портрет слова», куда входят его скетчи (коллокации на основе базовых синтаксических связей), контекстуально близкие слова, выявленные на основе расчета семантических векторов, однокоренные слова в корпусе. Описания методов, использованных для подготовки «Портрета слова», приводятся в разделах 3.2 и 3.3. Подробное описание пользовательских возможностей, реализованных в новых корпусных инструментах, было представлено в статье (Савчук и др., 2024).

Целевая аудитория НКРЯ

Старая платформа: Платформа ориентирована на подготовленного пользователя-лингвиста, который использует корпус как источник материала для лингвистических исследований.

Новая платформа: Новая платформа ставит своей задачей расширить аудиторию пользователей, в том числе привлекая менее подготовленных пользователей, не работавших ранее с языковым корпусами. Корпус существует в мобильной версии, имеет богатейшую документацию, логика интерфейса минимизирует усилия пользователя по получению информации. Подробно идеология обновления интерфейса новой корпусной платформы представлена в разделе 2.2.

Ниже будут более подробно рассмотрены три аспекта технической реализации новой корпусной платформы. Во-первых, это концептуально новые подходы к архитектуре корпуса, корпусному ядру и веб-интерфейсу. Во-вторых, разработанные сервисы для корпусного анализа данных. В-третьих, нейросетевые модели, использованные для разметки данных.

2 Корпусная платформа нового поколения: примененные подходы и решения

Разработка новой платформы НКРЯ включала в себя перестройку на уровне «бэкенда» — архитектуры индексации, поиска и статистической обработки корпусных данных, а также концептуальное обновление на уровне «фронтенда», т.е. пользовательского интерфейса, с помощью которого пользователь взаимодействует с корпусами НКРЯ. Для взаимодействия между корпусной архитектурой и веб-интерфейсом был создан новый инструмент для работы с корпусами, позволяющий взаимодействовать через скрипты, использующие API, без необходимости пользовательского интерфейса. Ниже мы последовательно рассмотрим подходы и решения для каждого из этих направлений.

2.1 Общая архитектура системы

К корпусной платформе нового поколения предъявлялись требования не только соответствия современным стандартам сервисов, предоставляемых крупными лингвистическими корпусами, но и обеспечения гибкости для последующей модификации и развития в соответствии с перспективными подходами, которые могут возникнуть в будущем. На момент разработки имелась существенная неопределенность, в целом характерная для ИТ-проектов: у нас не было представления о полном функционале, который в будущем потребует поддержки в НКРЯ. С развитием корпусных технологий возникают новые потребности, и корпусная платформа должна быть готова к их реализации. Поэтому требовалось организовать программную систему так, чтобы в будущем по возможности облегчить добавление нового функционала. Для этого мы стремились обобщить различные требования (уже имеющиеся или же потенциально возможные) к функционалу в однородные с точки зрения технической реализации группы. Таким образом, перед корпусной платформой ставилась задача поддержки не конкретных видов функционала, а целых функциональных групп; конкретные виды функционала рассматривались как представители этих групп. Например, вместо поддержки конкретного списка возможных атрибутов, приписываемых каждому токену в тексте, вводилось понятие «атрибута, приписанного к токену», и программировались универсальные алгоритмы, рассматривающие список атрибутов заданного типа как параметр.

Такой подход потребовал унификации корпусных данных: атрибуты одинакового типа обрабатываются одними и теми же алгоритмами, а значит должны быть единообразно представлены. Унификация была достигнута либо путем изменения разметки исходных текстов, либо за счет подключения небольших модулей-конвертеров, что позволяло избежать переобучения лингвистов-разметчиков, готовящих тексты корпуса.

Структурно программная система разделена на три независимые части. Вычислительное ядро реализует универсальный функционал для целой функциональной группы, лингвистическое ядро обеспечивает поддержку конкретных функций, используя реализованный функционал вычислительного ядра, а интерфейсный модуль осуществляет взаимодействие лингвистического ядра с пользователями. Таким образом, например, добавление новой функциональности, касающейся приписанных к токенам атрибутов, выполняется в вычислительном ядре и влияет одновременно на все атрибуты, и таким образом может не затронуть или незначительно затронуть лингвистическое ядро, а поддержка новых атрибутов легко реализовывается в лингвистическом ядре и не требует изменений вычислительных алгоритмов.

Такое разделение позволило использовать в корпусной платформе разные вычислительные ядра в зависимости от размера и структурной сложности корпуса. Так, например, для больших корпусов применяется вычислительное ядро, построенное на базе поисковой системы ElasticSearch, в то время как для сложно структурированной разметки небольшого Синтаксического корпуса лучше подошло вычислительное ядро на базе реляционной базы данных MySQL.

Выделение отдельного интерфейсного модуля важно, поскольку подходы к построению графических интерфейсов пользователя быстро меняются, а отдельный модуль легче заменить.

Были выделены следующие виды разметки, поддерживаемой платформой:

Структурная единица, реализуемая вычислительным ядром	**Структурная единица, реализуемая лингвистическим ядром**	**Приписываемые к структурной единице атрибуты**
токен	слово	лемма, грамматические параметры, морфемное членение, семантические, орфоэпические и др. параметры (в случае снятой омонимии) словоформа, а также вспомогательные параметры: повтор лексемы, знаки препинания до и после слова, начало/конец предложения, слово с заглавной буквы (в случаях как снятой, так и не снятой омонимии)
разбор	вариант снятия омонимии	лемма, грамматические параметры, морфемное членение, семантические, орфоэпические и др. параметры (в случае неснятой омонимии)
сегмент	предложение / фраза в устных корпусах / грамота в корпусе «Берестяные грамоты»	в настоящее время нет атрибутов, но предусмотрена их поддержка при необходимости
текст	письменный текст / устный текст / выровненные между собой тексты в параллельных корпусах (=текст+перевод) / текст+последовательность аудио- или видеофрагментов в мультимедийных корпусах	мета-атрибуты: тип текста, жанр, тематика и т.д.
фрагмент (последовательность подряд идущих сегментов)	абзац, строфа, реплика, клипотекст[16], зона выравнивания (в параллельных корпусах)	говорящий (для реплик и клипотекстов), вычисленный возраст аудитории (в детском корпусе), речевые акты (приписываются к клипотексту)
подмножество токенов внутри сегмента	клауза, группа, микросинтаксическая конструкция	тип клаузы, вид микросинтаксической конструкции, лемма микросинтаксической конструкции (заложено в платформу, но в настоящее время поддерживается ограниченно)

[16] В мультимедийный корпусах клипотекст — минимальная единица, состоящая из отрывка видео или аудио и соответствующего ему текста, а также набора жестов и речевых действий.

Структурная единица, реализуемая вычислительным ядром	**Структурная единица, реализуемая лингвистическим ядром**	**Приписываемые к структурной единице атрибуты**
последовательность токенов, пересекающая границы сегментов	стихотворная строка	метр, количество стоп/иктов/слогов, клаузула, схема
ребро графа токенов внутри сегмента	синтаксическая связь, лексико-функциональная связь	направление связи, тип связи, вспомогательный предлог (в лексико-функциональной связи)
меж-токены (фантомы)	эллидированное слово[17]	как у токенов, но отсутствует словоформа
выравнивание фрагментов	выравнивание текста с переводом в параллельных корпусах, выравнивание текста с видео/аудио в мультимедийных корпусах	нет атрибутов
соответствие токенов внутри выровненных фрагментов	«пословное выравнивание» (спроектировано, но в данный момент не используется)	нет атрибутов
внутритокенная разметка	ударение, морфема, шрифтовое выделение	атрибуты морфологического разбора, тип шрифтового выделения

Таблица 1: Виды разметки, поддерживаемые НКРЯ

В дальнейшем при возникновении новых атрибутов, попадающих в один из выделенных классов, не потребуется изменения вычислительного ядра. Вся необходимая функциональность будет в этом случае обеспечена лишь не очень значительными изменениями лингвистического ядра. Ниже мы покажем на трех примерах, каким образом вычислительный модуль позволяет гибкую настройку нового лингвистического функционала.

Пример 1. Организация поиска в мультимедийном корпусе

В корпусной платформе нового поколения тексты, состоящие из слов, и аннотации жестикуляции в видеозаписи, состоящие из отдельных жестов, представляются при помощи одного и того же программного механизма. Таким образом, выровненный с текстом видеоряд внутренне представляется тем же способом, что и два параллельных текста, выровненных между собой. Это позволяет переносить функциональные возможности, разработанные для параллельных корпусов, и на мультимедийные. Так, для параллельных корпусов программная платформа нового поколения поддерживает поиск по запросу, условия которого накладываются на тексты на обоих языках в выровненной паре. Это позволяет, например, искать такие английские предложения, содержащие слово *cat*, перевод которых содержит слово *кошка* (а не *кот*). При этом условия на каждом языке могут накладываться не только на отдельные слова, но и на их сочетания (например, на расстояния между словами) (Сичинава 2022). Новый запрос обеспечения возможности поиска по комбинации жестов в мультимедийном корпусе (с возможностью комбинации условий

[17] Эллидированному слову в тексте приписываются все атрибуты, которые есть у обычного слова, за исключением собственно словоформы — чтобы по ним можно было осуществлять поиск.

на жесты с условиями на текст) оказался фактически аналогичным поиску по условиям, накладываемым на слова на двух языках (с точностью до замены слов одного из языков на жесты). Таким образом, реализация поиска по комбинации жестов потребовала лишь небольших изменений в лингвистическом ядре.

Пример 2. Поиск по совпадению и различию атрибутов токена

В настоящее время система поддерживает только поиск по совпадению некоторых атрибутов одного слова в тексте с соответствующими атрибутами соседнего слова. Это делается при помощи дополнительных приписанных к токену атрибутов «повтор лексемы», «повтор числа» и т.д. Однако для многих лингвистических исследований принципиальна важна возможность поиска по совпадению атрибутов слов, не идущих в тексте подряд.

Наш анализ показал, что если интерпретировать совпадение двух атрибутов графовой связью между двумя токенами, то условие на то, что у слова А некий атрибут совпадает с соответствующим атрибутом слова Б, может быть представлено как условие на графовую связь аналогично поисковому запросу по расстоянию или по заданной синтаксической связи. Таким образом, для выполнения поставленной задачи потребуется дополнить поисковый индекс разметкой графовых связей отдельного типа, отражающей совпадение атрибутов между токенами. После этого потребуются лишь небольшие изменения лингвистического ядра, чтобы возможность такого поиска был включена в корпусную систему. Такие изменения уже включены в план работ на будущее.

Пример 3. Разметка сложности текста для читателей разного возраста

При создании корпуса «От 2 до 15» для отдельных частей произведения размечался показатель сложности текста в терминах предполагаемого возраста читателя (Morozov, Glazkova, Iomdin 2022). Такое приближение часто используется в прикладных исследованиях, посвященных созданию алгоритмов автоматической оценки сложности текста. В этом случае читательский опыт респондентов одного возраста считается схожим, а верхняя граница сложности устанавливается на уровне последних классов школы или младших курсов университета. В ситуации с корпусом «От 2 до 15» возрастная разметка была экспериментально собрана таким образом, чтобы определить наиболее популярные художественные произведения у носителей языка различного (школьного) возраста. Считался порог, при котором 50% респондентов к этому возрасту прочли книгу.

Авторами исследования была разработана нейросетевая модель, позволяющая предсказывать сложность текста. На вход модель получает фрагмент текста (один или несколько абзацев), на выходе возвращает предполагаемый возраст читателя.

Перед разработчиками платформы встала задача организации поиска по атрибуту уровня читательского опыта. Анализ показал, что атрибут необходимо приписывать к фрагментам текста (а нарезку текста на фрагменты организовать в соответствии с разметкой). Таким образом, путем несложных изменений в лингвистическом ядре может быть реализован поиск[18], аналогичный поиску по условиям на говорящего, реализованному в устном корпусе.

Приведенные примеры показывают, каким образом базовый список единиц, поддерживаемых вычислительным ядром (см. выше Таблицу 1), не только обеспечивают уже существующую функциональность поиска, но и открывают гибкие возможности для ее расширения по мере развития корпусной специальной разметки и запросов на новые исследовательские возможности.

2.2 Автоматизированное взаимодействие корпусной платформы с другими лингвистическими системами

В настоящее время корпуса, доступные для пользователей через сеть Интернет, являются одним из важнейших инструментов корпусной лингвистики. Их создатель берет на себя задачи по подготовке и хранению корпусных данных, снабжению их инструментами поиска и лингвистического анализа, предоставлению удобного интерфейса для доступа к этим инструментам, поддержания и своевременного обновления программного и аппаратного обеспечения, необходимого

[18] В настоящий момент разметка фрагментов моделью недоступна на корпусной платформе. Фрагментам приписана сложность всего текста, что является временным и упрощенным решением. Полноценная разметка фрагментов планируется к внедрению на корпусную платформу в первой половине 2025 г.

для функционирования системы. Все это существенно снижает порог вхождения для пользователей, позволяя им приступить к работе без развертывания корпуса и инструментов к нему на собственном персональном компьютере. Однако такой подход сужает спектр возможных исследований, поскольку позволяет применять к корпусным данным только набор инструментов, реализованный в интерфейсе веб-сайта. Частично преодолеть это ограничение можно за счет предоставления пользователю возможности автоматизировать выполнение запросов к корпусу. Благодаря такой возможности пользователь может строить сложные комбинации из многочисленных однотипных запросов, решая задачи, недоступные ему через интерфейс (например, повторить один и тот же запрос для тысячи различных лемм). Автоматизация обеспечивает меньше возможностей, чем развертывание корпуса и инструментов к нему на собственном компьютере, однако не требует существенных ресурсов на стороне пользователя и позволяет избежать проблемы с авторскими правами, возникающей при открытой публикации всех текстов корпуса.

Автоматизация запросов к корпусной платформе осуществляется при помощи программного интерфейса приложений (англ. Application Programming Interface, API). Разработанный для нужд НКРЯ API обеспечивает возможность выполнения произвольных запросов, доступных через интерфейс. Однако в настоящее время API не доступен стороннему пользователю, а используется только самим графическим интерфейсом системы. Такой подход позволил нам отделить реализацию интерфейса пользователя от непосредственно поискового сервера (поисковый запрос полностью формируется на стороне бэкенда и передается на сторону фронтенда при помощи API).

Таким образом, в настоящее время нельзя утверждать, что API решает задачу автоматизации. Однако такое использование API подтверждает его универсальность: поскольку любая операция с корпусом преобразуется интерфейсом в запрос к API, можно сделать вывод, что API обеспечивает все необходимые возможности.

После определенной доработки планируется сделать API общедоступным, а также разработать библиотеку на языке Python для реализации наиболее популярных сценариев использования корпуса через API. Выбор языка Python обусловлен его популярностью среди специалистов по компьютерной лингвистике.

В качестве базового протокола для передачи данных в рамках API мы использовали протокол сериализации структурированных данных Protocol Buffers[19] (ProtoBuf).

Можно выделить следующие виды запросов к API, соответствующие основным видам использования корпусной платформы:

1. запрос основных статистических данных о корпусах НКРЯ;
2. запрос данных о конфигурации конкретного корпуса: доступных в корпусе видах выдачи и их параметрах, возможных настройках отображения, сортировки и группировки выдачи;
3. запрос набора доступных в конкретном корпусе поисковых форм (как, например, поиск точных форм или поиск коллокаций), а также их состава — набора доступных для задания полей и их иерархии;
4. запрос типов атрибутов, имеющихся в корпусе, и возможных значений для тех атрибутов, которые выбираются из списка;
5. поисковый или аналитический запрос с указанием желаемого типа представления результата (например, конкорданс, KWIC или частотность) из числа доступных в корпусе;
6. запрос «Портрета» конкретного слова с указанием списка типов выдачи из числа доступных в корпусе;
7. запрос «Портрета» конкретного корпуса.

Для каждого вида запроса разработаны форматы сообщений, которыми обмениваются клиент и сервер. Использование протокола ProtoBuf позволяет автоматически проверять соответствие формата сообщения требуемому, что снижает вероятность ошибки.

[19] https://protobuf.dev/

2.3 Новая концепция интерфейса корпусной платформы, ориентированная на широкий круг пользователей

Как было показано во вступительной обзорной части, метод корпусного анализа вышел далеко за пределы собственно академических лингвистических исследований. НКРЯ активно используется исследователями из смежных гуманитарных дисциплин, преподавателями, писателями и переводчиками. Кроме того, поиск в Корпусе не всегда связан с профессиональной деятельностью пользователя; он может служить инструментом для удовлетворения общего интереса к русскому языку и культуре. Одной из ключевых задач при разработке новой корпусной платформы стало изменение дизайна и функциональности интерфейса таким образом, чтобы доступ к данным корпуса не требовал специальной лингвистической подготовки, устранялись существующие барьеры и Национальный корпус становился доступным для всех, кто проявляет к нему интерес.

Доступность корпусной платформы для широкого круга пользователей предполагает выполнение следующих требований:

- возможность использовать платформу с разнообразных устройств: настольного компьютера, планшета, смартфона;
- поддержка иноязычных пользователей, использующих Корпус для изучения русского языка;
- возможность получения всего имеющегося спектра информации по минимальному запросу;
- визуальная наглядность отображаемых результатов;
- уменьшение количества лишних действий пользователя;
- развитая система контекстных подсказок и руководств и анонсов, помогающая пользователю ориентироваться в программной системе.

Рассмотрим, как новый интерфейс корпусной платформы решает перечисленные задачи.

2.3.1 Адаптация интерфейса под пользовательские устройства

По данным SimilarWeb[20], на январь 2025 около 70% пользователей интернета пользуются им с мобильных устройств: смартфонов или планшетов. Доля мобильных пользователей постоянно растет. Не являются исключением и пользователи НКРЯ.

Перед разработчиками платформы нового поколения стояла задача, с одной стороны, обеспечить аналогичность интерфейса при доступе с различных устройств, а с другой — учитывать особенности каждого вида устройств при построении интерфейса для них. Для этого новая версия сайта поддерживает набор стилевых таблиц (в соответствии с технологией CSS) для различных размеров экрана, соответствующих наиболее распространенным классам мобильных и настольных устройств — от самого миниатюрного смартфона до настольного компьютера с хорошим разрешением экрана. Разработанное веб-приложение реализует подход «первичности мобильной версии» (*англ.* mobile first), в соответствии с которым самые первыми в списке вариантов стилевых таблиц находятся таблицы для самых миниатюрных мобильных устройств (обладающих не только самыми маленькими размерами экрана, но и наименьшими вычислительными ресурсами). Это позволяет им прекратить ресурсоемкий для них дальнейший перебор вариантов. Пользователю автоматически открывается версия, наиболее подходящая для размера устройства, с которого он выходит в интернет (Рис. 1).

[20] https://www.similarweb.com/ru/platforms/

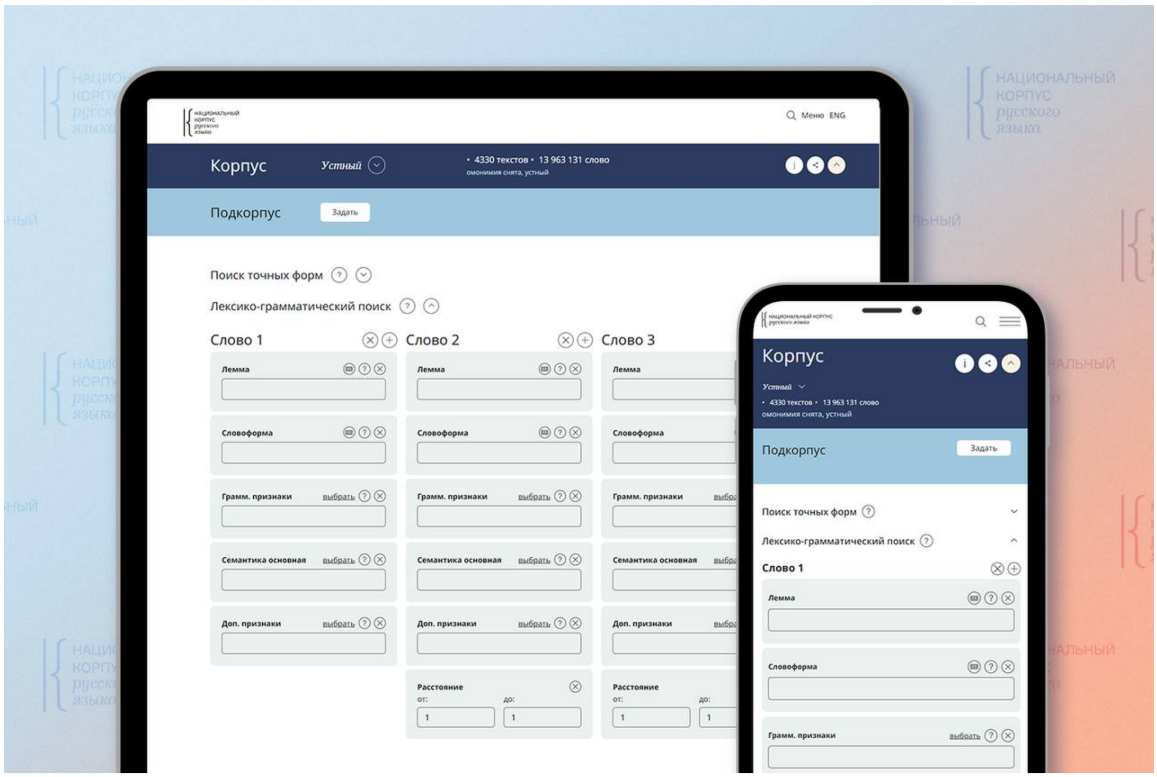

Рис. 1: Слева — версия сайта для ПК и планшета, справа — оптимизированная для мобильных устройств.

2.3.2 Поддержка иноязычных пользователей

Иностранные пользователи, использующие Корпус для изучения и исследования русского языка, заинтересованы в получении примеров и контекстов на русском языке. В то же время им может быть сложно выбрать настройки и другие параметры поиска, пользуясь русскоязычной терминологией. Поэтому для пользователей сайта, основным языком которых не является русский, реализована возможность переключить язык интерфейса на английский.

2.3.3 Получения всего имеющегося спектра информации по минимальному запросу

В то время как для опытных пользователей, хорошо знакомых с функционалом системы, естественно максимально конкретизировать запрос, заранее понимая, в каком виде ожидаются результаты, новые пользователи Корпуса, а также пользователи, недостаточно понимающие весь функционал системы, предпочитают наглядно увидеть все возможные результаты по своему запросу, чтобы выбрать, какие из них им интересны. Это позволяет снять барьер на вхождение и дать широкой аудитории инструменты для решения собственных задач. Для профессиональной аудитории привычно и ценно самостоятельно строить сложные поисковые запросы, анализировать большое количество данных, иметь возможность дополнительно обработать выдачу и сделать собственные научные выводы. В то же время у новой аудитории задача часто другая — быстро получить простой ответ на свой вопрос. Благодаря модульности и расширяемости платформы, нам удается на основе одного и того же внутреннего инструментария строить разный интерфейс для профессионалов и широкой аудитории.

Поскольку все содержательные лингвистические операции (поиск и анализ результатов) выполняются интерфейсом не напрямую, а через API, открывается возможность использовать одни и те же вызовы API в различных местах интерфейса. Приведем несколько примеров.

В 2022 году в НКРЯ появился сервис «Обзор возможностей», который дает пользователям представление о ключевых возможностях, доступных в НКРЯ, знакомит с общими принципами

устройства интерфейса, показывает, какие виды результатов можно получить, информирует о типичных ошибках при конструировании поисковых запросов. При этом используются те же самые вызовы API, что и при соответствующих запросах в основном интерфейсе этих корпусов.

Сервис «Портрет слова» также позволяет не конструировать несколько запросов к поисковому и другому функционалу и затем самостоятельно комбинировать их результаты, а делает это автоматически. Пользователю нужно лишь ввести начальную форму слова, после чего в визуально компактной и понятной форме сервис представит разнообразную информацию для всех имеющихся разборов заданной леммы. Пользователь увидит скетчи слова (как список коллокаций для основных синтаксических отношений), все грамматические формы слова (без необходимости искать и сравнивать разборы в разных примерах) и так далее.

Из «Портрета слова» налажен переход в полный функционал поиска и наоборот, из результатов поиска, кликнув на разбор любого слова, можно перейти к его «Портрету». Такие перекрестные ссылки, позволяющие пользователю переходить из сервиса в сервис, не теряя контекст своего запроса, — это еще один пример того, как новый интерфейс помогает пользователям осваивать возможности корпуса.

2.3.4 Визуальная наглядность отображаемых результатов

Графическое представление результатов дает возможность быстро и эффективно донести до пользователя сложную для восприятия информацию. Так, например, барометр частотности в «Портрете слова» позволяет одним взглядом оценить, насколько частотной является лемма.

С помощью круговых, столбчатых диаграмм, географических карт и графиков в интерфейсе «Портретов корпусов» НКРЯ подается информация о структуре и составе корпусов. В «Портретах подкорпусов» можно с помощью сравнительных диаграмм проанализировать, насколько пользовательский подкорпус отличается от корпуса в целом.

Для того чтобы экспертная аудитория могла делать более глубокие и обоснованные научные выводы, важно учитывать особенности механизмов, с помощью которых проводились расчеты. Принципиально новый подход, реализованный в новом интерфейсе, состоит в том, что ограничения примененных методов анализа данных не только описываются в руководстве пользователя, но и визуализируются сразу при выдаче. Так, в виде выдачи «Частотность» показываются доверительные интервалы для рассчитанной частотности (Рис. 2). При отображении графиков указываются временные границы, за пределами которых данных слишком мало для достоверных выводов. Под графиком отображается тепловая шкала, описывающая количество текстов, в которых найдены результаты, в разные периоды времени, позволяющая оценить, насколько рост частотности слова является случайным выбросом в конкретном тексте или же объективно наблюдаемым явлением (Рис. 3).

Слово 1 Словоформа	← [1..2] →	Слово 2 Лемма	Вхождения	Доля	Доля	ipm	Конкорданс
какая	1	разница	2	50%	[15%, 85%]	28.21	Примеры
какая	1	жалость	1	25%	[4.56%, 69.94%]	14.11	Примеры

Рис. 2: Вид выдачи «Частотность»

Рис. 3: Вид выдачи «График»

Еще одним примером доступной визуализации является облако похожих слов, реализованное в функционале «Портрета слова». «Похожие слова» отображаются в виде облака тегов, в котором размер букв и удаленность слов друг от друга характеризуют степень близости контекстов употребления слов (Рис. 4). Для морфемного разбора использована наглядная нотация, принятая в школьном преподавании русского языка (Рис. 5). Такие визуализации стали возможны благодаря внедрению инструментов статистического анализа и портретирования (см. разделы 3.1 и 3.2).

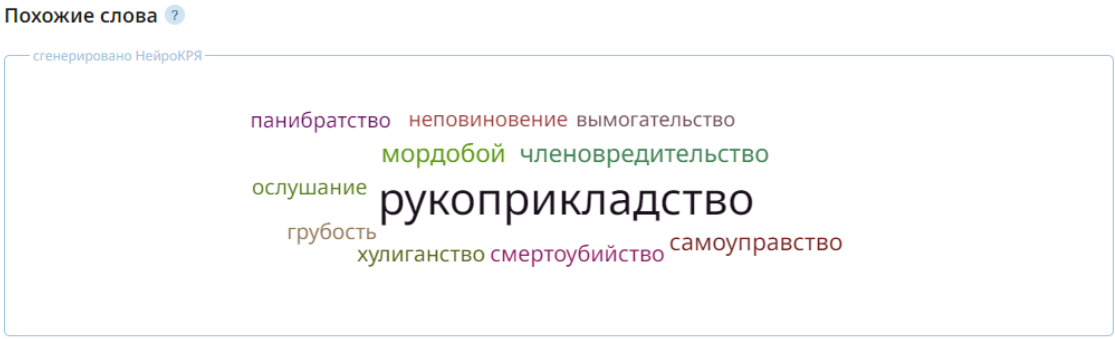

Рис. 4: Виджет «Похожие слова»

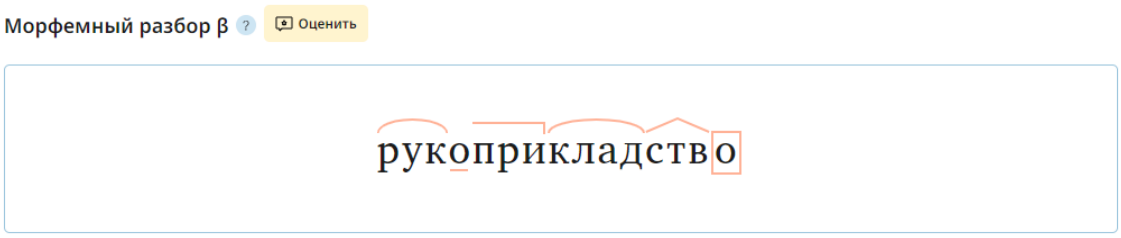

Рис. 5: Виджет «Морфемный разбор»

В нескольких сценариях поиска по Корпусу элементы интерфейса были намеренно перегруппированы по сравнению со старой версией корпусной платформы. Особенно заметна перегруппировка в интерфейсе задания условий лексико-грамматического поиска. Группы условий на искомые слова в словосочетании теперь расположены в одну строку слева направо (Рис. 6 и 7).

Такой подход визуально более интуитивен для пользователей, поскольку в тексте слова также обычно располагаются на одной строке друг за другом.

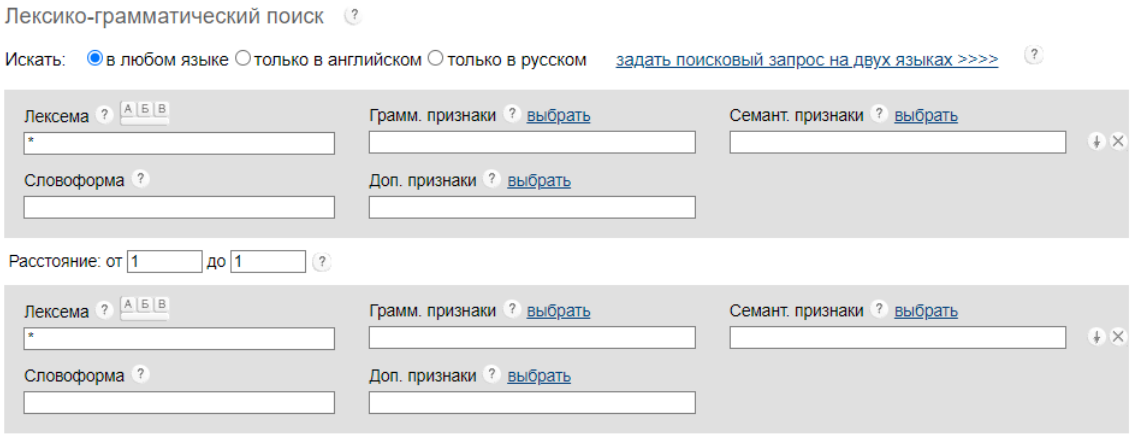

Рис. 6: Старое расположение условий лексико-грамматического поиска

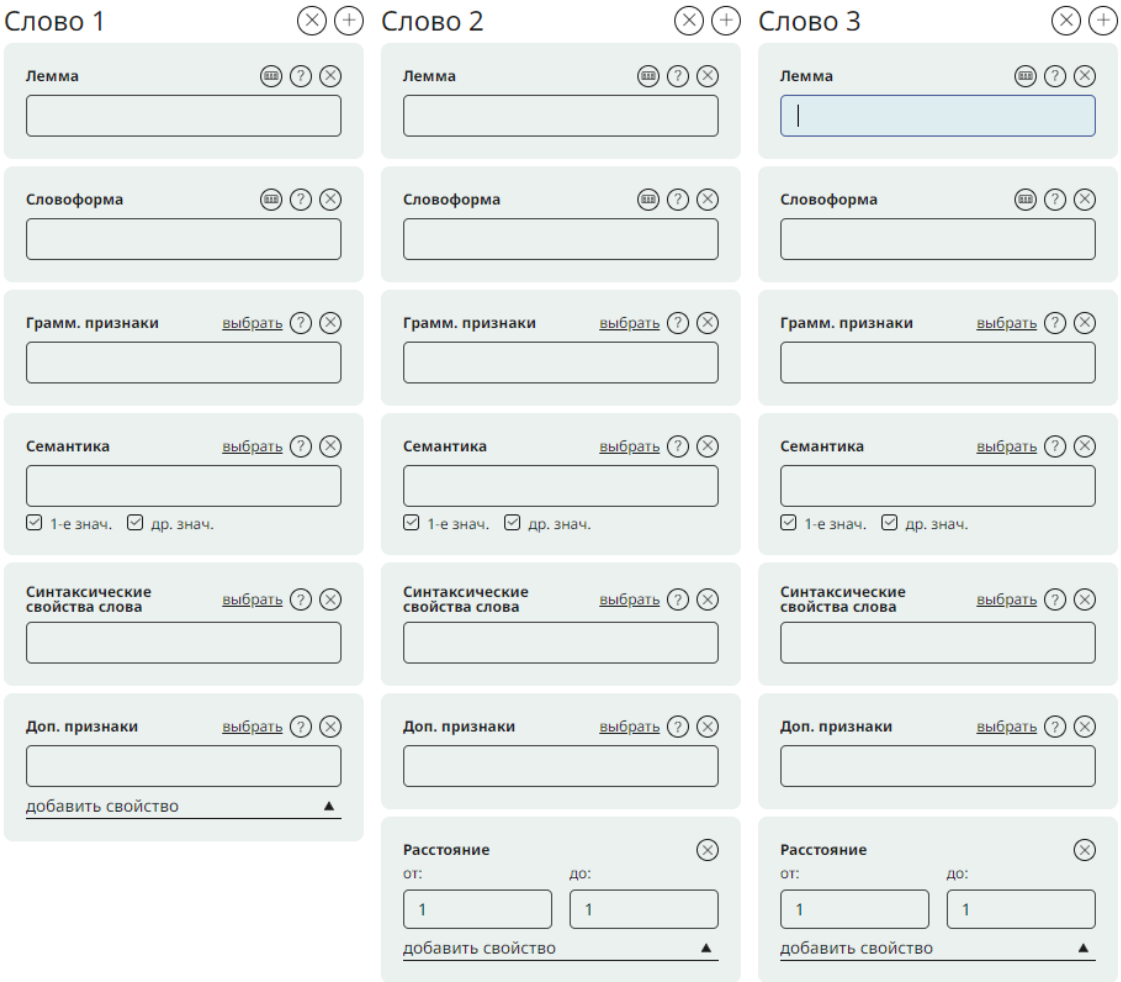

Рис. 7: Новое расположение условий лексико-грамматического поиска

Перегруппировка затронула и отображение результатов поиска в параллельных корпусах (Рис. 8 и 9). Теперь оригинальный фрагмент располагается слева, а переводы справа (можно переключаться между разными переводами). Это позволяет разместить на одном экране ПК больше примеров. Для мобильных устройств реализовано переключение с помощью слайдера, что более привычно для пользователей смартфонов.

Рис. 8: Старое расположение результатов поиска в параллельных корпусах

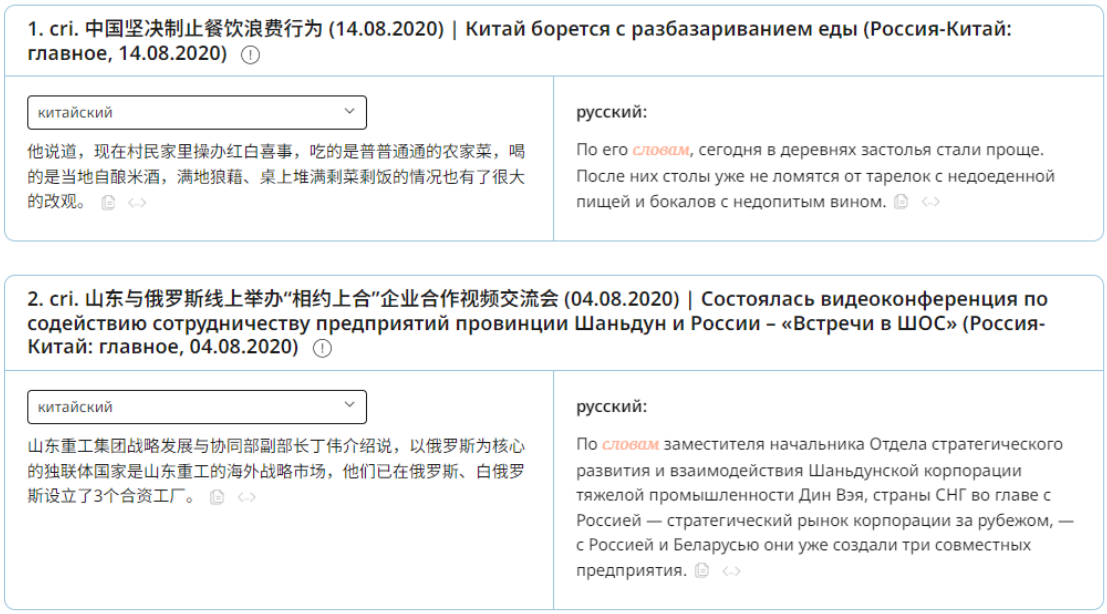

Рис. 9: Новое расположение результатов поиска в параллельных корпусах

2.3.5 Уменьшение количества лишних действий пользователя

Любой пользователь, вне зависимости от квалификации, совершает ряд действий в интерфейсе системы при каждом обращении к корпусу. Такие действия не должны отнимать много усилий, а напротив, должны быть максимально быстрыми и очевидными.

В новом интерфейсе непосредственно с главной страницы организован доступ к поиску по любому из корпусов, а также выведены ссылки для доступа к другому функционалу, который часто используется.

Существенная экономия достигается за счет отказа от перезагрузки веб-страницы при каждой операции в интерфейсе. Определенные достижения в этом направлении были уже в старом интерфейсе: некоторые сложные операции в нем выполнялись без перезагрузки страницы. Новый интерфейс изначально спроектирован так, чтобы уменьшить количество перезагрузок страниц.

Частичное изменение содержимого страницы в нем реализуется через асинхронный запрос к серверу с последующим переписыванием части страницы после получения ответа. Обработчики таких запросов на стороне сервера легковесны благодаря использованию таких же асинхронных запросов к API (см. раздел 2.2).

Все возможные на сегодняшний день пользовательские настройки поисковой выдачи собраны в едином меню «Настройки», что позволяет пользователю быстро их найти.

Выбор настроек запоминается в браузере пользователя и применяется для следующих поисковых запросов. Аналогично запоминаются предпочтительный вид поиска в каждом корпусе (соответствующая форма поиска будет всегда показываться открытой), вид выдачи, который будет открываться по умолчанию, а также режим отображения (или скрытия) подробной информации о запросе в шапке корпуса.

При постоянной работе с поиском по корпусу для удержания контекста важно всегда иметь перед глазами запрос, с которым работаешь. Полезным нововведением является отображение в шапке корпуса не только информации о параметрах искомого слова, но и о параметрах заданного пользователем подкорпуса. Возможность в любой момент вернуться к форме и откорректировать любые параметры сокращает усилия в сравнении с заданием параметров заново.

Для обмена результатами исследований, в том числе при публикации в научных журналах, теперь можно воспользоваться короткими ссылками на запрос и кнопкой «Скопировать пример», с помощью которой в буфер обмена помещается информация о примере и его выходных данных.

2.3.6 Система контекстных подсказок и анонсов, руководство пользователя

Для того чтобы менее подготовленные пользователи могли быстрее научиться пользоваться новым функционалом, в новом интерфейсе поддерживается регулярно обновляемое руководство пользователя, доступен поиск по руководству.

В разделе «Совет дня» «Обзора возможностей» регулярно размещается подробная информация о наиболее интересных нововведениях.

Актуальность вспомогательной информации, размещенной на сайте Корпуса, поддерживается с помощью системы управления контентом, которая позволяет в онлайн режиме структурировать, тегировать и редактировать онлайн анонсы, статьи, контекстные подсказки и руководство пользователя.

Корпуса НКРЯ различаются по данным, типам разметки и функциональным возможностям запросов и анализа. Тем не менее, общая концепция интерфейса, ориентированная на удобство, доступность и прозрачность взаимодействия пользователя с сервисами, сохраняется для всех корпусов. Важным является единый стандарт интерфейса, применяемый во всех корпусах, который обеспечивает для пользователя интуитивную легкость перехода между ними и расширяет доступность специализированных ресурсов. Этот аспект особенно ярко проявляется в новых сервисах анализа данных корпуса, например, таких как «Портрет слова». Хотя часть инструментов еще доступна не для всех корпусов, унифицированный шаблон помогает пользователю ориентироваться в информации и одновременно включает в себя перспективы дальнейшего развития сервисов в специализированных корпусах. Описание технологических аспектов, лежащих в основе развития новых сервисов, будет представлено в следующем разделе.

3 Инструменты анализа корпусных данных

3.1 Основные направления развития статистико-аналитической компоненты НКРЯ

Современные методы корпусной лингвистики в наибольшей степени ориентированы на использование количественного анализа распределения языковых единиц. Именно поэтому, как показано в разделе 1.1.3, современные корпусные платформы не ограничиваются лишь конкордансами для отображения словоупотреблений, но включают дополнительные сервисы, которые позволяют систематизировать, обобщать и статистически оценивать результаты анализа корпусных данных. Инструменты квантитативного корпусного анализа могут быть применены уже на этапе поиска, как, например, при поиске по коллокациям или же использоваться для дополнительного исследования результатов поискового запроса. Кроме того, квантитативный анализ может быть

проведен для всего корпуса в целом или для выбранного подкорпуса. В данном разделе мы рассмотрим категории аналитических инструментов, которые вошли в лингвистическое ядро программной платформы НКРЯ нового поколения.

3.1.1 Инструменты статистической характеризации корпусов и подкорпусов

Статистические инструменты этого типа позволяют строить портрет корпуса и подкорпуса, получать статистическое распределение текстов по значениям мета-атрибутов и строить диахронические графики.

Поскольку корпусная платформа позволяет рассматривать произвольный набор текстов, заданный пользовательскими условиями, как подкорпус, количество гипотетически возможных подкорпусов очень велико. А значит, статистические характеристики подкорпусов не могут быть предварительно рассчитаны на этапе индексации. Все расчеты выполняются вычислительным ядром непосредственно при обработке запроса. Это возможно, поскольку для осуществления расчетов не требуются сами тексты, достаточно лишь их заголовков и метаданных, что существенно снижает объем обрабатываемой информации.

В то же время, статистическая информация о корпусе в целом может быть для ускорения вычислена и сохранена в момент индексации. Если в будущем будут выявлены фиксированные подкорпусы, статистика по которым востребована существенно чаще остальных, то в целях ускорения вычисление статистик по ним также может быть перенесено на этап индексации.

3.1.2 Инструменты статистической характеризации результатов поиска словосочетаний

Статистические характеристики этого типа позволяют строить распределение, удовлетворяющее поисковым условиям словоформ и лемм и получать наиболее распространенные в результатах поиска n-граммы. Также этот инструмент используется для вывода информации о частотности и формах в «Портрете слова».

Это наиболее ресурсоемкие вычисления, выполняемые непосредственно в момент запроса, поскольку количество результатов поискового запроса может быть очень велико, а в экстремальных случаях — даже превышать размер самого корпуса (в случае, если одно и то же слово входит в несколько разных словосочетаний, удовлетворяющих условию поиска). При отображении результатов поиска вычисления могут быть прекращены, как только нужное количество примеров сформировано, но при подсчете статистики должны быть учтены все результаты или их репрезентативная подвыборка. Так, в случае если количество результатов поиска превышает миллион, в качестве такой выборки рассматривается случайное подмножество в миллион результатов и производится расчет относительных показателей только на основе них. При этом абсолютные показатели получаются из относительных нормированием на полное количество результатов поиска. Как использование для вычислений случайной подвыборки, так и малый размер всей выборки могут вести к существенному падению точности получаемых результатов, поэтому система рассчитывает статистические доверительные интервалы, которые отображаются напротив вычисленных значений.

Для ускорения вычислений в оперативной памяти поддерживается полный индекс отдельных атрибутов слов. Это существенно повышает требования к аппаратному обеспечению (на каждый атрибут требуется несколько гигабайт оперативной памяти), но резко ускоряет расчеты.

Также в будущем возможно применение кэширования результатов расчетов. Это может значительно увеличить скорость в ситуации, когда пользователи, переключаясь, делают одинаковый запрос несколько раз за небольшой промежуток времени (например, переключаются туда-обратно между разными экранами в интерфейсе).

Особой подкатегорией этой категории является статистика по метаатрибутам примеров, встретившихся в выборке. Фактически, это статистика по подкорпусу текстов, отобранных по запросу, при этом каждый текст в ней участвует с весом, равным числу найденных в нем вхождений.

3.1.3 Инструменты статистической характеризации лемм

Инструменты этой категории позволяют, например, находить похожие слова, то есть слова, встречающиеся в одинаковых контекстах. В виджете «Похожие слова» отображаются ближайшие семантические ассоциаты слова; коэффициент близости слов подсчитывается с помощью моделей

дистрибутивной семантики, построенных на актуальных материалах основного корпуса НКРЯ (см. об этом подробнее в разделе 3.3). Вычисления характеристик этого типа требует предрасчетов на этапе индексации текстов с сохранением информации, привязанной к каждой лемме. В момент пользовательского запроса происходит статистический расчет на основе сохраненной информации, а не самих текстов корпуса, что критично снижает вычислительную сложность.

3.1.4 Статистические коллокации

Статистические коллокации для произвольного запроса могут быть вычислены только в момент пользовательского запроса, но скетчи (заранее фиксированные для каждой части речи наборы коллокаций с учетом синтаксических связей) эффективнее вычислять на этапе индексации и сохранять для каждой леммы, поскольку количество лемм в корпусе ограничено и составляет не более нескольких сотен тысяч (с порогом встречаемости хотя бы 3 раза в 3 различных текстах). Для остальных лемм такая информация не сохраняется.

Таким образом, в вычислительном ядре алгоритмы реализации указанных статистических инструментов подразделяются на:

- алгоритмы, выполняемые разово в процессе индексации: результат работы такого алгоритма сохраняется в базе данных и готов к использованию при обработке запроса пользователя;
- алгоритмы, выполняемые в процессе пользовательского запроса на основе текстов корпуса и/или показателей, вычисленных и сохраненных в процессе индексации;
- алгоритмы, выполняемые в процессе пользовательского запроса на основании случайного подмножества результатов поиска. Результаты работы такого алгоритма являются приблизительными, поэтому они применяются в случае, когда точное вычисление в процессе запроса невозможно из-за ограничений на время ожидания.

В целом, можно заключить, что программная платформа НКРЯ нового поколения реализует широкий спектр аналитических инструментов обработки корпусных данных. Эффективность их реализации обеспечена за счет предварительных вычислений на этапе индексации текстов, использования (при необходимости) приближенных вычислений по рандомизированной подвыборке и эффективных по времени доступа, но затратных по памяти механизмов кэширования.

В рамках рассмотрения аналитических инструментов новой корпусной платформы НКРЯ мы уделим особое внимание двум нейросетевым моделям, которые были разработаны специально для сервиса «Портрет слова». Первая — модель словообразовательного разбора в «Портрете слова». Информация о внутренней структуре слова, получаемая в результате работы модели позволяет не только отобразить его морфемный состав, но связать слово с однокоренными словами. Пользователь может, таким образом, с помощью одного клика перейти с портрета исходного слова на портрет его однокоренного. Семантическая информация о слове обеспечивается с помощью векторной модели дистрибутивной семантики, подсчитанной для разных корпусов. Виджет «Похожих слов» таким образом обеспечивает связность исходного слова и его квазисинонимов и ассоциатов. Такой переход также реализован в интерфейсе. Ниже будет рассмотрена каждая из моделей.

3.2 Модель словообразовательного разбора

Помимо широко используемых во многих корпусах видов разметки, таких как, например, разметка морфологических свойств слов, в НКРЯ встречается и специализированная разметка. Одним из видов такой разметки является словообразовательная, то есть разметка морфем, из которых состоит слово, и их типов. Словообразовательная разметка востребована как для лингвистических исследований (Гришина и др., 2009), так и для обучения русскому языку. Многие из орфографических правил, вызывающих наибольшие затруднения, связаны именно с морфемной структурой слова — например, правописание безударных гласных в корне или правописание приставок[21]. В НКРЯ словообразовательная разметка присутствует в двух корпусах: Основном и Обучающем. Важно отметить, что в НКРЯ разметка применяется к лемме слова без учета формы

[21] https://yandex.ru/company/researches/2016/ya_spelling

или контекста конкретного словоупотребления. С точки зрения архитектуры системы морфемный разбор является атрибутом, приписанным структурной единице «разбор». Для каждого разбора указан список морфем, их тип (приставка, корень, интерфикс, суффикс, окончание или постфикс) и линейная позиция в слове.

В основе разметки словообразовательной структуры в Основном корпусе лежит специально разработанный для корпуса словарь морфемного анализа **Morphodict-K**, где по состоянию на май 2023 года даны разборы для 75 тыс. лексем (310 тысяч неуникальных морфем). Этот словарь составлялся на основании идеологии «Словаря морфем русского языка» А. И. Кузнецовой и Т. Ф. Ефремовой (Кузнецова, Ефремова 1986). Принципы этой идеологии — значительная (хотя и не максимальная) дробность выделения морфем и соотносимость с другими лексемами аналогичного строения. Поэтому морфемное деление в разметке корпуса не совпадает с принятым, например, в школе. В исконных словах могут выделяться морфемы, даже если слово без них употребляется маргинально (*у-лыб-а-ть-ся*, ср. *у-смех-а-ть-ся*) или если мотивированность этимологии слова для современного носителя неочевидна (*на-сек-ом-ое*, *вос-точ-н-ый*). В иностранных словах заимствованные основы членятся (например, *ре-волюц-и-я*, *квит-анци-я*), если усматривается семантическое и структурное соответствие между ними и лексемами похожего строения (ср. *э-волюц-и-я*, *рас-квит-а-ть-ся*). Разбираются в том числе и служебные части речи, а также имена собственные и производные от них.

Разметка морфем в Обучающем корпусе опирается на разработанный на основе «Морфемно-орфографического словаря» А. Н. Тихонова (Тихонов 2002), словарь морфемного анализа **Morphodict-T**. Этот словарь содержит около 100 тыс. лексем. Морфемный состав слова в Morphodict-T определяется в соответствии с практикой морфемного анализа в средней школе. При этом используется более жесткий подход к определению того, какие смысловые связи являются прозрачными в современном языке, и, как правило, выделяется меньшее число морфем, чем в Основном корпусе: например, указанные выше слова анализируются как *улыб-а-ть-ся*, *насеком-ое*, *восточ-н-ый*, *революци-я*, *квитанци-я*. В «Портрете слова», представленном в Обучающем корпусе, дается морфемное строение только слов, относящихся к знаменательным частям речи, — нарицательным существительным, прилагательным, глаголам и наречиям.

На Рис. 10 и Рис. 11 видна разница между морфемным разбором в Основном корпусе, построенном на основе словаря Тихонова, и разбором в Обучающем корпусе, построенном на основе словаря Кузнецовой и Ефремовой, в виджетах «Портрета слова» этих корпусов.

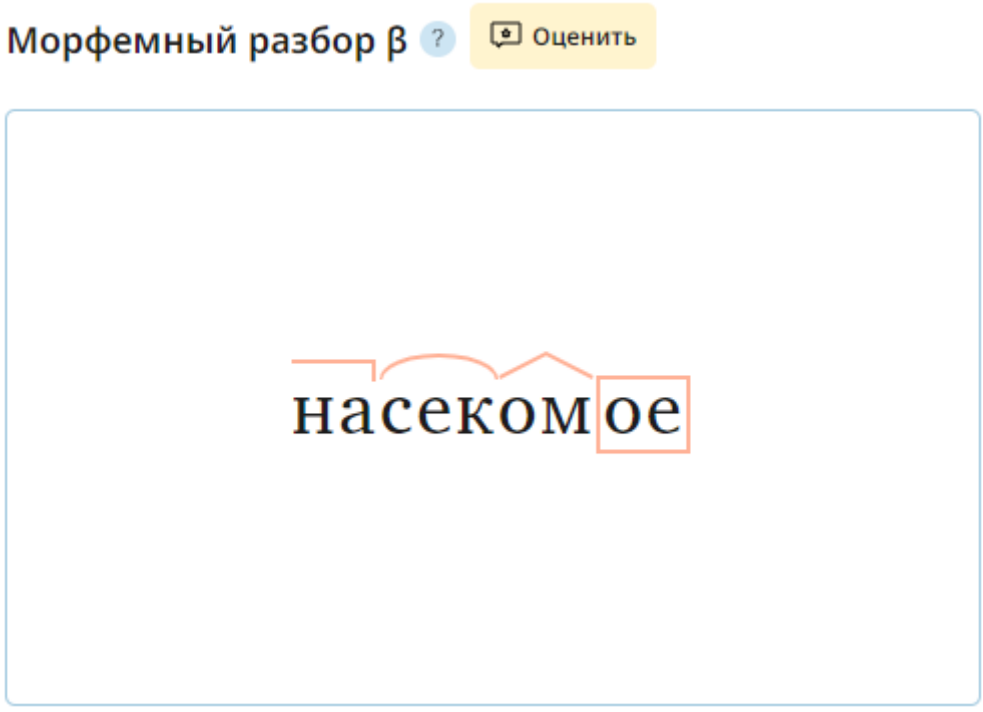

Рис. 10: Основной корпус. Морфемный разбор на основе словаря Кузнецовой и Ефремовой

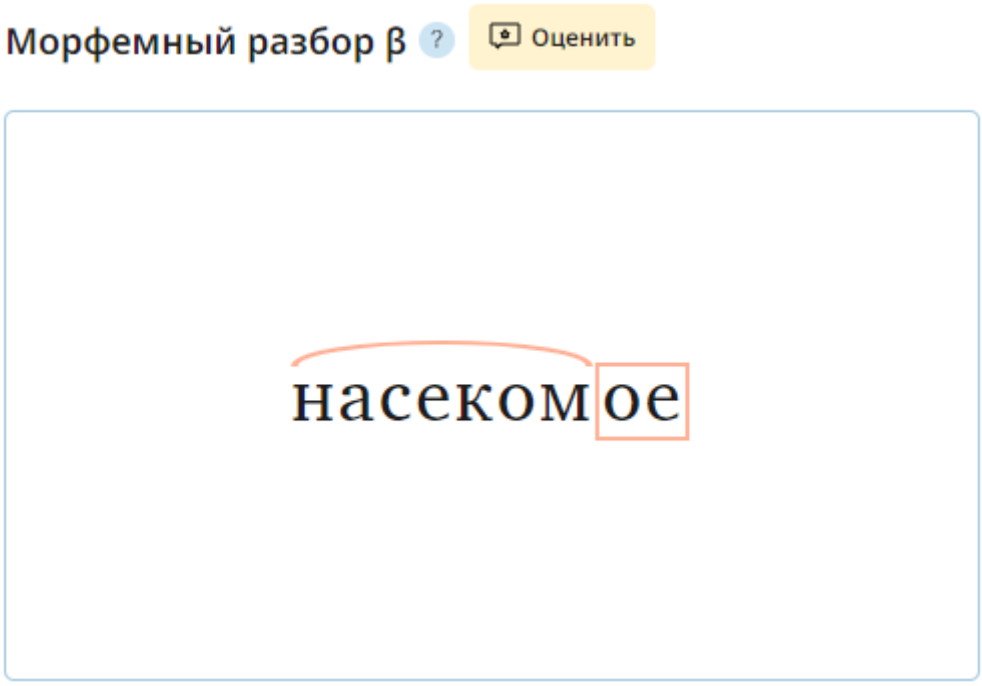

Рис. 11: Обучающий корпус. Морфемный разбор на основе словаря Тихонова

Из описанного выше следуют две основные проблемы словообразовательной разметки в НКРЯ. Во-первых, словари Morphodict-K и Morphodict-T сравнительно малы по отношению к многообразию всех лемм Основного и Обучающего корпусов. В совокупности в этих корпусах содержится более 300 тыс. уникальных лексем (с порогом встречаемости хотя бы 3 раза в 3 различных текстах), то есть имеющаяся ручная разметка далека от полноты. Во-вторых, существование различных, противоречащих друг другу подходов к морфемному членению заметно осложняет автоматическое пополнение словарей. При этом нельзя утверждать, что авторы конкретного словаря строго придерживаются единого для этого словаря алгоритма морфемного членения; в частных случаях применяются локальные решения, не удовлетворяющие описанному алгоритму (Иомдин 2019).

Несмотря на описанные проблемы, выборку достаточного размера можно использовать для обучения алгоритмов автоматического морфемного анализа. Как и во многих других областях обработки естественного языка, использование методов машинного обучения может обеспечивать высокое качество результатов. Так, в 2018 году был представлен алгоритм генерации морфемных разборов на базе ансамбля сверточных нейронных сетей (Sorokin, Kravtsova 2018), который показал значительный прирост качества по сравнению с ранее существовавшими алгоритмами построения морфемных разборов. Мы проанализировали качество работы предложенного алгоритма при его обучении на словарях Morphodict-T и Morphodict-K при помощи кросс-валидации по пяти выборкам и пришли к выводу, что модель, обученная на Morphodict-K, показывает значительно более высокие результаты. Особенно различается качество автоматической разметки по доле полностью верных разборов, что может свидетельствовать о более высокой внутренней согласованности данных в словаре Morphodict-K. Все полученные результаты приведены в Таблице 2.

Метрика	Morphodict-T	Morphodict-K
F-мера для границ морфем	98,09	98,66
Точность (precision) для границ морфем	97,79	98,58
Полнота (recall) для границ морфем	98,38	98,74

Метрика	Morphodict-T	Morphodict-K
Доля слов с верно определенными границами морфем (без учета их типа)	96,61	97,40
Доля полностью верных разборов	88,49	90,82

Таблица 2: Сравнение доли верных разборов в Morphodict-T и Morphodict-K

В то же время использованная нами модель не лишена недостатков. Исследование (Garipov, Morozov, Glazkova 2023) показало значительное снижение качества при тестировании на словах, содержащих корни, не встретившиеся в обучающей выборке, что может свидетельствовать о недостаточной обобщающей способности алгоритма. В дальнейшем мы планируем подробнее изучить возможные способы устранения этого недостатка.

В настоящий момент модель, полученная в результате обучения на Morphodict-K, интегрирована в сервис «Портрет слова» в Основном корпусе НКРЯ: из 314 935 различных лемм, представленных в сервисе, для 255 821 леммы разбор сгенерирован моделью (в интерфейсе корпуса это отображается пометой «сгенерировано НейроКРЯ» на рамке виджета). На сегодняшний день мы собираем отклики пользователей о качестве генерации и готовимся включить доработанную модель в поисковые возможности Основного корпуса. Параллельно с этим мы изучаем возможность добавления автоматических разборов и в Обучающий корпус.

Обе модели размещены в открытом доступе и доступны для выгрузки в разделе «Нейросетевые модели» на сайте НКРЯ.[22]

На Рис. 12 продемонстрирован разбор слова *эстетика* в Основном корпусе, порожденный моделью НейроКРЯ (происхождение разбора указано в рамке виджета).

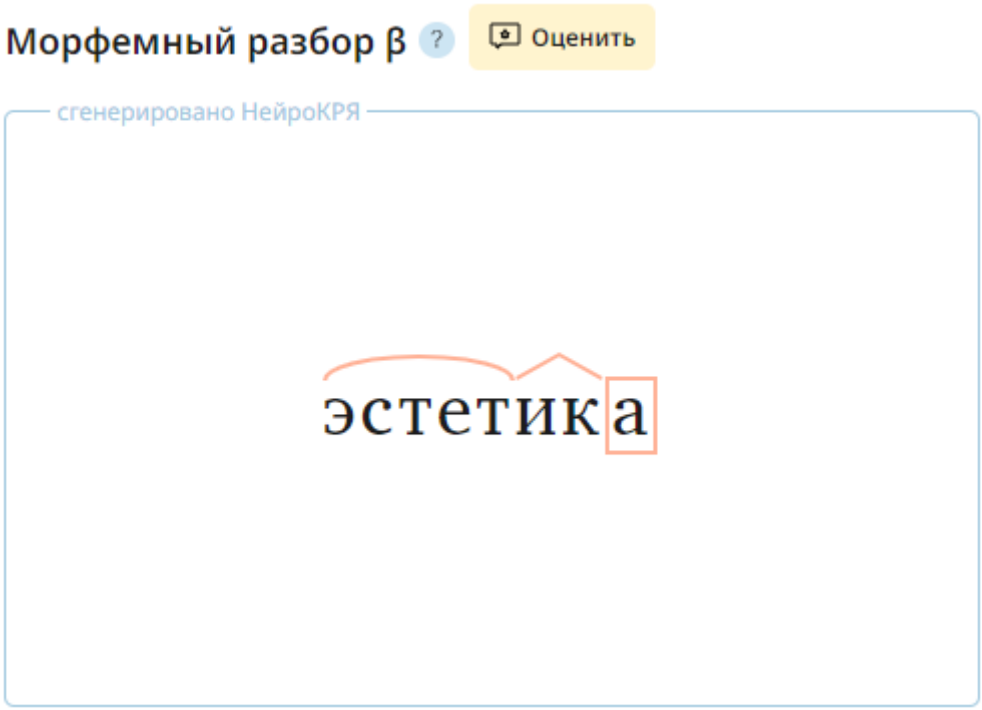

Рис. 12: Основной корпус: морфемный разбор, сгенерированный НейроКРЯ

[22] https://ruscorpora.ru/license-content/neuromodels

3.3 Векторные модели в «Портрете слова» (сервис «Похожие слова»)

Так как НКРЯ содержит весьма разнообразные по домену (типу, тематике, жанру и т.д.) и времени создания корпусы, одни и те же слова могут употребляться в них в несовпадающих значениях и наборах контекстов. Для того чтобы обнаружить и визуализировать особенности использования слов в различных корпусах, могут быть использованы модели векторного представления слов.

Использование для представления слов и текстов многомерных векторов повсеместно встречается в обработке естественного языка. В таких задачах, как классификация текстов, внутритекстовая разметка, генерация текста, конвертация слов в наборы чисел (многомерные вектора) является первым этапом работы с текстом. Наиболее простые алгоритмы получения таких представлений, например, one-hot кодирование, фактически никак не учитывают семантику кодируемого, тогда как современные языковые модели, например, BERT (Devlin et al., 2019), опираются не только на семантику кодируемого слова, но и на конкретный контекст его употребления. Промежуточным звеном между этими подходами являются различные алгоритмы построения статических эмбеддингов, например, CBoW и Skip-gram (Rehurek, Sojka 2011). Эти алгоритмы опираются на дистрибутивную гипотезу: предполагается, что слова, регулярно употребляющиеся в похожих контекстах, имеют схожую семантику. Полученные таким образом векторные представления позволяют, в том числе, оценивать семантическую схожесть между словами через косинусное расстояние между их векторами, что, в свою очередь, может быть использовано для поиска слов-ассоциатов.

Мы построили модели семантических векторов для существительных, глаголов, прилагательных и наречий для Основного, Обучающего, Газетных корпусов, Древнерусского, Старорусского, корпуса «От 2 до 15» и корпуса «Русская классика». При построении модели использовался алгоритм CBOW. Статические векторные модели на базе Основного корпуса строились и раньше, например, в работе (Kutuzov, Kuzmenko 2017). Однако в ходе нашей работы для Основного корпуса была обучена модель с использованием лемм, сгенерированных моделью RuBic (подробнее о модели см. раздел 4.1). Это позволило существенно уменьшить количество ошибочно сгенерированных или неправильно токенизированных лемм среди похожих слов. Сравнивая похожие слова для Основного и Старорусского корпусов, можно отслеживать семантические дрейфы (изменения значений слов со временем), а сравнивая Обучающий, корпус «Русская классика» и корпус Центральных СМИ, можно исследовать, например, журналистские штампы или употребление слов в разных типах дискурса. Приведем примеры, иллюстрирующие разницу ассоциатов для слов *игра*, *нива* и *трубить* в разных корпусах (Рис. 13–18):

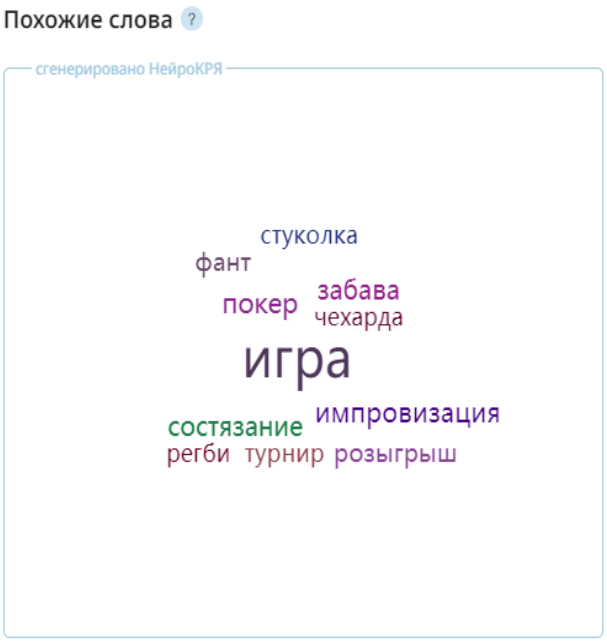

Рис. 13: Похожие слова для слова *игра* в Основном корпусе

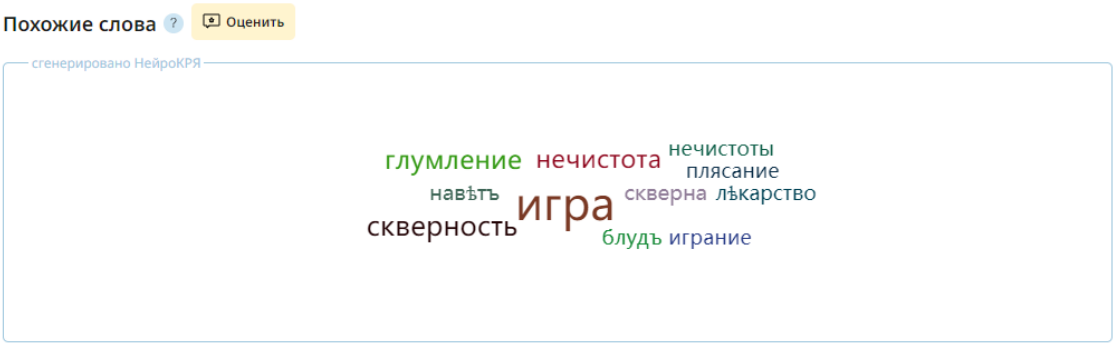

Рис. 14: Похожие слова для слова *игра* в Старорусском корпусе

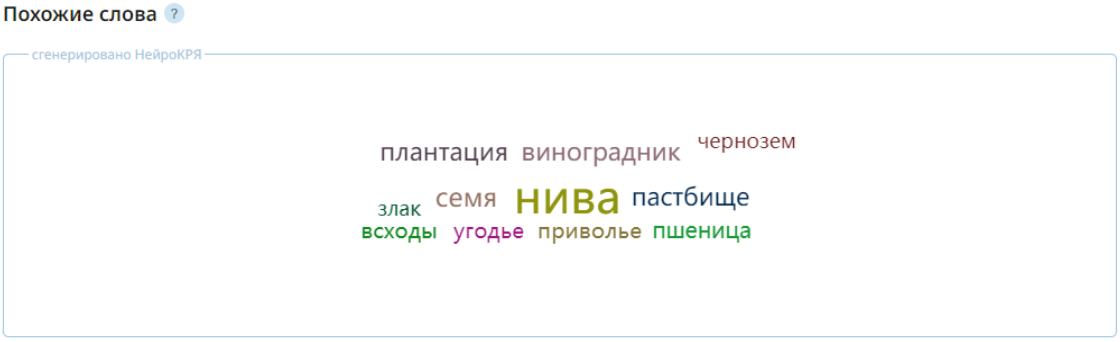

Рис. 15: Похожие слова для слова *нива* в Обучающем корпусе

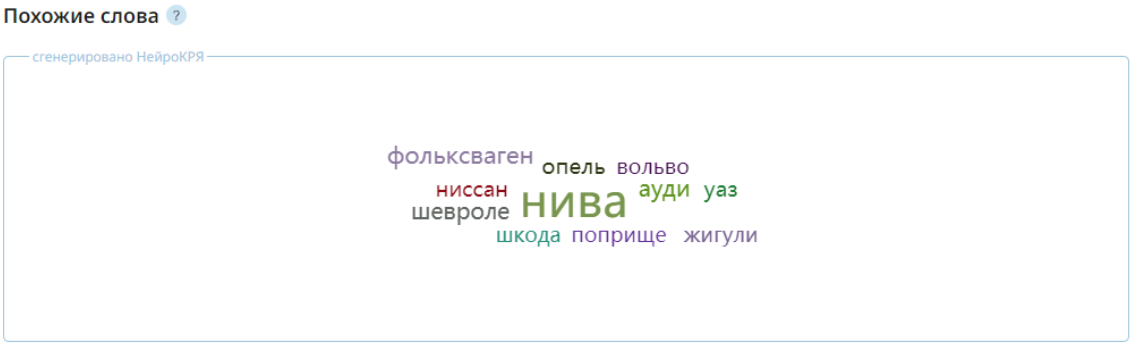

Рис. 16: Похожие слова для слова *нива* в корпусе Центральных СМИ

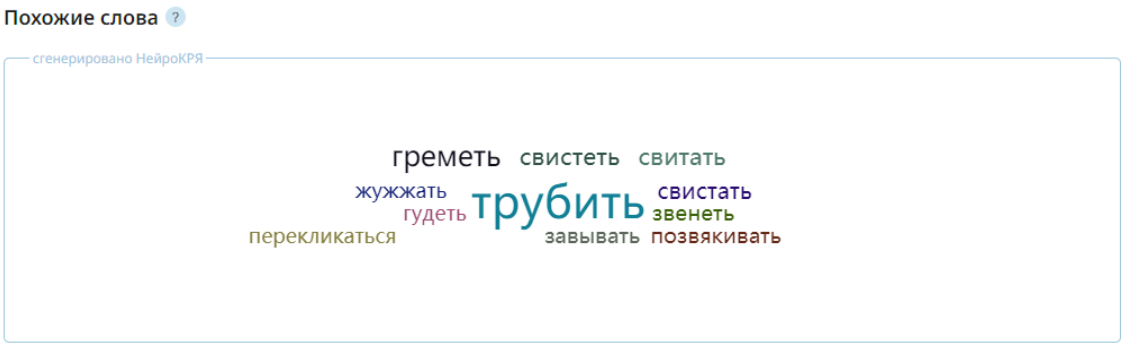

Рис. 17: Похожие слова для слова *трубить* в корпусе «Русская классика»

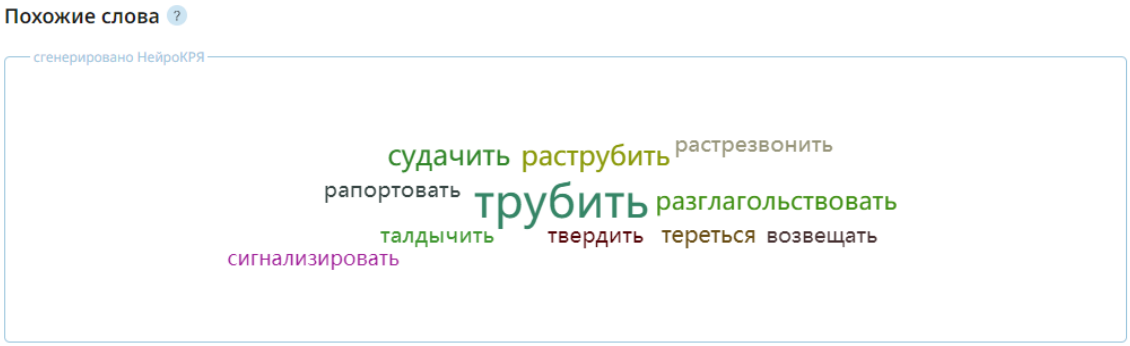

Рис. 18: Похожие слова для слова *трубить* в корпусе Центральных СМИ

Все используемые на сегодняшний день модели имеют одинаковую архитектуру (CBoW из библиотеки (Rehurek, Sojka 2011)) и схожие параметры обучения: окно размером 5, порог встречаемости 5-10 в зависимости от корпуса. Модель, обученная на Основном корпусе, показала корреляцию 0.367 (по Спирмену) на выборке RuSimLex365 (Kutuzov, Kunilovskaya 2018) на Корпусе региональных СМИ — 0.227.

Все семь моделей размещены в открытом доступе на странице «Нейросетевый модели» сайта НКРЯ[23] и могут быть использованы для научных и прикладных исследований.

4 Нейросетевые модели в разметке данных и метаданных НКРЯ

Наиболее ресурсно затратным этапом при подготовке корпуса является этап корпусной разметки. Это касается как собственно внутритекстовой лингвистической разметки, так и разметки текстовых метаданных. Алгоритмы автоматической разметки морфологии показывали недостаточно высокое качество, поэтому центральным решением старой платформы НКРЯ был отказ от снятия омонимии — словоформе приписывались все возможные разборы, выбор релевантного разбора оставался на стороне пользователя. Как уже было сказано выше в (1.1.3), такая неопределенность блокировала развития статистических и аналитических сервисов, поскольку омонимичные формы или разборы существенно зашумляют любые подсчеты. Применение интеллектуальных моделей для разметки данных позволяет значительно ускорить включение текстов в корпус при сохранении высокого качества лингвистической разметки. Ниже будут последовательно представлены модели, которые сейчас используются при подготовке данных НКРЯ. Это, во-первых, нейросетевая модель морфосинтаксической разметки Rubic, а во-вторых, комплекс моделей для разметки метаданных: жанров в корпусе «Социальные сети» и ключевых слов в текстах Газетного корпуса. Использование этих моделей уже изменило экосистему НКРЯ, открыв новые возможности для значительного облегчения наиболее трудоемкого этапа подготовки корпусных данных.

4.1 Нейросетевая модель морфосинтаксической разметки для русского языка Rubic

Задача, которую решает нейросетевая модель Rubic, — это автоматическая разметка текста, а именно: лемматизация, морфологическая характеристика для всех токенов, включая определение части речи, построение дерева синтаксической зависимости предложения. Каждая из этих операций предполагает разрешение омонимии. Таким образом, Rubic представляет собой альтернативу морфологическому анализатору MyStem (Зобнин, Носырев, 2015), ранее применявшемуся для обработки текстов НКРЯ и основанному на грамматическом словаре, и, кроме того, выполняет синтаксическую разметку. При разработке модели Rubic'a ставились задачи улучшения обработки «несловарных» разборов (например, *ажник, летось, сподтишка*), просторечных и грамматически аномальных форм и конструкций (например, *силов*, *хоцца, подумашь, ефту*), словоформ, записанных в нестандартной орфографии (в том числе петровской эпохи, в дореволюционной орфографии, а также в советской орфографии до реформы 1956 года). Модель также должна корректно обрабатывать архаичные формы из церковнославянского языка (например, *бысть, быша, многая лета*) согласно соглашениям, принятым для исторических корпусов НКРЯ.

[23] https://ruscorpora.ru/license-content/neuromodels

4.1.1 Принципы работы Rubic

Архитектура Rubic'a (Lyashevskaya et al., 2023) основана на архитектуре модели qbic, победившей в соревновании по обработке русского языка GramEval2020 (Anastasyev 2020). Используется однослойный LSTM-энкодер, комбинирующий векторизованное представление слов, получаемые из BERT-подобной модели (в текущей реализации, sberbank-ai/ruBert, предобученный на 30 Гб данных) и морфологические пометы, приписываемые анализатором PyMorphy2. Полученное представление анализируется тремя декодерами, выполняющими задачи классификации для выбора: а) части речи и грамматических признаков, б) леммы и в) дерева зависимостей. Rubic обучается в мультизадачном режиме, то есть веса классификаторов (а-в) определяются независимо друг от друга.

Для обучения модели использовались специально подготовленные обучающие данные на основе корпусов СинТагРус, UD-Taiga (Droganova, Zeman 2018; Droganova, Lyashevskaya, 2018) и НКРЯ (Lyashevskaya et al., 2020). Они охватывают тексты различных временных эпох и жанров (проза, газетные тексты, поэзия, социальные сети, википедия) общим объемом свыше 2,4 миллиона токенов. Разметка обучающего корпуса текстов проверена вручную. Все данные приводятся в расширенном формате морфологической и синтаксической разметки UD-ext (Lyashevskaya 2019). Такой подход, как уже говорилось выше в (1.1.2), следует глобальной тенденции стандартизации корпусной разметки, и в частности, унификации морфо-синтаксических тегов, обеспечивающую возможность кросс-языковых исследований. Выбранный формат UD-ext ставит в соответствие наборы частеречных тегов и морфологических признаков, используемые при разметке текстов в НКРЯ и в русских корпусах Universal dependencies, см. Таблицу 3.

UPOS	ADJ, ADP, ADV, ADVPRO, ANUM, AUX, CCONJ, COM, DET, INIT, INTJ, NOUN, NUM, PARENTH, PART, PRED, PREDPRO, PRON, PROPN, PUNCT, SCONJ, SYM, VERB, X
FEAT (основные)	Abbr, Animacy, Aspect, Case, Clitic, Degree, Gender, Mood, Number, Person, Tense, Transit, Variant, VerbForm, Voice
FEAT (лексические признаки)	NameType, NumForm, NumType, Poss, Polarity, PronType, Reflex
FEAT (другие дополнительные)	Anom, Hyph, InflClass, Foreign, Typo

Таблица 3: Формат UD-ext

Чтобы улучшить обработку предложений, написанных заглавными буквами, а также с использованием буквы «Е» и разного рода кавычек и отточий, применяется аугментация данных объемом 3200 предложений (40 тысяч словоформ).

Rubic работает с текстовыми данными, представленными в формате CoNLL-U. При подготовке текстов НКРЯ используется отдельная модель токенизации (см. ниже). На этапе предобработки для улучшения качества работы модели токены, состоящие из смеси кириллических букв и других символов (знаков ударения, диакритик и разного «шума»), очищаются специальным модулем. Кроме того, набор правил premodern2modern приводит тексты, представленные в старой орфографии разных периодов, к современной орфографии. Поскольку размеченные данные поступают в формате xml и на индексацию также уходят данные в формате xml, предусмотрены инструменты конвертации форматов xml -> CoNLL-U и обратной конвертации CoNLL-U -> xml.

Целевая синтаксическая разметка представляется в формате CONLL-U, который затем сохраняется в синтаксисе XML, принятом в НКРЯ. Это позволяет, с одной стороны, сохранить подход к синтаксической разметке, принятый в Universal Dependencies, а с другой — использовать морфологическую и семантическую разметку в стандарте НКРЯ.

4.1.1.1 Токенизатор

Практически любой анализ текста начинается с его разбиения на фрагменты (токенизации). В рамках задачи снятия морфологической неоднозначности наиболее подходящими размерами фрагментов являются предложения и отдельные слова. В простейшем случае для разбиения текста на предложения можно воспользоваться разбиением по набору знаков препинания (например, точка, вопросительный и восклицательный знаки), а для разбиения на слова — по пробелам. Однако в реальности такой подход не может дать высокого качества как из-за использования знаков препинания в других целях (например, в инициалах), так и из-за опечаток (например, пропущенные пробелы). Так как качество разбиения критически важно для последующего анализа, мы провели ряд экспериментов по выбору оптимальной модели токенизации. Нами были протестированы предобученные алгоритмы udpipe (Straka, Hajic, Straková 2016), razdel[24], spacy[25], nltk (Bird, Loper, Klein 2009), pymorphy2 (Korobov 2015), MyStem[26], rusenttokenize[27], а также их комбинации. Тестирование проводилось на специально подготовленной выборке сложных случаев (выборка GOLD). Лучший результат продемонстрировало сочетание алгоритма razdel для сегментации на предложения и spacy для сегментации на слова. Полные результаты тестирования приведены в Таблице 4.

		Сегментация на предложения					
		razdel	*spacy*	*nltk*	*rusenttokenize*	*udpipe*	*mystem*
Токенизация	*razdel*	F1-s: 0.5562 F1-w: 0.8946	F1-s: 0.4577 F1-w: 0.8946	F1-s: 0.3299 F1-w: 0.8946	F1-s: 0.4545 F1-w: 0.8946	F1-s: 0.4532 F1-w: 0.8946	F1-s: 0.3441 F1-w: 0.8946
	spacy	**F1-s: 0.5562** **F1-w: 0.9272**	F1-s: 0.4577 F1-w: 0.9273	F1-s: 0.3299 F1-w: 0.9271	F1-s: 0.4545 F1-w: 0.9272	F1-s: 0.4132 F1-w: 0.9272	F1-s: 0.3441 F1-w: 0.9272
	nltk	F1-s: 0.5562 F1-w: 0.9060	F1-s: 0.4577 F1-w: 0.9032	F1-s: 0.3299 F1-w: 0.9050	F1-s: 0.4545 F1-w: 0.9041	F1-s: 0.4132 F1-w: 0.9041	F1-s: 0.3441 F1-w: 0.9041
	pymorphy	F1-s: 0.5562 F1-w: 0.8734	F1-s: 0.4577 F1-w: 0.8734	F1-s: 0.3299 F1-w: 0.8734	F1-s: 0.4545 F1-w: 0.8734	F1-s: 0.4132 F1-w: 0.8735	F1-s: 0.3441 F1-w: 0.8735
	udpipe	F1-s: 0.5562 F1-w: 0.8387	F1-s: 0.4577 F1-w: 0.8387	F1-s: 0.3299 F1-w: 0.8387	F1-s: 0.4545 F1-w: 0.8387	F1-s: 0.4132 F1-w: 0.8387	F1-s: 0.3441 F1-w: 0.8387
	MyStem	F1-s: 0.5562 F1-w: 0.8720	F1-s: 0.4577 F1-w: 0.8720	F1-s: 0.3299 F1-w: 0.8720	F1-s: 0.4545 F1-w: 0.8720	F1-s: 0.4132 F1-w: 0.8720	F1-s: 0.3441 F1-w: 0.8720

Таблица 4. Результаты тестирования различных алгоритмов сегментации и токенизации по двум метрикам: F1-s (по предложениям) и F1-w (по отдельным словам), аналогичным использованным на соревновании CoNLL 2018[28]

[24] https://github.com/natasha/razdel
[25] https://github.com/explosion/spaCy
[26] https://yandex.ru/dev/mystem/
[27] https://github.com/deeppavlov/ru_sentence_tokenizer
[28] https://universaldependencies.org/conll18/

Анализ допускаемых алгоритмом ошибок и создание специальных эвристических алгоритмов постобработки позволило достичь качества 0.95 по метрике F1. При этом не удалось обнаружить эвристик, которые бы позволили значительно улучшить результат сегментации на предложения (F1=0.55). В связи с этим было принято решение обучить на имеющихся данных собственную модель для токенизации. В качестве архитектуры была выбрана модель Stanza (Qi P. et al. 2020), разработанная Stanford NLP Group на базе рекуррентной нейронной сети LSTM, показавшая хорошие результаты в аналогичной задаче для английского языка. Эта модель является двухслойной. Первый слой состоит из одномерной сверточной нейронной сети и слоя BiLSTM. Выходные данные CNN объединяются со скрытыми состояниями BiLSTM и передаются на второй слой BiLSTM. В качестве функции потерь используется кросс-энтропия. Мы обучили модель со стандартными параметрами обучения на открытых датасетах Тайга (Shavrina, Shapovalova 2017) и СинТагРус[29], а также на внутренних данных из корпусов прозы XX–XXI веков, поэзии, корпусов со старой орфографией XVIII века, а также текстах новостей XXI века. Полученная модель на тестовой выборке продемонстрировала схожее качество по отдельным словам и значительно лучшее качество по предложениям (F1=0.63). Для дополнительного сравнения новой модели с предыдущим решением была подготовлена расширенная тестовая выборка, схожая по составу с реальными данными (выборка TEST). На материале этой выборки обученная модель значительно превзошла лучший из предобученных алгоритмов, достигнув метрики F1=0.95 по предложениям и 0.99 по отдельным словам (Таблица 5).

	razdel + spacy + эвристики	stanza
Выборка GOLD	F1-s: 0.556 F1-w: 0.952	F1-s: 0.637 F1-w: 0.944
Выборка TEST	F1-s: 0.943 F1-w: 0.992	F1-s: 0.956 F1-w: 0.996

Таблица 5. Сравнение работы комбинированного подхода (сегментация на предложения при помощи razdel, токенизация при помощи spacy и дополнительные эвристики постобработки) и обученной на данных корпуса модели Stanza. Сравнение проводится по двум метрикам: F1-s (по предложениям) и F1-w (по отдельным словам).

Модель токенизатора доступна для скачивания и размещена на странице «Нейросетевые модели» сайта НКРЯ.[30]

4.1.1.2 Классификатор морфологических признаков

После сегментации текста на предложения, а предложений — на слова, начинается этап приписывания каждому выделенному слову морфологических признаков. Классификатор морфологических признаков работает на принципе полного морфологического тега, иными словами, входом и выходом модели служит набор, состоящий из частеречного и грамматических тегов вида «NOUN|Animacy=Anim|Case=Nom|Gender=Fem|Number=Sing» (ср. разбор формы *кошка*). В отличие от классификаторов, в которых каждая грамматическая категория определяется независимо, данное решение позволяет избежать потери части разбора (например, одушевленности, в случае если вероятность ее определения в контексте низкая) и свести требуемый набор грамматических помет для частеречных классов и подклассов к грамматическому стандарту корпуса.

4.1.1.3 Лемматизатор

Размеченное морфологическими пометами слово далее попадает в лемматизатор. В основе лемматизации лежат правила преобразования словоформы в лемму вида «удалить последовательность символов в конце строки длины N, удалить последовательность символов в начале строки

[29] https://ruscorpora.ru/page/corpora-datasets/
[30] https://ruscorpora.ru/license-content/neuromodels

длины M > добавить последовательность D в конце строки > применить маску капитализации / декапитализации». Используются преобразования, встретившиеся в обучающих данных более 3 раз, чтобы исключить влияние несистемных опечаток и другого шума. В зависимости от объема обучающих данных, такой подход дает от 1000 до 2000 правил (ср. наблюдение «о менее 1000 классов лемматизации» в работе Michurina et al., 2021). Модуль лемматизации получает на вход словоформу с меткой ее части речи, определенной морфологическим модулем. Для улучшения качества лемматизации предусмотрена возможность сравнения наиболее вероятных гипотез лемм с данными словаря, составленного вручную по данным существующих корпусов с ручной разметкой и словарей.

4.1.1.4 Разметка синтаксического дерева

Наконец, на последнем этапе разметки происходит анализ синтаксического дерева каждого предложения. При построении гипотез синтаксических деревьев используется подход Т. Дозата и К. Маннинга (Dozat, Manning, 2016) на основе глубокого биаффинного внимания для определения пар связанных словоформ и метки синтаксического отношения.

4.1.2 Результаты

Качество работы модели Rubic'a оценивалось на коллекции тестовых данных, представляющих разные сферы употребления языка, см. Таблицу 6.

	fiction	news	poetry	social	wiki
Часть речи	0.9922	0.9893	0.9923	0.9777	0.9808
Леммы	0.9930	0.9923	0.9846	0.9848	0.9780
Морфологические признаки	0.9591	0.9517	0.9654	0.9528	0.9423
Неименованные синтаксические связи	0.9599	0.9563	0.9106	0.9296	0.9457
Именованные синтаксические связи	0.9530	0.9425	0.8942	0.9153	0.9231

Таблица 6: Результаты работы Rubic’a на тестовом множестве

Модель хорошо справляется с определением частей речи, морфологическим разбором некоторых грамматических категорий, таких как сравнительная степень, переходность, вид. При лемматизации высокое качество обработки модель демонстрирует для слов продуктивных парадигм. Результаты синтаксического парсинга показывают, что модель чаще всего правильно анализирует большинство часто употребляемых конструкций — вводные и сочиненные конструкции, предложные и атрибутивные группы и т.п. Также хорошо определяются дальние связи (например, субъект на расстоянии 5-10 слов от предиката).

Критическим образом на качество анализа влияют ошибки в токенизации исходных данных, попадающих на вход Rubic’a.

Как правило, ошибки при автоматической разметке одновременно наблюдаются и в лемматизации, и в морфологических признаках, и в дереве синтаксических зависимостей. Это может быть связано со структурой предложения, например, из-за недостаточности контекста для одиночных слов в клаузе (ср., например, предложение прямой речи «- *Бушую».*), при наличии оборванных словоформ или перед многоточиями. Анализатор сталкивается с трудностями при определении части речи, морфологических и синтаксических признаков на границе предложений. Например, в начале предложения правильно определяется категория существительного, но возникает

ошибка в недостаточном/избыточном определении классов имен собственных (*Надежда, Воробьиха*). В той же позиции начала предложения может неправильно определиться категория одушевленности, что влечет за собой также ошибку и в синтаксическом разборе:

Белены объелись или выпили лишнее!
белена NOUN _ Animacy=Anim|Case=Nom|Gender=Masc|Number=Plur 2 nsubj

При морфологическом анализе ошибки возникают из-за наличия в текстах авторских искажений и скандирования (*рр-рр-а-а-аз, стооооой* и т.п.), аномальных, но при этом частотных форм (*испужамшись*) слов с нестандартной капитализацией (*варЮ*), а также содержащих кавычки и букву «Е».

Синтаксический анализатор достаточно уверенно определяет зависимости в пределах клаузы (в именных группах, группах глагола, предикативов и наречий), между прямой речью и клаузой, сопровождающей прямую речь, между клаузами в сложносочиненных предложениях. Ошибки возникают при определении вершины в безглагольных клаузах, в клаузах с эллипсисом вершины, в определении обращений и дискурсивных элементов, не выделенных пунктуационно. Наблюдаются ошибки в том случае, когда две клаузы связаны отношением сочинения VS. паратаксиса («*Упал, упал человек, тонет!*» — *упал, упал* (бессоюзное сочинение глаголов, имеющих общий субъект)) должно быть связано как conj, но Rubic размечает как parataxis (при этом *тонет* правильно размечается как conj)).

С точки зрения зависимости качества работы модели от жанра текста можно сказать, что несколько большую трудность вызывают тексты соцсетей, энциклопедические, математические тексты, поэзия. Очевидной причиной ошибок являются нестандартные синтаксические конструкции и пунктуация этих текстов, эллипсис и оборванные высказывания, редкие имена собственные, использование в данных жанрах предложений большой длины. Кроме того, можно заметить, что при обработке таблиц и списков литературы, транслируемых токенизатором в одно предложение, синтаксический модуль пытается трактовать отношения между словоформами как предикативные и именные зависимости, а не как «плоские» отношения типа списков.

В целом, анализ качества работы модели показывает, что самым слабым звеном среди модулей является лемматизатор. Часто ошибочно лемматизируются короткие слова с выпадающим гласным (*мох, пес* и т.п.), существительные с чередованием в конце слова (*котенок — котята* и т.п.), глаголы с чередованиями в корне (*жать, плыть, выть, петь, скрести* и т.п.), аббревиатуры (г. — *город/господин/грамм*), степени сравнения прилагательных и наречий на «по-» (*потише — тихий, тио, тико)*. Для улучшения качества лемматизации корпуса на этапе постобработки используются эвристики для замены ошибочных форм по спискам, составленным вручную для форм разных частей речи (около 50 тысяч правил).

4.1.3 Перспективы

Rubic планируется улучшить модулем, позволяющим делать альтернативные (правильные) морфологические разборы, включая часть речи и лемму. Для этого прежде всего необходимо разработать формат добавления альтернативных разборов в обучающие данные, а именно — изменения и дополнения в формат CoNLL-U, который используется Rubic'ом. Отметим, что некоторые возможности представления альтернативных разборов уже есть в текущей версии обучающих данных для Rubic'a. Так, уже разработан формат представления альтернативных разборов, таких как разбор формы *расположенный* как причастия и прилагательного. Кроме того, леммы, не соответствующие литературному варианту русского языка (*удилом* (ед. ч. от sg. tt.), *шеколад* — дневники А.К. Гладкова) могут быть связаны с литературным вариантом леммы через вспомогательные таблицы, аналогичные тем, которые используются в панхроническом поиске (подробнее о принципах работы панхронического поиска см. в статье (Савчук и др., 2024).

Как уже говорилось выше в (4.1.2), качество работы лемматизатора, используемого в Rubic'e, требует дальнейших улучшений. Для повышения эффективности работы нейросети можно было бы предложить несколько другой алгоритм лемматизации. В настоящий момент распределение ответов лемматизатора следует, условно, закону Парето: доля правильных вариантов составляет

от 50 до 95%, остальное — неправильные ответы разного рода, убывающие по частоте. Улучшенная модель лемматизации могла бы быть более чувствительна к словоформам непродуктивных парадигм, а также могла бы учитывать не только длину, но и буквенный состав «псевдоокончаний», лучше разрешать омонимию слов с пересечением парадигм (*белка*/*белок*, *стрелка*/*стрелок*). Можно также предложить выделить в отдельный модуль обработку аббревиатур. Это связано с тем, что у аббревиатур другое распределение операций лемматизации, но при этом их определение все же чувствительно к контексту. Кроме того, использование аббревиатур при обучении ухудшает качество лемматизации обычных слов.

Возможно дополнение структуры Rubic'a алгоритмом обработки искаженных и просторечных форм (*гавагите, г'азог'ву, бегат, стоооой, по-до-жди-и-и-и* и т.п.), которые часто встречаются в корпусных данных, подлежащих разметке.

Качество работы Rubic'a может быть повышено за счет целевого пополнения всех обучающих текстовых множеств для охвата языковых явлений и жанрового разнообразия в объеме, достаточном для обучения. Кроме пополнения обучающих данных, необходима и чистка обучающего корпуса, исправление опечаток, ошибок разметки, расстановки знаков препинания. В частности, это актуально для разрешения омонимии «*причастие-прилагательное*», «*предикатив-наречие-прилагательное*». Также при работе по улучшению обучающих данных необходимо уделить внимание редким синтаксическим конструкциям и длинным предложениям, при автоматическом разборе которых Rubic часто допускает ошибки.

4.2 Разметка жанров в корпусе «Социальные сети»

Корпус «Социальные сети» содержит тексты из открытых интернет-источников и включает в себя записи в блогах и сообщения в мессенджерах (подробнее о балансе источников корпуса «Социальные сети» см. в статье (Савчук и др., 2024)). Поскольку понятие «социальные сети» в этом случае трактуется максимально широко, а также в связи с большим объемом корпуса (почти 160 млн словоупотреблений), появилась необходимость в автоматической разметке жанров для систематизации текстов корпуса.

Особенность разметки жанров в случае корпуса «Социальные сети» заключается в том, что на начальном этапе в корпусе отсутствовала ручная разметка жанров. Поэтому первым этапом работы стала экспертная оценка выборки текстов и формирование списка широко представленных жанров. Для этого была сформирована выборка, содержащая около трех тысяч текстов, выбранных случайным образом. В результате экспертной оценки были выделены 13 жанров, наиболее широко представленных в выборке. К ним были добавлены жанры «Биография», «Гороскоп» и «Интернет-рейтинг» как содержащие довольно характерные тексты. Тексты прочих жанров были объединены в класс «Неопределенная категория». Список жанров и распределение текстов в выборке по результатам оценки экспертов представлены в Таблице 7 в столбцах «Жанр» и «Количество текстов, размеченных экспертами» соответственно.

Распределение текстов по жанрам по результатам экспертной оценки случайной выборки текстов получилось неравномерным. К наиболее широко представленным классам (жанрам) — «Анонс | объявление», «Неопределенная категория» и «Информационное сообщение» — относятся 1019, 671 и 343 текста соответственно, то есть больше двух третей от объема рассмотренной выборки. Неравномерное распределение жанров затрудняет обучение моделей для автоматической разметки категорий текстов, поэтому с целью выравнивания количества примеров в классах в набор обучающих данных для ряда жанров были добавлены дополнительные тексты. Дополнительные тексты были преимущественно получены из текстов Основного и Регионального корпусов. Для жанра «Интернет-рейтинг» были использованы тексты корпуса «Социальные сети», содержащие характерные для данного жанра фразы. Итоговое распределение текстов в наборе данных для обучения модели автоматической разметки жанров представлено в столбце «Всего текстов» в Таблице 7.

Жанр	Количество текстов, размеченных экспертами	Количество дополнительных текстов	Всего текстов	Источник дополнительных текстов
Анекдот	78	153	231	Основной корпус
Анонс | объявление	1019	-	1019	-
Биография	2	454	456	Основной корпус, Региональный корпус
Вопрос	36	-	36	-
Гороскоп	8	204	212	Основной корпус, Региональный корпус
Инструкция | совет | рекомендация	81	-	81	-
Интернет-рейтинг	8	253	261	Тексты корпуса «Социальные сети», подобранные по ключевым словам
Информационное сообщение	343	-	343	-
История	44	192	236	Тексты, вручную подобранные экспертами
Неопределенная категория	671	-	671	-
Отзыв | рецензия	215	-	215	-
Оценка	34	150	184	
Поздравление	45	674	719	Основной корпус, Региональный корпус
Поэзия	145	-	145	-
Прецедентный текст	119	-	119	-
Рецепт	93	223	316	Основной корпус, Региональный корпус
Итого	2941	2303	5244	

Таблица 7: Набор данных для разметки жанров в корпусе «Социальные сети»

Для разметки жанров была выбрана предварительно обученная модель RuRoBERTa [31] (Zmitrovich et al., 2023), повторяющая архитектуру модели RoBERTa для англоязычных текстов (Liu et al., 2019). Модель использует токенизацию по принципу Byte-level BPE (Gage 1994). Для предварительного обучения были использованы тексты Википедии, а также коллекции новостных и художественных текстов, текстов веб-ресурсов и русскоязычных субтитров. RuRoBERTa показывает высокие результаты в задачах классификации текстов на русском языке (в частности, в рамках последних соревнований Dialogue Evaluation (Golubev, Rusnachenko, Loukachevitch 2023). Для автоматической разметки жанров модель была дообучена на собранном наборе данных с использованием следующих параметров: скорость обучения — 5e-6, количество эпох обучения — 3, максимальная длина входной последовательности — 256 токенов. При обучении жанрам назначались веса в зависимости от их доли в наборе данных.

Качество модели было проверено с помощью десятикратной перекрестной проверки (кроссвалидации для десяти фолдов). Перекрестная проверка выполнялась следующим образом. Выборка текстов корпуса «Социальные сети», размеченная экспертами, была десять раз разбита на обучающую и тестовую подвыборки по принципу скользящего окна. К обучающей подвыборке были добавлены дополнительные тексты, после чего на дополненной обучающей подвыборке выполнялось дообучение модели с использованием параметров, указанных выше. Тестирование модели выполнялось на тестовой выборке. Таким образом, в результате перекрестной проверки были получены десять значений показателей качества модели на контрольных подмножествах данных. Итоговая оценка качества представляет собой среднее арифметическое этих значений. Значение F-меры с макроусреднением составило 54,42% для 16 жанров, доля правильных ответов (accuracy) составила 71,16%.

Для итоговой разметки жанров в корпусе «Социальные сети» был использован ансамбль из трех моделей, дообученных на наборе данных, состоящем из текстов, размеченных экспертами, и дополнительных текстов. Объединение предсказаний моделей в ансамбле осуществлялось по принципу усреднения предсказанных вероятностей жанров (soft voting). На этапе постобработки предсказаний модели все сверхкороткие тексты были перенесены в класс «Неопределенная категория». Пороговое значение, являющееся критерием для определения коротких текстов, составляет 40 токенов. Такое значение было получено в результате эмпирического анализа текстов корпуса. Количество токенов определяется с помощью токенизатора модели RuBERT (Kuratov 2019).

4.3 Разметка ключевых слов в Корпусе региональных СМИ

Корпус Региональных СМИ в основном состоит из коротких информационных текстов, опубликованных в газетах различного уровня (Савчук 2015). Каждый текст, как правило, посвящен единственной теме. Для описания тематики и упрощения поиска текстов в корпусе выполнена автоматическая разметка ключевых слов. Одно ключевое слово может состоять из одного существительного в именительном падеже в единственном или множественном числе (*праздник, переломы*) либо из двусловного сочетания (биграммы) с главным словом-существительным (*таяние снега, обычные дни*).

Извлечение ключевых слов из текстов Региональных СМИ выполнено с помощью библиотеки RuTermExtract[32]. Алгоритм, лежащий в основе RuTermExtract, представляет собой адаптированную для русского языка версию алгоритма TermExtract[33]. Он построен на анализе морфологических характеристик слов и словосочетаний и набора правил для извлечения ключевых слов. Для морфологического анализа в русскоязычной версии используется библиотека PyMorphy2 (Korobov 2015).

На этапе предобработки текстов биграммы, заключенные в кавычки, были объединены символом «_», чтобы алгоритм рассматривал их как униграмму (например, *«Комсомольская правда»* -> *«Комсомольская_правда»*) и не разделял на слова. Предобработанные тексты подавались на вход алгоритму RuTermExtract со следующими параметрами: максимальное количество извлеченных ключевых слов — 20; параметр nested, позволяющий извлекать ключевые слова, лежащие

[31] https://huggingface.co/ai-forever/ruRoberta-large
[32] https://github.com/igor-shevchenko/rutermextract
[33] https://pypi.org/project/topia.termextract

внутри других ключевых слов, — True. С помощью алгоритма для текстов были получены первичные списки ключевых слов в нижнем регистре. К полученным ключевым словам применялась несколько шагов постобработки:

1. замена символа «_», добавленного в биграммы на этапе предобработки, пробелом;
2. удаление ключевых слов, состоящих из трех и более слов;
3. удаление полных и кратких имен в соответствии со списком личных имен;
4. удаление однокоренных униграмм с помощью модели Morphodict-K (см. Раздел (3.2.));
5. проведение нормализации словосочетаний на основе списка правил с помощью библиотеки PyMorphy2 (Korobov 2015);
6. обработка некоторых распространенных ошибок (например, такой ошибкой является постановка второго слова в форму генетива в некоторых именованных сущностях: «*юрий лужкова*» -> «*юрий лужков*»);
7. удаление ключевых слов таким образом, чтобы длина списка составляла не более 15 ключевых слов.

5 Заключение

В статье представлено описание обновленной платформы НКРЯ с технологической точки зрения. Это обновление является важнейшим этапом 20-летнего развития Национального корпуса русского языка. Следует подчеркнуть, что речь идет не об отдельных нововведениях, а о внедрении комплексного подхода, основанного на идеологических принципах, соответствующих современным практикам и стандартам развития корпусных ресурсов, а также общим тенденциям цифрового развития общества. Эти принципы находят свое выражение в переходе к модульной и гибкой архитектуре корпусного ядра и веб-интерфейса, открытой для дальнейших изменений и масштабного пополнения корпусов; в разработке сервисов для анализа данных, которые позволяют исследователям переходить от ручного анализа примеров к количественному анализу, основанному на статистическом обобщении распределения лексических единиц; в интеграции технологий искусственного интеллекта в процесс подготовки корпусных данных; в создании собственных нейросетевых моделей и их размещении в открытом доступе; и, наконец, в ориентации на привлечение более широкой аудитории к работе с Национальным корпусом русского языка, расширение возможностей применения корпусных данных не только в лингвистических исследованиях, но и в педагогике, а также в качестве источника языковых данных для самых разных областей гуманитарного знания.

Литература

[1] Баранов А. Н. (2023). Инструментарий лингвистики в лингвистической экспертизе: корпусные технологии // Язык. Право. Общество. С. 54-58.
[2] Гришина Е. А. и др. (2009). О задачах и методах словообразовательной разметки в корпусе текстов // Полярный вестник (Тромсё), 2009, № 12. С. 5–25.
[3] Зобнин А. И., Носырев Г. В. (2015). Морфологический анализатор MyStem 3.0 // Труды Института русского языка им. В. В. Виноградова. 2015. № 6. С. 300–310.
[4] Иомдин Б. Л. Как определять однокоренные слова? // Русская речь. 2019. No 1. С. 109–115.
[5] Кузнецова А. И., Ефремова Т. Ф. (1986). Словарь морфем русского языка. Москва: Рус. яз., 1986.
[6] Савчук С. О. (2015). Корпус региональных газет России и зарубежья // Труды Института русского языка им. В. В. Виноградова. 2015. № 6. С. 163—193.
[7] Савчук С. О. и др. (2024). Национальный корпус русского языка 2.0: новые возможности и перспективы развития //Вопросы языкознания. Т. 2. С. 7-34.
[8] Сичинава Д. В. (2005). Национальный корпус русского языка: очерк предыстории // Национальный корпус русского языка: 2003—2005. М.: Индрик. С. 21—30.
[9] Сичинава Д. В. (2022). Корпус берестяных грамот как параллельный // Труды Института русского языка им. В. В. Виноградова. 2022. № 2 (32), 92-106.
[10] Тихонов А. Н. (2002). Морфемно-орфографический словарь. Москва: Астрель: АСТ, 2002.
[11] Aksan Y. et al. (2012). Construction of the Turkish National Corpus (TNC) // LREC. 2012. P. 3223-3227.
[12] Anastasyev D., (2020). Exploring pretrained models for joint morphosyntactic parsing of Russian // Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference “Dialogue”, volume 19. P. 1–12.

[13] Beeby A., Rodríguez Inés P. and Sánchez-Gijón P. (eds). (2009). Corpus Use and Translating. Corpus Use for Learning to Translate and Learning Corpus Use to Translate, Amsterdam: Benjamins.
[14] Biber D. (1993). Representativeness in corpus design // Literary and linguistic computing. Vol. 8. № 4. P. 243–257.
[15] Bird, St., Loper E., Klein E. (2009). Natural Language Processing with Python. O'Reilly Media Inc.
[16] Boulton A. (2011). Data-driven learning: the perpetual enigma // S. Goźdź-Roszkowski. Explorations across Languages and Corpora, Peter Lang. P. 563-580.
[17] Bowker L. (2018). Corpus linguistics is not just for linguists: Considering the potential of computer-based corpus methods for library and information science research // Library Hi Tech, 36(2). P. 358-371.
[18] Buchholz, S., Marsi, E. (2006). CoNLL-X shared task on multilingual dependency parsing. // Proceedings of the tenth conference on computational natural language learning (CoNLL-X). P. 149-164.
[19] Calzolari, N., McNaught, J., & Zampolli, A. (1996). EAGLES Final Report: EAGLES Editors' Introduction. Pisa, Italy, EAG-EB-EI.
[20] Chartrand, L. (2022). Modeling and corpus methods in experimental philosophy. Philosophy Compass, 17(6).
[21] Chen K. J. et al. (1996). Sinica corpus: Design methodology for balanced corpora //Proceedings of the 11th Pacific Asia Conference on Language, Information and Computation. P. 167-176.
[22] Coulthard, M. (1994). On the Use of Corpora in the Analysis of Forensic Texts', Forensic Linguistics: International Journal of Speech, Language and the Law 1(1). P. 27–43.
[23] Coulthard, M., Johnson, A. and Wright, D. (2017) An Introduction to Forensic Linguistics: Language in Evidence, London: Routledge.
[24] Curtotti M., Mccreath E. (2010). Corpus based classification of text in Australian contracts // Proceedings of the Australasian Language Technology Association Workshop.
[25] Davies, M. (2021). The coronavirus corpus: Design, construction, and use // International journal of corpus linguistics, 26(4). P. 583-598.
[26] De Marneffe, M. C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. Computational linguistics, 47(2). P. 255-308.
[27] Devlin J. et al. (2019). Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 2019. P. 4171—4186.
[28] Doval, I., Sánchez Nieto, M. T. (2019) Parallel Corpora for Contrastive and Translation Studies. New Resources and Applications. Amsterdam: John Benjamins
[29] Dozat T., Manning Ch. D. (2016). Deep Biaffine Attention for Neural Dependency Parsing https://arxiv.org/abs/1611.01734
[30] Droganova, K, Lyashevskaya O. (2018). Cross-tagset parsing evaluation for Russian // Digital Transformation and Global Society Third International Conference, DTGS 2018, St. Petersburg, Russia, May 30 – June 2, 2018, Revised Selected Papers, Part I / Ed. by Daniel A. Alexandrov, A. V. Boukhanovsky, A. V. Chugunov, Y. Kabanov, O. Koltsova. Issue 858. P. 380-390.
[31] Droganova, K, Lyashevskaya O., Zeman D. (2018). Data Conversion and Consistency of Monolingual Corpora: Russian UD Treebanks // Proceedings of TLT 2018 International Workshop on Treebanks and Linguistic Theories, 13-14 November 2018, Oslo, Norway. NEALT Proceedings Series. Linköping University Electronic Press, 2018. P. 52-65.
[32] Evert, S. (2008). Corpora and collocations // A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, article 58, Mouton de Gruyter, Berlin. P. 1212-1248.
[33] Firth, J. R. (1951/1957): Modes of meaning. In: Papers in Linguistics, 1934-1951. Oxford: Oxford University Press.
[34] Francis, W. N., Kučera H. (1982). Frequency Analysis of English Usage: Lexicon and Grammar. Boston: Houghton Mifflin.
[35] Gage F. (1994). A new algorithm for data compression. C Users J. 12, 2. P. 23–38.
[36] Garipov T., Morozov D. Glazkova A. (2023). Generalization Ability of CNN-Based Morpheme Segmentation // 2023 Ivannikov Ispras Open Conference (ISPRAS), Moscow, Russian Federation. P. 58-62.
[37] Geyken A. (2007). The DWDS corpus: A reference corpus for the German language of the 20th century // Collocations and idioms: Linguistic, lexicographic, and computational aspects. T. 23. P. 41.
[38] Golubev, A., Rusnachenko, N., Loukachevitch, N.V. (2023). RuSentNE-2023: Evaluating Entity-Oriented Sentiment Analysis on Russian News Texts. ArXiv, abs/2305.17679.
[39] Heffer, C. (2005) The Language of Jury Trial: A Corpus-Aided Analysis of Legal-Lay Discourse, Basingstoke: Palgrave.
[40] Ide, N. (1998). Corpus enconding standard: SGML guidelines for encoding linguistic corpora // LREC. P. 463-470.

[41] Ide, N. et al. (2017). Community standards for linguistically-annotated resources // Handbook of linguistic annotation. P. 113-165.
[42] Imran, M., Mitra, P., & Castillo, C. (2016). Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages. arXiv preprint arXiv:1605.05894.
[43] Johns, T. & P. King (Eds.), (1991), Classroom Concordancing // English Language Research Journal, 4.
[44] Kiyong L., Laurent R. (2010). Towards Interoperability of ISO Standards for Language Resource Management // ICGL 2010. Hong Kong, Hong Kong SAR China. 9p.
[45] Kopotev, M., et al. (2015). Online extraction of Russian multiword expressions // The 5th workshop on balto-slavic natural language processing. P. 43-45.
[46] Korobov M. (2015). Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts. P. 320-332.
[47] Kuratov, Y. (2019). Adaptation of deep bidirectional multilingual transformers for Russian language / Y. Kuratov, M. Arkhipov // Komp'juternaja Lingvistika i Intellektual'nye Tehnologii, Moscow. Vol. 18. Moscow, 2019. P. 333-339.
[48] Kutuzov, A., Kunilovskaya, M. (2018). Size vs. Structure in Training Corpora for Word Embedding Models: Araneum Russicum Maximum and Russian National Corpus. In: van der Aalst, W., et al. Analysis of Images, Social Networks and Texts. AIST 2017. Lecture Notes in Computer Science, vol 10716. Springer, Cham.
[49] 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.
[50] Leech, G. (1993). Corpus annotation schemes // Literary and linguistic computing, 8(4), P. 275-281.
[51] Liu, Y., et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv, abs/1907.11692.
[52] Lyashevskaya O. (2019). A reusable tagset for the morphologically rich language in change: A case of Middle Russian // Komp'juternaja Lingvistika i Intellektual'nye Tehnologii. P 422–434.
[53] Lyashevskaya O. et al., (2023). Disambiguation in context in the Russian National Corpus: 20 years later // Computational Linguistics and Intellectual Technologies Papers from the Annual International Conference "Dialogue" (2023). Issue 22. P. 307-318.
[54] Lyashevskaya O. N. et al. (2020). GRAMEVAL 2020 Shared Task: Russian Full Morphology and Universal Dependencies Parsing. Computational Linguistics and Intellectual Technologies Papers from the Annual International Conference "Dialogue" (2020) Issue 19, P. 553-569.
[55] Machálek, T (2020a). Word at a Glance: Modular Word Profile Aggregator // Proceedings of LREC 2020. P. 7011–7016.
[56] McCarthy, M. (2008). Accessing and interpreting corpus information in the teacher education context // Language Teaching, 41 (4). P. 563-574.
[57] McEnery, T., Hardie, A. (2012) Corpus Linguistics: Method, theory and practice. Cambridge: Cambridge University Press.
[58] McEnery, T., Wilson, A. (2001) Corpus Linguistics. An Introduction. Edinburgh: Edinburgh University Press.
[59] Morozov, D. A., Glazkova A. V., Iomdin B. L. (2022). Text complexity and linguistic features: Their correlation in English and Russian // Russian Journal of Linguistics 26 (2). P. 426–448.
[60] Newman, J., & Cox, C. (2021). Corpus annotation //A practical handbook of corpus linguistics. — Cham : Springer International Publishing, 2021. — C. 25-48.
[61] Nivre, J. et. Al. (2016). Universal dependencies v1: A multilingual treebank collection. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). P. 1659-1666.
[62] Poletto, F. et al. (2021). Resources and benchmark corpora for hate speech detection: a systematic review // Lang Resources & Evaluation 55. P. 477–523
[63] Rehurek, R., Sojka, P. (2011). Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2)).
[64] Reppen R. (2021). Building a corpus: what are key considerations? // O'Keeffe A., McCarthy M. (ed.). The Routledge handbook of corpus linguistics. Routledge, 2021. P. 13-20
[65] Roll U., Correia R. A., Berger-Tal O. (2018). Using machine learning to disentangle homonyms in large text corpora //Conservation Biology. V. 32. №. 3. P. 716-724.
[66] Schrauwen S. (2010). Machine learning approaches to sentiment analysis using the Dutch Netlog Corpus // Computational Linguistics and Psycholinguistics Research Center. P. 30-34.
[67] Shavrina T., Shapovalova O. (2017) TO THE METHODOLOGY OF CORPUS CONSTRUCTION FOR MACHINE LEARNING: «TAIGA» SYNTAX TREE CORPUS AND PARSER. in proc. of "CORPORA2017", international conference, Saint-Petersbourg.
[68] Shawar B. A., Atwell E. S. (2005). Using corpora in machine-learning chatbot systems //International journal of corpus linguistics. V. 10. №. 4. P. 489-516.

[69] Sorokin, A., Kravtsova, A. Deep Convolutional Networks for Supervised Morpheme Segmentation of Russian Language // Ustalov, D., Filchenkov, A., Pivovarova, L., Žižka, J. (eds) Artificial Intelligence and Natural Language. AINL 2018. Communications in Computer and Information Science, vol 930. Springer, Cham.
[70] Stefanowitsch, A. (2020). Corpus linguistics: A guide to the methodology. Berlin: Language Science Press, 2020.
[71] Straka, M., Hajic, J., & Straková, J. (2016). UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). P. 4290-4297.
[72] Wiedemann, G. (2013). Opening up to big data: Computer-assisted analysis of textual data in social sciences // Historical Social Research/Historische Sozialforschung. P. 332-357.
[73] Wray, A. (2013). Formulaic language // Language Teaching, 46. P. 316–334.
[74] Yang C., Lin K. H. Y., Chen H. H. (2007). Emotion classification using web blog corpora //IEEE/WIC/ACM International Conference on Web Intelligence (WI'07). IEEE, 2007. P. 275-278.
[75] Zanettin, F. (2013). Corpus methods for descriptive translation studies // Procedia-Social and Behavioral Sciences, 95. P. 20-32.
[76] Zmitrovich, D., et al. (2023). A Family of Pretrained Transformer Language Models for Russian. ArXiv, abs/2309.10931.

Logical stress and gesture synonymy in the cephalic channel

Evdokimova A. A.
Institute of Linguistics, Russian Academy of Sciences /
1 bld. 1 Bolshoy Kislovsky Lane, 125009 Moscow, Russia
arochka@gmail.com

Abstract

In this article, based on the material of the Russian-language reference subcorpus RUPEX and the Spanish-language monologue subcorpus CAFE (comunicación de los artistas flamencos españoles) annotated in ELAN, we analyze head gestures that mark logical stress and consider all the variations that occur in this position. When determining the position of coincidence of a gesture and stress, we took into account the principle described by Grishina about the anticipation or delay of gestures depending on the strategy chosen by the speaker. In 20% of cases, the gesture anticipates the logical stress, starting a little earlier, and in 13%, on the contrary, it occurs immediately after the stressed syllable, highlighted by the accent of the word. According to our research, some of these head gestures are synonymous with each other in other positions as well. Some gestures were chosen by the subjects according to the characteristic features of their cephalic portrait (for example, moving the head forward with a sideways tilt), while others turned out to be typical that highlight significant words in a monologue and are characteristic of both Russian and Spanish (for example, the pragmatic gesture Down). When compared with the data of the MUMIN corpus research group and the authors of the Spanish-language corpus collected from "spontaneous" speeches on talent shows, principles were developed for describing gestures in the cephalic channel in marked positions from the point of view of different functional approaches, taking into account the following factors: the influence of other kinetic channels, the presence/absence of a listener, and the superposition of gesture functions on top of each other. In the Russian language, in 73% of cases, the opposite movement is adjacent to or behind the gesture, which indicates the visualization of "emphatic tone curvature" or different types of "skid". In the remaining cases, an intensification of the gesture is observed by anticipating it with the same one, but of smaller amplitude. Testing the annotation principles on two corpora showed their effectiveness as a basis for developing automatic head gesture annotation.

Keywords: cephalic channel; logical stress; corpus RUPEX; corpus CAFE; gesture synonymy; gesture functions
DOI: 10.28995/2075-7182-2025-23-1043-1054

Логическое ударение и синонимия жестов в цефалическом канале

Евдокимова А. А.
Институт языкознания РАН /
125009, Москва, Большой Кисловский пер. 1 стр. 1
arochka@gmail.com

Аннотация

В статье на основе проанотированного в ELAN материала русскоязычного эталонного подкорпуса RUPEX и испаноязычного монологического подкорпуса CAFE (comunicación de los artistas flamencos españoles) проанализированы жесты головы, маркирующие логическое ударение, и рассмотрены все вариации, которые встречаются в этой позиции. При определении позиции совпадения жеста и ударения мы учитывали принцип, описанный Гришиной о предвосхищении или запаздывании жестов в зависимости от выбранной говорящим стратегии. Жест в 20 % случаев предвосхищает логический акцент, начинаясь чуть раньше, а в 13 % наоборот, происходит сразу после ударного слога, выделяемого акцентом слова. Согласно нашим исследованиям, некоторые из этих жестов головы синонимичны друг другу и в других позициях. Одни жесты были выбраны испытуемыми согласно характерным особенностям их цефалического портрета (например, выдвижение головы вперед с наклоном вбок), а другие оказались типичными, выделяющими значимые слова в монологе и характерны и для русского, и для испанского языков (например, прагматический жест Down). При сравнении с данными группы исследователей корпуса MUMIN и авторами испаноязычного корпуса, собранного из «спонтанного» речей на шоу талантов, были выработаны принципы описания жестов в цефалическом канале в

маркированных позициях с точки зрения разных функциональных подходов с учетом таких факторов: влияние других кинетических каналов, наличие/отсутствие слушателя, наслоение функций жестов друг на друга. На материале русского языка в 73 % случаев «соседним» или «за жест до» оказывается противоположное движение, что свидетельствует о визуализации «эмфатического искривления тона» или разных типов «заносов», а в оставшихся случаях наблюдается усиление жеста путем его предвосхищения таким же, но меньшей амплитуды. Апробация принципов аннотирования на двух корпусах показала эффективность их использования как основы для разработки автоматической разметки жестов головы.

Ключевые слова: цефалический канал; логическое ударение; корпус RUPEX; корпус CAFE, синонимия жестов, функции жестов.

1 Введение

Продолжая исследования жестов головы, проведенные на материале корпуса RUPEX[1], в рамках которых была разработана методика аннотирования цефалического канала [22, 34] и были классифицированы некоторые из жестов головы [25, 26], мы хотим обратиться к двум взаимосвязанным сюжетам. Первый, важный для дальнейшей автоматической разметки такого рода жестов, это принципы выявления универсальных жестов, их описание, визуализация и создание электронного словаря. Второй — апробация разработанной методики на нерусскоязычном материале и проверка ее универсальности. Поскольку каждый из этих сюжетов достаточно объемен, для сокращения рассматриваемого материала и ограничения его одинаковыми условиями мы выбрали те случаи, когда жестовое поведение подчеркивает логический акцент.

2 Условия исследования

Определим те параметры исследования, из которых будем исходить.

2.1 Методология

Суть выбранного метода аннотирования движений головы [22, 26, 34] сводится к следующему. В программе Elan (https://archive.mpi.nl/tla/elan) весь поток видео размечается по видимым глазом движениям, где под движением мы понимаем однонаправленное действие (вниз, вверх, влево, вправо и т.п.) с заданной траекторией и протяженностью. Далее весь этот массив распределяется по слоям, согласно разработанным нами принципам [22, 34] и получает в зависимых слоях теги, характеризующие направление, тип движения, сопутствующий контекст, физические характеристики. С опорой на совокупность данных полученных слоев конкретизируются функции каждого из движений, и они собираются в более крупные группы (жесты, жестовые кластеры и т.п.) с уточнением их коммуникативных функций, границ и возможных наложений друг на друга. Мультимодальных работ с анализом цефалического канала на материале испанского языка не так много [12, 21] и они не содержат детального описания применяемой методики аннотирования. По этой причине движения головы предлагается аннотировать по принципам, разработанным на материале мультимодального корпуса RUPEX и тем самым апробировать универсальность этой методики для данных другого языка. При анализе вербального канала для корпуса RUPEX мы опирались на разметку, выполненную коллегами и представленную на сайте, а также на ее принципы [31]. Для испанского материала нами была выполнена на основе тонограмм, сделанных в программе Praat, собственная разметка, включившая в себя кроме слоев с разбивкой на слова и на ЭДЕ только слои с маркировкой просодических единиц.

2.2 Материал

Ключевым материалом анализа являлся мультиканальный корпус RUPEX («Рассказы и разговоры о грушах», подробнее см. сайт проекта www.multidiscourse.ru и [29]). Для исследования логического ударения было выбрано 2 из 3 записи «эталонного» подкорпуса – #04 и #22, длительностью 1 час 28 минут. Из 4 этапов каждой записи мы обратились к 2 и 4 этапу, поскольку они являются монологическими: Нарратор (N) рассказывает сюжет фильма для

[1] Более детально о проекте см. https://multidiscourse.ru/corpus/

Пересказчика (R), который фильма не видел; после уточняющего диалога R пересказывает Слушателю сюжет фильма. В качестве иноязычного материала использовался собранный нами испаноязычный корпус CAFE (comunicación de los artistas flamencos españoles), состоящий из видеоматериалов с перфомансов, спектаклей, интервью, импровизаций, мастер-классов и других выступлений артистов фламенко. На данном этапе исследования в этот корпус входит 405 видео различной длины от 0,5 минуты до 1,5 часов, из которых была составлена выборка в 17 видео с примерами спонтанной монологической речи артистов (длительностью 1 час 25 минут).

2.3 Подход к выделению и маркировке жестов в цефалическом канале

Функционально почти каждому жесту головы свойственна развитая омонимия и одно движение в зависимости от сопутствующих факторов (акценты в просодическом канале, семантика в вербальном, условия коммуникации и т.п.) может быть интерпретировано по-разному. Поэтому при аннотировании важно учесть все возможные наслоения смыслов и отразить их соответствующими тегами [22, 33]. Как было выявлено нашими коллегами [1, 2, 3, 5, 6, 9, 10, 11, 15, 16, 17, 18, 23] и нами [25, 26, 27] движения головы по своим функциям в коммуникации делятся на жесты прагматические (привлекающие внимание слушателя к деталям разговора, выражающие отношение говорящего к ним), указательные (указывающие на объект или субъект в том же пространстве или на их проекции), указательно-прагматические (например, Pragmatic center [25], с указанием на сопутствующие жесты рук или ног), изобразительные (образно показывающие предмет разговора, передающие его очертания или иные характеристики), регуляторные (регулирующие коммуникацию, позволяющие говорящему проверить реакцию слушающих), ритмические (отражающие ритмическую структуру дискурса) и аккомодаторы (служащие связками между жестами и используемые для корректировки позы головы) [26].

В статье, посвященной ритмическим жестам головы, бровей и рук, маркирующих акценты в испанском языке [12], представлен на материале интервью в программе «Operación Triunfo» корпус, где описание жестов в каждом из каналов было сосредоточено на анализе фаз жестов, приходящихся на акцент. Этот корпус хранится в сети: https://osf.io/m7tfr/?view_only=238699b07bc4429a9353ccccc8f56afa, и предложенная там таблица с разметкой материала показывает, что анализ спонтанного материала без маркировки типа и направления движений не дает адекватной картины происходящего. Более того, при просмотре видео, на основе которого выполнена разметка, выяснилось, что часть жестов, которые были приписаны голове, таковыми не являются. Они представляют собой перемещения (согласно нашей разметке Displacement), когда голова меняет координаты в пространстве за счет движений корпуса.

2.4 Функции жестов головы

Жесты головы, как они представлены в русском языке, делятся на две большие группы. В первой из них жесты имеют собственные названия: кивки (nods, jerks), повороты (turns), наклоны (tilts), мотания (shakes), вращения (rotations). Вторая группа описывается через их функциональный тип и/или вид движения (направление (down, up) или другие характеристики (slide)). Для полноты описания многие исследователи [5, 14, 15, 18, 23], предпочитают сочетать оба подхода, называя жесты по типу и конкретизируя характеристики движения, указывая функционал. Так появляются названия вида «прагматический вниз» или «указательный наверх». Однако, поскольку в цефалическом канале частотна функциональная синонимия, когда на один жест наслаиваются разные функции, обусловленные сопутствующими факторами: расположением слушающего по отношению к говорящему, местоположением рассматриваемого фрагмента в высказывании, семантикой слова, на которое приходится жест, нам представляется уместным в таких случаях при аннотировании перечислять все возможности. Например, в начале высказывания говорящий поворачивает голову направо, где сидит слушающий. С одной стороны, это регуляторный жест, с другой — прагматический, привлекающий внимание слушающего = «я начинаю». Если на это начало приходится и логическое ударение, то тогда такой жест будет еще и ритмическим, согласно терминологии, П. Экмана и У. Фризена в работе [7], а МакНилл в своей работе [19] называл их биениями (beats). В этой позиции такой поворот головы оказывается синонимичен жестам «голова вниз» или «кивок вниз», которые согласно исследованиям на материале разных языков [3, 11, 17, 23,], совпадают с эмфатически выделенными зонами текста и отмечают синтаксические границы

внутри высказывания или границу реплик в диалоге. Как показали исследования последних лет на материале русского языка [23, 25, 26, 27] многие из этих жестов имеют еще и прагматический характер.

2.5 Логическое ударение

Под логическим ударением мы будем понимать подвид фразового ударения, в том числе, и те случаи, когда оно совпадает с эмфатическим. На данном этапе исследования мы хотим проверить гипотезу, что жесты головы в монологической речи подчеркивают логическое ударение независимо от его типа как в русском, так и в испанском языках. Несмотря на тот факт, что испаноязычные исследователи отмечают, что в испанском языке ритмические жесты головы в сочетании с просодическими акцентами чаще приходятся на вторую часть высказывания, содержащую рему [12], мы сознательно не стали останавливаться на противопоставлении темы и ремы в нашем материале. Нам важнее посмотреть, насколько для цефалического канала сработает «manual McGurk effect» [4] и будет ли голова показывать подобно рукам направление тона при ударении. Ж. Кальбрис, анализируя речь Саркози в видеоинтервью на французском языке [6], подтверждает, что движения головы вверх или вниз воспроизводят восходящий и нисходящий тоны интонации, У. Хадар с коллегами указывают, что жесты головы на английском материале подчеркивают ударные слоги [9, 10], а Х. Граф с коллегами выделяет у кивков функцию подчеркивания просодии [8]. Е. А. Гришина отмечает, что жест головой вниз используется в русском языке для визуализации акцентных диакритик [23]. Поскольку Т.Е. Янко в своей статье пишет, что при эмфазе наблюдается падение тона при реме и подъем при теме, а им предшествуют противоположные движения, и этот феномен она называет «эмфатическим «искривлением» тона» [34], нам представляется интересным на материале цефалического канала в разбираемых корпусах посмотреть, будет ли визуализировано с помощью жестов это искривление при наличии эмфазы. А также будет ли наблюдаться что-то подобное в других случаях логического ударения, определяемых нами по выделенным релевантным изменениям тона. Фактически мы хотим определить для неэмфатических позиций, будут ли передаваться жестами происходящий на предударном слоге небольшой подъем частоты основного тона ("занос") или падение частоты ("обратный занос"), описанные в русской интонологии [13, 20, 24, 30, 32, 35]. Для испанского корпуса в программе Praat были построены тонограммы, которые позволяли верифицировать выделяемые акценты (рис. 1).

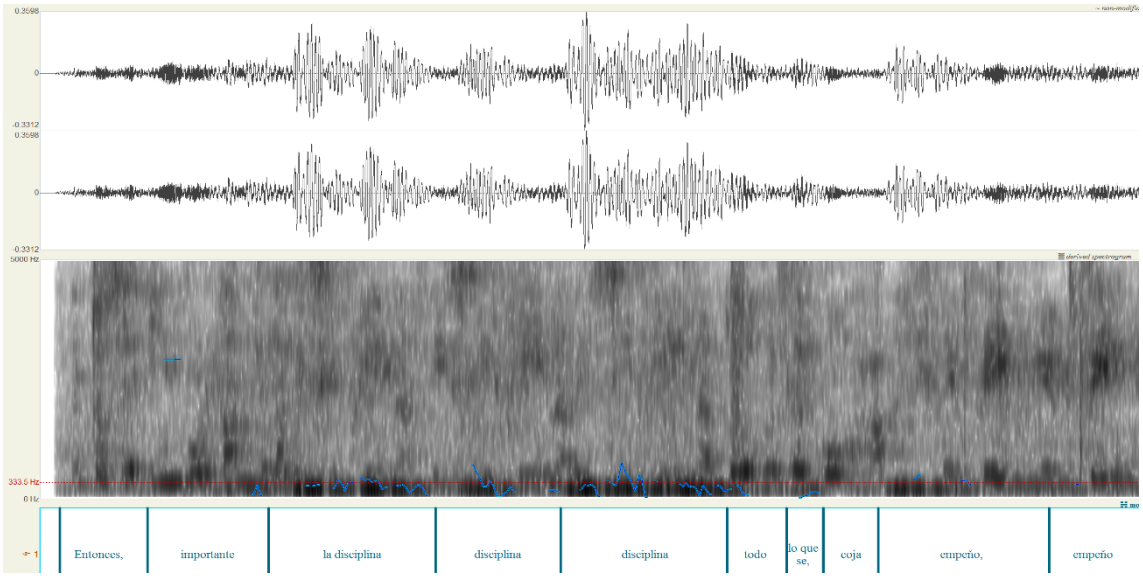

Рис. 1: Тонограмма предложения Кармен ла Телегоны: "Entonces, importante la disciplina, disciplina, disciplina, todo lo que se, coja, empeño, empeño"

3 Результаты

Выборка случаев с логическим ударением показала, что кроме лексем из акцентируемого класса слов (среди которых в подкорпусе RUPEX самое частотное «вот»), в разбираемых монологах, в основном акцентировались глаголы с семантикой движения и неодушевленные существительные («опорные точки текста», как груши или велосипед). По этой причине для последних двух категорий мы должны учитывать наслоения других функций жестов кроме ритмической и прагматической, связанных с семантикой, выраженной в вербальном канале.

Начнем наш анализ цефалического канала с распределения типов движений головы Нарратора из 22 записи.

Рис. 2: Распределение движений головы 22N

Как видно из графика (рис. 2), жесты в монологе Нарратора, совпадающие с логическим ударением, по направлению ориентированы большей частью к собеседнику (уже упомянутый Е.А. Гришиной [23] жест головой вниз и жест вперед). При анализе было выявлено, что они почти все оказались также прагматическими, в том числе указательно-прагматическими (Pragmatic center [25]) с указанием на жест рук. Некоторые из разбираемых жестов — изобразительные (как на глаголе движения «спускаться»). Несмотря на возникшее преобладание определенного типа движений, они разные по своим физическим характеристикам, например, длительность, амплитуда, скорость. Это наводит нас на мысль, что фиксация этих параметров и их дифференциация сможет в дальнейшем лучше расклассифицировать жесты такого рода и приблизить нас к автоматической разметке. Отдельного внимания заслуживает жестовый контекст, оказалось, что в 73% случаев «соседним» или «за жест до» оказывается противоположное движение. Например, перед поворотом головы влево, пришедшемся на ударение, оказывается поворот вправо, т.е. «занос», «обратный занос» и «эмфатическое искривление тона» дублируются жестами. В оставшихся 27%, что особенно характерно для движения головой вперед (Forward) или слайда влево (SlideLeft), наблюдается «жестовое усиление»: часто такое же по направлению движение предшествует основному, совпавшему с ударением, и иногда в рамках другого кинетического канала.

При сравнении с монологом Пересказчика в той же записи (рис. 3) жест головой вниз также значимо превышает другие варианты, однако, вторым по значимости оказывается наклон вправо (TiltRight).

Рис. 3: Распределение движений головы 22R

Поскольку положение корпуса Пересказчика таково, что при наклоне головы вправо у него улучшается обзор на тех участников, которые видели фильм, то в этих случаях происходит наслоение регуляторной функции на ритмическую, когда Пересказчик проверяет сказанное по реакции тех, кто ему рассказывал сюжет. Такое же распределение мы наблюдаем и в других записях эталонного подкорпуса, где также лидируют жесты головой вниз у участников обеих ролей. А, на втором месте, у Пересказчиков чаще всего встречаются жесты с наклоном или поворотом головы вправо. За исключением Нарратора из 4 записи (рис. 4), у которой регуляторные жесты (повороты головы влево или вправо, наклоны) оказались более частотными в разбираемых контекстах. А поворот влево, где сидел Слушатель, и вовсе обогнал по частотности движение головой вниз. В качестве фоновых жестов при наклонах и поворотах головы этот Нарратор часто использует движения корпусом вниз, что можно считать синонимом прагматического жеста головой вниз.

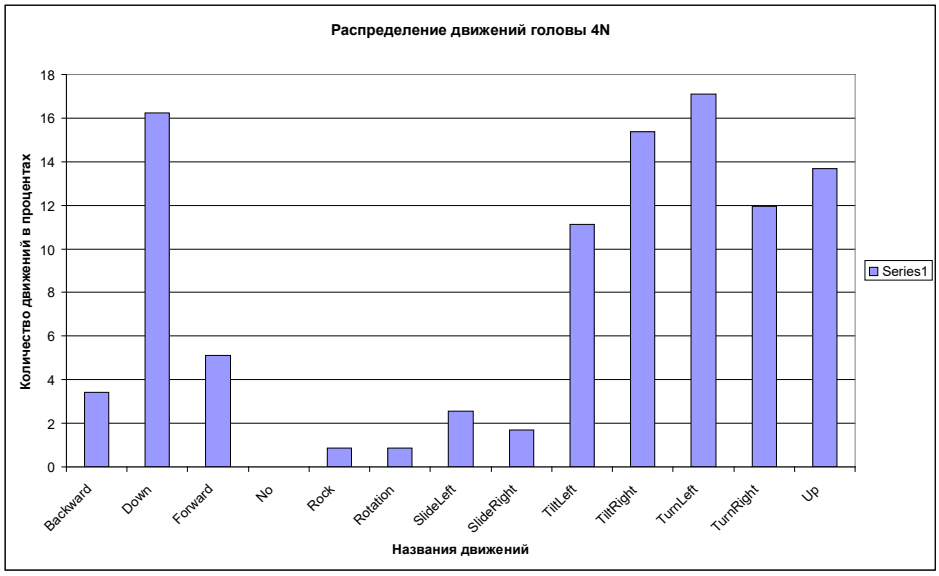

Рис. 4: Распределение движений головы 4N

Анализ распределения движений головы Пересказчика 4 записи (рис. 5) показал преобладание жеста головой вниз. Следующим по значимости оказалось движение головой наверх. Оба предпочитаемых жеста у этого испытуемого часто могли оказаться рядом, в соседних выделенных акцентом словах. Одно из движений фактически являлось иллюстрацией одного из видов «заноса» и использовалось для усиления второго движения. Такое же явление наблюдается и в паре поворот головы влево — поворот головы вправо.

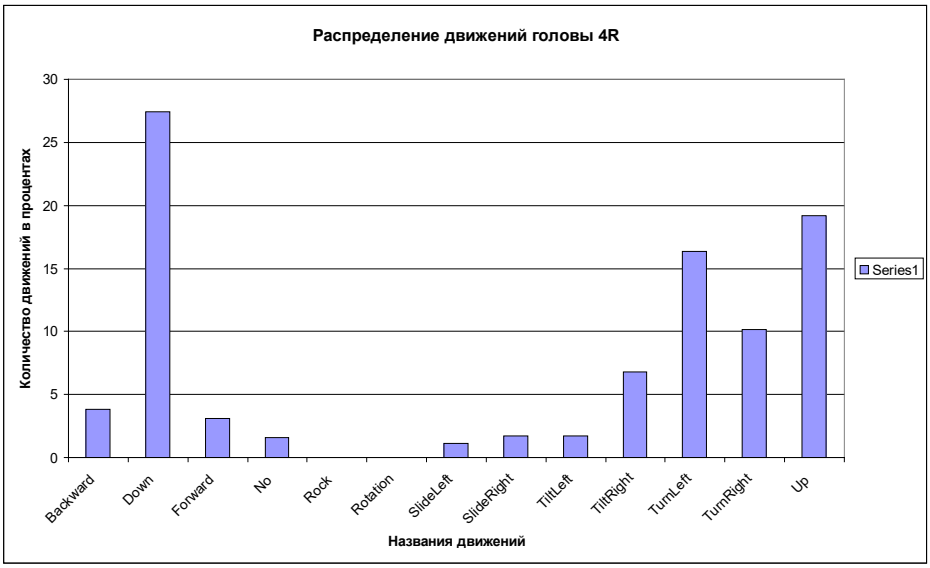

Рис. 5: Распределение движений головы 4R

Как было выявлено ранее на эталонном подкорпусе RUPEX, один из характерных моментов цефалического канала – это склонность к кластеризации жестов с другими типами движений [28]. Для Нарратора в 22 записи в 66 жестовых кластерах, пришедшихся на логическое ударение, лидирующее положение занимает движение головой вверх, которое в соединении с третьим по частотности движением головой назад составляет один из жестов дистанцирования [23]. Кластеров жестов у Пересказчика 22 записи — 40, с преобладанием жеста головой назад. У этого участника такое движение нехарактерно для его цефалического портрета, в отличие от Нарратора, поэтому использование им этого движения в рамках кластера жестов становится для нас маркированным и значимым для подчеркивания логического ударения.

Обратимся теперь к монологическому подкорпусу CAFE, содержащему спонтанные монологи, который отличается по своей структуре от RUPEX, так как видео, которые в него включены, короче по длительности (в среднем от 1 до 2,5 минут), хотя встречаются и более длинные, как беседы Рубена Ольмо (ок. 25-30 мин.). Поэтому для сбора материала о цефалическом портрете каждого из участников были добавлены 3-5 видеофрагментов, при этом из материалов интервью рассматривались только монологические отрывки. Это позволило учесть контекст привычных паттернов движений. Так, например, для Антонио Нахарро характерна очень небольшая амплитуда движений головы даже в интервью, так как он привык из-за танца с кастаньетами (рис. 6) максимально фиксировать голову, перемещая руки вокруг нее.

Рис. 6: Антонио Нахарро исполняет перформанс

Вследствие этой особенности, он и в своей хореографии [10], и в интервью (рис. 7) отдает предпочтение для маркировки логического ударения или акцента в музыке жесту головой вверх, а не вниз.

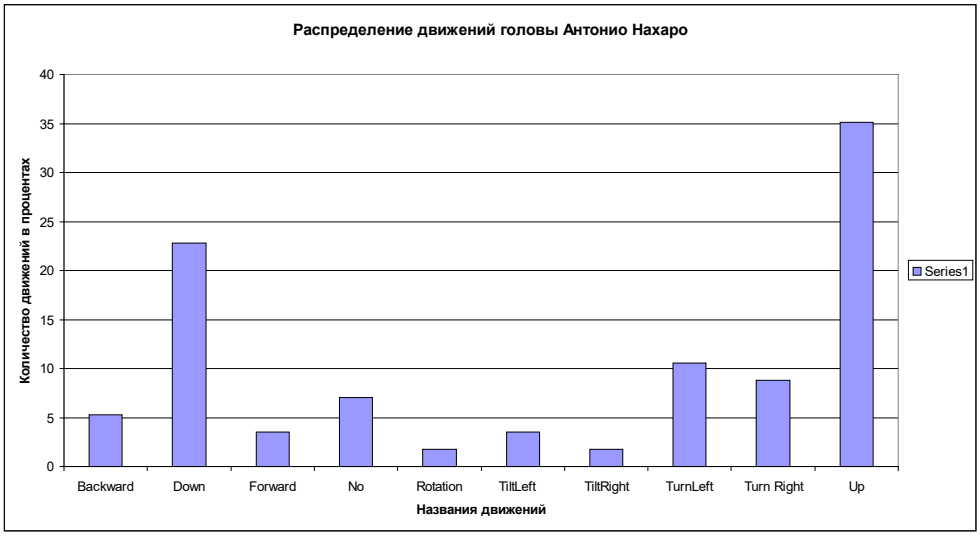

Рис. 7: Жесты головы в интервью и рилсах Антонио Нахарро

Иное мы наблюдаем при разметке фрагментов из интервью и занятий Кармен Ла Телегоны (рис. 8). Для ее цефалического портрета свойственна богатая жестикуляция головой с использованием как жестов, так и жестовых кластеров с большим числом фоновых движений корпуса. В интересующих нас контекстах она чаще использует прагматический и ритмический жест головой вниз, но может, в зависимости от места расположения собеседника, использовать регуляторные повороты головой в сторону собеседника. Надо отметить, что на проанализированных пяти ее видео, больше кластеров жестов наблюдается в ее рилсах, снятых во время занятий и бесед с учениками, чем в официальных интервью.

Рис. 8: Жесты головы в интервью и рилсах Carmen La Telegona.

Похожее распределение типов движений головы (рис. 9) мы получили при анализе цефалического поведения Рубена Ольмо в рамках его видео-бесед о фламенко с коллегами. Поскольку чаще его коллеги располагались от него справа, то регуляторный поворот головы вправо и наклон вправо для этих его видео характернее, чем соответствующие жесты влево.

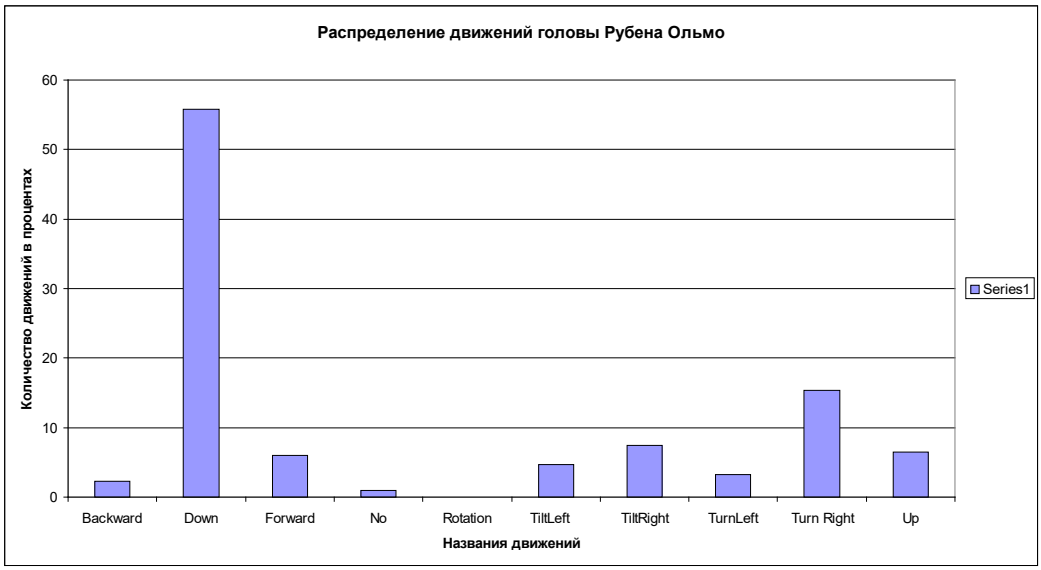

Рис. 8: Жесты головы в интервью Рубена Ольмо

У остальных участников, включенных в корпус CAFE, преобладает использование ритмического жеста "голова вниз", который иногда сопровождается фоновыми жестами прагматического характера и движениями корпуса назад и вперед. Использование кластеров жестов не так частотно как в русскоязычном корпусе, возможно, из-за бОльшего контроля движений тела вследствие профессии, выбранных в корпус испанцев. Использование предвосхищающего движения в противоположном направлении для увеличения амплитуды жеста, сопровождающего логическое ударение, также характерно для испанских артистов, что особенно, заметно в монологах Кармен Ла Телегоны и Рубена Ольмо.

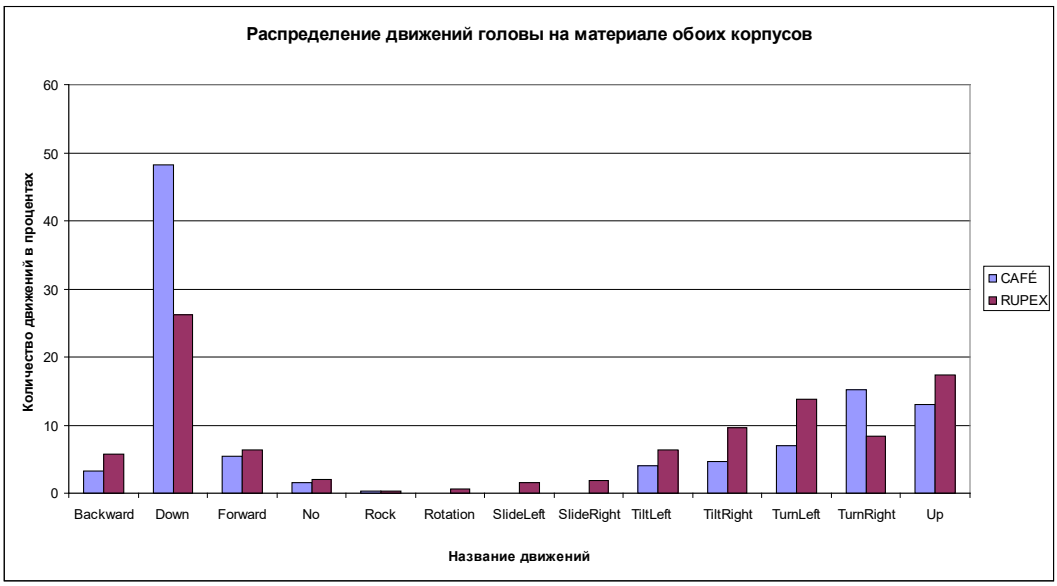

Рис. 9: Жесты головы в выборках из обоих корпусов, CAFE и RUPEX.

4 Заключение

Таким образом, в материале обоих корпусов значимо преобладают ритмические и прагматические жесты головой вниз (рис. 9) независимо от расположения говорящего по отношению к слушающим. В случаях, когда для говорящего важна реакция слушающего, а тот сидит не напротив него, но сбоку (как в случае Пересказчиков в RUPEX), вторыми по значимости оказываются жесты с дополнительной регуляторной функцией (повороты и наклоны головы в сторону собеседника). Фоновые жесты, повторяющие по направлению движения корпуса или другого кинетического канала, также сопровождают логическое ударение, часто в таких случаях входя в жестовые кластеры. Разнообразие вариантов распределения жестов в жестовых кластерах характеризует цефалический портрет говорящего. Однако, одиночные жесты в этих контекстах унифицированы независимо от цефалического портрета испытуемого и кроме ритмической функции сочетают и в русском, и в испанском языках прагматическую, указательно-прагматическую (Pragmatic center на жесты рук), регуляторную и для некоторых контекстов изобразительную. На разветвленность омонимии жестов в этих случаях оказывает влияние семантика лексем, с которыми они совпадают в вербальном канале. Для русского корпуса это глаголы движения и значимые для повествования неодушевленные существительные, для испанского корпуса упоминаемые коллеги, неодушевленные существительные и глаголы, характеризующие танец. Если говорить о совпадении, оно часто бывает неполным, жест примерно в 20% случаев предвосхищает логический акцент, начинаясь чуть раньше, а в 13% наоборот, начинается сразу после ударного слога, выделяемого акцентом слова. На материале русского языка в 73% случаев «соседним» или «за жест до» оказывается противоположное движение, что свидетельствует о визуализации движения тона («заносов» разного типа и «искривления тона») не только при эмфазе. В оставшихся случаях наблюдается усиление жеста путем его предвосхищения таким же по направлению и одной из коммуникативных функций, но обычно меньшей амплитуды. Если учесть все замеченные тенденции и проверить, есть ли корреляция между амплитудой и длительностью жеста и количеством накладывающихся функций и значений, то можно будет разработать алгоритм автоматической разметки функций жестов головы.

References

[1] Allwood J., Cerrato L., Dybkjaer L., Jokinen K., Navarretta C., Paggio P. The MUMIN multimodal coding scheme. Access mode: https://cst.dk/mumin/workshop200406/MUMIN-coding-scheme-v1.3.doc.

[2] Allwood J., Cerrato L., Jokinen K., Navarretta C., Paggio P. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena // Language Resources and Evaluation 41(3-4) — 2007. — P. 273-287

[3] Boholm M., Allwood J. Repeated head movements, their function and relation to speech // Proceedings of LREC workshop on multimodal corpora advances in capturing coding and analyzing multimodality. — Valetta, 2010.— P. 6–10.

[4] Bosker H. R., Peeters D. Beat gestures influence which speech sounds you hear. // Proceedings of the Royal Society B: Biological Sciences. — 288. — 2021. — doi: 10.1098/rspb.2020.2419.

[5] Calbris G. Contribution à une analyse sémiologique de la mimique faciale et gestuelle française dans ses rapports avec la communication verbale, 4 vol. (1. Expérimentation, 2. Taxinomie, 3. Synthèse, 4. Annexe illustrative), thèse d'État, 1983. 1478 p., 17 microfiches.

[6] Calbris G. La tête de Nicolas Sarkozy, ou les fonctions des gestes de la tête durant l'énonciation // Mots. Le langages du politique. № 86. — 2008.— P. 98–118. Access mode: https://journals.openedition.org/mots/14002.

[7] Ekman P., Friesen W. V. The repertoire of nonverbal behavior: categories, origins, usage and coding // Semiotica. V. 1. Iss. 1. — 1969.— P. 49–98.

[8] Graf H. P., Cosatto E., Strom V., Fu Jie Huang. Visual prosody: Facial movements accompanying speech // Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FGR'02). — Washington (DC), 2002. — P. 396–401.

[9] Hadar U., Steiner T. J., Grant E. C., Clifford Rose F. Kinematics of head movement accompanying speech during conversation. // Human Movement Science 2 — 1983 — P. 35-46.

[10] Hadar U., Steiner T. J., Clifford Rose F. Head movement during listening turns in conversation. // Journal of Nonverbal Behavior 9(4) — Winter 1985 — P. 214-228.

[11] Ishi C. T., Ishiguro H., Hagita N. Analysis of inter- and intra-speaker variability of head motions during spoken dialogue // Göcke R., Lucey P., Lucey S. (eds). AVSP-2008 Proceedings, International Conference on Auditory-Visual Speech Processing 2008, September 26–29, 2008; ISCA Archive Tanga-looma Wild Dolphin Resort, Moreton Island, Queensland. — P. 37–42.

[12] Jiménez-Bravo M., Marrero-Aguiar V. Multimodal prosody: gestures and speech in the perception of prominence in Spanish. // Frontiers in Communication. — Vol. 9. — 2024. — P. 01-20. — Access mode: https://www.frontiersin.org/journals/communication/articles/10.3389/fcomm.2024.1287363.

[13] Keijsper C. E. Comparing Dutch and Russian pitch contours. Russian Linguistics 7 — 1983 — P. 101 — 154.

[14] Kendon A. Some relationships between body motion and speech. An analysis of an example // Studies in Dyadic Communication, A. Siegman, B. Pope éd., — Elmsford, New York, Pergamon Press, 1972. — P. 177-210.

[15] Kousidis S., Malisz Z., Wagner P., Schlangen D. Exploring Annotation of Head Gesture Forms in Spontaneous Human Interaction // TiGeR 2013, Tilburg Gesture Research Meeting. — Access mode: https://tiger.uvt.nl/pdf/papers/kousidis.pdf.

[16] Krahmer E., Swerts M. The effects of visual beats on prosodic prominence: acoustic analyses, auditory perception and visual perception. Journal of Memory and Language 57 — 2007 — P. 396–414 — doi:10.1016/j.jml. 2007.06.005.

[17] McClave E. Linguistic functions of head movements in the context of speach. // Journal of Pragmatics 32 — 2000. — P. 855-78. Access mode: https://web.media.mit.edu/~cynthiab/Readings/McClave-99.pdf.

[18] McClave E. Cognitive universals: evidence from head movements in the context of speech // IIe ISGS Conference, Lyon, 15-18 juin 2005, Interacting Bodies. Abstracts. — 2005 — P. 128.

[19] McNeill D. Hand and Mind. What Gestures Reveal about Thought. — Chicago: 1992.

[20] Odé S. Russian intonation: a perceptual description. — Amsterdam: Brill, 1989.

[21] Pan R., García-Díaz J. A., Rodríguez-García M. Á., Valencia-García R. Spanish MEACorpus 2023: A multimodal speech–text corpus for emotion analysis in Spanish from natural environments. // Computer Standards & Interfaces, Volume 90. — 2024. — P. 103856. — https://doi.org/10.1016/j.csi.2024.103856. Access mode: https://www.sciencedirect.com/science/article/pii/S0920548924000254.

[22] Sukhova N.V., Evdokimova A.A. Annotating the cephalic channel // Fedorova Olga V., Kibrik Andrej A. (eds.) The MCD Handbook: A practical guide to annotating multichannel discourse. — To appear.

[23] Гришина Е. А. Русская жестикуляция с лингвистической точки зрения (корпусные исследования). — М.: Издательский дом ЯСК, 2017.

[24] Дурягин П.В. Интонация частного вопроса в русском языке: экспериментальное исследование источников вариативности // Русский язык в научном освещении 1 — 2021 — С. 137–177. — https://doi.org/10.31912/rjano-2021.1.6.

[25] Евдокимова А.А. Новые типы прагматических жестов головы — Pragmatic center и Pragmatic away // Лингвистика и методика преподавания иностранных языков. — 2020 — Т. 1 (12) — С. 136–148.

[26] Евдокимова А. А. К вопросу о методике выделения функциональных типов жестов в цефалическом канале в рамках мультиканального анализа // Когнитивная наука в Москве: новые исследования. Материалы конференции 23–24 июня 2021. Под редакцией Е.В. Печенковой, М.В. Фаликман, А.Я. Койфман. — М.: ООО Буки-Веди, 2021 — С. 515–520.
[27] Евдокимова А.А. Жесты головы в перформансах хореографа Антонио Нахарро. Постановка проблемы. // Язык-Музыка-Жест: информационные перекрестки (LMGIC-2024). Сборник материалов международной научной конференции, Санкт-Петербург, 18-20 Апреля 2024 г. / Ред. Эйсмонт П., Алексеева-Нилова Т. — СПб.: Скифия-принт. — 2024. — С. 36–38.
[28] Евдокимова А.А., Николаева Ю.В. Кинетические кластеры и их функциональные типы // Труды Института русского языка им. В.В. Виноградова. — 2022. — № 2. — С. 184–199.
[29] Кибрик А. А. Русский мультиканальный дискурс. Часть II. Разработка корпуса и направления исследований, Психологический журнал, Т. 39 (2) — 2018 — С. 79–90.
[30] Кодзасов С. В. Исследования в области русской просодии. — М.: ЯСК, 2009.
[31] Коротаев Н.А. «Рассказы и разговоры о грушах»: принципы вокальной аннотации (Версия 10.01.2019). Доступ: http://multidiscourse.ru.
[32] Оде С. Заметки о понятии тональный акцент на примере русского языка. // Проблемы фонетики. / Ред. Касаткина Р.Ф. — Москва: Наука, 2007. — С. 237–249.
[33] Сухова Н.В., Евдокимова А.А. "Рассказы и разговоры о грушах": аннотирование цефалических движений. (Версия 14.12.2018). Доступ: http://multidiscourse.ru.
[34] Янко Т.Е. Просодические средства эмфазы // Фонетика и нефонетика. К 70-летию Сандро В. Кодзасова. / Ред. А.В.Архипов, В. Федорова и др. — М.: Языки славянских культур, 2008. — С. 658-668.
[35] Янко Т.Е. Интонационные стратегии русской речи в сопоставительном аспекте. — М.: Языки славянских культур, 2008.

Backtranslation Invariance Boosts Effectiveness of Non-English Prompts

Kurtukova A.
NTR Labs, Tomsk, Russia;
Tomsk State University of Control Systems and Radioelectronics, Tomsk, Russia
akurtukova@ntr.ai

Kozachenko A.
NTR Labs,
Tomsk, Russia
akozachenko@ntr.ai

Abstracts

We present an approach to improving non-English prompts based on backtranslation invariance (the semantics of the prompt should not change after automatic translation to English and back). It improves prompts in non-English languages for a variety of Large Language Models (LLMs), including GPT-4-o, Llama-3.1, and Mixtral8x7B. We evaluate the approach for Russian and Finnish languages. In the benchmark of removing commas from a sentence, the proposed approach achieved an accuracy increase of 42% for Russian and 54% for Finnish compared to non-invariant prompts (LLaMA). In the benchmark of counting commas, accuracy increase of 19% for Russian and 11% for Finnish (GPT).

Keywords: Prompt engineering, large language models, translation, Backtranslation Invariance
DOI: 10.28995/2075-7182-2025-23-1055-1065

1 Introduction

As the large language models (LLMs) productized and became a go-to tool in various fields including natural language processing (NLP), translation and text generation, it became apparent that the performance of these models varies across languages [1-3]. Research shows that multilingual LLMs are more effective for the English language, which is due to prevalence of English in the data used for training [4-6].

Using LLM for non-English languages may lose effectiveness due to the peculiarities of specific languages. Inflectional languages (such as Russian) and agglutinative languages (such as Finnish) have specific features that create certain difficulties when working with LLM [7]. In such languages, the morphological structure can vary significantly depending on the context [8, 9], which requires the model to have a deep understanding of syntactic and semantic relationships [10]. Unlike English, in inflectional languages the diversity of forms can lead to increased ambiguity and complicate information processing [11]. In addition, many linguistic features such as word order, case use, and agreement can be ignored or misinterpreted by models that do not have sufficient experience working with specific language groups [12]. This calls into question the universality of approaches to prompting and processing texts in languages other than English.

The hypothesis of this study was that a specific formulation of a prompt in a non-English language provides LLM results comparable in quality to those obtained using English-language prompts. To achieve this goal, we hypothesized that it is important that the original prompt has invariance to backtranslation through English.

2 Prior research

The paper [2] is devoted to assessing the impact of non-English prompts on the effectiveness of a recommender system based on LLM. In the study, the authors considered both the out-of-the-box model and the T5 model retrained on multilingual prompts. The experiments were conducted for English, Turkish, and Spanish. Various prompting techniques were considered. The effectiveness of LLM was

assessed using the HitRate and NDCG metrics on open datasets: ML1M, LastFM, and Amazon-Beauty. The results obtained by the authors showed that the use of non-English prompts has a negative impact on the effectiveness of LLM as part of a recommender system. However, retraining LLM on multilingual prompts ensured uniform effectiveness of recommendations regardless of language. The efficiency measured by HitRate@10 for English decreased from 0.0679 to 0.0523, for Spanish from 0.0551 to 0.0505, and for Turkish increased from 0.0505 to 0.0523. The efficiency measured by NDCG@10 for English decreased from 0.0370 to 0.0288, for Spanish increased from 0.0297 to 0.0302, and for Turkish from 0.0269 to 0.0288.

The paper [3] presents an approach to qualitative and quantitative evaluation of LLM capabilities to work with multilingual data. The authors' approach is based on forward and backward translations and is tested for solving common sense reasoning and pun detection problems in order to determine the type of bilingualism demonstrated by LLM. To evaluate the approach, the authors used the GSM8K (primary school math problems) and CommonsenseQA (multiple-choice logical questions) and WebQuestions (question-answer pairs) datasets translated into French, Spanish, German, Japanese and Chinese. GPT-4 was used as a model. The obtained results demonstrated a significant difference in the efficiency of LLM for English and non-English languages. When solving math problems, prompting in English was on average 10% better than in other languages, and when solving logical problems - 15%. On WebQuestions, the results between European and English languages were approximately equal, while the effectiveness of LLM in Japanese and Chinese was lower by 16% and 28%, respectively.

The authors of the paper [4] considered the problem of transferring the capabilities of efficient generation and execution of LLM instructions to non-English languages. The study was based on the application of both out-of-box and pre-trained models of the LLaMA family (LLaMA, LLaMA2, LLaMA Chinese, etc.) to 4 standard benchmarks: C-Eval, MMLU, AGI-Eval, and GAOKAO-Bench. As part of the experiments, the authors evaluated the metrics of accuracy, fluency, informativeness, logical coherence, and harmlessness on the LLM-Eval dataset, which includes various educational tasks. The metrics were evaluated in 14 different languages: Chinese, Arabic, Vietnamese, etc. Based on the results obtained, the authors came to the following conclusions:

- expanding the vocabulary does not provide any advantages when training on data volumes of tens of billions of tokens;
- pretraining can improve the quality of responses, but does not always lead to a significant increase in the model's knowledge level;
- improving the LLM's ability to understand non-English languages is achieved at the expense of losing its initial understanding of English.

The authors of the paper [5] presented their MindMerge approach based on using the output data of one multilingual model trained for the translation task as input data for a multilingual LLM. This approach provided correct translation of not only prompts, but also user queries themselves. The efficiency of the approach was evaluated on the MGSM and MSVAMP datasets containing samples in Russian, Japanese, Spanish, and other languages. The difference in the efficiency of LLM with and without the approach reached 7.6% on average for all languages and 8% for low-resource languages.

Summarizing, several key conclusions can be drawn:

- The use of prompts in non-English languages negatively impacts the effectiveness of LLMs. Models predominantly trained on English data may distort information when processing other languages unless additional training is conducted.
- Retraining models on multilingual datasets improves their performance for non-English languages, highlighting the importance of tailoring LLM to specific languages to achieve uniform performance.
- Improvements in LLM's ability to understand non-English languages may come at a cost in English comprehension, indicating potential limitations of current approaches to training multilingual models.

3 An approach to prompt engineering based on Backtranslation invariance

We suggest an approach that is based on the hypothesis that the Backtranslation invariance of the prompt into English has a positive effect on the quality of LLM answers. We have first empirically observed this phenomenon when using compound sentences in prompts in Russian. Despite the fact that the prompt rules in Russian were formulated correctly and unambiguously, some of them were ignored by the LLM. To resolve this conflict, we decided to resort to machine translation of the prompt into English, and then from English to the original language. We observed that as a result of this manipulation some rules lost their original meaning as intended in the source language, and these were exactly the rules that were ignored by the LLM (likely, due to the impossibility of unambiguous interpretation).

The illustration of the approach to prompt engineering based on Backtranslation invariance is presented in Fig. 1. The approach includes:

- A cycle. Creating and modifying the original prompt or part of the prompt in language X to achieve invariance to backtranslation.
- Using machine translation from language X to English on the original prompt.
- Using machine translation on the resulting English prompt for backtranslation to language X.

If the prompt obtained after steps 2 and 3 matches the prompt from step 1, the cycle exits. Otherwise, the cycle repeats again, starting from step 1.

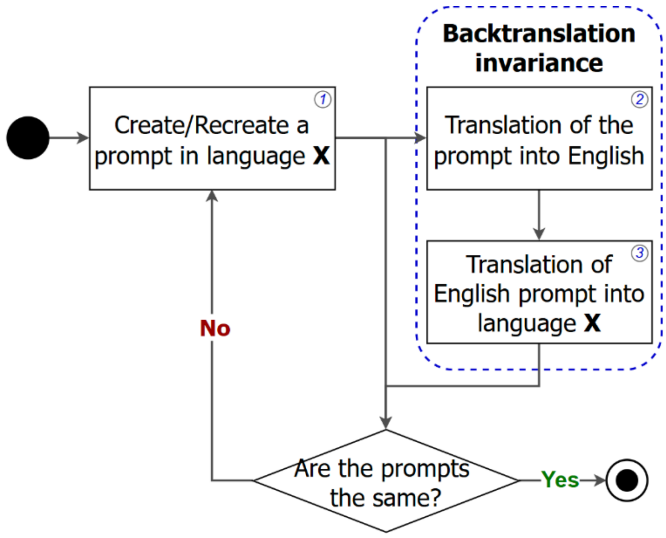

Figure 1: UML diagram of the approach

4 Dataset and sets of prompts

4.1 Test task

We created a small benchmark dataset to check the rule-following in different languages with ambiguous wording of the prompt. We created a small benchmark dataset to check the rule-following in different languages with ambiguous wording of the prompt. The first simple task was to remove commas from the text of a sentence if the word "I" was present. The second, more difficult task was to count the number of commas in the sentence.

The dataset was based on the prose of Russian literary classics (Tolstoy, Chekhov, etc.). The total number of works amounted to 297, with a total of 182,476 non-empty lines. Text segments (chunks) were filtered based on several criteria:

- A length of at least 10 words;
- Absence of Latin characters and punctuation marks indicating direct or indirect speech;
- Presence of at least one comma in the chunk – the number of such chunks was limited by 90% of the total dataset size;

- Absence of commas in the line – the number of such chunks was limited by 10% of the total dataset size;
- Presence of at least one pronoun “I” – the number of such chunks was limited by 70% of the total dataset size;
- Absence of the pronoun “I” – the number of such chunks was limited by 30% of the total dataset size.

The total number of chunks extracted was 15,436. To reduce computational load, 248 chunks were selected conforming to ratio requirements and taking into account that some chunks met multiple conditions simultaneously.

4.2 Prompt sets for English language

The English prompts were developed as a benchmark to show that the least ambiguous non-English prompts should achieve quality metrics comparable to the English prompts.

The first prompt for removing all commas from the text when the condition is met (En1):
“I will provide you with a sentence. Please rewrite it. Please answer in Russian only.
If the word ‘I’ is in the sentence, remove all commas from the text.
There should be nothing in the answer except for the rewritten sentence.
*Sentence: **{sample}**”*

The second prompt is for counting the number of commas in the text (hereinafter – En2):
“I will provide you with a sentence.
Count the number of commas in the text of the sentence and write only the number of commas.
*Sentence: **{sample}**”.*

4.3 Prompt sets for non-English languages

In this study, an elementary prompt was developed to perform the task of counting commas in a sentence. The elementary prompt (Ru1-1, Fi1-1), whose wording is transparent, was intentionally made difficult for the LLM to understand.

This study tested the invariance of back-translation of English cues. The potential effectiveness of this approach was hypothesized based on the fact that a large amount of English-language data is used to train multilingual LLMs, while the proportion of data in some non-English languages may be insignificant.

Table 1 presents the prompts for two languages other than English (Russian, Finnish). The wording of the first prompt with a conditional construction is presented with the prefix 1-, and the second prompt without a conditional construction is presented with the prefix 2- in the column **“**Substring **“**{prompt}**””**.

The second prompt meaning changes significantly when backtranslated through English, but for a native Russian speaker the differences between the formulations are not significant.

We calculate the word error rate (WER) metric for translation from the original language into English using the reference English text. The WER was also estimated for backtranslation through English to the original language. The obtained WER values were summed up for further analysis. The results obtained are presented in Table 2. The hypothesis of invariance to back translation suggests that if the final absolute value of WER between different prompts exceeds a certain threshold, then the probability of high-quality text processing by a LLM with such a prompt will be reduced.

Language	Original test prompt	Substring “{prompt}”	Notation
Russian	Я напишу тебе предложение. Перепиши его. Пожалуйста, отвечай на русском языке. {prompt} В ответе должно быть только переписанное предложение. Предложение: {sample}	Если внутри предложения есть слово "я", то удали все запятые из текста.	Ru1-1
		Обнаружив внутри текста предложения "я" в качестве слова, нельзя переписывать запятые из данного текста.	Ru1-2
	Я напишу тебе предложение. Перепиши его. Пожалуйста, отвечай на русском языке. {prompt} Предложение: {sample}	Ты обязан подсчитать количество запятых в тексте предложения и написать только количество запятых.	Ru2-1
		Посчитай запятых объём в тексте, написанном далее, предоставив для меня одну эту цифру только.	Ru2-2
Finnish	Kirjoitan sinulle ehdotuksen. Kirjoita se uudelleen. Vastaus venäjäksi. {prompt} Vastaus ei saa sisältää muuta kuin uudelleenkirjoitetun lauseen. Tarjous: {sample}	Jos lauseen sisällä on sana "mina", poista tekstistä kaikki pilkut.	Fi1-1
		Jos löydät sanan "I" lauseen tekstistä, et voi poistaa pilkkuja kyseisestä tekstistä.	Fi1-2
	Kirjoitan sinulle ehdotuksen. {prompt} Tarjous: {question}	Arvioi lauseessa olevien erottimien määrä ja merkitse vain pilkkujen määrä.	Fi2-1
		Laske, kuinka monta erotinmerkki koko tekstissä on, ja ilmoita pilkkua lukumäärä - tämä on pyyntö.	Fi2-2

Table 1: Prompts in non-English languages

Notation	Translation into English	Eng Tr. - Eng Orig. WER	Backtranslation	Src. Orig. - Src. Backtr. WER	WER Sum.
Ru1-1	If the word "I" is inside the sentence, then remove all commas from the text	0.21	Если слово "я" находится внутри предложения, то уберите все запятые из текста	0.5	0.71
Ru1-2	If you find "I" as a word inside the text of a sentence, you cannot rewrite commas from this text	1.0	Если вы встретите "I" в качестве слова в тексте предложения, вы не сможете переписать запятые из этого текста	1.00	2.0
Ru2-1	You must count the number of commas in the text of the sentence and write only the number of commas	0.22	Вы должны подсчитать количество запятых в тексте предложения и написать только их количество	0.31	0.53
Ru2-2	Count the volume of commas in the text written below, providing me with this one figure only	0.61	Подсчитайте количество запятых в тексте, написанном ниже, и получите только эту цифру	0.71	1.32
Fi1-1	Estimate the number of separators in the sentence and mark only the number of commas.	0.33	Arvioi lauseessa olevien erottimien määrä ja merkitse vain pilkkujen määrä.	0.0	0.33
Fi1-2	If you find the word "I" in the text of a sentence, you cannot remove commas from that text	0.79	Jos löydät sanan "I" lauseen tekstistä, et voi poistaa pilkkuja kyseisestä tekstistä	0.0	0.79
Fi2-1	Estimate the number of separators in the sentence and mark only the number of commas	0.00	Arvioi lauseessa olevien erottimien määrä ja merkitse vain pilkkujen määrä	0.0	0.00
Fi2-2	Calculate how many separator characters there are in the whole text, and indicate the number of comma - this is a request	1.06	Laske, kuinka monta erotinmerkkiä koko tekstissä on, ja ilmoita pilkun määrä - Tämä on pyyntö	0.27	1.33

Table 2: Results of WER metrics calculation

5 Experimental setup and Results

We consider 3 models: GPT-4, Mixtral and LLaMA. Information on the parameters for each LLM is presented in Table 3. The selection of parameters was based on the experience of using these models in RAG systems.

LLM	Configuration
gpt-4o-mini [12]	temperature: 0.14, top_p: 0.95, max_tokens: 4000
Mixtral-8x7B-Instruct-v0.1 [13]	temperature: 0.15, n_predict: 6000, top_p: 0.95, min_p: 0.05, repeat_penalty: 1.2, presence_penalty: 1
Meta-Llama-3.1-70B-Instruct [14]	temperature: 0.14, n_predict: 4000, top_p: 0.95, min_p: 0.04, repeat_penalty: 1.095, frequency_penalty: 0.01, presence_penalty: 1.3

Table 3: LLM Configuration

The LLM performance metrics were evaluated on the dataset described in detail in Section 3. For Russian and Finnish, experiments were performed with 2 equal-meaning prompts, the first of which (Ru1-1 and Fi1-1) was obtained using the Backtranslation Invariance-based approach described in Section 2, and the second by replacing words with synonyms or polysemous words.

To calculate the accuracy metric, the reference dataset was modified by applying strict rules im-plemented through regular expressions. For the first rule, described in prompts 1-1, 1-2, and En1, respectively, commas were removed only if the sample contained the letter “I” as a separate word. For the second rule, described in prompts 2-1 through 2-2 for non-English languages and En1 for English, all commas in the samples were counted.

The result of the LLM's work on the first proposal was considered correct if:

- commas were removed or counted according to the rule described in the prompt;
- the answer contained Cyrillic or only a number, as required by the prompt;
- the words contained in the answer were not glued together by removing space characters.

The result of the LLM's work on the second proposal was considered correct if:

- the first number that came up matched the number of commas in the sentence;
- the answer contained less than 50 words.

The results obtained for the Russian language are presented in Fig. 2 and Fig. 3. The graph in Fig. 2 shows the accuracy metrics for the case requiring the removal of commas when the sample contains the independent word “I”. Prompt En1 is the reference result, Ru1-1 has a formulation obtained by the developed approach, prompt Ru1-2 has a complicated formulation of the condition. Similar to the case without a condition, GPT and LLaMA show a result identical to that obtained for the English-language prompt - 61%. The difference in the performance of prompts Ru1-1 and Ru1-2 is 47% for GPT, 42% for LLaMA and 34% for Mixtral.

The graph in Fig. 3 shows the accuracy metrics for the case requiring counting the number of commas in a sample. Prompt En2 is the reference result, Ru2-1 has the formulation obtained by the developed approach, and prompt Ru2-2 has a complicated formulation of the problem. The best result using the developed approach reaches 53% using GPT, which is 19% higher than the result obtained using an ambiguously formulated prompt.

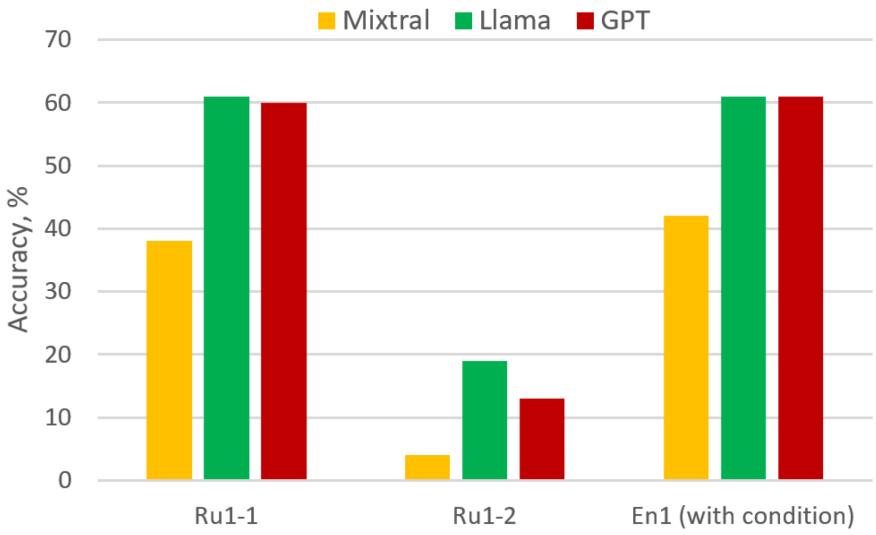

Figure 2: Accuracy metrics for Russian language (with condition)

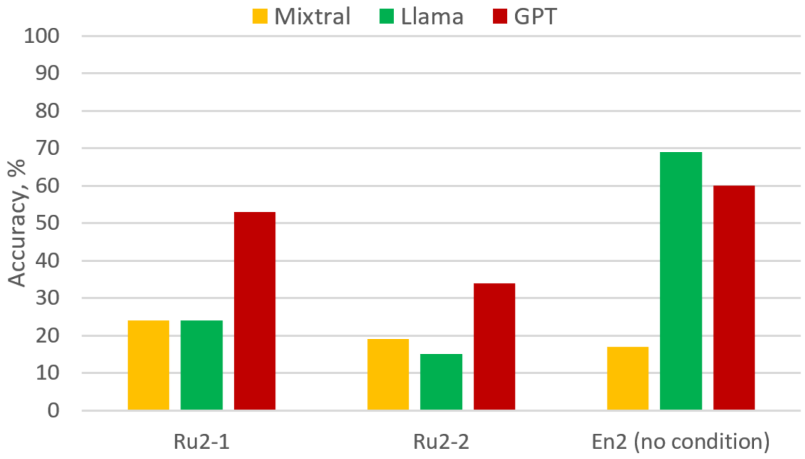

Figure 3: Accuracy metrics for Russian language (no condition)

The results obtained for Finnish are presented in Fig. 4 and 5. For the case with the condition (Fig. 4), GPT and LLaMA show the best result of 61% for both the reference prompt and the prompt obtained by the Backtranslation Invariance approach. The difference in accuracy between Fi1-1 and Fi1-2 reaches 54%. The graph in Fig. 5 shows the accuracy metrics for the case requiring counting the number of commas in a sample. Prompt Fi2-1 has the formulation obtained by the developed approach, and prompt Fi2-2 has a complicated formulation of the problem. The best result using the developed approach reaches 36% using GPT, which is 11% higher than the result obtained using an ambiguously formulated prompt.

It is important to note that the accuracy of Mixtral for Fi2-1 is slightly higher than that shown in the graph, this is due to the fact that, unlike other models that present results in digital format, its answer was presented in numerals.

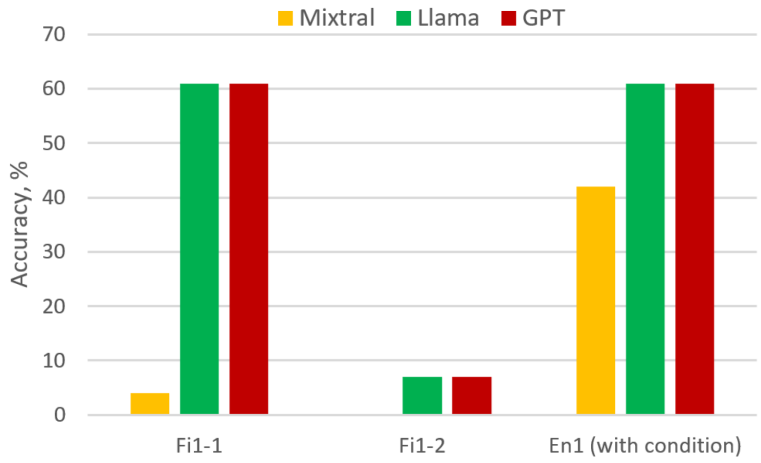

Figure 4: Accuracy metrics for Finnish (with condition)

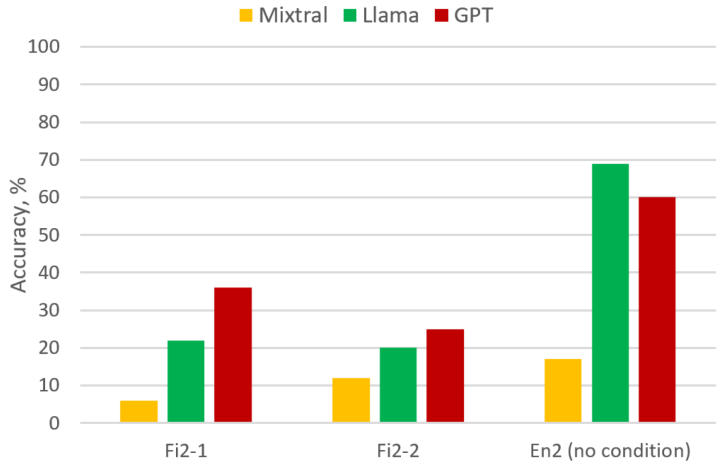

Figure 5: Accuracy metrics for Finnish (no condition)

According to the results of a series of experiments, an increase in the number of polysemous or synonymous words that may lose their original meaning when machine translated into English leads to a decrease in the ability of LLM to understand the context of the prompt and the tasks it contains. The obtained metrics also confirm that the approach based on backtranslation invariance allows achieving high efficiency of LLM, in some cases equal to that obtained when using an English-language prompt.

The key experimental results and comparison of the performance of LLM using the backtranslation invariance-based approach and without it are presented in Table 4.

Language	**LLM**	**Condition**	**Invariant prompt, %**	**Non-invariant prompt, %**	**Delta, %**
Russian	GPT	–	53	34	19
	LLaMA	+	61	19	**42**
Finnish	GPT	–	36	25	11
	LLaMA	+	61	7	**54**

Table 4: Summary of experimental results

6 Discussion

The results showed that the use of prompts with full back-translation invariance can positively affect the performance of language models for a given task. It is worth noting that only limited formulations of prompts were tested in this study. Although improvements in the performance of the language model were observed on specific examples, further research is needed to evaluate the scalability of the approach on more diverse datasets and tasks. It is important to study the applicability of the method to other languages, especially those where grammar and vocabulary features can affect the results of back-translation.

The current implementation of this approach requires significant manual effort. At the stage of prompt preparation, it is necessary to manually check their invariance using back-translation tools, such as Google and Yandex translators. This process can be labor-intensive and limit the scalability of the method. In the future, the approach can be automated by integrating with translator APIs. This will significantly speed up the process of checking the invariance of prompts. In addition, using synonym dictionaries to automatically substitute alternative formulations can simplify the creation of texts that are resistant to back translation. Automation of these processes will allow this method to be applied to a large number of diverse tasks.

7 Conclusion

In the course of the study, the particular effectiveness of Backtranslation Invariance for LLM prompting was established. The results of a series of experiments conducted for Russian and Finnish using three different LLMs showed that an unambiguous and transparent formulation of a non-English prompt allows achieving results comparable to those obtained with an English-language prompt. For each language, we evaluated the LLM with different formulations. The difference between the most confusing of them and the one obtained by the developed approach was 42% for Russian and 54% for Finnish. The best ability to understand the languages considered was demonstrated by the LLaMA-3.1 and GPT-4-o-mini models.

In future research, we plan to investigate Backtranslation Invariance in relation to languages other than English.

Acknowledgements

The authors would like to thank the anonymous reviewers who provided valuable feedback that significantly improved the article.

References

[1] Gabriel Nicholas, Aliya Bhatia. Lost in Translation: Large Language Models in Non-English Content Analysis // Computation and Language. – 2023. – Vol. arXiv:2306.07377. – Access mode: https://arxiv.org/abs/2306.07377.

[2] Makbule Gulcin Ozsoy. Multilingual Prompts in LLM-Based Recommenders: Performance Across Languages // Information Retrieval. – 2024. – Vol. arXiv: arXiv:2409.07604. – Access mode: https://arxiv.org/abs/2409.07604.

[3] Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, Grzegorz Kondrak. Don't Trust ChatGPT when Your Question is not in English: A Study of Multilingual Abilities and Types of LLMs // Computation and Language. – 2023. – Vol. arXiv:2305.16339. – Access mode: https://arxiv.org/abs/2305.16339.

[4] Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, Xuanjing Huang. LLaMA Beyond English: An Empirical Study on Language Capability Transfer // Computation and Language. – 2024. – Vol. arXiv:2401.01055. – Access mode: https://arxiv.org/abs/2401.01055.

[5] Zixian Huang, Wenhao Zhu, Gong Cheng, Lei Li, Fei Yuan. MindMerger: Efficient Boosting LLM Reasoning in non-English Languages // Computation and Language. – 2024. – Vol. arXiv:2405.17386. – Access mode: https://arxiv.org/abs/2405.17386.

[6] Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, Sadao Kurohashi. Beyond English-Centric LLMs: What Language Do Multilingual Language Models Think in? // Computation and Language. – 2025. – Vol. arXiv:2408.10811. – Access mode: https://arxiv.org/abs/2408.10811.

[7] Marion Di Marco, Alexander Fraser. Subword Segmentation in LLMs: Looking at Inflection and Consistency // Association for Computational Linguistics. – 2024. – Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. – P. 12050–12060. – Access mode: https://aclanthology.org/2024.emnlp-main.672.
[8] Garn Rimma. Interactive Russian Grammar: The Case System // Journal of the National Council of Less Commonly Taught Languages. – 2009. – Vol. 6. – P. 37–58.
[9] Timberlake Alan. A Reference Grammar of Russian // Cambridge University Press: Reference Grammars. – 510 p.
[10] Valentin Hofmann, Leonie Weissweiler, David Mortensen, Hinrich Schütze, Janet Pierrehumbert. Derivational Morphology Reveals Analogical Generalization in Large Language Models // Computation and Language. – 2024. – Vol. arXiv:2411.07990. – Access mode: https://arxiv.org/abs/2411.07990.
[11] Laurie Bauer. The function of word-formation and the inflection-derivation distinction // Words in their Places. A Festschrift for J. Lachlan Mackenzie. – 2004. – Access mode: https://www.wgtn.ac.nz/lals/about/staff/publications/Bauer-Infl-Deriv.pdf.
[12] Mostafa Abdou, Vinit Ravishankar, Artur Kulmizev, Anders Søgaard. Word Order Does Matter and Shuffled Language Models Know It // Association for Computational Linguistics. – 2022. – Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – Access mode: https://aclanthology.org/2022.acl-long.476.
[13] GPT-4o mini: advancing cost-efficient intelligence. – Access mode: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.
[14] Albert Q. Jiang et al. Mixtral of Experts // Machine Learning. – Vol. arXiv:2401.04088. – Access mode: https://arxiv.org/abs/2401.04088.
[15] Llama-3.1-70B-Instruct. – Access mode: https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct.

The methodology of multi-criteria evaluation of text markup models based on inconsistent expert markup

Alexander Levikin
MSU
s02210450@gmail.com

Ildar Khabutdinov
Antiplagiat Company
khabutdinov@ap-team.ru

Andrey Grabovoy
Antiplagiat Company
grabovoy@ap-team.ru

Konstantin Vorontsov
MSU Institute for Artificial Intelligence
vorontsov@mlsa-iai.ru

Abstract

A wide class of natural language processing tasks is solved using markup. At the moment, the vast majority of models and datasets rely on a simple markup structure containing only fragments and labels. Moreover, simple classification metrics such as F_1, Precision, Recall are used to evaluate the model's accuracy. The problem with such metrics is that they do not take into account all aspects of the markup structure and that they are applicable only under the assumption of the existence of an ideal markup. This paper proposes a more general and universal markup structure that allows solving complex problems and builds a methodology for multi-criteria evaluation of text markup models based on inconsistent expert markup. After that, the application of the constructed method is considered to assess the quality of the model obtained within the winning algorithm of the "READ//ABLE" competition, which focused on building an effective essay markup system. The results demonstrate that the new markup structure and evaluation approach provides a more comprehensive and accurate assessment of model performance, addressing the limitations of traditional metrics by accounting for complex markup scenarios and expert inconsistencies.

Keywords: multi-criteria assessment methodology, inconsistent text markup.
DOI: 10.28995/2075-7182-2025-23-1066-1080

Методика многокритериального оценивания моделей разметки текста по несогласованным экспертным разметкам

Александр Левыкин
МГУ
s02210450@gmail.com

Ильдар Хабутдинов
Антиплагиат
khabutdinov@ap-team.ru

Андрей Грабовой
Антиплагиат
grabovoy@ap-team.ru

Константин Воронцов
Институт ИИ МГУ
vorontsov@mlsa-iai.ru

Аннотация

С использованием разметок решается широкий класс задач обработки естественного языка. На данный момент подавляющее большинство моделей и датасетов опираются на простую структуру разметки, содержащую лишь фрагменты и метки. Более того, для оценки точности модели используются простые метрики классификации, такие как F_1, Precision, Recall. Проблема таких метрик в том, что они не учитывают все аспекты структуры разметки, и в том, что они применимы лишь в предположении существования идеальной разметки. В данной работе описывается более общая и универсальная структура разметки, позволяющая решать комплексные задачи, и строится методика многокритериального оценивания моделей разметки текста по несогласованным экспертным разметкам. После чего рассматривается применение построенного метода для оценки качества модели, полученной в рамках конкурса "ПРО//ЧТЕНИЕ", целью которого являлось создание эффективной системы разметки эссе. Результаты показали, что новые структура разметки и подход к оценке обеспечивают более полную и точную оценку эффективности модели, устраняя ограничения традиционных метрик за счет учета сложных сценариев разметки и несогласованности действий экспертов.

Ключевые слова: методика многокритериального оценивания, несогласованные разметки.

1 Introduction

1.1 Motivation and Contribution

The field of Natural Language Processing encompasses a diverse range of tasks aimed at extracting, analyzing, and utilizing information from textual data. One fundamental approach to solving these tasks is through the use of markup systems. Markup enables the identification, classification and annotation of textual elements. These include detecting manipulation in news articles (Ott et al., 2011), identifying evaluative language in texts (Wiebe, 2000), classifying documents by topic or category (McCallum and Nigam, 1998), determining sentiment polarity (Pang et al., 2002), recognizing emotions expressed in text (Alrasheedy et al., 2022) and automating the evaluation of students' essays (Khabutdinov et al., 2024). By systematically tagging textual fragments with appropriate labels a structured representation is created.

Current approaches to markup evaluation often rely on datasets and metrics that vary in complexity and scope. For instance, the MultiCoNER (Malmasi et al., 2022; Fetahu et al., 2023) dataset uses the BIO scheme to annotate tokens within sentences, with quality metrics like Precision, Recall, and F1 scores (Buckland and Gey, 1994; Kawata and Kikui, 2019) for both tagging and mention detection. However, this approach fails to account for partially matching spans, assumes a single reference markup, and lacks support for complex tasks such as linking fragments or adding multiple tags and comments.

Another example is the RURED (Gordeev et al., 2020) dataset, which includes named entities and their relations within texts. It employs metrics such as Cohen's Kappa (Sim and Wright, 2005) to measure inter-annotator agreement. While this dataset supports links between fragments, it does not accommodate overtexts, fragment combinations, or multiple tags for a single fragment or link.

Both approaches highlight limitations in existing systems, particularly in handling complex structures, multi-annotator scenarios, and nuanced evaluation criteria. Addressing these gaps is critical for advancing markup systems and their applications.

In order to identify more complex and composite language techniques, such as multistep manipulation, it becomes necessary to take into account the connection between fragments and group them. In addition, there is a desire to add comments and overtexts to the selected fragments. All these wishes are taken into account in the built markup structure described in the next chapter. After building the model, the question arises about evaluating the quality of its work. The difficulty of evaluation lies in the fact that there is no ideal, absolutely correct markup, but only a set of expert markups that differ slightly from each other. Therefore, when evaluating a model, we are not talking about its quality, but only about its consistency and similarity with experts. This article will propose a methodology for multi-criteria evaluation of text markup models based on inconsistent expert markup.

1.2 Method validation

We validate our approach at the "READ//ABLE" competition. The competitive task is to overcome a given technological barrier by building an algorithm for marking up Unified State Exam (USE) essays[1]. According to the procedure for conducting the final partitioning stage, it overcomes the technological barrier if the partitioning algorithm solves the competition problem and its average accuracy of algorithmic partitioning on the final sample is not worse than the average accuracy of expert partitioning calculated from expert partitions obtained under time-constrained conditions. We collaborate with the competition winner to explore the markup approach to the solution. The model architecture consists of various components to detect factual, logical, grammatical and speech (lexical violation) errors, as well as to highlight meaning blocks. Below is a brief description of the architecture.

Nowadays, one of the most effective open-source models in the Grammatical Error Correction task for the English language is the GECToR (Omelianchuk et al., 2020) model. For the grammar checker they have adapted the GECToR architecture for Russian and named it accordingly — RuGECToR (Khabutdinov et al., 2024). The choice of the architecture is due to the fact that it is easy to interpret and does not require a large amount of training data. The RuGECToR model is also utilised to check punctuation

[1] `https://fipi.ru`

compliance, despite the fact that punctuation compliance is not examined as part of the competition.

In order to verify compliance with speech norms, they use both rule-based and transformer-based (Vaswani et al., 2017) models. The classical model detects repetitions and tautologies in adjacent sentences, while the BERT-based (Devlin et al., 2019; Yang et al., 2019b; He et al., 2020; Warner et al., 2024; Liu et al., 2019) model finds more complex errors by classifying tokens.

The fact checker implements a pipeline for automated fact verification in text, combining document retrieval, segment extraction, and claim classification. The pipeline first uses an Anserini-based (Yang et al., 2017) search engine to retrieve relevant documents for a given claim. Extracted documents are processed by a Sentence Transformer (Reimers and Gurevych, 2019) to identify the most relevant segments based on cosine similarity between embeddings of the query and text passages. As knowledge bases they use collections of historical and literary documents, Wikipedia and news history. The BERT model for sequence classification then evaluates the relationship between the query and the retrieved segments.

The text logic error checker combines several approaches, the results of which are then aggregated.

The first approach has two main steps: candidate search and candidate classification. The candidate search starts with the comparison of candidate-reference pairs, where a candidate is a sentence in which a logical error is possible, and the reference is a fragment with which a logic violation occurs. Each pair is passed to the Question Answering (Yang et al., 2019a) BERT-based model input to refine the boundaries of the beginning and the end of the fragment. Then candidate-reference pairs with refined boundaries are fed to the input of the candidate classifier to get an error code or information that there is no logical error.

The second approach finds logical errors in the division of text into paragraphs, identifying cases where two paragraphs should be merged because they are logically related. Using the BERT model to evaluate the connectedness of paragraphs, the algorithm checks whether they can be merged without losing meaning.

The third approach also uses BERT-like models to predict the probability of logical succession between sentences in order to detect different types of errors. The first model evaluates the relationship between sentences using the Next Sentence Prediction (Shi and Demberg, 2019) task, and if the probability of logical succession between two sentences is low, it marks it as a logical sequence violation. The second model analyses the violation of causality between two sentences by binary classification.

The checker for meaning block detection in essays operates in several distinct stages. First, the input text is segmented into sentences. Next, the embeddings are passed through a BERT-base model to generate contextualized token representations. These representations are processed by a Conditional Random Field layer (Lafferty et al., 2001), which assigns a semantic label to each token based on its context and the subject-specific model. Predicted labels are aggregated to form contiguous spans representing meaning blocks. In the final stage, a post-processing step aligns the detected spans with sentence boundaries.

The evaluation results demonstrate that the algorithm achieves markup quality comparable to human annotators, particularly in overall consistency and tagging accuracy. While human annotators outperform the algorithm in fragment text consistency, the algorithm excels in maintaining consistent tagging and producing cohesive markup when fragments are aggregated. These findings highlight the algorithm's strengths in systematic tasks and suggest areas for improvement, particularly in nuanced text selection, to further align its performance with human capabilities.

2 Problem Statement

A generalized markup structure is proposed for consideration, which has the following form.

The L markup is a set of markup elements:

$$L = \{E_1, .., E_n\}, \tag{1}$$

where E_i is a markup element.

The markup element E is a triple:

$$E = (\{F_1, .., F_m\}, \{O_1^E, .., O_k^E\}, \{t_1^E, .., t_l^E\}), \tag{2}$$

where F_i is a fragment, O_i^E is a overtext, t_i is a tag (label).

Fragment F is a triple:

$$F = (s, f, \{t_1^F, .., t_v^F\}, \{O_1^F, .., O_q^F\}), \tag{3}$$

where $s, f \in \mathbb{R}$ are the beginning and end of the selected fragment, t_i is the tag, O_i^F is the text.

The O overtext is a two:

$$O = (C, \{t_1^O, .., t_p^O\}), \tag{4}$$

where C is a comment string, t_i is a tag. The superscripts F, E, or O emphasize that the tag/overtext refers to a fragment, markup element, or overtext, respectively.

A tag (the same as a label) t is an element of the tag dictionary, $t \in T$. The T tag dictionary is a set of words and phrases organized into a structure and used in markup, $T = \{t_1, .., t_n\}$.

To evaluate the consistency of the resulting markup, it is necessary to build a mapping $C(L_1, L_2)$:

$$C(L_1, L_2) : L_1 \times L_2 \longrightarrow [0, 1]. \tag{5}$$

3 Proposed Method

The comparison of the two markups' similarity is based on the F-measure:

$$F(A, B) = \frac{2|A \cap B|}{|A| + |B|}, \tag{6}$$

where A, B are some sets.

Let L_1, L_2 be markups of the same document. The consistency $C(L_1, L_2)$ of markups L_1, L_2 is a weighted average of criteria, each evaluating a part of the markup:

$$C(L_1, L_2) = \sum_{i=1}^{n} w_i \cdot C_i(L_1, L_2), \tag{7}$$

where $w_1 + ... + w_n = 1$, and $w_i \geq 0$. Each criterion assesses the similarity of certain markup components, such as overtexts, fragments, tags, etc. Moreover, each criterion can be composite and itself be a weighted average of its criteria.

Since some components of the markup (markup elements, fragments, overtexts) have a composite structure, to compute the similarity of sets consisting of them, it is necessary to establish an accordance between an object from one set $x_i \in X$ to an object from another set $y_j \in Y$, i.e., find the most similar objects from the two sets and associate them:

$$A_{X,Y} = \{(i_1, j_1), .., (i_q, .., j_q)\}, \tag{8}$$

where each index pair (i, j) means that element x_i is associated with element y_j. The best accordance, i.e., one that maximizes the consistency of the two sets, is called optimal accordance:

$$A_{X,Y}^{opt} : C(X, Y; A_{X,Y}^{opt}) = \max_{A_{X,Y}} C(X, Y; A_{X,Y}) \tag{9}$$

Note that in formula (7), the consistency $C(L_1, L_2)$ of markups L_1, L_2 is computed with the optimal accordance of markup elements, fragments, and overtexts: $C(L_1, L_2) \equiv C(L_1, L_2; A^{opt})$. Further in formula notations, the dependence on the optimal accordance will be omitted. Similarly, the consistency of fragments and overtexts in the final consistency is calculated with their optimal accordance. Thus, in the process of computing the consistency of sets of objects with a composite structure, the task of finding the optimal accordance is solved.

In markup tasks, there is a set of documents $D = \{d_1, .., d_n\}$, and each document d contains markups: $d = \{L_1, .., L_m\}$. It is assumed that each document contains algorithmic markup L^{alg} and several

expert markups. To evaluate the model, the following quantities are introduced: Average accuracy of algorithmic markup (AMA):

$$\text{AMA} = \frac{1}{|D|}\sum_{d\in D}\frac{1}{|d|-1}\sum_{\substack{L\in d,\\ L\neq L^{alg}}} C(L^{alg}, L) \tag{10}$$

Average accuracy of expert markups (EMA):

$$\text{EMA} = \frac{1}{|D|}\sum_{d\in D}\frac{2}{(|d|-1)(|d|-2)}\sum_{\substack{L_1,L_2\in d\\ L_1,L_2\neq L^{alg}}} C(L_1, L_2) \tag{11}$$

Relative accuracy of algorithmic markup (RMA):

$$\text{RMA} = \frac{\text{AMA}}{\text{EMA}} \tag{12}$$

If RMA ≥ 1, then it can be stated that the algorithmic markups are consistent with expert markups at least as well as expert markups are consistent with each other; in other words, it can be said that the markup algorithm works no worse than experts.

3.1 Consistency of markup elements accordance

This criterion calculated as the F-measure - the ratio of elements for which accordance was established — they "found a match" in the optimal accordance from the adjacent set:

$$C_A(L_1, L_2) = \frac{2\cdot|A_L|}{|L_1|+|L_2|}, \tag{13}$$

where $A_L \equiv A^{opt}_{L_1,L_2}$ is the optimal accordance of markup elements L_1, L_2.

3.2 Tags consistency

This criterion is F-measure for sets of tags of markup elements:

$$C_T(L_1, L_2) = \frac{1}{|A_L|}\sum_{(i,j)\in A_L}\frac{2\cdot|T_i\cap T_j|}{|T_i|+|T_j|}, \tag{14}$$

where $T_i = \{t^E_1, .., t^E_n\}$ is the set of tags of markup element E_i.

3.3 Overtexts consistency

This criterion calculated as a weighted average of criteria 3.3.1, 3.3.2, 3.3.3, averaged over all pairs from the accordance of markup elements:

$$C_O(L_1, L_2) = \frac{1}{|A_L|}\sum_{(i,j)\in A_L}\sum_{k=1}^{N} w_k\cdot C^O_i(E_i, E_j), \tag{15}$$

where $A_M \equiv A^{opt}_{L_1,L_2}$ is the optimal accordance of markup elements.

3.3.1 Consistency of overtexts accordance

This criterion calculated as F-measure - the proportion of overtexts for which accordance was established — they "found a match" in the optimal accordance from the adjacent set):

$$C^O_1(E_1, E_2) = \frac{2\cdot|A_O|}{|E^O_1|+|E^O_2|}, \tag{16}$$

where $A_O \equiv A^{opt}_{E^O_1,E^O_2}$ is the optimal accordance of overtexts of markup elements E_1, E_2. E^O_i is the set of overtexts of markup element E_i.

3.3.2 Consistency of overtexts texts

This criterion is the average F-measure of similarity between overtexts texts (as bags of words) of overtexts:

$$C_2^O(E_1, E_2) = \frac{1}{|A_O|} \sum_{(i,j) \in A_O} \frac{2 \cdot |C_i^* \cap C_j^*|}{|C_i^*| + |C_j^*|}, \tag{17}$$

where C_i^* is the representation of overtext comment O_i as a bag of words. Additionally, lemmatization and conversion to lowercase are performed.

3.3.3 Consistency of overtexts tags

This criterion is F-measure for sets of tags of overtexts:

$$C_3^O(E_1, E_2) = \frac{1}{|A_O|} \sum_{(i,j) \in A_O} \frac{2 \cdot |T_i \cap T_j|}{|T_i| + |T_j|}, \tag{18}$$

where $T_i = \{t_1^O, .., t_n^O\}$ is the set of tags of overtext O_i.

3.4 Fragments consistency

This creterion is composite and is calculated as the weighted average of criteria 3.4.1, 3.4.2, 3.4.3, 3.4.4, averaged over all pairs from the accordance of markup elements:

$$C_F(L_1, L_2) = \frac{1}{|A_L|} \sum_{(i,j) \in A_L} \sum_{i=1}^{N} w_k \cdot C_i^F(E_i, E_j), \tag{19}$$

where $A_L \equiv A^{opt}_{L_1, L_2}$ is the optimal accordance of markup elements, $N = 4$ in this case.

3.4.1 Consistency of fragments accordance

This criterion is calculated as F-measure - the proportion of fragments for which accordance was established — they “found a match” in the optimal accordance from the adjacent set:

$$C_1^F(E_1, E_2) = \frac{2 \cdot |A_F|}{|E_1^F| + |E_2^F|}, \tag{20}$$

where E_i^F is the set of fragments of markup element E_i, $A_F \equiv A^{opt}_{E_1^F, E_2^F}$ is the optimal accordance of fragments of markup elements E_1, E_2.

3.4.2 Consistency of fragments texts

This criterion is F-measure for selected text segments - ratio of doubled intersection length to the sum of their lengths:

$$C_2^F(E_1, E_2) = \frac{1}{|A_F|} \sum_{(i,j) \in A_F} \frac{2 \cdot |U_i \cap U_j|}{|U_i| + |U_j|}, \tag{21}$$

where $U_i = [s, f]_i$ is the selected text of fragment F_i.

3.4.3 Consistency of fragments tags

This criterion is F-measure for sets of tags of fragments:

$$C_3^F(E_1, E_2) = \frac{1}{|A_F|} \sum_{(i,j) \in A_F} \frac{2 \cdot |T_i \cap T_j|}{|T_i| + |T_j|}, \tag{22}$$

where $T_i = \{t_1^F, .., t_v^F\}$ is the set of tags of fragment F_i.

3.4.4 Consistency of overtexts fragments

This criterion is computed absolutely analogous to criterion 3.3.

3.5 Consistency of union of fragments texts

This criterion is F-measure for text fragments obtained by merging all fragments of a markup element:

$$C_{UF}(L_1, L_2) = \frac{1}{|A_L|} \sum_{(i,j) \in A_L} \frac{2 \cdot |U_i^* \cap U_j^*|}{|U_i^*| + |U_j^*|}, \tag{23}$$

where U_i^* is the union of texts (selected segments) of fragments of markup element E_i: $U_i^* = [s, f]_1 \cup ... \cup [s, f]_n$.

3.6 Consistency of union of fragments tags

This criterion is F-measure for two sets of tags, each obtained by merging sets of tags of fragments of its markup:

$$C_{UF}(L_1, L_2) = \frac{1}{|A_L|} \sum_{(i,j) \in A_L} \frac{2 \cdot |T_i^* \cap T_j^*|}{|T_i^*| + |T_j^*|}, \tag{24}$$

where $T_i^* = T_1 \cup ... \cup T_n$ is the union of tag sets of markup element E_i.

4 Experiments

The methodology described above was used in practice to evaluate the model built within the winning algorithm of the "READ//ABLE" competition. In the first subsection, we describe the structure of the competition, the markup and fields of the document. In the second subsection we describe how we generalized the described markup above for this competition. In the third subsection, we discuss the obtained results.

4.1 READ//ABLE description

The "READ//ABLE" competition is a technological challenge organized by the National Technology Initiative (NTI) in Russia. Launched in 2019, its goal is to stimulate the development of machine learning approaches capable of creating artificial intelligence systems that deeply understand text meaning and analyze cause-and-effect relationships across a wide range of topics.

The "READ//ABLE" competition is dedicated to the examination of USE essays for five school subjects: history, russian, english, literature and social. The competition's technological barrier involves developing a robust software system that can identify errors in academic essays, matching the performance of a human specialist within a limited time frame. Participants are tasked with creating intelligent systems that detect errors in essays of up to 12,000 characters, with a processing time of no more than 60 seconds per essay.

In December 2022, the Russian company "Antiplagiat" was declared the winner. Their solution demonstrated a quality level of 100.14% compared to human experts, earning them the prize of 100 million rubles.

The competition has been conducted in multiple cycles, with each cycle comprising qualification and final trials. Additionally, several satellite contests focusing on specific sub-tasks have been held to support teams in developing comprehensive solutions.

In this subsection we want to describe the components of the algorithm, as well as the data structure.

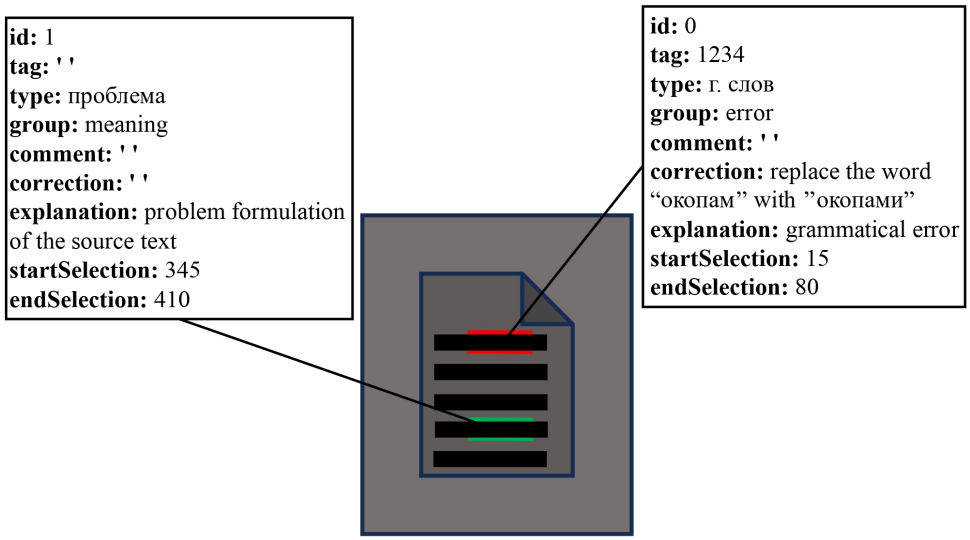

Figure 1: A description of the document markup from the competition “READ//ABLE” after execution of the algorithm.

Field Name	Description
id	Unique fragment number
group	“error” or “meaning”; type “error” indicates a localized error, type “meaning” evaluates reasoning blocks
type	Indicates the type of error or meaning block
tag	A string of unique letters or numbers linking related fragments; may be absent if localized
comment	Details the error if not present in the error classifier; otherwise, left empty
correction	Provides a corrected version of the fragment without errors
explanation	A detailed commentary applying to the highlighted fragment
startSelection	Fragment start position
endSelection	Fragment end position

Table 1: Description of fields for annotation of document fragments.

The document before evaluation contains two fields: meta information and essay text. After the essay has been checked, the criteria and document markup fields are added to the document. Fig. 1 shows an example of document markup. Table 1 describes the fields of the markup fragment.

The final grade is automatically calculated from the obtained markup according to the USE criteria.

Segmentable errors can be categorised into four general types — grammatical, speech (inappropriate or redundant use of words in context), logical and factual errors. It is also an additional task to segment the meaning blocks.

The main stages of the Essay Checking System are depicted in Fig. 2. Essay Checking System receives the essay document, when user send it for evaluation. Then it goes to Entrypoint — the main component, that routes the document to the checkers. The Essay checking system consists of five checkers, each of which is responsible for a specific task.

Most of the algorithms were trained on data that was provided by the competition organisers. The data were marked up by the USE experts. The internet was also parsed for essay texts to train unsupervised approaches.

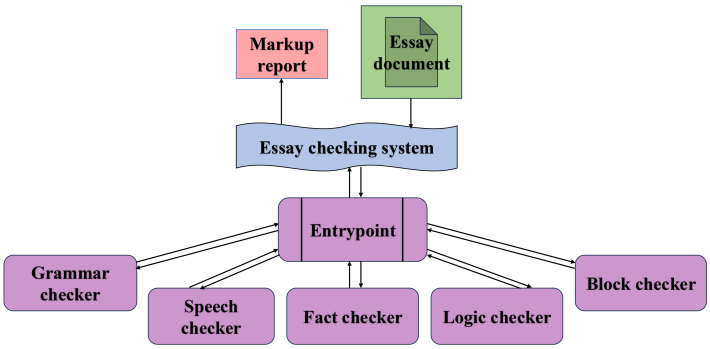

Figure 2: A figure depicting the process of producing document markup. The picture shows the main Entrypoint component, which sends the text of the document to the appropriate checkers to find errors or select meaning blocks, and then aggregates their results.

4.2 Evaluation metrics

The pairwise accuracy $M(X, Y)$ of annotation X relative to annotation Y is calculated as the weighted average of seven metrics $M_1(X, Y), \ldots, M_7(X, Y)$ with weights $w_1, \ldots, w_7$:

$$M(X, Y) = \frac{\sum_{i=1}^{7} w_i M_i(X, Y)}{\sum_{i=1}^{7} w_i}$$

Weights w_i determine the significance of each metric, with $w_i = 0$ excluding a metric.

Essay Score Prediction Accuracy

Measures the match between essay scores derived from annotations X and Y:

$$M_1(X, Y) = \left(1 - \frac{\sum_i |K_i(X) - K_i(Y)|}{\max K}\right) \cdot 100\%$$

Fragment Detection Accuracy and Recall

Evaluates matching fragments in annotations X and Y. Let us introduce a set D of fragment pairs (i, k) such that each $x_i \in X$ corresponds to at most one y_k and each $y_k \in Y$ corresponds to at most one x_i:

$$\text{Precision} = \frac{|D|}{n}, \quad \text{Recall} = \frac{|D|}{m}, \quad M_2(X, Y) = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Code Prediction Accuracy

Proportion of matched fragments in document D with identical codes:

$$M_3(X, Y) = \frac{1}{|D|} \sum_{(i,k) \in D} [\text{type}(x_i) = \text{type}(y_k)]$$

Subtype Prediction Accuracy

Proportion of matched fragments in document D with identical error subtypes:

$$M_4(X, Y) = \frac{1}{|D|} \sum_{(i,k) \in D} [\text{subtype}(x_i) = \text{subtype}(y_k)]$$

Fragment Localization Accuracy

Measures overlap using the Jaccard index:

$$M_5(X,Y) = \frac{1}{|D|} \sum_{(i,k)\in D} \frac{|x_i \cap y_k|}{|x_i \cup y_k|}$$

Correction Accuracy

Proportion of fragments with identical corrections:

$$M_6(X,Y) = \frac{1}{|D|} \sum_{(i,k)\in D} [\text{correction}(x_i) = \text{correction}(y_k)]$$

Explanation Accuracy

The average expert judgement of explanation accuracy across all markup fragments that have explanations. This is the only metric based not on comparison with the markup, but on experts' evaluations. Experts score each explanation in the tested algorithmic markup from 0 to 5 points. The total score is made up of answers to the following questions regarding the given explanation:

1. It is most likely to be understood by the author of the essay.
2. It correctly explains the essence of the error or gives a relevant reference to the source.
3. It leaves no opportunity for appeal.
4. It refers to the text of the work and specifically to the highlighted fragment
5. It solves the pedagogical problem and helps to avoid similar mistakes in the future.

If the examiner considers that the fragment is not an error or does not require an explanation, then it is expected to give zeros in all questions, and the mark for this explanation should be zero. The explanation in the expert markup automatically gets maximum. In order to reduce labour costs, expert checking of explanations is only carried out during the Final Tests.

$$M_7(X,Y) = \frac{\text{Expert Score}}{\text{Maximum Score}} \cdot 100\%$$

Optimistic Accuracy

Optimistic relative pairwise accuracy of the algorithmic markup of a single essay, when compared to the entire set {E} of expert markups of that essay:

$$M_{\text{opt}}(A, \{E\}) = \frac{\max_E M(A,E)}{\min_{E,E'} M(E,E')} \cdot 100\%$$

Average Accuracy

The average relative pairwise accuracy of the algorithmic markup of one essay, when compared to the entire set {E} of expert markups of that essay:

$$M_{\text{avg}}(A, \{E\}) = \frac{\text{avg}_E M(A,E)}{\text{avg}_{E,E'} M(E,E')} \cdot 100\%$$

Overall Relative Accuracy (OTAR)

Combines optimistic and average accuracy using parameter H:

$$\text{OTAR} = \frac{H \cdot \text{avg}_E M(A,E) + (1-H) \cdot \max\limits_E M(A,E)}{H \cdot \text{avg}_{E,E'} M(E,E') + (1-H) \cdot \min\limits_{E,E'} M(E,E')} \cdot 100\%$$

The prerequisite for winning the competition is overcoming the technological barrier of OTAR > 100%.

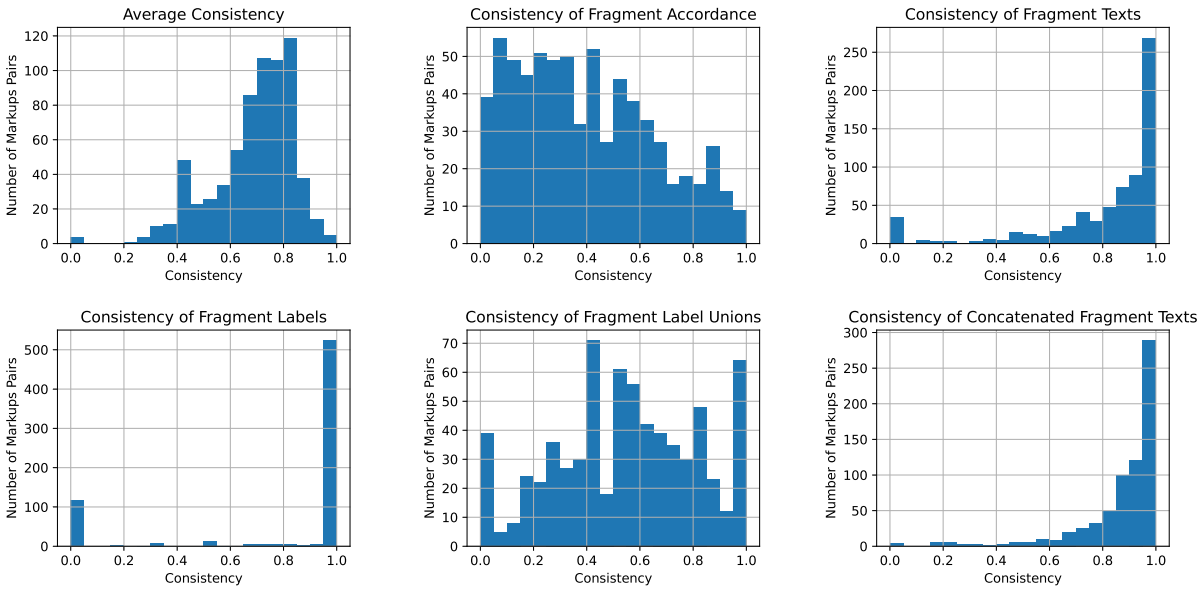

Figure 3: Consistency criteria for pairs of annotations made by annotators

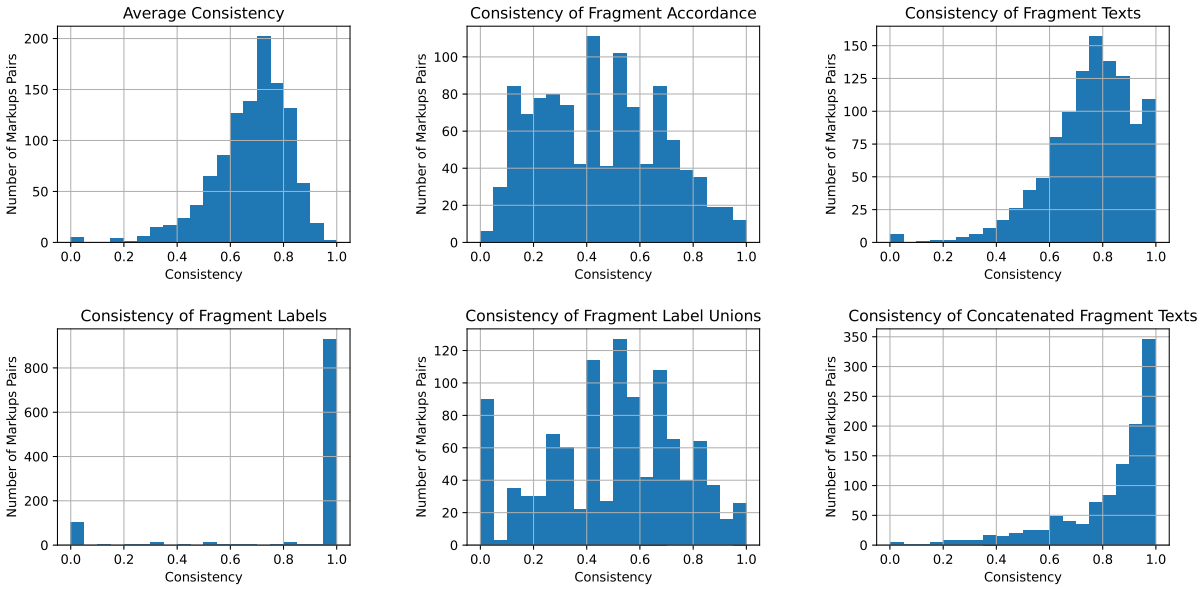

Figure 4: Consistency criteria for pairs of annotations, one made by the algorithm and the other by a human

4.3 READ//ABLE winning algorithm evaluation

The markup structure within the competition was simple, as the markup contained only one markup element, which consisted of fragments with tags, and the overtexts left by the model were not considered since the annotators did not leave them. In section 4.2, we showed a special case of approbation of the developed metrics from section 3. These metrics were applied to validate the competition. The winning system scored an accuracy of 100.14% compared to the average markup of the USE expert — the OTAR metric.

A set of 500 documents was considered, with 1595 annotations of these documents made within the “READ//ABLE” competition. Among them, 1005 annotations were made by annotators, and 500 were

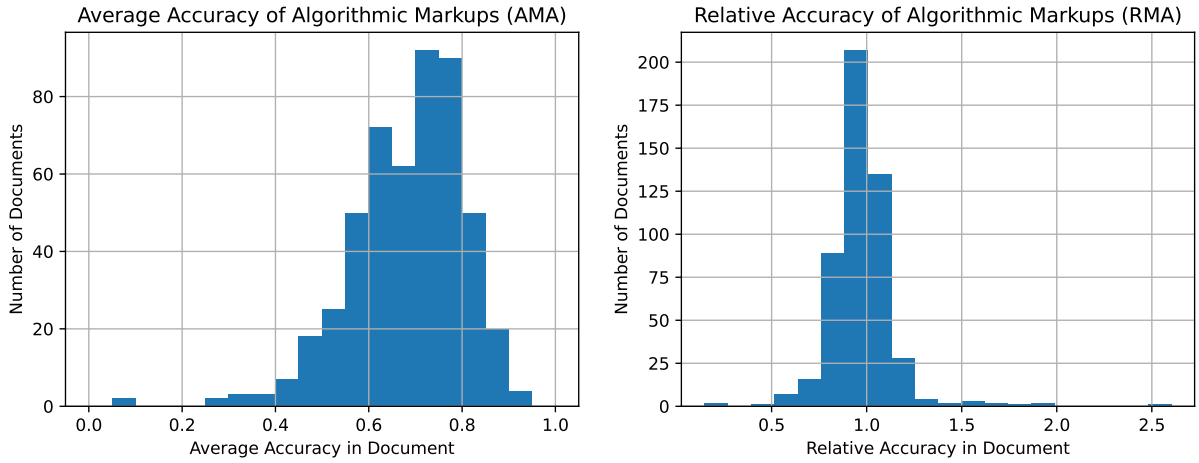

Figure 5: AMA and RMA metrics for the model obtained within the “READ//ABLE” competition

made by the model. Histograms of each of the consistency criteria and their average value are presented in Fig. 3 and Fig. 4. Fig. 5 presents the histogram of AMA and RMA metrics.

4.4 Results discussion

The evaluation of algorithmic markup compared to human annotators reveals several key insights into the model’s consistency, accuracy, and potential limitations. In order to conclude about the consistency of the algorithm with the experts and the experts with each other we examine the histograms in Fig. 3 and Fig. 4.

The histogram for "Consistency of Fragment Texts" shows that human annotations exhibit higher clustering around higher consistency values, whereas the algorithm demonstrates more significant variation. This suggests that while the algorithm is effective, there are cases where it either extracts incorrect fragments or fails to identify certain errors consistently.

Comparing human annotations with the algorithm’s output in terms of fragment tags, we observe that the algorithm achieves a relatively high level of accuracy. The histogram for "Consistency of Fragment Labels" suggests that the algorithm’s tagging process aligns well with human annotators, though minor discrepancies exist. Specifically, human annotators tend to be more rigid in their label selection, while the algorithm exhibits greater variability. One explanation is that the algorithm relies on probabilistic methods or learned patterns rather than strict rule-based tagging. While this allows it to generalize well, it also introduces occasional misclassifications.

The histogram for "Consistency of Fragment Label Unions" indicates that both the algorithm and human annotators demonstrate significant variability. It suggests that the algorithm performs on par with human annotators, indicating that its generalization capability is relatively strong. This is an encouraging result because it demonstrates that even though the algorithm is not perfect at fragment identification on a case-by-case basis, its overall trend aligns with human judgments.

In the "Consistency of Fragment Accordance" histogram, we notice a relatively wide distribution, with many of instances having low consistency values. This suggests that even human annotators exhibit notable differences in how they mark fragments, meaning that essay annotation is inherently subjective. This result emphasizes the need to use consistency metrics in markup tasks, since in the case of standard metrics, it is not obvious what to use as ground truth.

In Fig. 5 the RMA histogram shows that the model’s accuracy is centered around 1, meaning that it agrees with human annotators on average as much as they agree with each other. However, the presence of some extreme cases where the RMA deviates significantly suggests that there are specific instances where the algorithm either outperforms or underperforms compared to human annotators. In cases where

RMA > 1, the algorithm likely follows more rigid, rule-based logic that leads to greater consistency than human annotators, who may be influenced by subjectivity. In cases where RMA < 1, the algorithm struggles with contextual nuances that humans naturally interpret more accurately.

Fig. 5 shows that the AMA values are concentrated around higher accuracy levels, there is still some distribution toward lower values, indicating cases where the algorithm struggles. These outliers likely correspond to edge cases where human judgment plays a significant role, such as ambiguous errors or unconventional phrasing in essays.

The results show that the algorithm achieves human-comparable annotation accuracy, though inconsistencies in fragment selection highlight the subjectivity of human markup. While the model performs well in structured error detection, it struggles with context-sensitive cases and hierarchical relationships between errors. Future improvements should focus on refining contextual understanding and integrating expert feedback to enhance annotation consistency. We also see the need to use consistency metrics as essay evaluation is very subjective, and if standard NLP metrics are used, it is not clear what counts as true.

5 Conclusion

Evaluating the quality of textual markup in tasks involving subjective and context-dependent annotations remains a significant challenge in Natural Language Processing. Standard evaluation metrics often fail to capture the nuanced differences between human and algorithmic annotation, especially when multiple valid interpretations are possible. This is crucial in such applied tasks as essay evaluation, personalized writing feedback, grammar and style correction, and intelligent tutoring systems.

In this study, we introduced a multi-criteria evaluation method for assessing markup consistency and quality, which allows for detailed comparison between human annotators and automated algorithms, taking into account all the features of its generalized structure. This method was applied to the "READ//ABLE" competition, providing valuable insights into the strengths and weaknesses of the evaluated language model.

Our analysis revealed that the algorithm demonstrates quality comparable to human annotators in fragment tagging consistency and overall markup cohesion. However, the model exhibited lower performance in fragment text consistency, suggesting that while it excels in systematic and structural tasks, there is room for improvement in handling the subtleties of text extraction.

The possibility of including/excluding additional criteria and changing the weighting coefficients ensures the adaptability of the evaluation method to a wide range of markup structures and tasks. This makes it a robust tool for assessing algorithms in various Natural Language Processing problems.

References

Mashary Alrasheedy, Ravie Muniyandi, and Fariza Fauzi. 2022. Text-based emotion detection and applications: A literature review. P 1–9, 10.

Michael K. Buckland and Fredric C. Gey. 1994. The relationship between recall and precision. *J. Am. Soc. Inf. Sci.*, 45:12–19.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. // Jill Burstein, Christy Doran, and Thamar Solorio, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, P 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Besnik Fetahu, Zhiyu Chen, Sudipta Kar, Oleg Rokhlenko, and Shervin Malmasi. 2023. MultiCoNER v2: a large multilingual dataset for fine-grained and noisy named entity recognition. // Houda Bouamor, Juan Pino, and Kalika Bali, *Findings of the Association for Computational Linguistics: EMNLP 2023*, P 2027–2051, Singapore, December. Association for Computational Linguistics.

Denis Gordeev, Adis Davletov, A. Rey, G. Akzhigitova, and G. Geymbukh. 2020. Relation extraction dataset for the russian. P 348–360, 01.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *CoRR*, abs/2006.03654.

Naotaka Kawata and Genichiro Kikui. 2019. Mention detection method for entity linking. // *2019 IEEE International Conference on Big Data, Cloud Computing, Data Science & Engineering (BCD)*, P 41–45.

I. A. Khabutdinov, A. V. Chashchin, A. V. Grabovoy, A. S. Kildyakov, and U. V. Chekhovich. 2024. Rugector: Rule-based neural network model for russian language grammatical error correction. *Program. Comput. Softw.*, 50(4):315–321, July.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. // *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, P 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. cite arxiv:1907.11692.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. // Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, *Proceedings of the 29th International Conference on Computational Linguistics*, P 3798–3809, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Andrew McCallum and Kamal Nigam. 1998. A comparison of event models for naive bayes text classification. // *AAAI Conference on Artificial Intelligence*.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. // Jill Burstein, Ekaterina Kochmar, Claudia Leacock, Nitin Madnani, Ildikó Pilán, Helen Yannakoudakis, and Torsten Zesch, *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, P 163–170, Seattle, WA, USA → Online, July. Association for Computational Linguistics.

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*, HLT '11, P 309–319, USA. Association for Computational Linguistics.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. // *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, P 79–86. Association for Computational Linguistics, July.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. // *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11.

Wei Shi and Vera Demberg. 2019. Next sentence prediction helps implicit discourse relation classification within and across domains. // Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, P 5790–5796, Hong Kong, China, November. Association for Computational Linguistics.

Julius Sim and Chris C Wright. 2005. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, 85(3):257–268, 03.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. // *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, P 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.

Janyce Wiebe. 2000. Learning subjective adjectives from corpora. // *AAAI/IAAI*.

Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the use of lucene for information retrieval research. // *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, P 1253–1256, New York, NY, USA. Association for Computing Machinery.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019a. End-to-end open-domain question answering with BERTserini. // Waleed Ammar, Annie Louis, and Nasrin Mostafazadeh, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, P 72–77, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le, 2019b. *XLNet: generalized autoregressive pretraining for language understanding*. Curran Associates Inc., Red Hook, NY, USA.

System of interjection reactions to maintain communication by a companion robot

Alice S. Luria
Lobachevsky State University of Nizhni Novgorod
a_lurija@mail.ru

Artemiy A. Kotov
Kurchatov Institute National Research Center, Russian State University for the Humanities
kotov@harpia.ru

Abstract

The article presents a classification of interjections developed to describe a wide range of interjection reactions in a multimodal corpus, as well as aimed to be applied by companion robots that communicate through speech and gestures. As part of the study, an experiment was conducted to test part of the developed classification. In the experiment, the F-2 robot used interjections (and accompanying gestures) from the developed classification, as well as, for comparison, automatically synthesized interjections. The accuracy of human recognition of the illocutionary force of interjections was assessed. The conducted experiment showed that the level of understanding of the interjection reactions developed within this work is higher than the level of understanding of automated reactions. The results of the experiment confirmed the effectiveness of the developed classification of interjections in the framework of communication between a robot and a user.

Key words: communicative function, emotional-intentional reaction, system of interjections, semantics
DOI: 10.28995/2075-7182-2025-23-1081-1090

Система междометных реакций для поддержания коммуникации роботом-компаньоном

Лурия А. С.
Нижегородский государственный университет имени Н. И Лобачевского
a_lurija@mail.ru

Котов А. А.
Национальный исследовательский центр «Курчатовский институт», Российский государственный гуманитарный университет
kotov@harpia.ru

Аннотация

В статье приводится классификация междометий, разработанная для описания широкого спектра междометных реакций в мультимодальном корпусе, а также ориентированная на прикладное применение роботами-компаньонами, поддерживающими коммуникацию с помощью речи и жестов. В рамках исследования проведён эксперимент по проверке части разработанной классификации. В эксперименте робот Ф-2 использовал междометия (и сопровождающие их жесты) из разработанной классификации, а также — для сравнения — автоматически синтезированные междометия. Оценивалась точность распознавания человеком иллокутивной цели междометий. Проведённый эксперимент показал, что уровень понимания разработанных нами междометных реакций выше уровня понимания автоматизированных реакций. Результаты эксперимента подтвердили эффективность разработанной классификации междометий в рамках коммуникации робота и человека.

Ключевые слова: коммуникативная функция, эмоционально-интенциональная реакция, система междометий, семантика

1 Введение

Одной из центральных задач, актуальных для процесса моделирования коммуникативного поведения робота-компаньона, является обеспечение его естественности. Приоритет проблемы имитации роботом коммуникативного поведения человека экспериментально обусловлен — в ряде исследований установлена корреляция степени вовлечённости человека в коммуникацию с роботом и наличия такой имитации [1; 2]. Основу моделирования естественного коммуникативного поведения робота составляет разработка жестово-мимического и интонационного компонентов устного высказывания, одна из главных функций которых — трансляция эмоций и выражение коммуникативных функций. Междометия могут быть ключевыми элементами коммуникации для робота: по некоторым подсчётам междометия составляют до 49% высказываний человека и обеспечивают как эмоциональное взаимодействие, так и передачу смысла [3]. Возможность трансляции определённого эмоционального состояния роботом обеспечивает эффективность коммуникации с человеком [2].

В рамках данной работы мы называем междометиями широкий круг вокальных эмоциональных реакций и средств выражения коммуникативных функций (иллокутивных целей). В качестве составляющих междометной реакции рассматриваем собственно междометия, сопровождающие их жесты и элементы мимики. Традиционный взгляд на междометия как единицы языка предполагает наличие у них существенной омонимии. Так, междометие *А!* может выражать согласие, понимание, боль, азарт и т.д. [10]. В готовом виде такое описание междометий не может быть приложено к моделированию речевого поведения роботов. В рамках данной работы мы предлагаем классификацию междометных реакций и проверяем часть этой классификации в эксперименте с роботом-компаньоном.

2 Теоретические предпосылки исследования

При разработке системы междометных реакций мы опирались на теорию базовых эмоций [4; 5] и на теорию эмоций авторства A. Ortony, G.L. Clore и A. Collins, классифицирующую эмоции по типу эмоциогенного события («триггера») [6]. В целом, мы рассматривали междометия как реакцию на некоторый триггер, которая выражает (а) эмоцию, (б) коммуникативную реакцию, (в) иллокутивную цель или (г) этикетный ответ. Это разделение соответствует классификации грамматико-семантических групп В.В. Виноградова: эмоциональные междометия, реагирующие реплики (междометия), императивные междометия и этикетные междометия [7]. Каждая междометная реакция может быть описана с помощью: (а) типичного триггера (стимула), (б) вербального и невербального сегмента для воспроизведения на роботе (включая текстовое описание поведения и ссылку на вхождения междометия в корпус МУРКО), (в) метаязыковой клаузы, например, ‘я удивлён’, ‘я рад’, ‘я отказываюсь делать это’ — аналогично метаязыку А. Вежбицкой [8].

Междометия указанных четырёх классов делятся на подклассы. Так эмоциональные междометия рассматриваются нами как реакции на триггер, соответствующие одной из четырёх базовых эмоций: радость, печаль, страх и интерес.

Реагирующие и императивные междометия могут содержать в качестве центрального семантического компонента:

- указание на эмоцию — в этом случае используются только две общие эмоции: общая позитивная и общая негативная эмоции (например, ‘я оцениваю это негативно, поэтому я не согласен’, ‘я оцениваю это позитивно, поэтому прошу тебя это сделать’);
- интенциональный компонент смысла (иллокутивная цель) — в этом случае междометия делятся на подклассы в зависимости от коммуникативной функции, формы выражения интенционального значения (прямая или косвенная), роли в структуре диалога (вопрос — подтверждение, просьба — отказ и др.).

В ходе разработки системы междометий мы описали 163 случая коммуникативного употребления междометий на основе мультимодального подкорпуса Национального корпуса русского языка [9].

Группы междометий	**Описание групп междометий**	**Подгруппы междометий**
Эмоциональные междометия	Вокальные жесты, коррелирующие с ментально-эмоциональным состоянием человека	Соответствие междометия эмоции: ● радость (*Ура!* = ‘у меня получилось’) ● печаль (*Ох* = ‘мне жаль’) ● страх (*Ой* = ‘мне страшно’) ● интерес (*Она* = ‘мне интересно’)
Реагирующие реплики	Реплики для оценки высказывания собеседника (согласие, несогласие, сомнение) и для выражения социальных эмоций	● Эмоциональные (*ничего себе!* = ‘я приятно удивлён твоему высказыванию’) ● Интенциональные (*ага, точно* = ‘то, что ты говоришь, очень вероятно’, *угу* = ‘мне бы хотелось, чтобы ты продолжал говорить’)
Императивные междометия	Сигналы, направленные на побуждение (*да-да*, *марш!* и др.) к коммуникации, поддержание коммуникации (*так*, *ну-ну* и др.), прерывание (*тсс! стоп!* и др.) или корректировку действия или высказывания (*эй! брось ты!*)	● Инициирующие (*тсс!* = ‘я хочу, чтобы ты молчал’) ● Ответные (*а?* = ‘я хочу, чтобы ты повторил’)
Этикетные междометия	Формулы речевого этикета, служащие для побуждения к коммуникации (*удачи!* и др.), прерывания (*извиняюсь!* и др.) или корректировки (*пожалуйста*, *простите* и др.) действия или высказывания, а также для этически приемлемого выражения отношения к действию или коммуникативному высказыванию (*спасибо!* и др.)	● Интенциональные (*простите!* = ‘я прерву вас, я не согласен’) ● Эмоциональные: - Общие позитивные (*Мерси!* = ‘я рад, что вы меня заметили’) - Общие негативные (*Спасибо-спасибо!* = ‘мне это не нужно’)

Таблица 1. Предлагаемая грамматико-семантическая классификация междометных реакций

Междометные реакции, входящие в группы реагирующих реплик, императивных междометий и этикетных междометий делятся на подгруппы в зависимости от их функции. Междометия с доминирующим эмоциональным компонентом в своём выражении в наибольшей степени зависят от интонационных характеристик и от их невербального сопровождения [10], например, междометие *А!* будет обладать разными фонетическими характеристиками для каждой из выбранных базовых эмоций. Таким образом, фонетическая и мультимодальная информация будет критичной для правильного воспроизведения эмоционального междометия на роботе. В ходе корпусного исследования мы описали невербальные характеристики таких междометий, а именно показатели интенсивности и динамику основного тона, установленные с помощью программы «Speech Analyser». Описанные интонационные характеристики междометных реакций в целом соответствуют интонационным характеристикам эмоций в соответствии с классификацией зон эмоциональных состояний Н.И. Витт [11].

Междометия с доминирующим интенциональным компонентом семантики более тесно связаны с речевым сегментом. Например, междометия *стоп!* или *мерси!* рассматриваются нами как императивное и этикетное междометия, но во многих языковых описаниях эти слова классифицируются как полнозначные слова русского языка.

3 Описание эксперимента

С целью проверки разработанной нами междометной системы был проведён эксперимент, направленный на оценку понимания междометных реакций (из предлагаемой классификации) в сравнении с пониманием тех же языковых единиц, воспроизведённых с помощью системы синтеза текста «Yandex Cloud». Невербальное сопровождение междометных реакций соответствовало разработанной нами системе и в случае автоматизированного воспроизведения реакции, и в случае неавтоматизированного воспроизведения реакции.

Для эксперимента были выбраны междометия *ага*, *угу* и *точно*. Их выбор обусловлен частотностью употребления и многозначностью — данным языковым единицам соответствуют несколько междометных реакций. Согласно нашей классификации, выбранные реакции соответствуют грамматико-семантической группе реагирующих реплик. Для автоматизированных реплик *ага* и *точно* был выбран невербальный паттерн согласия и подтверждения, для автоматизированной реплики *угу* — паттерн несогласия и отрицания. Неавтоматизированным репликам соответствовали следующие значения и выразительные средства:

- Междометия *ага* и *угу*. Реакции, соответствующие междометиям, в рамках эксперимента выступили в значениях согласия и несогласия. Трансляторами значения согласия выступали: символический жест краткого вертикального кивка головы, нейтральное мимическое выражение, ровное движение тона голоса и краткость при произнесении. Трансляторами значения несогласия (в данном случае междометия *ага* и *угу* семантически сходны с междометием несогласия *ещё чего!;* наиболее частотный контекст употребления междометий в МУРКО: выражение ироничного согласия, подразумевающего под собой отказ, в ответ на просьбу или предположение собеседника) — ярко выраженный вертикальный кивок головы, мимическое выражение сведённых к переносице бровей и восходяще-нисходящая интонация.
- Междометие *точно*. Реакции, соответствующие данному междометию, в рамках эксперимента выступили в значениях подтверждения-припоминания, подтверждения-уточнения и подтверждения предварительно запрошенного ответа собеседника. Значению подтверждения-уточнения соответствовали невербальные знаки и интонационные характеристики аналогичные реакции подтверждения, выражаемой междометиями *ага* и *угу*, значению подтверждения-припоминания — лёгкое отклонение головы назад, приподнятые брови, распахнутые глаза и нисходящее движение тона голоса, значению подтверждения предварительно запрошенного ответа собеседника — плавный вертикальный кивок головы, приподнятые брови и восходящее движение тона.

Для неавтоматизированных реакций в ходе эксперимента мы оценивали отклонения в распознавании заданных функций. В рамках оценивания интерпретаций автоматизированных междометных реакций нами учитывались соответствие или несоответствие интерпретаций вышеуказанным паттернам и частотность ответов для конкретной интерпретации. Для автоматизированных реакций мы оценивали ошибку как отклонение от варианта, заданного невербальным паттерном.

Указанные интонационные параметры каждого из перечисленных значений неавтоматизированных междометных реакций соответствуют эмоциональной зоне удивления, входящей в классификацию Н.И. Витт [11]. Интонационные параметры, сгенерированные «Yandex Cloud» в среднем отличаются от предлагаемых нами параметров большей интенсивностью, частотой колебаний и длительностью произнесения реплики (см. Рисунки 1 и 2).

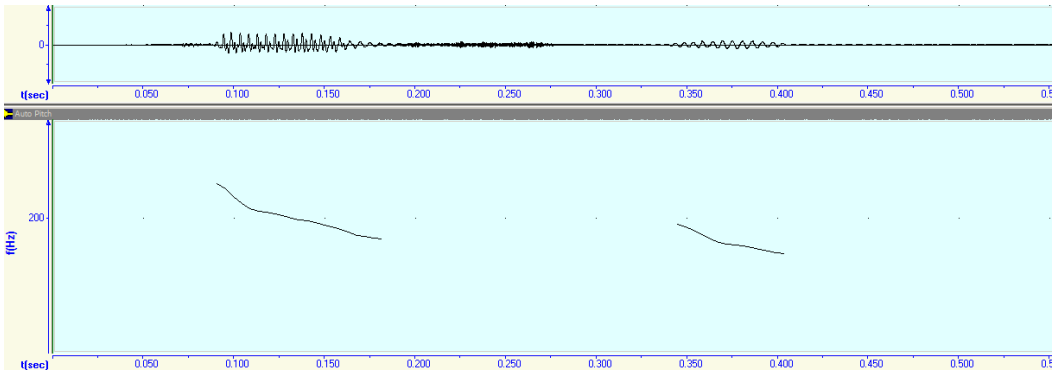

Рисунок 1. Интонационные параметры неавтоматизированного междометия *точно*, выражающего согласие

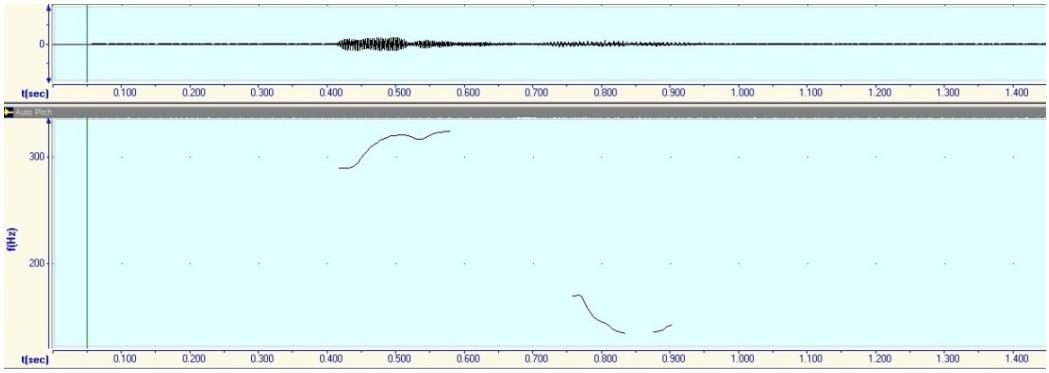

Рисунок 2. Интонационные параметры автоматизированного междометия *точно,* выражающего согласие

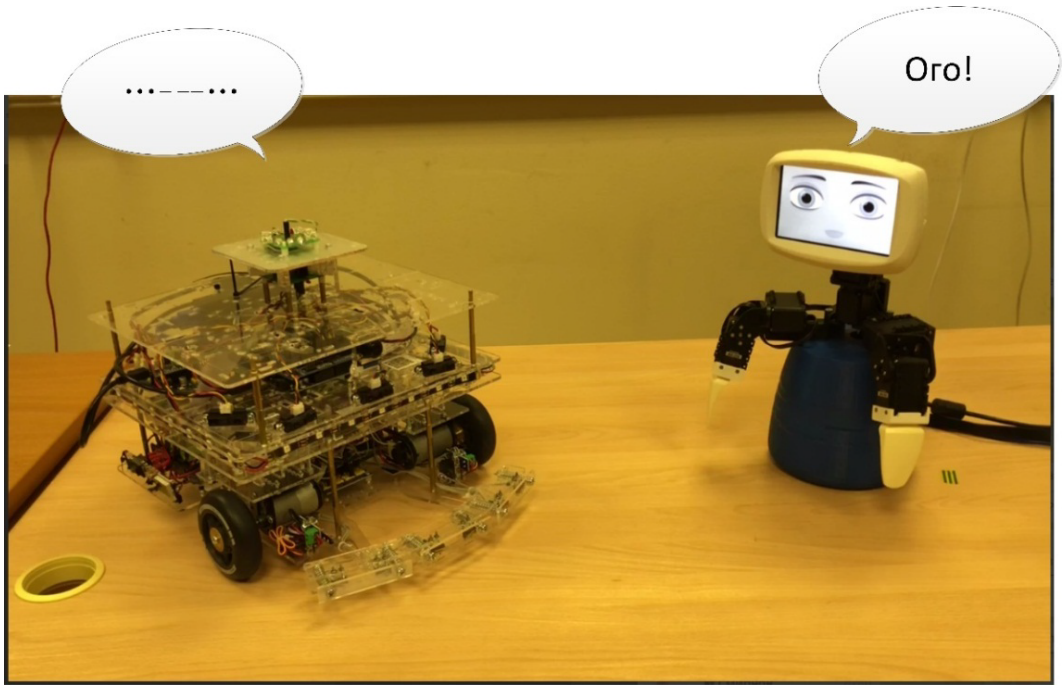

Рисунок 3. Кадр стимульного видео: неантропоморфный робот (слева) сообщает что-то на азбуке Морзе, робот Ф-2 отвечает междометной реакцией (выноски с текстом добавлены для иллюстративных целей)

Выполнение роботом выбранных для эксперимента междометных реакций было обеспечено за счёт описания невербальных элементов этих реакций на языке Behavior Markup Language. С целью воспроизведения звуковой составляющей междометных реакций были сделаны записи междометий, входящих в МУРКО, диктором с опорой на интонационные параметры, установленные в ходе корпусного исследования. Соответствие аудиозаписей, сделанных диктором, необходимым интонационным параметрам было проверено с помощью программы «Speech Analyser».

Для эксперимента был записан ряд видеофрагментов. В кадре неантропоморфный робот озвучивал некоторое высказывание с помощью азбуки Морзе: сообщал последовательность долгих и кратких сигналов (при этом испытуемые не были знакомы с азбукой Морзе). Робот Ф-2 в ответ демонстрировал междометную реакцию, включая произнесение междометия. Задача испытуемых состояла в том, чтобы оценить междометную реакцию робота Ф-2 по видеозаписи «диалога роботов» (см. Рисунок 3). Использование азбуки Морзе для исходной реплики заставляет испытуемых оценивать реакцию робота Ф-2 без учёта семантики или интонации реплики-стимула.

В рамках проведённого эксперимента было опрошено 30 человек, возраст испытуемых варьировался от 23 до 70 лет. Реципиентам предлагалось посмотреть видеозаписи и после каждой видеозаписи ответить на вопросы об эмоциональной составляющей и интенциональности междометной реакции робота Ф-2, а также предлагалось сделать предположение о коммуникативной функции входящей реплики (установление семантических связей между стимулом и реакцией позволяет получить более точное представление о понимании междометной реакции). Оценка эмоциональной составляющей семантики междометных реакций производилась в соответствии с классификацией эмоций на положительные и отрицательные. Для оценки интерпретации интенции междометной реакции реципиентам был задан вопрос «О чём робот сообщает собеседнику?», для предположения относительно семантики входящей реплики была озвучена просьба «Предположите, чем могла быть спровоцирована такая реакция робота?». Ответы на вопросы о коммуникативной функции междометной реакции и стимуле междометной реакции озвучивались испытуемыми в свободной форме.

4 Результаты исследования

С целью обеспечения наглядности представим результаты эксперимента в виде таблиц. В первом столбце перечислены междометия, соответствующие исследуемым междометным реакциям. Каждая строка таблицы, начинающаяся с междометия, отображает ряд данных, соответствующих этому междометию. Последний столбец таблицы содержит в себе данные о количестве ошибок в строке, соотносящейся с междометием, стоящим вначале строки (данные отображаются в процентах). Среди иллюстрируемых таблицами данных соответствие каждого междометия знаку эмоциональной составляющей междометной реакции (таблицы 2, 3, 4) и соответствие коммуникативной функции (таблица 5). Данные в табличном формате приводятся и для автоматизированных, и для неавтоматизированных реакций.

4.1 Восприятие эмоционального компонента междометных реакций

		Ага (+)	Ага (−)	Не выявлено	% ошибок
Неавтом. реакции	Ага (согл.)	27	3	0	10%
	Ага (несогл.)	3	23	4	23%
Автом. реакции	Ага (согл.)	17	6	5	37%

Таблица 2. Показатели понимания эмоционального компонента междометных реакций *ага*

		Угу (+)	Угу (−)	Не выявлено	% ошибок
Неавтом. реакции	Угу (согл.)	26	3	1	13%
	Угу (несогл.)	7	22	1	27%
Автом. реакции	Угу (несогл.)	5	20	5	33%

Таблица 3. Показатели понимания эмоционального компонента междометных реакций *угу*

		Точно (+)	Точно (−)	Не выявлено	% ошибок
Неавтом. реакции	Точно (согл. С ответом)	27	1	2	10%
	Точно (согл.)	24	4	2	20%
	Точно (воспомин.)	25	2	3	17%
Автом. реакции	Точно (согл.)	12	10	8	60%

Таблица 4. Показатели понимания эмоционального компонента междометных реакций *точно*

Доля ответов, указывающих на верную интерпретацию эмоционального компонента семантики неавтоматизированных междометных реакций, от общего количества интерпретаций составила 83%, аналогичные показатели для автоматизированных реакций составили 38 %. При этом количество ответов, свидетельствующих о невозможности оценить неавтоматизированную междометную реакцию как положительную или отрицательную, составила 6 %, свидетельствующих о невозможности оценить автоматизированную междометную реакцию — 20 %.

Вариативность оценки эмоционального компонента семантики, как положительного или отрицательного, каждой из междометных реакций не обладает яркой выраженностью. Наименее низкие показатели ошибочной интерпретации эмоциональной составляющей семантики (10%) соотносятся с междометными реакциями согласия *ага* и *точно*, наиболее высокие показатели (27 %) — с междометной реакцией несогласия *угу.*

Оценка эмоционального компонента семантики каждой из автоматизированных междометных реакций сводится к следующим показателям: междометие согласия *ага* — 37% ошибочных интерпретаций, междометие несогласия *угу* — 33%, междометие согласия *точно* — 60%.

4.2 Восприятие коммуникативных функций междометных реакций

Неавтоматизированные междометные реакции								
	Ага (согл.)	Ага (несогл.)	Угу (согл.)	Угу (несогл.)	Точно (согл. с ответом)	Точно (согл.)	Точно (воспом.)	% ошибок
Ага (согл.)	30	0						0%
Ага (несогл.)	7	23						23%
Угу (согл.)			30	0				0%
Угу (несогл.)			10	20				33%
Точно (согл с ответом.)					30	0	0	0%
Точно (согл.)					2	27	1	10%
Точно (вспом.)					0	3	25	10%
Автоматизированные междометные реакции								
Ага (согл.)	28	2						7%
Угу (несогл.)			26	4				87%
Точно (согл.)					0	16	14	53%

Таблица 5. Показатели понимания коммуникативных функций автоматизированных и неавтоматизированных междометных реакций

Доля ответов, содержащих в себе верную интерпретацию коммуникативных функций неавтоматизированных междометных реакций, от общего количества интерпретаций составила 88%, аналогичные показатели для автоматизированных реакций составили 53%. Наиболее низкие показатели понимания коммуникативных функций были продемонстрированы реципиентами в рамках интерпретации автоматизированной междометной реакции *угу*, выполняющей функцию несогласия, и автоматизированной междометной реакции *точно*, выполняющей функцию подтверждения. Интерпретация междометной реакции отрицания большинством реципиентов была представлена формулировками «согласие сквозь раздражение» и «вынужденное согласие». Данные формулировки свидетельствуют о влиянии на интерпретацию испытуемыми автоматизированной реакции невербального паттерна и интонации. Мы можем предположить, что за счёт присутствующих средств выражения эмоций, фокус внимания реципиентов, предложивших указанные интерпретации, был смещён с интенциональной составляющей семантики на эмоциональную. Междометная реакция подтверждения наиболее часто понималась испытуемыми как вопрос-уточнение.

Наименьшие показатели понимания коммуникативных функций неавтоматизированных междометных реакций (доля ошибочных интерпретаций — 33 %) относятся к междометной реакции несогласия *угу*, наиболее высокие показатели (доля ошибочных интерпретаций — 0 %) — к междометным реакциям согласия *угу* и *ага* и к междометной реакции согласия с ответом *точно*.

4.3 Моделирование контекста междометных реакций

Наиболее частотным предположением (28 ответов из 30) относительно семантики реплики-стимула, ответом на которую выступили реакции согласия и подтверждения *ага* и *угу*, является формулировка «просьба сделать что-нибудь». Реплики-стимулы, ответом на которые являлись реакции отрицания и несогласия, представленные теми же междометиями, чаще всего интерпретировались как «угроза» (5 ответов из 30), «приказ» (8 ответов из 30), «просьба сделать что-нибудь» (7 ответов из 30) или «ложное обещание» (5 ответов из 30). Примечательно, что частотность интерпретации реплик-стимулов как «угрозы» или «приказа» прослеживается в рамках анализа реципиентами ответных автоматизированных реакций, а частотность интерпретации этих реплик как «ложного обещания» или «просьбы сделать что-нибудь» — в рамках ответных неавтоматизированных реакций. В связи с этим можно отметить, что автоматизированная реакция аналогично неавтоматизированной реализует коммуникативную функцию побуждения к действию, демонстрируя при этом более экспрессивную форму выражения. Значительное преобладание интерпретаций реплик-стимулов как обозначающих побуждение к действию соотносится с интерпретацией коммуникативных функций как «согласие» и «несогласие». Данное соотношение может быть объяснимо восприятием робота исключительно как исполнителя распоряжений человека [12] или же связано с интонационными параметрами междометий *ага* и *угу*, среди которых отмечается краткость произнесения, также характерная, согласно результатам нашего корпусного исследования, для междометий согласия (*хорошо*, *ладно*, *есть!* и т.д.). Эта семантическая особенность устной реализации междометий *ага* и *угу* требует дополнительных исследований.

Наиболее частотными предположениями относительно семантики реплики-стимула, реакцией на которую выступало междометие *точно*, являются ответы «вопрос-уточнение» и «утверждение», что не противоречит ответам опрашиваемых относительно коммуникативных функций.

5 Выводы

Согласно данным проведённого эксперимента, показатели уровня понимания разработанных нами междометных реакций выше, чем показатели уровня понимания междометных реакций, сгенерированных автоматически. Установленное соотношение справедливо и для результатов интерпретации реципиентами коммуникативной функции, и для результатов интерпретации эмоционального компонента семантики, и для результатов распознавания контекста реакции.

Полученные экспериментальные данные указывают на эффективность разработанной нами системы междометий, которая подтверждается высокими показателями верной интерпретации двух компонентов семантики неавтоматизированных междометных реакций (88 % — доля ответов, указывающих на верную интерпретацию когнитивных функций, 83 % — доля ответов, указывающих на верную интерпретацию эмоционального компонента семантики) и свидетельствует о возможности применения данной междометной системы с целью обеспечения коммуникации робота и человека. Выявленные в ходе нашего исследования особенности восприятия междометных реакций свидетельствуют о необходимости проведения исследований, направленных на:

- разработку системы естественных интонаций для речи робота с целью обеспечения корректной трансляции эмоционального компонента семантики произносимых им реплик;
- выявление особенностей восприятия робота человеком с целью установления коммуникативных компонентов, характерных для робота и влияющих на семантику произносимых им реплик.

Литература[1]

[1] Breazeal C., Scassellati B. Robots that imitate humans // Trends in Cognitive Sciences. — 2002. — V. 6(11). — Pp. 481-487.

[2] Leite I., Pereira A., Martinho C., Paiva A. Are Emotional Robots More Fun to Play With? / I. Leite, A. Pereira, C. Martinho, A. Paiva // Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication — Munich: Technische Universität München, 2008. — Pp. 77–82.

[3] Campbell N. Extra-Semantic Protocols; Input Requirements for the Synthesis of Dialogue Speech // Affective Dialogue Systems. ADS 2004. Lecture Notes in Computer Science. — Vol 3068. –Springer, Berlin, Heidelberg. — https://doi.org/10.1007/978-3-540-24842-2_22

[4] Izard K. E. Psychology of Emotions [Psikhologiya ehmocij]. — St.-Petersburg: Piter, 2006.

[5] Ekman P. Emotions in the human face. — Cambridge, 1982.

[6] Clore G. L., Ortony A. Cognition in emotion: Always, sometimes, or never? // Cognitive neuroscience of emotion. — NewYork: Oxford University Press, 2000. — Pp. 24–61.

[7] Vinogradov V. V. Russian language. Grammatical teaching about the word. [Russkij yazyk. Grammaticheskoe uchenie o slove]. Moscow, 1947.

[8] Wierzbicka A. Interpretation of emotional concepts [Tolkovanie ehmocional'nykh konceptov] // Language, Culture, Cognition [Yazyk. Kul'tura. Poznanie]. Russian Dictionaries, 1996. Pp. 326–375.

[9] Luria A. S. Development of a system of interjections and accompanying gestures for communication between a companion robot and a human: dis. ... mag. phil. sciences. [Razrabotka sistemy mezhdometij i soprovozhdayushchikh zhestov dlya kommunikacii robota-kompan'ona i cheloveka: dis. … mag. fil. nauk]. — Kazan’, 2022.

[10] Sharonov I. A. Interjections in speech, text and dictionary [Mezhdometiya v rechi, tekste i slovare]. — Moscow, RSUH, 2008.

[11] Witt N. B. Speech and Emotions [Rech' i ehmocii]. — Moscow: Moscow State Linguistic University, 1984.

[12] Jordan D. Robots [Roboty]. — Moscow.: Tochka, 2017.

[1] References, Scopus version

Литература[2]

[1] Breazeal C., Scassellati B. Robots that imitate humans // Trends in Cognitive Sciences. — 2002. — V. 6(11). — Pp. 481-487.

[2] Leite I., Pereira A., Martinho C., Paiva A. Are Emotional Robots More Fun to Play With? / I. Leite, A. Pereira, C. Martinho, A. Paiva // Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication — Munich: Technische Universität München, 2008. — Pp. 77–82.

[3] Campbell N. Extra-Semantic Protocols; Input Requirements for the Synthesis of Dialogue Speech // Affective Dialogue Systems. ADS 2004. Lecture Notes in Computer Science. — Vol 3068. –Springer, Berlin, Heidelberg. — https://doi.org/10.1007/978-3-540-24842-2_22

[4] Изард К. Э. Психология эмоций. — Санкт-Петербург: Питер, 2006.

[5] Ekman P. Emotions in the human face. — Cambridge, 1982.

[6] Clore G. L., Ortony A. Cognition in emotion: Always, sometimes, or never? // Cognitive neuroscience of emotion. — NewYork: Oxford University Press, 2000. — Pp. 24–61.

[7] Виноградов В. В. Русский язык. Грамматическое учение о слове. — М., 1947.

[8] Вежбицкая, А. Толкование эмоциональных концептов // Язык. Культура. Познание. Русские словари, 1996. — С. 326–375.

[9] Лурия А. С. Разработка системы междометий и сопровождающих жестов для коммуникации робота-компаньона и человека: дис. … маг. фил. наук. — Казань, 2022.

[10] Шаронов И. А. Междометия в речи, тексте и словаре. — М.: РГГУ, 2008.

[11] Витт Н.В. Речь и эмоции. — М.: МГПИИЯ, 1984.

[12] Джордан Д. Роботы. — М.: Точка, 2017.

[2] References, РИНЦ version

Precedent texts of the corpus “one speech day” and comic passe-partout of everyday communication

Peresypkina X. A.
SPbSU / Saint Petersburg, Russia
xeniaperesypkina@mail.ru

Bogdanova-Beglarian N. V.
SPbSU / Saint Petersburg, Russia
n.bogdanova@spbu.ru

Abstract

One of the phenomena of modern communication should be recognized as “comic passe-partouts” (CP)—special linguacultural units present in the mental lexicon of native speakers, actively functioning in colloquial speech and illustrating both the frequency of use of ready-made units and all kinds of constructions peculiar to the speakers and the regularity of spontaneous speech creation. CPs arise on the basis of precedent texts, reflect their inherent ability to modifications, have, as a rule, a structure of construction and are realized in speech (oral and oral-written) in many variants, cf.: “Slovo patsana. Krov’ na asfal’te”: *Slovo brevna. Shchepki na asfal’te ulitsy Rusa*; *Slovo kotana. Sherst’ na divane*; “Vostok — delo tonkoe”: *Dizajn — delo tonkoe, dazhe esli rech’ pro prostoj ulichnyj znak, vyvesku ili nadpis’*; *Mda… zagorat’ na plyazhe s napil’nikom v ruke — delo tonkoe*. The phenomenon of CP is an example of the “unlocking” of the semantics of the precedent text and the functioning of the “construction” created on its basis in a conditionally infinite number of contexts. The existence of this new speech phenomenon cannot be ignored in any description of modern communication, especially in various applied aspects of linguistics, as well as in the creation of automatic systems for processing natural speech or artificial intelligence.

Keywords: precedent text, comic passe-partout, speech corpus, everyday speech, stable unit
DOI: 10.28995/2075-7182-2025-23-1091-1099

Прецедентные тексты корпуса «один речевой день» и комические паспарту повседневной коммуникации

Пересыпкина К. А.
СПбГУ / Санкт-Петербург, Россия
xeniaperesypkina@mail.ru

Богданова-Бегларян Н. В.
СПбГУ / Санкт-Петербург, Россия
n.bogdanova@spbu.ru

Аннотация

Одним из явлений современной коммуникации надо признать «комические паспарту» (КП) — особые лингвокультурные единицы, присутствующие в ментальном лексиконе носителей языка, активно функционирующие в разговорной речи и иллюстрирующие как свойственную носителям частоту использования готовых единиц и всякого рода конструкций, так и регулярность спонтанного речетворчества. КП возникают на базе прецедентных текстов, отражают характерную для них способность к модификациям, имеют, как правило, структуру конструкции и реализуются в речи (как устной, так и устно-письменной в социальных сетях) во множестве вариантов, ср.: «Слово пацана. Кровь на асфальте»: *Слово бревна. Щепки на асфальте улицы Руса*; *Слово котана. Шерсть на диване*; <Восток — дело тонкое> => *Дизайн — дело тонкое, даже если речь про простой уличный знак, вывеску или надпись*; *Мда… загорать на пляже с напильником в руке — дело тонкое*. Феномен КП — пример «размыкания» семантики прецедентного текста и функционирования «конструкции», созданной на его основе, в условно бесконечном количестве контекстов. Существование этого нового речевого явления нельзя игнорировать в любом описании современной коммуникации, особенно в различных прикладных аспектах лингвистики, а также при создании автоматических систем обработки естественной речи или искусственного интеллекта.

Ключевые слова: прецедентный текст, комическое паспарту, речевой корпус, повседневная речь, устойчивая единица.

1 Введение

Спецификой современной коммуникативной ситуации является наличие в ней большого количества устойчивых, чаще неоднословных, речевых единиц, особого рода «стабилизовавшихся» построений, в которых черты собственно грамматические выступают в неразрывном единстве с чертами лексико-фразеологическими (*Шведова* 1960: 7). Это объясняется тем, что «говорящий, находясь в условиях непринужденного, неподготовленного общения, стремится упростить и облегчить свое “речевое поведение”, поэтому он легко и часто прибегает к готовым языковым формулам, в том числе всякого рода клише, шаблонам, стереотипам» (*Земская и др.* 1981). Ср. также: «Наличие в речевой практике говорящих на любом языке большого числа устойчивых повторяющихся выражений — факт сам по себе хорошо известный. Феномены такого рода, определяемые как “идиомы”, “устойчивые сочетания”, “речевые формулы”, “речевые штампы”, “клише”, находят в этом качестве определенное место в любом описании языка» (*Гаспаров* 1996: 121). Высокая частотность таких устойчивых неоднословных единиц (УНЕ) в нашей коммуникации, а также свойственное им разнообразие ставят перед исследователями задачу их выявления и систематизации.

В отечественной лингвистике это привело, в частности, к созданию специализированных баз данных, главные из которых — *Русский Конструктикон* (https://constructicon.github.io/russian) и *Прагматикон* (https://pragmaticon.ruscorpora.ru). Объектом фиксации в этих базах являются не привычные для корпусной лингвистики токены и не отдельные слова, а более крупные единицы (конструкции и дискурсивные формулы), находящиеся на стыке между словарем и грамматикой и фактически стирающие эту границу.

Помимо упомянутых баз данных, интерес к УНЕ привел к появлению и целого ряда конкретных исследований, в ряду которых можно отметить систематизацию УНЕ, построенную на материале корпуса русского языка повседневного общения «Один речевой день» (ОРД) (https://ord.spbu.ru). Отличительной чертой этого проекта стало обращение исключительно к устному повседневному дискурсу (*Отчет РНФ* 2024).

В ходе двух этапов ручного аннотирования корпусного материала, а также экспертной проверки его результатов удалось составить типологию УНЕ, включающую 8 классов (*Bogdanova-Beglarian et al.* 2024): коллокации разного типа (фразеологизированные, нефразеологизированные и даже окказиональные, которые воспринимаются носителями языка именно как коллокации, несмотря на единичность употребления (*копеечку стоит небезумную, вошь в ухо залетела*)), конструкции (в духе Грамматики конструкций), речевые формулы и формы-идиомы, неоднословные прагматические маркеры (*это самое, скажем так*) и прецедентные тексты (ПТ) или их фрагменты. Именно последние стали объектом внимания в настоящем исследовании, в фокусе которого — особый тип модификации прецедентных текстов в повседневной речи носителей языка и особенности бытования соотносимых с ним контекстов. О сложностях выявления, систематизации и перевода единиц такого рода писали неоднократно (см., например: *Беликов и др.* 2021), поэтому, полагаем, еще один взгляд на проблему лишним не будет.

2 Прецедентные тексты как класс устойчивых неоднословных единиц

На выборке в 1 млн словоформ из корпуса ОРД в ходе ручного аннотирования было выделено 8055 УНЕ, в том числе 3002 уникальные формы (*Отчет РНФ* 2024). Упорядоченный по частоте встречаемости список классов УНЕ представлен в таблице.

Ранг	Класс УНЕ	Кол-во	%
1	ПМ	2487	30,88
2	НК	2121	26,33
3	РФ	1673	20,77
4	ФК	800	9,93
5	ИД	447	5,55
6	КС	407	5,05
7	ПТ	101	1,25
8	ОК	19	0,24

Таблица. Ранжированный список классов устойчивых неоднословных единиц в корпусе ОРД

Из таблицы видно, что класс ПТ, включающий 101 единицу, занимает в этой типологии предпоследнее место, «опережая» только окказиональные коллокации. Однако функционирование единиц этого типа УНЕ в нашей повседневной коммуникации перспективно для изучения в самых разных исследовательских целях. Так, интересны «разные пути анализа этого материала: с точки зрения того социума, которому понятны (или непонятны) эти тексты, с учетом языка-источника (русский или иностранный) и текста-источника (кинофильм, книга, анекдот, стихи, песни и т. п.), характера произведенных с текстом модификаций и степени его цельности/идиоматичности, а также способа его введения в устную речь и реакции на него обоих коммуникантов — как говорящего, так и слушающего» (*Богданова-Бегларян* 2018: 19). Можно рассматривать ПТ как источник возникновения в языке новых идиом, что, в частности, подтверждает их способность заменяться словом/выражением-идентификатором (*Это я удачно зашел!* = ‘Мне повезло!’) (*она же* 2020). Кроме того, нет сомнения, что ПТ как «пограничные составные элементы Лексикона» (*Беликов и др.* 2021: 44) представляют собой «класс наиболее трудных для понимания и перевода на другие языки единиц устного дискурса, <…> которые отражают не только национальную, но и поколенческую ментальность носителей языка. <…> Тексты, содержащие такие фрагменты, могут служить своеобразным тестом на знание русского языка и русской культуры, составляют особый разряд в классе безэквивалентной лексики, сложной для перевода на другие языки, а также требуют комментария не только в иностранной, но зачастую и в русской аудитории» (*Богданова-Бегларян* 2025).

По замечанию А.В. Батулиной, «для современной массовой лингвокультуры характерно широкое использование различных приемов игры со словом, едва ли не самым распространенным из которых в публицистике и в разговорной речи является трансформация прецедентных текстов. Разного рода лексико-семантические и лексико-синтаксические модификации лежат в основе образования “паремий нового типа” (термин Т.Г. Никитиной), к которым относятся, напр., антифразы (трансформации загадок, пословиц, фразеологических единиц (ФЕ)), футбольные речевки и антипословицы» (*Батулина* 2011: 214). Все подобные единицы, по мнению автора, как и ПТ, «присутствуют в языковой способности современной языковой личности как *живые и активные единицы* (курсив наш. — *К. П., Н. Б.-Б.*)» (*там же*). Это и объясняет не ослабевающий в лингвистике интерес к материалу такого типа.

В настоящей статье представлен еще один взгляд на модификации прецедентных текстов в современной коммуникации.

3 Прецедентные тексты в корпусе ОРД: некоторые количественные данные

В ходе анализа аннотированного подкорпуса ОРД, в числе прочего, для каждого класса УНЕ был составлен частотный список соответствующих единиц. В классе ПТ наиболее частотными, встретившимися по два раза (по 1,98 % от общего количества в материале исследования, все остальные ПТ единичны), оказались следующие 6 единиц: ХОРОШО СИДИМ, ВРЕМЯ ПОКАЖЕТ, ВРЕМЯ СОБИРАТЬ КАМНИ, ЕЩЁ НЕ ВЕЧЕР, НА СЕБЯ ЛЮБИМУЮ и ПОЗДНО ПИТЬ БОРЖОМИ. Хорошо видна не самая высокая на фоне других классов УНЕ частотность ПТ в нашей повседневной речи: ср., например, самый частотный прагматический маркер в массиве УНЕ — ЭТО САМОЕ (12,55 %), самая частотная конструкция ДЕЛО В ТОМ ЧТО (10,57 %), самая частотная форма-идиома НИ ФИГА (4,92 %) или речевая формула ВСЁ РАВНО (4,24 %). При этом, против всякого ожидания, фразеологические коллокации (ФК) по частоте заметно проигрывают прецедентным текстам, ср. первые 4 самых частотных ФК нашей коммуникации: НЕ ДАЙ БОГ (1,13 %), ДАЙ БОГ (0,88 %), С УМА СОЙТИ (0,63 %), ЧЁРТ-ТЕ ЧТО (0,63%).

Приведем еще некоторые данные про ПТ в нашей речи (*Отчет РНФ* 2024):

- общее количество уникальных фраз — 95;
- общее количество фраз с единичной частотой — 89 (93,68 % от общего числа), что указывает на разнообразие материала;
- средняя частота в корпусе — 1,05 %;
- ПТ (как и окказиональные коллокации) минимально представлены в речи всех групп говорящих, во всех коммуникативных ситуациях и во всех местах общения, особенно в формальных, таких как поликлиника и офис;

- в речи молодежи до 24 лет ПТ практически отсутствуют;
- ПТ редки в речи обеих гендерных групп, но мужчины используют их чуть чаще (1,39 %) по сравнению с женщинами (1,20 %);
- более высокая доля ПТ (2,11%) в речи говорящих с низким уровнем речевой компетенции (УРК) по сравнению с другими группами (средний и высокий УРК);
- доля ПТ минимальна во всех группах говорящих, упорядоченных по уровню образования, однако наибольшая их встречаемость наблюдается у людей с незаконченным высшим образованием (2,20 %);
- в профессиональных группах ПТ встречаются крайне редко и только у представителей некоторых профессий: творческие работники, работники сферы обслуживания или рабочие.

Приведенные данные могут быть использованы в лингвистических исследованиях для изучения современных прецедентных текстов, а также полезны для анализа культурного кода и межтекстовых отсылок в разговорной речи. Обратим внимание, что, по данным корпуса ОРД, собранного более 10 лет назад, *в речи молодежи до 24 лет ПТ практически отсутствуют*, что вступает в некоторое противоречие с результатами проведенного в настоящей работе анализа комических паспарту, которые не только создаются на базе конкретных ПТ, но и сами функционируют «прецедентно», т. е. являются «(1) значимыми для той или иной личности в познавательном и эмоциональном отношениях, (2) имеющими сверхличностный характер, т. е. хорошо известными и широкому окружению данной личности <…>, и, наконец, такими, (3) обращение к которым возобновляется неоднократно в дискурсе данной языковой личности» (*Караулов* 2010: 216). Комические паспарту выявляются не в последнюю очередь именно в речи современной молодежи, что заставляет задуматься и о возможном сущностном изменении источников прецедентных текстов, и о специфике происходящих с ними трансформаций.

4 Комические паспарту как новый речевой феномен: понятие и специфика

Отличительные черты современного состояния коммуникативной ситуации реализуются в массе неоднородных явлений, одним из которых следует признать феномен *комических паспарту*, наблюдаемый в последнее время.

Новое речевое явление, названное нами «феноменом комических паспарту», возникает в повседневной коммуникации (не только устной, но и устно-письменной, свойственной общению в Интернете[1]) на основе прецедентных текстов и представляет собой особый тип их модификации, требующий трансконцептуального подхода при исследовании. Под КП понимаются лингвокультурные единицы, имеющие прецедентный генезис, существующие в когнитивном пространстве носителей языка в качестве инвариантов и реализующиеся в повседневной речевой практике во множестве вариантов (<Денег нет, но вы держитесь> → *Ноутов нет, но вы держитесь; Щуки нет, но вы держитесь. Поехали на рыбалку, а попали в какой-то рассказ Стивена Кинга «Дети кукурузы», судя по тому, какое пугало из распятой вороны нам попалось*). Так, термином «комические паспарту» было решено обозначать и (1) варианты, и (2) инварианты, так как последние за редкими исключениями не присутствуют в речи, являясь результатом метаязыковой рефлексии, и (3) феномен в целом. Следует подчеркнуть, что существительное «инвариант» в данном случае может иметь две трактовки: во-первых, условный «инвариант-инвариант» (<X нет, но вы держитесь>): *Кушать нет, но вы держитесь; чистых нет но вы держитесь*; во-вторых, условный «инвариант-вариант» (например, <N_2 нет, но вы держитесь>): *Потому что пока, например, «ферритина нет, но вы держитесь» (не уверена, что это дословная цитата), апатия и вялость не уйдут, как бы корректно не были подобраны антидепрессанты; По пути встречающиеся немногочисленные экипажи, лишь уныло разводили руками — рыбы нет, но вы держитесь* (см. первые публикации на эту тему: *Пересыпкина* 2024; *Пересыпкина, Богданова-Бегларян* 2024). Комические паспарту (ядерная лексема, вошедшая в составной термин, была выбрана из ряда синонимов («клише»,

[1] В виртуальной устно-письменной коммуникации «происходит выработка такой системы речевого общения, которая диктует специфические, игровые, экспериментальные алгоритмы порождения и употребления знака, позволяющие коммуникантам выйти за пределы стереотипов “строгого” речевого узуса» (*Гридина, Талашманов* 2019: 31).

«рамка», «формула», «шаблон») в связи с имманентно присущей ей образностью и широким коннотативным полем) в их соотношении с ПТ из корпуса ОРД и стали объектом внимания в настоящем исследовании.

КП как «прецедентные инварианты разной степени упорядоченности и актуализирующие их в процессе повседневной коммуникации варианты» (*Пересыпкина* 2024: 75) — <Цвет настроения синий> (исходный ПТ/КП) → *Цвет настроения индиго*; *Цвет настроения “Цирк”*; *Цвет настроения — любой!* (варианты ПТ/КП) — сближаются с понятием *конструкции* в рамках Грамматики конструкций (*Fillmore et al*. 1988); или с понятием «синтаксического фразеоида», под которым подразумевается «своеобразная конструкция-колодка, в которую подставляется значимый элемент» (*Коган* 2018: 44): «<это> X, Карл», «это X здорового человека, а это X курильщика», «ничего не X только X — X» (*там же*: 44-45). Вышеприведенный пример ПТ/КП вполне можно представить как конструкцию (КС) или синтаксический фразеоид: <Цвет настроения X>[2].

Феномен КП определенно «больше, чем прецедентный текст» (*Пересыпкина* 2024: 75), это актуальный многокомпонентный речевой факт, синтезирующий «лудический», «прецедентный» и «комический» комплексы и на данный момент теоретически не вписанный в научную парадигму (*Пересыпкина, Богданова-Бегларян* 2024: 767), что и делает его столь заманчивым для дальнейшего анализа.

Показалось, в частности, любопытным посмотреть, не стали ли (а если стали, то как именно) 6 самых частых ПТ из корпуса ОРД (топ-6) основой для языковой игры, для создания различных трансформаций, в том числе — модификаций определенного типа, комических паспарту.

5 Прецедентные тексты русской повседневной речи как возможная основа для возникновения комических паспарту

Для проверки выдвинутой гипотезы был произведен поиск модификаций всех единиц из списка топ-6 ПТ по различным сайтам, в том числе в газетном подкорпусе НКРЯ (https://ruscorpora.ru) — и результат не заставил себя ждать. Пять из шести единиц частотного списка ОРД обнаружили свое бытование в различных трансформированных вариантах. Исключением стало только выражение ПОЗДНО ПИТЬ БОРЖОМИ (фрагмент фразы ПОЗДНО ПИТЬ БОРЖОМИ, КОГДА ПОЧКИ ОТКАЗАЛИ/ОТВАЛИЛИСЬ, которую приписывают Козьме Пруткову; означает ‘упущенный момент: поздно что-либо предпринимать, когда сложившаяся ситуация не поддается исправлению’) — вероятно, в связи с высокой степенью его идиоматичности.

Остальные 5 ПТ обнаружили функционирование в модифицированном виде, и некоторые трансформации могут быть свидетельством постепенного формирования или бытования комических паспарту на их основе.

С учетом способности становиться базой для комических паспарту эти 5 ПТ удалось разбить на две группы: (1) ПТ = КС = КП и (2) ПТ = КС ≠ КП. Из предложенных схем (моделей) видно, что практически все ПТ могут быть представлены как конструкции, с наличием в их структуре постоянных (якорных) компонентов и переменных слотов, но не все могут интерпретироваться как КП (прецедентность + комизм, возникающий вследствие нарушения ожидания и отклонения от нормы, сопровождающих лудическую лингвокреативную деятельность). Однозначно материал ПТ на эти группы, конечно, не делится[3], но с некоторой натяжкой предложенную систематизацию ПТ на этом основании провести все же оказалось возможным.

Так, к группе (1) были отнесены следующие прецедентные тексты, в которых налицо и конструкция-паспарту, и комический эффект, связанный с принципиальным семантическим разрывом с изначальным ПТ.

[2] Ср. еще ряд близких понятий: *фразеосхема*, где «значения опорных слов <…> оказываются сдвинутыми» (*Шмелев* 1976: 134), *синтаксическая идиома* (*Кайгородова* 1999: 10), *синтаксическая фразема* (*Иомдин* 2006), и некоторые др.

[3] Это вполне объяснимо: естественная диффузность спонтанной (как устной, так и устно-письменной) речи «порой становится причиной целого ряда возможных интерпретаций одного контекста; однако эта черта, как отмечают лингвисты, не может повлиять на достоверность аргументированных выводов, сделанных на материале устной речи, и, следовательно, не может стать препятствием на пути исследователя» (*ПМ* 2021: 312-313).

- ХОРОШО СИДИМ <Хорошо X/V_{1Pl}> (около 20 вариантов): *Магазин продуктов «**Хорошо Едим**»; **ХОРОШО ГУДИМ**! Омский студент играет русский рок-н-ролл... на гуслях; **Хорошо стоим**: к чему готовиться московским автомобилистам этим летом*; *Мы вот **хорошо тонем**: ((Скоро квакать начнём!* и т. д.
- ВРЕМЯ СОБИРАТЬ КАМНИ <Время Inf_1 (и время Inf_2)> (более 10 вариантов; употребляется и полная форма ПТ, и ее первый фрагмент): ***Время спать, и время просыпаться***; ***Время читать, и время бросать книги в костер***; ***Время собирать деньги***: *в 2019-м россияне накопят на четверть больше*; *озабоченные домочадцы, виляющие от подноса к подносу: **время разбрасывать комплименты, время собирать бутерброды***; и т. д.

Видно, что данные ПТ трансформируются семантически разнообразно, но формально практически однообразно: слоты заполняются ограниченным количеством словоформ. В связи с этим особенно интересным становится соотношение понятий «индивидуальное» — «коллективное», «условно окказиональное» — «условно узуальное», «частотное» — «нечастотное» и др.

К группе (2) были отнесены те прецедентные тексты, которые, даже в ранге конструкций, все же не могут расцениваться как комические паспарту, так как в них компоненты реализуются в своих основных значениях.

- ВРЕМЯ ПОКАЖЕТ <X покажет> (6 употреблений; высокая степень идиоматичности выражения — в некоторых случаях сохраняется семантика будущности, а в других контекстах актуализируется словарное значение глагола *показать*): ***Дело покажет***. *Три струны; «**День покажет**». Как рассказывать истории беженцев; «Как по-вашему, это шейка бедра?» — спросила я. Та же непроницаемость: «**Рентген покажет**»; **Снег покажет**; **СЫР покажет**; Терять было нечего. **Будущее покажет**, насколько она права.*
- ЕЩЁ НЕ ВЕЧЕР <Еще не X> (3 употребления; высокая степень идиоматичности выражения — сохраняется семантика будущности, однако заданного прецедентным текстом значения ‘есть шанс, еще не все потеряно’ варианты не реализуют): ***Ещё не зима***. *Ноябрь любого сведёт с ума; **Ещё не лето**..; Моё гетто не спит. Yofu. **Ещё не ночь**.*
- НА СЕБЯ ЛЮБИМУЮ <На себя X-ую> (более 10 вариантов; высокая степень идиоматичности выражения — в большинстве случаев сохраняется изначальная положительная семантика (*родная*, *ненаглядная*, *милая*), а в контекстах с прилагательными *безмозглая*, *зябкая*, вероятно, появляются сторонние коннотации): *В любви главное не тянуть одеялку **на себя, ненаглядную**, а суметь отдать эту одеялку тому, кого любишь; это же трата времени **на себя родную**; Она зябла у зеркала, читая письмо, и иногда взглядывала **на себя, зябкую***; и т. д.

Таким образом, некоторые трансформации, создаваемые на основе наиболее частотных прецедентных текстов корпуса ОРД, могут быть расценены как свидетельства формирования комических паспарту, в которых «размыкается» семантика изначальных ПТ, а «идиоматичность» приобретает иной, «конструктивный», характер.

6 Заключение

Предложенный анализ показывает еще один способ бытования прецедентных текстов в современной коммуникации, а также еще один возможный путь их описания и систематизации. При таком подходе ПТ рассматривается как конструкция, в составе которой выделяются якорные компоненты и переменные слоты, при этом такая конструкция, при соблюдении некоторых условий, является комическим паспарту, в котором прецедентность сочетается с комизмом, а основным критерием выделения становится семантическая и в некоторых случаях формальная «удаленность» от изначального прецедентного текста. Одновременно такой подход к анализу материала устной и устно-письменной речи, составляющих все разнообразие современной

коммуникации, позволяет осветить еще одну грань нового коммуникативного феномена, каким являются комические паспарту.

Возможным представляется соотношение с рассмотренными понятиями КС и КП понятия *мема*, под которым понимается «неформальная и юмористическая коммуникативная единица: идея, символ или образ, которые быстро распространяются от человека к человеку, а также в интернете» (https://ru.wikipedia.org), однако это требует отдельного исследования.

Комические паспарту, как мы постарались показать в этой статье, представляют собой особый речевой феномен, в котором синтезируется «лудическое», «прецедентное» (шире — «интертекстуальное») и «комическое», они наглядно иллюстрируют конкретные процессы и тенденции, организующие актуальный русский разговорный дискурс. Взгляд на комические паспарту с точки зрения прецедентности способен, вне всякого сомнения, вывести исследование на уровень когнитивистики, психо- и социолингвистики, а также оказаться полезным во многих прикладных аспектах, таких как лингводидактика, практика перевода, речевое портретирование (индивида или социума), лингвистическая экспертиза и т. п.

Перспективной представляется и дальнейшая нюансировка границ и свойств феномена комических паспарту, детальное сопоставление явления с понятиями «конструкция», «коллокация», «фразеосхема», «идиома».

Благодарность

Исследование выполнено при поддержке гранта РНФ «Структура и функционирование устойчивых неоднословных единиц русской повседневной речи».

Список источников

[1] *Википедия* [Электронный ресурс]. URL: https://ru.wikipedia.org.
[2] *НКРЯ* — Национальный корпус русского языка [Электронный ресурс]. URL: https://ruscorpora.ru.
[3] *ОРД* — Корпус русского языка повседневного общения «Один речевой день» [Электронный ресурс]. URL: https://ord.spbu.ru.
[4] *Прагматикон* [Электронный ресурс]. URL: https://pragmaticon.ruscorpora.ru.
[5] *Русский Конструктикон* [Электронный ресурс]. URL: https://constructicon.github.io/russian.

Список использованной литературы

[1] *Батулина А.В.* О лексикографическом представлении антипословиц в «Прикольном словаре» В.М. Мокиенко, Х. Вальтера // Проблемы истории, филологии, культуры. — 2011, № 3 (33). — С. 214-217.
[2] *Беликов В.И., Верещагина А.Д., Селегей В.П.* На границах лексикона: дифференциальные корпусные исследования паремийного фонда РЯ // Труды международной конференции «Корпусная лингвистика-2021». 1-3 июля 2021 г., Санкт-Петербург / Отв. ред. *В.П. Захаров*. — СПб.: Скифия-принт, 2021. — С. 44-55.
[3] *Богданова-Бегларян Н.В.* Прецедентный текст в устной повседневной речи: возможности лингвистического описания // Вестник Бурятского гос. ун-та. Филология. — 2018, № 2. Т. 2. — С. 19-25.
[4] *Богданова-Бегларян Н.В.* Прецедентные тексты как источник новых идиом // Новое в русской и славянской фразеологии / *А. Архангельская* (ред.). — Olomouc: Univerzita Palackého v Olomouci, 2020. — С. 424-428.
[5] *Богданова-Бегларян Н.В.* Прецедентный текст как «непереводимая игра слов»: функционирование в русской повседневной речи и проблемы понимания и преподавания в иностранной аудитории // IX Международная научная конференция «Современные проблемы славянской филологии: Форма и смысл. К 130-летию со дня рождения В. Шкловского» (11-12 ноября 2023 года), гос. университет Чжэнчжи, Тайбэй (Тайвань). — 2025. — В печати.
[6] *Гаспаров Б.М.* Язык, память, образ. Лингвистика языкового существования. — М.: Новое литературное обозрение. Вып. IX / Ред. вып. *И. Прохорова*. 1996. — 352 с.
[7] *Гридина Т.А., Талашманов С.С.* Языковая игра в современной интернет-коммуникации: метаязыковой аспект // Политическая лингвистика. — 2019. № 3 (75). — С. 31-37.
[8] *Земская Е.А., Китайгородская М.В., Ширяев Е.Н.* Русская разговорная речь. Общие вопросы. Словообразование. Синтаксис. — М.: Наука, 1981. — 276 с.

[9] *Иомдин Л.Л.* Многозначные синтаксические фраземы: между лексикой и синтаксисом // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2006». — М.: РГГУ, 2006. — С. 202-206.
[10] *Кайгородова И.Н.* Проблема синтаксической идиоматики (на материале русского языка). — Астрахань: Изд-во Астраханского гос. пед. ун-та, 1999. — 249 с.
[11] *Караулов Ю.Н.* Русский язык и языковая личность. — М.: Изд-во ЛКИ, 2010. — 264 с.
[12] *Коган Е.С.* О статусе некоторых устойчивых единиц в речи социальной микрогруппы // Вестник Пермского ун-та. Российская и зарубежная филология. — 2018. Т. 10. Вып. 3. — С. 42-51.
[13] *Отчет РНФ* — Структура и функционирование устойчивых неоднословных единиц русской повседневной речи. Коллективный проект / Авторы: *Н.В. Богданова-Бегларян, О.В. Блинова, М.В. Хохлова, Т.Ю. Шерстинова, А.Д. Базаржапова, Т.Л. Колосовская, Т.И. Попова, Д.А. Стойка* / Рук. *Н.В. Богданова-Бегларян.* — СПб., 2024. — 109 с. (машинопись).
[14] *Пересыпкина К.А.* Больше, чем прецедентный текст: комические паспарту как явление современной коммуникации // Социо- и психолингвистические исследования. — 2024. Вып. 12. — С. 75-78.
[15] *Пересыпкина К.А., Богданова-Бегларян Н.В.* Комические паспарту как новое явление современной коммуникации // Коммуникативные исследования. — 2024. Т. 11. № 4. — С. 758-772.
[16] *ПМ* — Прагматические маркеры русской повседневной речи: словарь-монография / Сост., отв. ред. и автор предисловия *Н.В. Богданова-Бегларян*. — СПб.: Нестор-История, 2021. — 520 с.
[17] *Шведова Н.Ю.* Очерки по синтаксису русской разговорной речи. — М.: Изд-во Академии наук СССР, 1960. — 378 с.
[18] *Шмелев Д.Н.* Синтаксическая членимость высказывания в современном русском языке. — М.: Наука, 1976. — 155 с.
[19] *Bogdanova-Beglarian, N.V., Blinova, O.V., Khokhlova, M.V., Sherstinova, T.Yu., Popova, T.I.* Multiword Units in Russian Everyday Speech: Empirical Classification and Corpus-Based Studies // XXVIth International Conference “Speech and Computer”, SPECOM-2024. Proceedings. Part 1. Belgrade, Serbia. November 25-28, 2024 / *A. Karpov, V. Delic* (eds.). LNAI 15299. — Springer, 2024. — Pp. 187-200.
[20] *Fillmore Ch.J., Kay P., O’Connor M.C.* Regularity and Idiomaticity in Grammatical Constructions: The Case of Let alone // Language. — 1988. No 64 (3). — Pp. 501-538.

References

[1] Batulina, A.V. (2011), O leksikograficheskom predstavlenii antiposlovits v “Prikol’nom slovare” V.M. Mokienko, H. Val’tera [Lexicographic Presentation of Anti-proverbs in the ‘Cool Dictionary’ by V.M. Mokiyenko and H. Walter]. *Problemy istorii, filologii, kul’tury* [Problems of History, Philology, Culture], No. 3(33), pp. 214-217. (in Russian).
[2] Belikov, V.I., Vereshchagina, A.D., Selegey, V.P. (2021), Na granicakh leksikona: differencial’nye korpusnye issledovaniya paremijnogo fonda RYa [On the Borders of the Lexicon: Differential Corpus Studies of the Paremiological Fund of the Russian Language]. *Trudy mezhdunarodnoj konferencii “Korpusnaya lingvistika-2021”. 1-3 iyulya 2021 g., Sankt-Peterburg* [Proceedings of the International Conference “Corpus Linguistics-2021”. July 1-3, 2021, St. Petersburg. Ed. by V.P. Zaharov], St. Petersburg, Skifiya-Print publ., pp. 44-55. (in Russian).
[3] Bogdanova-Beglarian, N.V. (2018), Pretsedentnyj tekst v ustnoj povsednevnoj rechi: vozmozhnosti lingvisticheskogo opisaniya [Precedent Text in Oral Everyday Speech: Possibilities of Linguistic Description]. *Vestnik Buryatskogo gos. un-ta. Filologiya* [Bulletin of the Buryat State University. Philology], Vol. 2. No. 2, pp. 19-25. (in Russian).
[4] Bogdanova-Beglarian, N.V. (2020), Pretsedentnye teksty kak istochnik novykh idiom [Precedent Texts as a Source of New Idioms]. *Novoe v russkoj i slavyanskoj frazeologii* [New in Russian and Slavic phraseology. Ed. by A. Arkhangelskaya], Olomouc, Univerzita Palackého v Olomouci publ., pp. 424-428. (in Russian).
[5] Bogdanova-Beglarian, N.V. (2025), Pretsedentnyj tekst kak “neperevodimaya igra slov”: funktsionirovanie v russkoj povsednevnoj rechi i problemy ponimaniya i prepodavaniya v inostrannoj auditorii [Precedent Text as ‘Untranslatable Wordplay’: Functioning in Russian Everyday Speech and Problems of Understanding and Teaching in a Foreign Audience]. *IX Mezhdunarodnaya nauchnaya konferentsiya “Sovremennye problemy slavyanskoj filologii: Forma i smysl. K 130-letiyu so dnya rozhdeniya V. Shklovskogo”* [IX International Scientific Conference ‘Modern Problems of Slavonic Philology: Form and Meaning. To the 130th anniversary of the birth of V. Shklovsky’ (11-12 November 2023), Zhengzhi State University], Taipei (Taiwan). (In print). (in Russian).
[6] Gasparov, B.M. (1996), Yazyk, pamyat’, obraz. Lingvistika yazykovogo sushchestvovaniya [Language, Memory, Image. Linguistics of Linguistic Existence]. *Novoe literaturnoe obozrenie* [New Literary Review], Moscow, Iss. IX, Ed. by I. Prokhorova, 352 p. (in Russian).

[7] Gridina, T.A., Talashmanov, S.S. (2019), Yazykovaya igra v sovremennoj internet-kommunikatsii: metayazykovoj aspect [Language Game in Modern Internet Communication: Metalinguistic Aspect]. *Political Linguistics*, No. 3 (75), pp. 31-37. (in Russian).
[8] Zemskaya, E.A., Kitajgorodskaya, M.V., Shiryaev, E.N. (1981), *Russkaya razgovornaya rech’. Obshchie voprosy. Slovoobrazovanie. Sintaksis* [Russian Colloquial Speech. General Questions. Word-Formation. Syntax], Moscow, Nauka publ., 276 p. (in Russian).
[9] Iomdin, L.L. (2006), Mnogoznachnye sintaksicheskie frazemy: mezhdu leksikoj i sintaksisom [Multivalued Syntactic Phrasemes: Between Lexicon and Syntax]. *Komp’yuternaya lingvistika i intellektual’nye tekhnologii. Trudy mezhdunarodnoj konferencii “Dialog 2006”* [Computational Linguistics and Intelligent Technologies. Proceedings of the International Conference “Dialogue 2006”], Moscow, RGGU, pp. 202-206. (in Russian).
[10] Kaygorodova, I.N. (1999), *Problema sintaksicheskoj idiomatiki (na materiale russkogo yazyka)* [The Problem of Syntactic Idiomaticity (based on the material of the Russian language)], Astrakhan, Astrakhan State Pedagogical University Press, 249 p. (in Russian).
[11] Karaulov, Y.N. (2010), *Russkiy yazyk i yazykovaya lichnost’* [Russian Language and Linguistic Personality], Moscow, LKI Publishing House, 264 p. (in Russian).
[12] Kogan, E.S. (2018), O statuse nekotoryh ustojchivyh edinic v rechi social’noj mikrogruppy [On the Status of some Stable Units in the Discourse of a Small Social Group]. *Vestnik Permskogo un-ta. Rossijskaya i zarubezhnaya filologiya* [Perm University Herald. Russian and Foreign Philology], Vol. 10. Iss. 3, pp. 42-51. (in Russian).
[13] Otchet RSF — *Struktura i funktsionirovanie ustojchivykh neodnoslovnykh edinic russkoj povsednevnoj rechi. Kollektivnyj proekt* [RSF Report — Structure and Functioning of Stable Multi-Word Units of Russian Everyday Speech. Collective Project] (2024), Avtory: Bogdanova-Beglarian, N.V., Blinova, O.V., Khokhlova, M.V., Sherstinova, T.Yu., Bazarzhapova, A.D., Kolosovskaya, T.L., Popova, T.I., Stoyka, D.A., Ruk. [Head] N.V. Bogdanova-Beglarian, St. Petersburg, 109 p. (typescript). (in Russian).
[14] Peresypkina, X.A. (2024), Bol’she, chem pretsedentnyj tekst: komicheskie paspartu kak yavlenie sovremennoj kommunikatsii [More than Just a Precedent Text: Comic Passe-Partout as a Phenomenon in Modern Communication]. *Socio- i psiholingvisticheskie issledovaniya* [Socio- and psycholinguistic studies]. Iss. 12, pp. 75-78. (in Russian).
[15] Peresypkina, X.A., Bogdanova-Beglarian, N.V. (2024), Komicheskie paspartu kak novoe yavlenie sovremennoj kommunikatsii [Comic Passe-Partout as a New Phenomenon of Modern Communication]. *Kommunikativnye issledovaniya* [Communicative Studies]. Vol. 11. no. 4, pp. 758-772. (in Russian).
[16] PM — *Pragmaticheskiye markery russkoy povsednevnoy rechi: slovar’-monografiya* / Sost., otv. red. i avtor predisloviya N.V. Bogdanova-Beglarian [*Pragmatic Markers of Russian Everyday Speech: Dictionary-Monograph* / Comp., editor and author of the preface N.V. Bogdanova-Beglarian]. St. Petersburg: Nestor-History, 2021. 520 p. (in Russian).
[17] Shvedova, N.Yu. (1960), *Ocherki po sintaksisu russkoj razgovornoj rechi* [Essays on the Syntax of Russian Colloquial Speech], Moscow, Academy of Sciences of the USSR publ., 378 p. (in Russian).
[18] Shmelev, D.N. (1976), *Sintakicheskaya chlenimost’ vyskazyvaniya v sovremennom russkom yazyke* [Syntactic Partibility of the Statement in Modern Russian], Moscow, Nauka, 155 p. (in Russian).
[19] Bogdanova-Beglarian, N.V., Blinova, O.V., Khokhlova, M.V., Sherstinova, T.Yu., Popova, T.I. (2024), Multiword Units in Russian Everyday Speech: Empirical Classification and Corpus-Based Studies. *XXVIth International Conference “Speech and Computer”*, SPECOM-2024. Proceedings. Part 1. Belgrade, Serbia. November 25-28 2024. A. Karpov, V. Delic⊂ (eds.). LNAI 15299. Springer, pp. 187-200.
[20] Fillmore, Ch.J., Kay, P., O’Connor, M.C. (1988), Regularity and Idiomaticity in Grammatical Constructions: The Case of Let alone. *Language*. No. 64 (3), pp. 501-538.

Readability assessment of written Adyghe using a baseline approach

Uliana Petrunina
Center for Language and Brain
HSE University
Moscow, Russia
upetrunina@hse.ru

Nina Zdorova
Center for Language and Brain
HSE University, Institute of Linguistics RAS
Moscow, Russia
nzdorova@hse.ru

Abstract

The study introduces a cross-linguistic approach extending the English-based Flesch Reading Ease formula for the assessment of Adyghe texts' readability level. The method relies on the corpus-based analysis of Adyghe shallow linguistic features, i.e. syllable length, word count, and sentence length. It allows to adjust the Flesch formula in accordance with these features by means of natural language processing (NLP) and corpus data analysis. Preliminary results showed that the adapted formula could overall adequately differentiate texts according to their complexity levels although it lacked precision in distinguishing between texts belonging to the same complexity range. The approach can be easily extended to other typologically different minority languages subject to their corpora size and availability.

Keywords: readability, text complexity, Flesch formula, shallow features, low-resource language, minority language, Adyghe

DOI: 10.28995/2075-7182-2025-23-1100-1109

Оценка сложности текстов на адыгейском с использованием типового подхода

Ульяна Петрунина
Центр языка и мозга
НИУ ВШЭ Москва, Россия
upetrunina@hse.ru

Нина Здорова
Центр языка и мозга НИУ ВШЭ,
ИЯз РАН, Москва, Россия
nzdorova@hse.ru

Аннотация

В настоящем исследовании представлен кросс-лингвистический метод адаптации англоязычной формулы Flesch Reading Ease для оценки уровня сложности адыгейских текстов. Метод опирается на корпусный анализ особенностей адыгейской лексики: размера слога, длины предложения и количества слов. Он позволяет скорректировать формулу Флеша в соответствии с лингвистическими особенностями адыгейского с помощью инструментов обработки естественного языка и анализа корпусных данных. Предварительные результаты показали, что адаптированная формула достаточно приемлемо различает тексты по уровню сложности, хотя ей не хватает точности в различении текстов, принадлежащих к одному и тому же диапазону сложности. Данный подход может быть с легкостью адаптирован на другие типологически различные малые языки в зависимости от объема и доступности корпусных данных.

Ключевые слова: формула Флеша, сложность текстов, адыгейский, малые языки, удобочитаемость текста, базовые лингвистические характеристики слова, малоресурсные языки

1 Introduction

Readability measure serves to estimate the complexity level of a given text comprehended by a reader. Readability level is dependent on the complexity of linguistic content, style quality, readability of print and reference to the reader (Bamberger, 2000, see also Rottensteiner,2010). As Dale and Chall (1949) put it, readability is "the sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers have with it" (23). Readability measures are commonly used in education including book publishing and language-learning applications, health care

and marketing domains (Antunes and Lopes, 2019, Madrazo Azpiazu and Pera, 2020, 644-645), as well as in psycholinguistics, e.g. L1/L2 reading comprehension (Baldwin, 1977, Xia et al., 2016).

Readability measures have been applied to assess text complexity in high-resource languages such as English (Aluisio et al., 2010), Spanish (Anula, 2007, Drndarević et al., 2013, Spaulding, 1956), German (Amstad, 1978, Hancke et al., 2012, Naderi et al., 2019), Swedish (Tillman and Hagberg, 2014), French and Portuguese (François and Miltsakaki, 2012), Russian (Karpov et al., 2014, Krioni et al., 2008, Oborneva, 2005a,b, Reynolds, 2016), Polish (Broda et al., 2014), Czech (Bendová, 2021), Chinese (Chen et al., 2011), Arabic (Al-Ajlan et al., 2008), and many others. Only few studies cover the topic of readability measures for low-resource languages such as Basque (Gonzalez-Dios et al., 2014), Sesotho (Sibeko, 2023), Bengali (Chakraborty et al., 2021), and Bangla (Islam et al., 2012).

To our knowledge, no such measures have been adapted for the minority languages of Russia including Adyghe, Bashkir, Tatar, Udmurt, and so forth. The paper aims to develop a cross-linguistic baseline approach that extends an English-based readability measure for Adyghe, a polysynthetic low-resource language, and makes it applicable to other minority languages. Adyghe's rich verb morphology and complex syllable structure provide sufficient input for the analysis of shallow features, improving the accuracy of complexity level estimation. The study's findings will be used to implement a text readability tool adapted to other low-resource languages, in addition to supporting research-based tasks in psycholinguistic experiments and second language instruction. Namely, it will be applied to prepare materials for tests diagnosing reading skills in elementary school students (e.g. the KARASIK test, Parešina, 2022; see also the Standardized Assessment of Reading Skills, Kornev, 1997). The implementation of the readability formula for Adyghe is available on GitHub.[1]

The paper is structured as follows: In Section 2, we overview extant readability metrics and focus on the adaptations of the Flesch Reading Ease. Section 3 presents the corpus data used in the experiments, NLP methods for collecting and processing it for the purpose of segmentation and syllable and sentence length estimates. Section 3.2 describes the adjustment of coefficients for Adyghe and the grading scores produced by the newly adapted formula; Section 3.3 provides error analysis. In Section 4, we discuss the implications of the obtained results and in Section 5, we conclude upon the experiment and outline future directions in the development of the formula.

2 Background

2.1 Readability metrics overview

Classic readability metrics are typically based on syllable-length, word-length and word-frequency variables given as parameters, complying with the assumption that longer sentences and longer words increase text complexity (Bailin and Grafstein, 2001, Sydes and Hartley, 1997). These variables consist of shallow linguistic features, namely, averages of words per sentence, syllables or letters per word, proportions of part-of-speech tags or out-of-simple-vocabulary words in text. Metrics computing scores based on a syllable and word level include the Flesch-Kincaid and Flesch Reading Ease (Flesch, Rudolph, 1948, Kincaid et al., 1975), Simple Measure of Gobbledygook (SMOG; McLaughlin, 1969), Gunning FOG Index (Gunning, 1971). Metrics based on word length or word frequency variables are applied to estimate syntactic complexity for the purpose of text simplification following the assumption that a text written for early readers contains more frequent words and shorter sentences (Chall and Dale, 1995, Crossley et al., 2008). This group of metrics includes the Automated Readability Index (ARI; Kincaid and Delionbach, 1973), Coleman Liau (Coleman, 1971), Läsbarhetsindex (LIX; Björnsson, 1983), Dale-Chall formula (Chall and Dale, 1995).

Compared to classic readability metrics, readability assessment in NLP approaches is classification-based and computes text cohesion and complexity on linguistic, discourse, and concept-based levels (Crossley et al., 2008, Dell'Orletta et al., 2012), such as the Coh-Metrix tool (Graesser et al., 2004). These approaches make estimates of text coherence using language models, parse tree-based predictors, computer probability and so on (François and Miltsakaki, 2012).

[1] `https://github.com/ulp16/FRE-ady`.

2.2 Flesch formula and its adaptation

The Flesch Reading Ease (FRE) formula is one of the most used classic readability formulae, which is also applied in numerous readability assessment tools. Because the formula relies on surface text features such as syllable-to-token and token-to-sentence ratios (Bendová, 2021), it has been adapted to a number of languages, including German (Amstad, 1978), Italian (Franchina and Vacca, 1986), Russian (Oborneva, 2005a,b, 2006), and Czech (Bendová, 2021).

FRE was developed by Rudolf Flesch (1948) for grading standard English reading material within the American education system, covering the range from approximately fourth grade to college graduate level with scores from 0 (unreadable) to 100 (very easy to read; Klare, 1969, see also DuBay, 2004). The FRE scores are calculated using Equation 1:

$$FRE_{\text{english}} = 206.835 - 1.015 \times ASL - 84.6 \times ASW \quad (1)$$

where *206.835*–a constant which delimits the ordinal FRE scale boundaries from 1 to 100,
ASL–Average Sentence Length based on number of words, and
ASW–Average number of Syllables per Word.
ASL and ASW coefficients are easily measured and are transparent to interpretation (Lanka and Pēks, 2013, 228). For English, FRE scores between 90–100 correspond to easy texts for junior students, 60–70 for school leavers and 0–30, for people with higher education. Equation 2 illustrates Oborneva's FRE formula:

$$FRE_{\text{russian}} = 206.835 - 1.3 \times ASL - 60.1 \times ASW \quad (2)$$

where 1.3 and 60.1 are adjusted coefficients calculated by multiplying the ratios of ASL and ASW in English and Russian by the original coefficients 1.015 and 84.6, respectively. The Russian ASL and ASW were determined on the basis of six million words from about 100 literary Russian-English fictions and dictionaries (Oborneva, 2005a,b, 2006). Kupriyanov et al. (2023) pointed out that Oborneva's formula was developed on fiction texts and therefore provided overestimated results for other types of texts. FRE for Russian was found to be the most suitable formula for evaluating texts in both Russian and Latvian; it was able to distinguish readability levels between Latvian texts written by 11$^{\text{th}}$ grade students and Physics textbooks, thereby demonstrating the latter's greater complexity (Lanka and Pēks, 2013, 233).

In our study, we followed the method of adapting an English FRE to another language drawing on Oborneva (2005a,b) because it preserves the FRE grading scale and uses easily available shallow textual properties as correction coefficients to adapt the formula to another language.

3 Experiment

3.1 Data and tools

We retrieved approximately 100000 lines of Adyghe (plain) texts by using API queries on the Adyghe corpus[2] provided by the Python *lingcorpora*[3] package. For English data, necessary as the basis for formula adjustment, we accessed the Brown Corpus, a one-million word electronic corpus of English texts such as news, reviews, editorial, fiction and so on,[4] using Natural Language Toolkit (NLTK).[5] To obtain counts for shallow features including the number of sentences, words, syllables and their averages from the English data we used in-built NLTK functions. For the Adyghe subcorpus we used a custom Python script for syllabification taking into consideration Adyghe syllable structure (Moroz, 2019) and characteristics of each letter in the the Adyghe alphabet including triple (e.g. *шIy*, *шъу*, *хъу*) and double letters (e.g. *Iy*, *шI*, *шъ*). Figure 1 illustrates an excerpt from the script output containing a list of tokens, their syllable structure and counts. For example, the adverb *ӀакIоу* 'together' is composed of two plosives (marked as O) *Ӏ* and *кI* and three vowels *а*, *о* and *у* (marked as V), making a total of two syllables. The

[2]The corpus is a closed pilot version of the Adyghe corpus which consists of press, (non-)fiction, and blog texts available at `http://web-corpora.net/AdygheCorpus/search/`

[3]`https://lingcorpora.github.io/lingcorpora.py/html/index.html`

[4]A complete list of genre is available at `http://icame.uib.no/brown/bcm-los.html`

[5]`https://www.nltk.org/api/nltk.corpus.html`

	token	syll	#
0	ащ	V0	1
1	дакIоу	0V0V	2
2	иIэнатIэкIи	V0VSV0V0V	5
3	лъагъэкIуатэ	0V0V0V0V	4

Figure 1: An output sample with estimates of Adyghe syllables for each token.

syllable structure of *ɔакIоу* is therefore OVOV. We then used the language non-specific *Lexicon Count* and *Sentence Count* functions to calculate text statistics for shallow features in Adyghe via the *textstat* library.[6]. Estimates for shallow features in Adyghe and English corpora are given in Table 1. It indicates that, despite a roughly similar number of words in both Adyghe and English corpora, English sentences are on average 1.299 times longer and English syllables are 0.862 times longer than their corresponding Adyghe sentences and syllables.

Feature	Adyghe	Brown
sent	104298	57340
word	1640756	1161192
syll	4829110	1260859
avg syll	2.89	2.49
avg sent len	16.74	21.75

Table 1: Counts of sentences, words, syllables *syll*, average syllable number *avg syll* and sentence length *avg sent len*.

The relation and effect size of Adyghe and English data were assessed statistically using non-parametric[7] tests via R (R Core Team, 2021). The Mann-Whitney ranks test (Kilgarriff, 2001) showed a significant difference between the Adyghe syllable/sentence length groups and their English counterparts (p < 2.2e-16 for both). The Glass's Rank Biserial Coefficient showed a small but meaningful positive difference between Adyghe and English sentence length samples (rg = 0.228) and a small but significant negative difference between Adyghe and English syllable length (rg = -0.18). The results confirm the statistical soundness of the data in Adyghe and English used for the formula adjustment.

3.2 Coefficient adjustment

We computed correction coefficients for the FRE formula by multiplying ratios of English to Adyghe averages for ASL and ASW (1.299 and 0.862) by the original coefficients 1.015 and 84.6, respectively. Preliminary testing of Adyghe preschooler texts (described below) resulted in an overly high score of 145. Although FRE scores exceeding 100 are technically possible, the text under analysis with the FRE score of 121.22 should consist of sentences with a single one-syllable word (Diamond Jr and Levy, 1994, Shneyderman et al., 2021, 2022). To prevent FRE scores from exceeding the scale boundaries, we reduced the English FRE constant from 206.835 to 150.835 by ensuring that the FRE score for the preschool texts corresponded to the range of 100.0–90.0. The constant was adjusted similarly to Amstad's (1978) adaptation of German FRE. The Amstad FRE relies on the adjusted weight of ASW measure and constant delimitating the scale as average word length in German tends to be higher that in English, see eq. 3.

$$FRE_{\text{german}} = 180 - ASL - (58.5 \times ASW) \tag{3}$$

The Amstand FRE was shown to provide good indication of sentence complexity in German texts alongside with neural-based models (Anschütz and Groh, 2022). Equation 4 illustrates the adjusted FRE for-

[6]The default English implementation is available at `https://pypi.org/project/textstat/`

[7]We chose these tests due to non-parametric distribution of all the samples diagnosed by the Anderson-Darling normality test (p < 2.2e-16).

mula for Adyghe:

$$FRE_{\text{adyghe}} = 150.835 - (1.32 \times ASL) - (72.93 \times ASW) \quad (4)$$

where 150.835–a corrected constant,
1.32–a corrected coefficient for ASL,
0.86–a corrected coefficient for ASW.

We then selected five samples of Adyghe texts from educational resources: fictions/poems for preschoolers (Апиш et al., 2017) and 5th grade learners of Adyghe (Apiš and Udžuxu, 2014), scholarly texts for 11th grade learners (Мамий et al., 2011), abstracts from the scientific conference proceedings aimed at higher education audience (Kesebeževa et al., 2021), and articles from the Adyghe newspaper (Адыгэ псалъэ №39, 2025) targeted at a wide range of age groups. The scores computed for each sample using the adapted Adyghe FRE formula and their text length[8] are given in Table 2. As is shown in

Target level	Text length	Exp FRE range	Ady FRE	Interpretation	Source
preschool	9397	100.0–90.0	90.48	very easy to read	anthology
5th grade	7648	100.0–90.0	91.32	very easy to read	textbook
11th grade	10099	60.0–70.0	66.51	standard language	handbook
higher education	13661	70.0–80.0	76.82	fairly easy to read	abstracts
unspecified	9500	70.0–80.0	75.24	fairly easy to read	newspaper

Table 2: Length of a text sample, expected FRE score ranges, observed FRE scores and reading interpretation for Adyghe texts written for preschoolers, learners in the 5th and 11th grades, higher education audience, newspaper readers.

Table 2, the adjusted FRE formula classified the preschool and 5th grade texts according to the expected complexity range of 100.0–90.00 as "very easy to read". However, the texts suitable for the 5th grade were scored higher in readability than the texts for preschoolers (91.32 versus 90.48). The FRE formula rated the 11th grade texts as written in "standard language" based on the 60.0–70.0 range,[9] while the abstracts and newspaper articles scored on a higher readability level of 70.0–80.0 as "fairly easy to read". The FRE scores did not appear to have been significantly impacted by differences in text length across the samples.

3.3 Error analysis

The text statistics shown in Table 3 offer some explanation for the above-mentioned FRE scores.

Texts	AvgSentLen	AvgSylLen
11th grade	15.36	2.89
newspaper	14.71	2.54
abstracts	10.17	2.71
preschool	7.08	2.19
5th grade	4.8	2.31

Table 3: Average sentence *AvgSentLen* and syllable length *AvgSylLen* observed in the evaluated samples sorted by *AvgSentLen*.

The 11th grade texts contained on average the longest sentences and words (in syllables), followed by the newspaper articles and conference abstracts. In contrast, the preschool and 5th grade texts had on average the shortest sentences and smallest number of syllables in words. Such differences in sentence and syllable length among these samples were mostly explained by their paragraph and sentence structure. First, although several paragraphs overlapped between the two samples, the preschool texts were taken from a monolingual textbook and the 5th-grade texts from a bilingual (Adyghe-Russian) textbook.

[8] Text length is a number of tokens in each sample.
[9] The score corresponds to the US 8th and 9th-grade levels.

Second, the preschool texts comprised six large paragraphs of prose texts, over 70 poems[10] and several dialogues, whereas the 5th grade texts contained mostly prose including dialogues with two- to four-word sentences, one-word exclamations (e.g. *АмкІышъ!* 'Nightingale!') and two- to three-word questions (e.g. *Хэта зэныбджэгъухэр?* 'Who [are] friends?'). Finally, most sentences in the 11th grade sample and newspaper articles tended to be long and complex, while sentences in the abstracts were relatively shorter.

4 Results and discussion

The above findings show that the FRE formula with adjusted coefficients and constant rated scholarly Adyghe texts roughly in the expected complexity range distinguishing between highly readable and standard texts. Without correcting the FRE constant, the formula produced overrated scores surpassing the limit of 100, e.g. 145 for the preschool texts. The formula also did not capture fine-grade differences between the preschool and 5th grade texts, on the one hand, and the scientific abstracts and the 11th grade texts, on the other. Instead, it placed the 5th grade texts and scientific texts higher on the readability scale than those suitable for preschoolers and students in the 11th grade. While the scholarly texts scored satisfactorily on the FRE scale, both the newspaper articles and conference abstracts were ranked as similar, quite readable texts suitable for school students.

Variance in the FRE readability rankings can be accounted by several factors: First, the monolingual preschool texts are structurally more complex than the 5th grade texts for bilingual learners, see Table 3. Second, the FRE does not consider syntactic structure of a sentence and lexical semantics of a word including neologisms, terminology, learned words, borrowings, stylistic devices and so forth. It is therefore unclear whether the FRE is relevant for rating verses as their syntactic structure and lexicon properties are often stylistically motivated including comma-separated sentences spread over several lines and/or words used figuratively. Redish (1981) argues that readability (Flesch) formulas are limited to prose texts whereas poems should be evaluated using the Dale-Chall formula based on a vocabulary list of acceptable words taking into consideration nonce-words and acronyms. Newspaper articles and conference abstracts should also be assessed for readability separately from standard academic textbooks and fiction/non-fiction prose since their straightforward sentence structure tends to be combined with lexically and/or semantically complex words.

5 Conclusion and future directions

In this paper, we have introduced a baseline approach that allows to grade Adyghe texts according to the FRE scale majorly ranking them in the expected readability ranges. Further empirical verification and statistical evaluation of the formula are needed to attain optimal results for grading written Adyghe. We intend to extend the approach for Buryat, Tatar, and Udmurt, using corpora APIs from the *lingcorpora* package.

As future work, we may potentially consider implementing FRE features in a classifier along the lines of *Textometer* (Laposhina and Lebedeva, 2021) or *Jasnopis* (Broda et al., 2014). The classifier could be enriched with features of distributional lexical similarity based on vector representations of word embeddings (see e.g. Anschütz and Groh, 2022, Martinc et al., 2021) and morphological information using a parser for Adyghe (e.g. *uniparser-grammar-adyghe*; Arkhangelskiy and Medvedeva, 2016).[11]

References

Amani A. Al-Ajlan, Hend S. Al-Khalifa, and AbdulMalik S. Al-Salman. Towards the development of an automatic readability measurements for Arabic language. In *2008 Third international conference on digital information management*, pages 506–511. IEEE, 2008.

[10] The poems ranged from two to 119 lines, with a one- to four-word lines, e.g.:
Аргъоир пэдыд,
ЫІпэ – мастэу мэлыд, [...]

[11] `https://github.com/timarkh/uniparser-grammar-adyghe`

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 fifth workshop on innovative use of NLP for building educational applications*, pages 1–9, 2010.

Toni Amstad. *Wie verständlich sind unsere Zeitungen? [How understandable are our newspapers?]*. Studenten-Schreib-Service, 1978.

Miriam Anschütz and Georg Groh. TUM Social Computing at GermEval 2022: Towards the Significance of Text Statistics and Neural Embeddings in Text Complexity Prediction. In *Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text*, pages 21–26, 2022.

Hélder Antunes and Carla Teixeira Lopes. Analyzing the adequacy of readability indicators to a non-English language. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10*, pages 149–155. Springer, 2019.

Alberto Anula. Tipos de textos, complejidad lingüıstica y facilicitación lectora. In *Actas del Sexto Congreso de Hispanistas de Asia*, pages 45–61, 2007.

F. N. Apiš and S. A. Udžuxu. *Adygejskij jazyk: 5 klass [The Adyghe language: the fifth grade]*. Kačestvo, 2014.

Timofey Arkhangelskiy and Maria Medvedeva. Developing Morphologically Annotated Corpora for Minority Languages of Russia. In *CLiF*, pages 1–6, 2016.

Alan Bailin and Ann Grafstein. The linguistic assumptions underlying readability formulae: A critique. *Language & Communication*, 21(3):285–301, 2001.

R. Scott Baldwin. Psycholinguistic Strategies as a Factor in Estimating the Readability of Written Texts. 1977.

Richard Bamberger. *Erfolgreiche Leseerziehung in Theorie und Praxis: mit besonderer Berücksichtigung des Projekts "Leistungs-und Motivationssteigerung im Lesen und Lernen unter dem Motto Lese- und Lernolympiade"*. Öbv & Hpt, 2000.

Klára Bendová. Using a parallel corpus to adapt the Flesch Reading Ease formula to Czech. *Jazykovedný časopis*, 72(2):477–487, 2021.

Carl-Hugo Björnsson. Readability of Newspapers in 11 Languages. *Reading Research Quarterly*, pages 480–497, 1983.

Bartosz Broda, Bartłomiej Nitoń, Włodzimierz Gruszczyński, and Maciej Ogrodniczuk. Measuring readability of Polish texts: Baseline experiments. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 573–580, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL `https://aclanthology.org/L14-1366/`.

Susmoy Chakraborty, Mir Tafseer Nayeem, and Wasi Uddin Ahmad. Simple or complex? Learning to predict readability of Bengali texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12621–12629, 2021.

Jeanne S. Chall and Edgar Dale. *Reability revisited: The new Dale-Chall readability formula*. MA: Brookline Books, Cambridge, 1995.

Yaw-Huei Chen, Yi-Han Tsai, and Yu-Ta Chen. Chinese readability assessment using TF-IDF and SVM. In *2011 International Conference on Machine Learning and Cybernetics*, volume 2, pages 705–710. IEEE, 2011.

Edmund B. Coleman. Developing a technology of written instruction: Some determiners of the complexity of prose. *Verbal learning research and the technology of written instruction*, pages 155–204, 1971.

Scott A. Crossley, Jerry Greenfield, and Danielle S. McNamara. Assessing text readability using cognitively based indices. *Tesol Quarterly*, 42(3):475–493, 2008.

Edgar Dale and Jeanne S. Chall. The Concept of Readability. *Elementary English*, 26(1):19–26, 1949.

Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. Genre-oriented readability assessment: A case study. In *Proceedings of the Workshop on Speech and Language Processing Tools in Education*, pages 91–98, 2012.

Arthur M Diamond Jr and David M Levy. The metrics of style: Adam smith teaches efficient rhetoric. *Economic Inquiry*, 32(1):138–145, 1994.

Biljana Drndarević, Sanja Štajner, Stefan Bott, Susana Bautista, and Horacio Saggion. Automatic text simplification in Spanish: A comparative evaluation of complementing modules. In *Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part II 14*, pages 488–500. Springer, 2013.

William DuBay. The principles of readability. *Impact Information*, 2004.

Flesch, Rudolph. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221, 1948.

Valerio Franchina and Roberto Vacca. Adaptation of Flesh readability index on a bilingual text written by the same author both in Italian and English languages. *Linguaggi*, 3:47–49, 1986.

Thomas François and Eleni Miltsakaki. Do NLP and machine learning improve traditional readability formulas? In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 49–57, 2012.

Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Haritz Salaberri. Simple or complex? Assessing the readability of Basque texts. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 334–344, 2014.

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2): 193–202, 2004.

Robert Gunning. *The technique of clear writing*. New York, McGraw-Hill Book Company, 1971.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of COLING 2012*, pages 1063–1080, 2012.

Zahurul Islam, Alexander Mehler, and Rashedur Rahman. Text readability classification of textbooks of a low-resource language. In *Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation*, pages 545–553. Waseda University, 2012.

Nikolay Karpov, Julia Baranova, and Fedor Vitugin. Single-sentence readability prediction in Russian. In *Analysis of Images, Social Networks and Texts: Third International Conference, AIST 2014, Yekaterinburg, Russia, April 10-12, 2014, Revised Selected Papers 3*, pages 91–100. Springer, 2014.

N. I. Kesebeževa, N. X. Kajtmesova, and Z. Ju. Šebzuxova, editors. *Soxranenie i razvitie jazykovogo nasledija v polikul□turnoj obrazovatel□noj srede [Maintenance and development of language heritage in polycultural educational environment*, 2021. Majkop.

Adam Kilgarriff. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):97–133, 2001.

J. Peter Kincaid and Leroy John Delionbach. Validation of the Automated Readability Index: A Follow-Up. *Human Factors: The Journal of Human Factors and Ergonomics Society*, 15:17 – 20, 1973. URL `https://api.semanticscholar.org/CorpusID:62344517`.

J. Peter Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom. Derivation of new readability formula for navy enlisted personnel. *Millington, TN: Navy Research Branch*, 1975.

George R. Klare. Automation of the Flesch "Reading Ease" Readability Formula, With Various Options. *Reading Research Quarterly*, 4:550, 1969. URL `https://api.semanticscholar.org/CorpusID:147693290`.

A. N. Kornev. *Narusheniya chteniya i pisma u detey. Uchebno-metodicheskoe posobie [Reading and writing impairments in children]*. Saint Petersburg: MiM, 1997.

N. K. Krioni, A. D. Nikin, and A. V. Filippova. Avtomatizirovannaja sistema analiza složnosti učebnyx tekstov [Automated system of analysis of scholarly texts]. *Vestnik Ufimskogo gosudarstvennogo aviacionnogo texničeskogo universiteta*, 11:101–107, 2008.

R. V. Kupriyanov, M. I. Solnyshkina, and P. A. Lekhnitskaya. Parametric Taxonomy of Educational Texts. *Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriya 2. Yazykoznanie [Science Journal of Volgograd State University. Linguistics]*, 22(6):80–94, 2023.

Maija Lanka and Ludis Pēks. Flesch Reading Ease Score as an Indicator for Selecting textbooks in Physics. In *Rural Environment Education Personality. Proceedings of the International Scientific Conference*, pages 227–234, 2013.

A. N. Laposhina and M. Yu. Lebedeva. Tekstometr: Onlain-instrument opredeleniya urovnya slozhnosti teksta po russkomu yazyku kak inostrannomu [Textometer: an Online Tool for Determining the Level of Complexity of a Text in Russian as a Foreign Language]. *Rusistika [Russian Language Studies]*, 2021.

Ion Madrazo Azpiazu and Maria Soledad Pera. Is cross-lingual readability assessment possible? *Journal of the Association for Information Science and Technology*, 71(6):644–656, 2020.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179, 2021.

G. Harry McLaughlin. SMOG Grading–a New Readability Formula. *Journal of Reading*, 12(8):639–646, 1969.

George Moroz. Adyghe syllable structure: From empirical data to generalizations. *Voprosy Jazykoznanija*, (2):82–95, 2019.

Babak Naderi, Salar Mohtaj, Kaspar Ensikat, and Sebastian Möller. Subjective assessment of text complexity: A dataset for German language. *arXiv preprint arXiv:1904.07733*, 2019.

Irina V. Oborneva. Avtomatizacija ocenki kačestva vosprijatija teksta [Automated assessment of text readability]. *Vestnik Moskovskogo gorodskogo pedagogičeskogo universiteta. Serija: Informatika i informatizacija obrazovanija*, (5):86–91, 2005a.

Irina V. Oborneva. Matematičeskaja model' ocenki učebnyx tekstov [Mathematical model of scholarly texts' evaluation]. *Vestnik Moskovskogo gorodskogo pedagogičeskogo universiteta. Serija: Informatika i informatizacija obrazovanija*, (4):152–158, 2005b.

Irina V. Oborneva. *Avtomatizirovannaya otsenka slozhnosti uchebnykh tekstov na osnove statisticheskikh parametrov [Automated assessment of the complexity of educational texts based on statistical parameters]*. PhD thesis, Moscow, June 2006.

E. A. Parešina. Razrabotka i standartizacija metodiki ocenki skorosti čtenija vslux i ponimanija pročitannogo u russkojazyčnyx mladšix škol'nikov [Development and standartization of methods for evaluating reading speed and comprehension in russian-speaking early grade students]. In *Sbornik tezisov VI Vserossijskoj naučnoj studenčeskoj konferencii NIU VŠÈ Nižnij Novgorod*, pages 195–197, 2022.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

Janice C. Redish. Understanding the limitations of readability formulas. *IEEE Transactions on Professional Communication*, PC-24(1):46–48, 1981.

Robert Reynolds. Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 289–300, 2016.

Sylvia Rottensteiner. Structure, function and readability of new textbooks in relation to comprehension. *Procedia-Social and Behavioral Sciences*, 2(2):3892–3898, 2010.

Matthew Shneyderman, Grace E Snow, Ruth Davis, Simon Best, and Lee M. Akst. Readability of online materials related to vocal cord leukoplakia. *OTO open*, 5(3):2473974X211032644, 2021.

Matthew Shneyderman, Ruth Davis, Grace Snow, Shumon Dhar, and Lee M. Akst. Zenker's diverticulum: readability and quality of online written education materials. *Dysphagia*, 37(6):1461–1467, 2022.

Johannes Sibeko. Developing Resources for Measuring Text Readability in Sesotho. In *CLARIN Annual Conference*, pages 120–132, 2023.

Seth Spaulding. A Spanish readability formula. *The Modern Language Journal*, 40(8):433–441, 1956.

Matthew Sydes and James Hartley. A thorn in the Flesch: Observations on the unreliability of computer-based readability formulae. *British Journal of Educational Technology*, 28(2):143–145, 1997.

Robin Tillman and Ludvig Hagberg. Readability algorithms compability on multiple languages, 2014.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. Text Readability Assessment for Second Language Learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA, June 2016. Association for Computational Linguistics. URL `https://aclanthology.org/W16-0502/`.

Адыгэ псалъэ №39. Адыгэ псалъэ/ АП №39 (24.789). `https://smikbr.ru/arhiv/2025/ap/04/02.pdf`, 2025. Accessed: 2018-04-08.

Ф. Н. Апиш, Б. Х. Лямова, and С. Р. Мешлок. *ЖъогъошIэт: адыгабзэм изэгъэшIэнкIэ кIIэлэцIыкIу IыгъыпIэхэм апае учебнэ-методическэ комплексым ихрестоматие / изэхэгъэуцон дэлэжьагъэхэр*. Качествэр, 2017.

Р. Г. Мамий, М. Н. Хачемизова, and Н. А. Хамерзокова. *Адыгэ литературэр: КIэлэегъаджэхэм апае IэпыIэгъу тхылъ*. Качество, 2011.

Towards the task of factuality assessment

Elena A. Suleymanova
A. K. Aylamazyan Program Systems Institute of RAS,
Pereslavl-Zalessky, Russia
yes2helen@gmail.com

Seda R. Momot
A. K. Aylamazyan Program Systems Institute of RAS,
Pereslavl-Zalessky, Russia
seda.egikian@gmail.com

Natalia A. Vlasova
A. K. Aylamazyan Program Systems Institute of RAS,
Pereslavl-Zalessky, Russia
nathalie.vlassova@gmail.com

Ilya N. Vozdvizhensky
A. K. Aylamazyan Program Systems Institute of RAS,
Pereslavl-Zalessky, Russia
vozdvin@yandex.ru

Abstract

Factuality concerns the extent to which the propositional content of a sentence conforms with the real world, according to the speaker. Factuality assessment is of practical interest for those NLP applications that have to do with textual information analysis: information extraction, information retrieval, text summarization, question-answering systems.

Existing practical approaches address the problem of factuality assessment using the "quantitative" measures of certainty and probability and thus have limited possibilities for annotating real texts. There is a need for a more appropriate model of factuality that could serve as the basis of the annotation scheme.

We suggest a model of factuality that makes use of speaker's cognitive attitudes as one of the parameters for discriminating between degrees of factuality. We also present a dataset that was manually annotated with factuality values in terms of the model.

Keywords: Keywords: factuality assessment; factuality model; annotating factuality in text; dataset annotated with factuality

DOI: 10.28995/2075-7182-2025-23-1110-1122

К задаче оценки субъективной достоверности

Сулейманова Е. А.
Институт программных систем им. А. К. Айламазяна РАН
Переславль-Залесский, Россия
yes2helen@gmail.com

Момот С. Р.
Институт программных систем им. А. К. Айламазяна РАН
Переславль-Залесский, Россия
seda.egikian@gmail.com

Власова Н. А.
Институт программных систем им. А. К. Айламазяна РАН
Переславль-Залесский, Россия
nathalie.vlassova@gmail.com

Воздвиженский И. Н.
Институт программных систем им. А. К. Айламазяна РАН
Переславль-Залесский, Россия
vozdvin@yandex.ru

Аннотация

Субъективная достоверность, или авторская квалификация сообщаемого по линии соответствия действительности, представляет практический интерес для тех приложений автоматической обработки текста, которые связаны с содержательным анализом информации: извлечение информации из неструктурированных источников, информационный поиск, аннотирование, реферирование, вопросно-ответные системы.

Существующие практические подходы к оценке субъективной достоверности оперируют исключительно «количественными» признаками уверенности и вероятности, что ограничивает возможности их применения для аннотирования реальных текстов. Актуальной остается задача построения более адекватной модели субъективной достоверности, которая могла бы лечь в основу аннотационной схемы.

В статье представлена модель субъективной достоверности, использующая в качестве одного из оснований для дифференциации значений противопоставление когнитивных установок говорящего. В терминах модели размечен датасет.

Ключевые слова: оценка субъективной достоверности; модель субъективной достоверности; аннотирование субъективной достоверности в тексте; датасет с разметкой субъективной достоверности

1 Введение

Субъективная достоверность (далее «СД») — аспект текста, отвечающий за квалификацию говорящим сообщаемого с точки зрения соответствия действительности.

Возможности современных нейросетевых технологий позволяют предположить, что решение задачи компьютерного анализа текста с т. зр. СД может быть в значительной степени сведено к подготовке релевантных задаче обучающих данных. Для аннотирования таких данных необходима теоретическая модель, которая, на наш взгляд, должна (1) охватывать всё разнообразие языковых средств, образующих функционально-семантическое поле СД, и (2) отвечать языковой интуиции.

Задача, подобная аннотированию СД, возникла за рубежом в русле работ по извлечению событий и известна под названием «оценка фактуальности[1]» (подробнее см. в разделе 2). В большинстве исследований используется шкала «certain-probable-possible», предложенная R. Saurí и J. Pustejovsky в 2008-2009 годах для корпуса FactBank [Saurí 2009]. Пытаясь применить эту шкалу на практике, мы регулярно сталкивались с одной и той же проблемой — отсутствием подходящего значения[2]. Идеи отечественных лингвистов (раздел 3) укрепили нас в мнении, что для описания значений СД «количественных» признаков уверенности и вероятности недостаточно. Рассмотрим пары примеров:

(1а) ***Я считаю**, что это сделал он.*
(1б) ***Я считаю**, что это следовало сделать.*

(2а) ***Мне кажется**, он пришел за объяснением.*
(2б) ***Мне кажется**, пришло время об этом поговорить.*

(3а) ***Едва ли** его за это похвалили.*
(3б) ***Едва ли** это заслуживает похвалы.*

(4а) ***Я не думаю**, что руководству известны причины произошедшего.*
(4б) ***Я не думаю**, что об этом лучше молчать.*

(5а) *Он* ***точно*** *дома*. [Я видел свет в окне] Ср. также *Он явно/определенно дома.*
(5б) *Он* ***точно*** *дома!* [Не сомневайся, я только что от него] Ср. **Он явно/определенно дома.*

(6а) [Я не был на собрании] ***Наверное/вероятно**, выбрали Петрова.* [Он самая подходящая кандидатура] Ср. **Вроде/кажется, выбрали Петрова*.
(6б) [Да, я был на собрании.] ***Вроде/кажется**, выбрали Петрова.* Ср. **Наверное/вероятно, выбрали Петрова*).

Сразу заметим, что степень «уверенности», выраженная показателями СД (они выделены жирным шрифтом), в каждой паре «а» и «б», очевидно, одинакова.

[1] Английский термин factuality следует, по-видимому, признать самым близким аналогом термина «субъективная достоверность».

[2] Из-за этого обстоятельства мы вынуждены были отказаться от сравнения нашего подхода с «мейнстримом» на датасете: непонятно, каким образом при подсчете коэффициента согласия учесть единодушное «затрудняюсь».

Во все парах пример «а» описывает случай, когда действительное положение дел (имеет ли место ситуация, выраженная подчеркнутым фрагментом) говорящему не известно и он высказывает предположение об этом, основываясь на косвенных данных. Во всех парах пример «а» допускает естественную интерпретацию в терминах вероятности, чего нельзя сказать о примерах «б».

В примерах (1б)-(4б) подчеркнутый фрагмент обозначает не ситуацию реального мира, а собственный взгляд говорящего, его субъективную точку зрения. В этих случаях показатель СД выражает степень уверенности, категоричности мнения; усмотреть в них вероятностную оценку либо нельзя, либо можно с очень большой натяжкой.

В примере (5б) говорящий доподлинно знает то, о чем сообщает. Наречие «точно» в данном случае не является маркером СД, это дискурсивный маркер настойчивой утвердительности (в ответ на сомнения собеседника; заметим, что примеры (5а) и (5б) различаются и интонационно).

Разница между (6а) и (6б) не так очевидна. Она в том, что в (6а) говорящий не знает, а только строит догадки. А в (6б) нет ничего похожего на предположение — это знание, которое говорящий осторожно квалифицирует как не совсем надежное. Различие в значениях СД между (6а) и (6б) подтверждается невозможностью взаимной замены в них показателей СД.

В следующем разделе мы попытаемся описать приведенные примеры в терминах подходов state-of-the-art.

Подход, который предлагается в настоящей статье, основывается на следующих положениях.

1. Значения СД могут иметь различную природу. Это относится как к случаям подчеркнутой достоверности (присутствие уверенности), так и к случаям проблематичной достоверности (отсутствие уверенности). И в тех, и в других случаях значение может быть вероятностным, а может и не быть таковым.
2. Природа (качественный тип) значения СД определяется взаимодействием трех факторов:
   – выбор говорящим маркера СД для оформления пропозиционального содержания,
   – характер самой пропозиции,
   – контекст (в т. ч. прагматический — условия речевого акта, в широком смысле).
3. Маркеры СД часто неоднозначны: один и тот же маркер способен обслуживать разные качественные типы значений СД. Для разрешения неоднозначности маркера важно принимать во внимание характер пропозиции и контекст.

Целью настоящей работы было проверить работоспособность подхода в целом. Построенная для этого модель в ее текущей версии имеет ряд ограничений:

а) рассматривается квалификация пропозиции только одним субъектом — говорящим[3];

б) рассматриваются только случаи, в которых оценка СД выражена лексическими средствами;

в) принятые по некоторым сложным и спорным случаям решения не могут считаться окончательными.

Модель была испытана на 510 реальных текстовых примерах с последующим подсчетом коэффициента согласия аннотаторов (каппа Флейса). Результаты разметки примеров в соответствии с предложенной моделью оформлены как общедоступный датасет.

2 Субъективная достоверность в компьютерной лингвистике

В компьютерной лингвистике анализом явлений, в той или иной степени имеющих отношение к СД, занимаются несколько направлений: оценка фактуальности (factuality assessment), аннотирование мнений (belief annotation and tagging), оценка достоверности/уверенности (certainty evaluation), анализ субъективности (subjectivity analysis) и установки говорящего (speaker's attitude), выявление субъективности (subjectivity detection). Эти исследования можно условно разделить на две группы.

Первая группа представлена работами по оценке фактуальности [Saurí, Pustejovsky 2012], [Lee 2015], [Lima 2020], [Gupta 2022], [Li 2024], [Rovera 2025]. Для этого направления интерес представляет то, как «событие» (event) соотносится с действительностью, с точки зрения субъекта. Оценка (значение фактуальности) представляет собой комбинацию признака «полярность» («+»,

[3] Случаям передачи чужой речи (чужой пропозиции) посвящено отдельное исследование.

«–») с некоторым количественным признаком *certainty* (достоверность, она же уверенность). Достоверные/уверенные случаи (*certain, CT*) противопоставляются недостоверным (*uncertain*), для оценки которых используются разные степени вероятности. В классической версии их две — *probable* (*PR*), и *possible* (*PS*). Иногда вместо этой шкалы используется непрерывная числовая шкала [–3;+3] [Lee 2015]. Примечательно, что к *certain* авторы концепции относят как имплицитно достоверные случаи (простые утверждения, презумпции, импликации), так и случаи с лексическими маркерами уверенности (*certain*, *sure*) и модальными показателями эпистемической необходимости (*must*, *have to*). Иными словами, «событие» ‘он дома’ в примерах *Он дома*; *Они (не) знают, что он дома*; *Я уверен, что он дома* и *Должно быть, он дома* получает одинаковое значение фактуальности (*CT*+)[4].

Что касается второй группы исследований, то в центре их внимания — субъективность как таковая. Это либо разделение всех пропозиций (предикатов, утверждений) на уверенное и неуверенное мнение (убеждение) — *committed belief* и *non-committed belief* [Diab 2009], [Prabkharan 2015], [García 2020], либо разделение высказываний на «субъективные» и «объективные» [Antici 2024]. Вопросы соотнесенности самой пропозиции с действительностью в подобных работах не затрагиваются, что не позволяет отнести используемые ими модели к полноценным моделям СД.

Оценка примеров из раздела «Введение» на шкале фактуальности и шкале оценки мнений выглядит следующим образом (вопросом помечены значения, которые приписаны за неимением более подходящих):

Пример	**Значение по шкале [Saurí 2009]**	**Значение по шкале [Diab 2009]**
1а	CT+*?*	committed belief
1б	CT+	committed belief
2а	PR+	non-committed belief
2б	PR+*?*	non-committed belief
3а	PR-	non-committed belief
3б	PR-*?*	non-committed belief
4а	PR-	non-committed belief
4б	PR-*?*	non-committed belief
5а	CT+	committed belief
5б	CT+*?*	committed belief
6а	PR+	non-committed belief
6б	PR+*?*	non-committed belief

Таблица 1: Разметка примеров по шкале фактуальности и шкале мнений (belief)

Легко видеть, что примеры «а» и «б» во всех парах получают одинаковые значения.

3 Теоретические предпосылки для альтернативного подхода к оценке субъективной достоверности

В отечественных теоретических работах по языкознанию можно найти много ценных наблюдений на интересующую нас тему.

[4] Значение фактуальности всегда привязано к «источнику» (source) — в данном случае это author.

[Яковлева 1994] считает, что при классификации показателей достоверности, кроме количественных признаков, нужно учитывать качество информации: характерная, полученная субъектом непосредственно (*кажется*, *как будто*, *вроде*) и нехарактерная, полученная на основе логического вывода (*наверно*, *видимо*).

[Булыгина, Шмелёв 1997] выделяют на этом основании два типа показателей недостоверности: показатели гипотетичности (у говорящего нет информации об истинности, он высказывает предположение, основываясь на логическом выводе или интуитивной догадке) и квазиассертивы (говорящий имеет непосредственный доступ к информации, но по каким-то причинам не совсем уверен в ее достоверности).

[Шмелёва 1984] усматривает в выражении семантики СД взаимодействие модусных категорий персуазивности (квалификация сообщаемого с т. зр. соответствия действительности) и авторизации (квалификация с т. зр. источников или способов получения информации).

[Дмитровская 1988], [Зализняк 1992] в сфере «знание-мнение» различают три типа установок: знание о верифицируемой пропозиции (знание), мнение о верифицируемой пропозиции (мнение-предположение) и мнение о неверифицируемой пропозиции (мнение-оценка). Мнение-оценка, как и знание, включает истинностную оценку подчиненной пропозиции, тогда как мнение-предположение — вероятностную. Мнение-оценку можно назвать «субъективным знанием», поскольку она имеет для говорящего статус субъективной истины.

[Разлогова 2005] для описания семантики вводных модальных слов предлагает иерархическую систему, в которой истинностные значения «Истина», «Ложь», «Неопределенность» распределяются по «логико-когнитивным состояниям» (они же «степени достоверности»): «Знание», «Уверенность», «Предположение».

4 Описание предлагаемой модели субъективной достоверности

4.1 Пропозиция как объект оценки

Пропозициональное содержание высказывания, или пропозиция, — это та часть смысла высказывания, которая «является естественным аргументом модальных операторов и предикатов пропозициональной установки» [Падучева 1985]. Сама пропозиция, выступающая объектом оценки СД, не содержится в предложении в явном виде.

Представление пропозиционального содержания

Чтобы сделать пропозицию объектом анализа, нужно ее каким-то образом «материализовать». Однозначного ответа на вопрос о том, как это сделать, не существует. Мы выбрали способ представления, который кажется довольно наглядным и при этом наиболее подходящим для рассуждений о СД: пропозиция представляется как значение придаточного предложения с союзом «то, что» (ср. у Куайна [Куайн 2000] пропозиция — это «абстрактный объект, мыслимый как то, что обозначается» ЧТО-придаточным (that-clause); о фактообразующем «то-что»-значении см. также [Арутюнова 1988]). Представленную в таком виде пропозицию можно будет естественным образом рассматривать как аргумент параметра, оценивающего ее с т. зр. СД.

Например:

(7) *Невероятно, чтобы <u>метеоритный дождь шел несколько тысячелетий</u>* — пропозиция, которая будет оцениваться с т. зр. СД, — ‘то, что метеоритный дождь шел несколько тысячелетий’

Отрицание в пропозиции

Чтобы не смешивать на этапе оценки СД отрицательную полярность пропозиции (наличие в ней отрицания) и ее ложность, отрицание считаем элементом пропозициональной структуры [Падучева 1985].

(8) *По-моему, <u>твоя песня не понравилась Алле</u>* — оценивается пропозиция ‘то, что твоя песня не понравилась Алле’

Модализованные пропозиции
Модальные слова считаем частью модализованной пропозиции. Исключение составляют операторы эпистемической возможности и необходимости, которые мы рассматриваем как эпистемические маркеры, ср. (9) и (10).

(9) *Поэтому не считаю, что <u>такие льготы должны быть одинаковыми во всех субъектах</u>* — оценивается пропозиция ‘то, что такие льготы должны быть одинаковыми во всех субъектах’, *должны* — показатель деонтической необходимости.

(10) *<u>Это</u> могло <u>быть связано с длительностью процедур госзакупок</u>* — оценивается пропозиция ‘то, что это было связано с длительностью процедур госзакупок’, *могло* — маркер эпистемической возможности.

Далее в примерах вместо восстановления оцениваемой пропозиции в явном виде ограничимся подчеркиванием содержащей ее части предложения.

Ограничение на способ выражения СД
В настоящей работе рассматриваются только те пропозиции, которые находятся в сфере действия лексически выраженных показателей СД.

Типы пропозиций
В интересах решаемой задачи релевантным представляется рассматривать пропозиции в плоскости «верифицируемость — неверифицируемость (оценочность)»:

- **Верифицируемые**. «Соотносятся с настоящим, прошлым или будущим положением дел в мире и в принципе могут быть верифицированы в момент произнесения соответствующего высказывания» [Дмитровская 1988]. Это прежде всего «событийные» пропозиции по [Шмелёва 1994] — они «портретируют» действительность;
- **Скорее верифицируемые**. «Логические» — «представляют результаты умственных операций и сообщают о некоторых установленных признаках, свойствах, отношениях» [Шмелёва 1994]: «анкетная» характеризация, идентификация, классификация (таксономическая идентификация), релятивные пропозиции;
- **Скорее неверифицируемые**. Рациональные (мотивированные) оценки — например, морально-этические;
- **Неверифицируемые**. Представляют свойства, оценку которых нельзя опровергнуть (хотя ее можно оспорить), — сенсорно-вкусовые, интеллектуальные, эмоциональные, эстетические (об оценочных суждениях см. [Дмитровская 1988]). Эти оценки не требуют мотивировки, поскольку «им не могут быть сопоставлены некоторые качества оцениваемого объекта [...] Оценка прямо проистекает из того ощущения, которое, независимо от воли и контроля, испытывает человек» [Арутюнова 1982].

4.2 Оцениваемый параметр — эпистемический статус

Для квалификации пропозиции с т. зр. СД мы ввели параметр *эпистемический статус* (ЭС). ЭС всегда привязан к субъекту. В настоящей работе, как уже говорилось, рассматриваются только те случаи, в которых ЭС отражает квалификацию пропозиции говорящим[5].

Качественным ядром ЭС выступает истинностное значение («истина» или «ложь»), отнесенное к одной из трех эпистемических категорий (когнитивных установок): не-мнение («знание») и два вида мнения — «полагание/оценка» и «гипотеза» (Рисунок 1). «Истина» и «ложь» в каждой категории приобретают специфический смысл, который отражается в их условных названиях: в категории «знание» мы имеем дело с «”объективной”[6] истиной» и «”объективной” ложью», в

[5] В случае передачи чужой речи один и тот же маркер может выражать эпистемические установки говорящего и другого субъекта. В таких случаях мы сейчас ограничиваемся установкой говорящего (см. пример с *соврал* в таблице 2). Примеры типа «я (тогда) думал, что Р», «я соврал, что Р», в которых говорящий-в-настоящем отделяет себя от субъекта речемыслительного акта (себя же в прошлом), приравниваются к случаям передачи чужой речи.
[6] «Объективный» взято в кавычки, поскольку означает здесь всего лишь «не имеющий признаков субъективного мнения».

категории «полагание/оценка» — с «субъективной истиной» и «субъективной ложью» и в категории «гипотеза» — с «гипотетической истиной» и «гипотетической ложью».

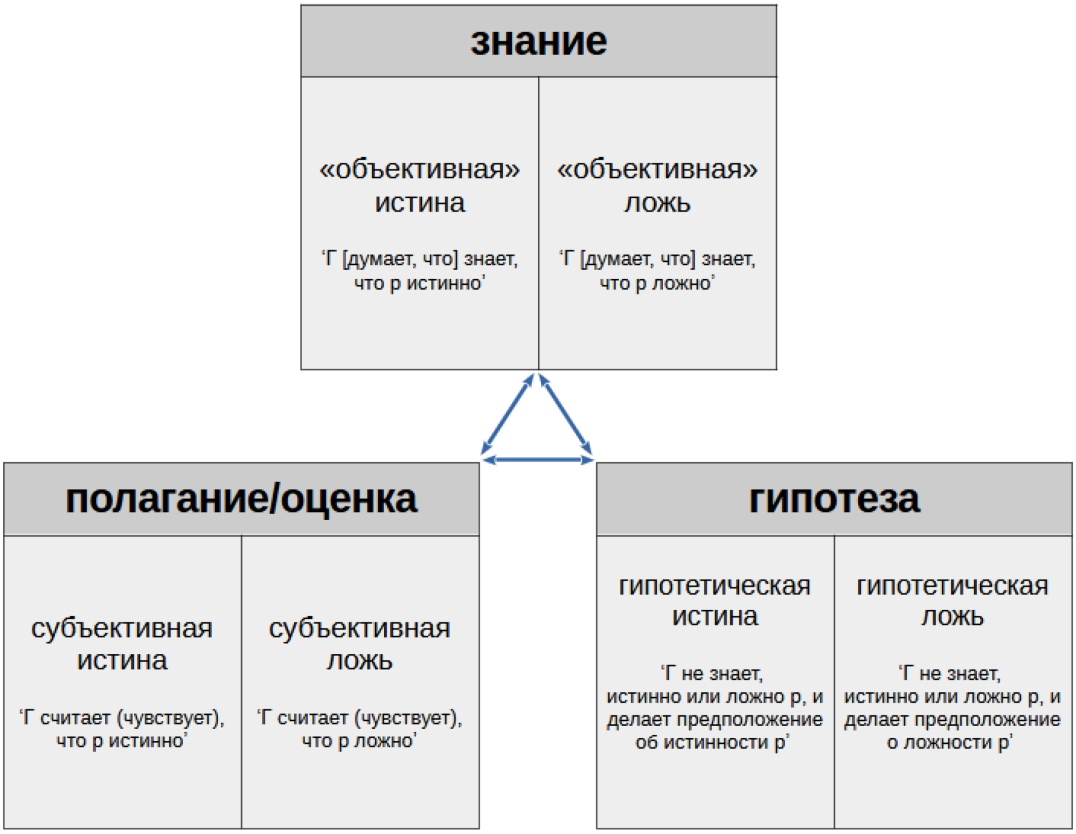

Рисунок 1: Когнитивные установки говорящего и соответствующие им истинностные значения

Количественные индексы

Гипотетические истина и ложь всегда имеют при себе индекс вероятности — одно из значений шкалы: «1», «весьма вероятно», «вероятно», «возможно».

Кроме того, истинностное значение в любой категории может быть охарактеризовано по признаку с обобщенной количественной семантикой (частные проявления которой — категоричность, уверенность, полнота, надежность) и значениями «сильн.», «слаб.».

4.3 Значения эпистемического статуса и принципы их присваивания

Категория «знание»: «“объективная” истина», «“объективная” ложь»

Главное, что отличает значения ЭС в этой категории, — отсутствие признаков собственного мнения говорящего. Информацию о положении дел, обозначенном пропозицией Р, говорящий получил, если можно так выразиться, «в готовом виде» — посредством чувственного восприятия (зрение, слух, обоняние), из собственного опыта, памяти, от других лиц или из сторонних источников и т. п. (примеры в таблице 2). Сама пропозиция обычно верифицируема, но не обязательно (ср., напр., *Я слышал, что это вкусно*). ЭС «сильная истина» или «сильная ложь» в категории «знание» пропозиция получает, если говорящий снабдил ее явным указанием на соответствие или несоответствие действительному положению дел. ЭС «слабая истина» или «слабая ложь» — если говорящий допускает, что его знания неполны или не совсем надежны (ошибка восприятия, несовершенство памяти и т. п.). Что касается нейтральных «объективных» истины и лжи, то они обычно выражаются имплицитно, однако их можно усмотреть и в некоторых лексически маркированных случаях.

«объективная» ИСТИНА Г знает (думает, что знает), что Р истинно	**количеств. признак (полнота / категоричность)**	**«объективная» ЛОЖЬ** Г знает (думает, что знает), что Р ложно
***Не отрицаю**, бывают моменты, когда фонари на улицах не горят из-за поломок.*	**нейтральная**	*Флавио не моргнув глазом **соврал**[7], что АЛОНСО самостоятельно провернул переговоры с МАКЛАРЕНОМ*
*Во-вторых, **насколько я понял**, люди, у которых врач заподозрил наличие зависимости, должны будут сдать анализ за свой счет и пройти тест.*	**слабая**	***Не слышал, чтобы** у перевозчиков возникали сложности с выплатой налогов**.***
*Когда **окончательно подтвердилось**, что выживших нет, это был не просто день официального национального траура – это было горе всей страны.*	**сильная**	***Неправда**, что русские генералы мечтали о броске к Атлантике, Средиземноморью, Ла-Маншу.*

Таблица 2: Примеры «объективных» ЭС пропозиции

Категория «полагание/оценка»: «субъективная истина», «субъективная ложь»
Пропозиция Р соотносится не с реальной действительностью, а с точкой зрения говорящего, с его собственной квалификацией некоторой ситуации или объекта. При этом говорящий не строит вероятностные догадки об истинности или ложности Р (как было бы в случае гипотезы, см. далее), а высказывает свое мнение об этом с большей или меньшей уверенностью, категоричностью. Пропозиция, имеющая статус «субъективная истина» или «субъективная ложь», не верифицируема, ее нельзя просто подтвердить или опровергнуть. Но с ней можно согласиться или не согласиться. Примеры — в таблице 3.

[7] Говорящий знает, что это не так.

субъективная ИСТИНА Г считает (чувствует), что Р истинно	**количеств. признак (уверенность / категоричность)**	**субъективная ЛОЖЬ** Г считает (чувствует), что Р ложно
***Полагаю**, что все ФАПы должны быть в прямой коммуникации и системе консультаций с краевыми*	**нейтральная**	***Я не согласен** с тем, что Дании уже пора открывать свои границы со Швецией, так как ситуация с коронавирусом в Швеции все еще не находится под контролем*
***Мне кажется**, ничего страшного не происходит*	**слабая**	***Не думаю**, что это нарушение прав человека.*
***Я уверен**, что с предателем надо поступать, как с изменником Родины.*	**сильная**	*"Я **решительно возражаю** против того, что это было соразмерное применение силы в той ситуации", - сказал Аррадондо.*

Таблица 3: Примеры субъективных ЭС пропозиции

Категория «гипотеза»: «гипотетическая истина», «гипотетическая ложь»
Говорящему не известно истинное положение дел, обозначенное пропозицией Р (или не может быть известно — судить об этом нам позволяет дискурсивный контекст, в т. ч. здравый смысл и другие фоновые знания). Говорящий может лишь высказать предположение об истинности или ложности Р, основанное на логическом выводе из имеющихся в его распоряжении данных или на интуитивной догадке (ср. определение гипотезы в [Булыгина, Шмелёв 1997]). Истинность или ложность Р в этом случае поддается оценке в терминах вероятности. Для количественной оценки вероятности мы используем четырехзначную шкалу: «1», «весьма вероятно», «вероятно», «возможно». Значение «возможно» интерпретируется не как минимальное значение вероятности, а как неопределенное ее значение, отличное от нуля. Исходим из того, что эпистемическая возможность предполагает ненулевую вероятность, а эпистемическая невозможность — нулевую. Для выражения невозможности (нулевой вероятности) «истины» используем «ложь» с вероятностью «1» (таблица 4).

Гипотетическая ИСТИНА	**Вероятность**	**Гипотетическая ЛОЖЬ**
Закручивать гайки для легальных интернет-видеосервисов бессмысленно, потому что нарушения происходят ***явно*** *не на их платформах.*	**«1»**	***Не может такого быть****, чтобы руководство не знало о существовании такой схемы.*
Простите мне пессимизм, но ***я практически уверен****, что судьба московского особняка купца Булошникова на Большой Никитской, 17, решена и никакие слушания (назначены на четверг, 17 января) ее уже не изменят.*	**«весьма вероятно»**	***Очень сомнительно****, что сам момент разрушения фигуры никто не видел, а значит, горожане не остановили порчу декоративной фигуры.*
Судя по всему*, переговорный процесс по конкретным проблемам продолжается.* *В него выстрелил Соколов, настоящее имя которого,* ***предположительно****, Вадим Красиков.* ***Наверное,*** *есть какие-то инвесторы, предприниматели, которые именно такие: построили – продали.*	**«вероятно»**	***Маловероятно****, что люди способны вырабатывать антитела, которые будут гарантировать их иммунитет к коронавирусу в течение долгого времени.* ***Я сомневаюсь****, что у мошенников есть доступ к этим базам.*
Нельзя исключать*, что некая общественная организация собирает информацию для подготовки соответствующей законодательной инициативы.* *Однако* ***существует вероятность****, что на самом деле найдены останки Марии дель Риччо — второй обитательницы монастыря, которая, как и Лиза Герардини, не была монахиней, но также удостоилась особого захоронения из-за своего благородного происхождения.*	**«возможно»**	***Не уверен****, что даже налоговые органы в полной мере осведомлены об их доходах.* ***Нельзя с уверенностью утверждать****, что Эци страдал при жизни от гастрита, ведь слизистая оболочка желудка не сохранилась.* .

Таблица 4: Примеры гипотетических ЭС пропозиции

Примечание. Для случаев, когда говорящий в той или иной форме отказывается брать на себя эпистемическую ответственность за пропозицию, предусмотрено особое значение ЭС — «неизвестно», например:

(11) ***Не берусь утверждать****, что Пригожин причастен и к указанному сайту в сети «ВКонтакте».*

5 Разметка датасета

Пригодность предложенной модели СД оценивалась путем разметки текстового материала, который представляет рассматриваемое явление в его многообразии, с последующим вычислением уровня согласия аннотаторов.

5.1 Принципы разметки субъективной достоверности

Разметке подлежат только лексически маркированные случаи выражения СД, удовлетворяющие следующим условиям:

1. Эксплицитность. СД оцениваемой пропозиции должна быть выражена явными лексическими средствами — словами и конструкциями, которые содержат в своем значении эпистемический компонент[8]. (В [Зализняк 1992] эпистемический компонент определяется как 'X знает, что P', 'X считает, что P', 'X считает, что P вероятно', 'X считает, что P возможно').

2. Перволичность. Субъектом эпистемической установки, выраженной маркером, является говорящий. Установки других субъектов — в случаях передачи чужой речи — сейчас не рассматриваются.

5.2 Датасет

С помощью описанного метода был размечен датасет из 510 текстовых фрагментов объемом одно-два предложения. В качестве источника использованы общий и газетные подкорпуса НКРЯ (https://ruscorpora.ru/).

Основной целью при отборе примеров был максимальный охват способов явного выражения СД. (Поскольку количественное распределение показателей в реальных текстах никак не учитывалось, о репрезентативности датасета мы говорить не можем.) Особое внимание уделялось потенциально неоднозначным показателям. Для каждого показателя (представителя группы синонимичных показателей) по возможности отбирались примеры, содержащие разные типы пропозиций.

Эпистемические показатели в плане выражения были разнообразны:

- предикаты внутреннего состояния, у которых эпистемический компонент является частью толкования (*я считаю…, думаю, что…, сомневаюсь, что…*) и синонимичные им конструкции (*есть сильные сомнения, что…*);
- предикаты восприятия (*я не видел, чтобы…*);
- модальные глаголы со значением эпистемической возможности и необходимости (*может, должен*);
- вводные слова (*возможно, по-моему, кажется*);
- наречия со значением оценки вероятности (*наверняка, маловероятно*);
- частицы (*едва ли, якобы*);
- эксплицитное указание на истинность или ложность ассоциированной пропозиции (*это правда, что…, информация о…. фейк*), а также частичный или полный отказ говорящего от эпистемической ответственности (*не берусь утверждать…*);
- указание на источник информации или на отсутствие такового (*доказано, что…, нет данных о том, что…*)

и др.

Разметка производилась в таблице, в столбцах которой аннотаторы для каждого примера записывали значения атрибутов ЭС (истинностное значение, установка, количественные индексы).

Датасет размещен в открытом доступе: https://github.com/VozdvIN/AIReC-EpistemicDataSet/releases

5.3 Оценка уровня согласия аннотаторов

Разметка датасета выполнялась тремя аннотаторами, имеющими лингвистическое образование. Для них была подготовлена подробная инструкция с содержательным описанием значений СД и показательными примерами.

Для оценки уровня согласия между аннотаторами был использован показатель Каппа Флейса [Fleiss 1971], далее KF. KF выражает степень отличия фактического распределения мнений аннотаторов относительно полностью случайного выбора значений.

Этот показатель рассчитывался дважды: для сокращенной и для полной версии разметки.

[8] Вопросы имплицитного выражения СД требуют отдельного рассмотрения.

При расчете KF для сокращенной разметки оценивалось согласие для пар «истинностное значение, категория» (например, «истина-знание», «ложь-полагание» и др., а также внекатегориальный статус «неизвестно»). «Количественные» признаки не учитывались. В этом случае KF=0,93.

При расчете KF для полной разметки оценивалось согласие по значению ЭС целиком, с учетом «количественных» показателей. Например, «ложь-полагание (нейтральная)» и «ложь-полагание слабая» в этом случае считались за несовпадающие значения. Для этого варианта KF=0,86.

В обоих случаях значение KF интерпретируется как «очень высокая степень согласия» [Fleiss 1981].

Типичные случаи расхождения аннотаторов в выборе категории имеют место, когда категориальная неоднозначность показателя соединяется с неочевидным характером пропозиции. Пример сложного выбора между субъективным и гипотетическим статусом:

(12) ***Я убежден*** *в том, что у России достаточно средств для того, чтобы не уступать в конкуренции.*

Пример размечен как «сильная истина-полагание/оценка» и «гипотетическая истина с вероятностью 1». Неясно, то ли автор-говорящий знаком с реальным положением дел (в данном случае — знает, какими средствами располагает Россия) и дает этому положению дел свою субъективную оценку (этих средств достаточно, чтобы…). То ли положение дел автору не известно, но он предполагает что оно таково, что он бы его оценил как достаточное для того, чтобы…). Вторая интерпретация сродни тому, что [Дмитровская 1988] называет оценкой-предположением («Я думаю, что книга интересная»).

Аналогичный пример — размечен как «гипотетическая ложь (вероятна)» и как «слабая ложь-полагание/оценка»:

(13) ***Мне не кажется****, что это станет серьезным подспорьем в пополнении бюджета* [Заголовок публикации: «Эксперт оценил пользу для бюджета от легализации продажи «красивых» номеров»]

Пока не решено, как лучше поступать в случаях, когда однозначно установить категорию значения не удается. Однако мы в любом случае исходим из того, что модель языкового явления должна давать возможность различать то, что в языке регулярно различается.

6 Заключение

В статье представлена модель субъективной достоверности, в которой для различения случаев маркированной достоверности предлагается выявлять когнитивную установку говорящего.

Модель была испытана при аннотировании датасета из 510 примеров тремя аннотаторами. Несмотря на высокие показатели согласия, в некоторых случаях решение давалось аннотаторам нелегко. Спорные случаи обсуждались, и по результатам обсуждения разметка могла быть частично исправлена.

Задачи на ближайшее будущее:

- систематизировать сложные и спорные случаи;
- рассмотреть возможность дифференциации значений внутри когнитивных установок;
- существенно увеличить объем размеченного материала;
- задействовать большее число аннотаторов.

References

[1] Antici F. et al. A corpus for sentence-level subjectivity detection on english news articles // Proceedings of the LREC-COLING 2024. P. 273–285. https://doi.org/10.48550/arXiv.2305.18034

[2] Diab M. et al. Committed belief annotation and tagging // Proceedings of the Third Linguistic Annotation Workshop (LAW III). 2009. — P. 68–73. https://aclanthology.org/W09-3012.pdf

[3] Fleiss J. L. Measuring nominal scale agreement among many raters // Psychological Bulletin 76 (5). 1971. — P. 378–382.

[4] Fleiss J. L. Statistical methods for rates and proportions. New York: John Wiley & Sons, 1985. — P. 38–46.

[5] Glòria Vázquez García, Ana María Fernández Montraveta. Annotating factuality in the TAGFACT corpus // Multiperspectives in analysis and corpus design. — 2020. P. 115–127.

[6] Ankita Gupta, Su Lin Blodgett, Justin H Gross, Brendan O'Connor. Examining Political Rhetoric with Epistemic Stance Detection // Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), November 7, 2022. — Pp. 89–104.

[7] Lee K. et al. Event detection and factuality assessment with non-expert supervision // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. —2015. — P. 1643–1648. https://yoavartzi.com/pub/lacz-emnlp.2015.pdf

[8] Chunyang Li, Hao Peng, Xiaozhi Wang, Yunjia Qi, Lei Hou, Bin Xu, and Juanzi Li. 2024. MAVEN-FACT: A large-scale event factuality detection dataset // Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA. — Association for Computational Linguistics. — P 11140–11158.

[9] Salvador Lima, Naiara Perez, Montse Cuadros, German Rigau. NUBES: A Corpus of Negation and Uncertainty in Spanish Clinical Texts // Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020). — Marseille, 11–16 May 2020. — P. 5772–5781. https://aclanthology.org/2020.lrec-1.708.pdf

[10] Minard A. L. et al. MEANTIME, the NewsReader multilingual event and time corpus // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). — 2016. — P. 4417–4422. http://www.lrec-conf.org/proceedings/lrec2016/pdf/488_Paper.pdf

[11] Marco Rovera, Serena Cristoforetti, Sara Tonelli. ModaFact: Multi-paradigm Evaluation for Joint Event Modality and Factuality Detection // Proceedings of the 31st International Conference on Computational Linguistics. — January 19–24, 2025. — P. 6378–6396.

[12] Saurí, R., Pustejovsky, J. FactBank: a corpus annotated with event factuality // Lang Resources & Evaluation 43, 227–268 (2009). https://doi.org/10.1007/s10579-009-9089-9

[13] Saurí R., Pustejovsky J. Are you sure that this happened? Assessing the factuality degree of events in text // Computational linguistics. — 2012. — 38(2) — P. 261–299.

[14] Арутюнова Н.Д. Аксиология в механизмах жизни и языка // Проблемы структурной лингвистики. — М.: Наука, 1984.

[15] Арутюнова, Нина Давидовна. Типы языковых значений [Текст] / Н. Д. Арутюнова; Оценка, событие, факт ; Отв. ред. Г. В. Степанов ; АН СССР, Ин-т языкознания. — М.: Наука, 1988.

[16] Булыгина Т. В., Шмелёв А. Д. Языковая концептуализация мира (на материале русской грамматики). — М.: Школа «Языки русской культуры», 1997. — 576 с.

[17] Дмитровская М. А. Знание и мнение: образ мира, образ человека // Логический анализ языка. Знание и мнение. — М.: Наука, 1988. — С. 6–18.

[18] Зализняк Анна А. Исследования по семантике предикатов внутреннего состояния. – Verlag Otto Sagner, München, 1992.

[19] Куайн У. В. О. Слово и объект. — М.: Логос, Праксис, 2000.

[20] Падучева Е.В. Высказывание и его соотнесенность с действительностью (Референциальные аспекты семантики местоимений) — М.: Наука, 1985.

[21] Разлогова Е. Э. Логико-когнитивные и стилистические аспекты семантики модальных слов: автореф. дис. ... д-р филол. наук. — Москва, 2005.

[22] Шмелёва Т.В. Смысловая организация предложения и проблема модальности // Актуальные проблемы русского синтаксиса / Под ред. К.В. Горшковой, Е.В. Клобуковой. — М., 1984. — С. 78–100.

[23] Шмелёва Т. В. Семантический синтаксис: текст лекций из курса «Современный русский язык» / Т. В. Шмелёва. — Красноярск: Красноярский государственный университет, 1994. — 46 с.

[24] Яковлева Е. С. Фрагменты русской языковой картины мира (модели пространства, времени и восприятия) / Е. С. Яковлева. — Москва : Гнозис, 1994. — 344 с. (Язык. Семиотика. Культура).

Interpretable approach to detecting semantic changes based on generated definitions

Tatarinov Maksim
HSE University
Nizhny Novgorod, Russia
tatarinovst0@gmail.com

Demidovsky Aleksandr
HSE University
Nizhny Novgorod, Russia
monadv@yandex.ru

Abstract

This paper investigates definition modeling as an approach to semantic change detection, which offers the advantage of providing human-readable explanations, unlike traditional embedding-based approaches that lack interpretability. Definition modeling leverages large language models to generate dictionary-like definitions based on target words and their contextual usages. Despite its potential, practical evaluations of this method remain scarce. In this study, FRED-T5 was fine-tuned using the Small Academic Dictionary for the task of definition modeling. Both quantitative and qualitative assessments of definition modeling's effectiveness in detecting semantic shifts within the Russian language were conducted. The approach achieved a Spearman's rank correlation coefficient of 0.815 on the Rushifteval task, demonstrating strong alignment with expert annotations and ranking among the leading solutions. For interpretability, a visualization algorithm was proposed that displays semantic changes over time. In the qualitative evaluation, our system successfully replicated manual linguistic analysis of 20 Russian words that had undergone semantic shifts. Analysis of the generated meanings and their temporal frequencies showed that this approach could be valuable for historical linguists and lexicographers.

Keywords: Semantic change, definition modeling, definition generation
DOI: 10.28995/2075-7182-2025-23-1123-1132

Интерпретируемый подход к детектированию семантических изменений слов на основе генерируемых определений

Татаринов Максим
НИУ ВШЭ
Нижний Новгород, Россия
tatarinovst0@gmail.com

Демидовский Александр
НИУ ВШЭ
Нижний Новгород, Россия
monadv@yandex.ru

Аннотация

В данной работе исследуется моделирование определений как подход к обнаружению семантических изменений, который имеет преимущество в виде понятных для человека объяснений, в отличие от традиционных подходов на основе векторных представлений, страдающих от недостатка интерпретируемости. Моделирование определений использует большую языковую модель для генерации словарных определений на основе целевых слов и их контекста. Несмотря на потенциал, практико-ориентированные оценки этого метода остаются ограниченными. В данном исследовании FRED-T5 была дообучена с помощью Малого академического словаря на задаче моделирования определений. Были проведены как количественные, так и качественные оценки эффективности моделирования определений в обнаружении семантических сдвигов в рамках русского языка. Подход достиг коэффициента ранговой корреляции Спирмена 0,815 в задаче Rushifteval, что демонстрирует сильное соответствие экспертным аннотациям, находясь среди лидирующих решений. Для интерпретируемости был предложен алгоритм визуализации, который отображает семантические изменения во времени. В качественной оценке наша система успешно воспроизвела ручной лингвистический анализ 20 русских слов, имевших семантическими сдвиги. Анализ сгенерированных значений и их временных частот показал, что этот подход может быть востребован для исторических лингвистов и лексикографов.

Ключевые слова: Семантические изменения, моделирование определений, генерация определений

1 Introduction

Static and contextual embeddings excel at capturing semantic relationships for detecting semantic change, but lack human-readable word descriptions. Advancements in recent research involve definition generation with language models, which offer more illustrative descriptions (Giulianelli et al., 2023; Fedorova et al., 2024). It could aid historical linguists and lexicographers in creating dictionaries and language history studies, such as Dobrushina and Daniel' (2018). However, the practical evaluation of this approach remains limited.

The primary objective of this study is to assess the effectiveness of language models in detecting semantic changes in words through the generation of definitions. It would use both quantitative metrics from a shared task and qualitative analysis by reproducing a linguistic analysis of words known to have undergone semantic shifts.

The paper is organized as follows: Section 2 reviews semantic change detection methods, evaluation methods for classifying errors in generated definitions and a strategy to acquire correct ones for comparison. Section 3 describes the proposed methodology. Section 4 presents the results and discusses their implications.

2 Related Work

2.1 Approaches to Semantic Change Detection

Semantic change is understood as change in the polysemy of a word over time. Although most solutions provide a quantitative measure of semantic change, such as a score or distance between vectors, to determine the extent of change, recently, a step towards a more explainable approach has been taken (Giulianelli et al., 2023; Fedorova et al., 2024).

There have been multiple approaches to semantic change detection:

Static Embeddings. Static embeddings provide a fixed representation of a word for the entire corpus. In the Shiftry (Kutuzov et al., 2020), Word2Vec (Mikolov et al., 2013) was utilized to examine semantic shifts by dividing the corpus by years to generate distinct word vectors for each period.

They need extensive data for stable representations, fail to differentiate multiple meanings of a word, and independently trained models produce incompatible vector spaces requiring alignment.

Contextual Embeddings. Contextual models such as BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018) generate different embeddings for a word depending on its context. Rachinskiy and Arefyev (2021) fine-tuned the XLM-R model to generate embeddings aligned with dictionary definitions. Arefyev et al. (2021) trained XLM-R on a large multilingual dataset and RuSemShift data.

The GlossReader approach showed limitations with culturally specific words and depended on predefined senses for visualization, while the DeepMistake method lacked visualization capabilities.

Definition Modeling. Definition modeling takes a target word with a usage example to generate a human-readable word definition based on context, akin to a dictionary entry (Giulianelli et al., 2023), unlike previous embedding approaches which produce abstract vector representations that are difficult to interpret.

Table 1: Example of Definition Modeling

Example Usage	He started to sleep poorly at night, waking up with a persistent headache.
Target Word	night
Generated Definition	The part of the day from sunset to sunrise.

Giulianelli et al. (2023) proposed using generated definitions as semantic embeddings for words, enabling semantic change detection. Fedorova et al. (2024) researched definition modeling for the task of semantic change detection finding it successful.

The main limitation of the approaches employing embeddings is their non-interpretability. The best case is the DeepMistake, whose visualization is limited to predetermined senses.

As for definition modeling, qualitative evaluation in Fedorova et al. (2024) is limited, as they leave "in-practice" evaluation for future research. Also, they used an unsupervised approach for an evaluation, while the proposed approach involved fine-tuning the vectorizer.

2.2 Classification of Errors in Generated Definitions

Studies by Huang et al. (2021) and Noraset et al. (2017) have proposed classifications for errors in generated definitions. Their work identified the following types:

Table 2: Types of Errors in Generated Definitions

Type	Russian Example	English Example (Translation)
Over-specification	кофе – горячий, горький напиток из жареных бразильских зерен	coffee – a hot, bitter beverage made from roasted Brazilian beans
Under-specification	капитан – член команды.	captain – team member
Self-referential	самосознание – состояние, при котором у человека присутствует самосознание	self-awareness – a state in which a person has self-awareness
Wrong Part of Speech	стекло – переместиться вниз, сбежать (о жидкости)	glass/spilt – to move down, escape (of a liquid)
Opposite Meaning	внутрь – ненаправленный в центр	inward – non-directed to the center
Close Semantics	машина – устройство с автоматическими функциями	machine – a device with automatic functions
Redundancy or Excessive Use of Generic Phrases	спутник – тот, кто совершает путь, путь вместе с кем-л.	companion – one who makes a journey, journey together with someone
Incorrectness	первый – следующий после всех остальных в списке предметов	first – next after all other items in the list
Correct	винодельня – заведение, помещение для изготовления вина	vineyard – establishment, premises for wine production

2.3 Acquiring correct definitions

Sternin and Rudakova (2017) outlines a method of generalizing dictionary definitions for determining correct semantic description of words, emphasizing the integration of diverse dictionary definitions to capture the full meaning. This procedure involves compiling all available definitions, differentiating meanings based on denotative principles, and synthesizing a unified semantic structure, with the final step organizing meanings from core to peripheral, accompanied by usage examples.

3 Proposed Approach

3.1 Fine-tuning LLM

A generative large language model M is trained on a dataset $D = \{(w_i, c_i, d_i)\}_{i=1}^{N}$, where each tuple contains a word w, its context c, and a corresponding definition d. The model learns to generate an accurate definition $\hat{d} = M(w, c)$ by minimizing the cross-entropy loss between its predicted token probabilities and the reference definitions:

$$L(M) = \sum_{i=1}^{N} \text{loss}(M(w_i, s_i), d_i), \tag{1}$$

3.2 Testing

Intrinsic evaluation is conducted using a test subset D_{test} of the dataset D to assess the quality of generated definitions $\hat{d}_j = M(w_j, c_j)$ compared to reference definitions d_j using string similarity metrics, defined as:

$$\text{metric} = \frac{1}{M} \sum_{j=1}^{P} \text{similarity}(\hat{d}_j, d_j) \tag{2}$$

where similarity measures the match between definitions, ranging from 0 (no similarity) to 1 (identical).

Extrinsic evaluation assesses the model's performance on a semantic change detection task with test set $S = \{(w_k, g_{k,(t_i,t_j)})\}_{k=1}^{Q}$, where w_k represents a target word, $g_{k,(t_i,t_j)}$ its gold semantic change score for the transition between periods t_i and t_j, and Q is the number of words in the test set.

For each word w_k in the test set, a set of usage contexts $U_{k,t} = \{u_{k,t,1}, u_{k,t,2}, \ldots, u_{k,t,n}\}$ is sampled from each time period $t \in \{t_1, t_2, t_3\}$ of the diachronic Russian National Corpus (Savchuk et al., 2024), where n is 100 or all if fewer available, in a similar way to Arefyev et al. (2021). For each period transition, the usages are paired, and definitions $\hat{d}_{k1}, \hat{d}_{k2}$ are generated by the model for each pair.

These definitions are then vectorized $\vec{d}_{k1}, \vec{d}_{k2}$ using a vectorizer V. The distance between the vectorized definitions $dist(\vec{d}_{k1}, \vec{d}_{k2})$ is calculated and converted to scores ranging from 1 (senses unrelated) to 4 (identical).

The mean values of the ratings for each word are compared with the gold scores from the task using Spearman's rank correlation.

3.3 Visualization

To illustrate semantic changes over time, generated definitions are transformed into vector representations using a vectorizer V.

A clustering algorithm C is then applied to group similar definitions.

For each cluster K_j, a prototypical definition $\hat{d}_{\text{proto}}$ is selected, which is defined as original definition whose vector $\vec{d}_{\text{proto}}$ is the closest to the center of the cluster (centroid).

Let $\vec{c_j}$ be the centroid of cluster K_j:

$$\vec{d}_{\text{proto},j} = \arg\min_{\vec{d} \in K_j} dist(\vec{d}, \vec{c_j}) \tag{3}$$

where $dist$ is a distance metric.

Bar charts are then created to display the frequency of different meanings over time.

3.4 Qualitative Analysis

A qualitative assessment begins with the selection of words known to have undergone semantic shifts based on existing linguistic research. Usage examples for these words are obtained from different time periods using a diachronic corpus. The trained model is applied to generate definitions for each word usage. The obtained definitions are compared with information from semantic descriptions of words, written based on Sternin and Rudakova (2017) method of generalizing dictionary definitions, and classified according to the error types in Table 2. Finally, changes in the frequency of meanings over time provided by the visualizations are examined and compared with historical usage data.

4 Results and Discussion

4.1 Model

FRED-T5-1.7B was chosen due to its performance in processing the Russian language (Zmitrovich et al., 2024). At the time of selection, it was the top performer on the RussianSuperGLUE benchmark (Shavrina et al., 2020), with a score of 0.762.

4.2 Training Data

FRED-T5-1.7B was trained on a dataset derived from "Small academic dictionary" (MAS) (Evgenyeva, 1981 1984).

The dataset was cleaned to remove usage labels, entries without usage examples or without informative definitions, such as *Состояние по знач. глаг. линять* [*State by the meaning of the verb "to shed"*],

and those that provided grammatical rather than lexical information, such as *наречие к причастию приглашающий* [*Adverb to the participle "inviting"*]. The resulting dataset of 122,350 entries was partitioned into training, development, and test sets with a 90%/5%/5% split.

Each entry was formatted and began with the word "Контекст" ["Context"] followed by a usage example, then the phrase "Определение слова" ["Word definition"], and the word itself.[1]

4.3 Evaluation Data

The *RuShiftEval* competition's test set (Kutuzov and Pivovarova, 2021) was utilized for evaluation. The task focuses on detecting semantic changes in Russian nouns across three historical transitions: RuShiftEval-1 (Pre-Soviet:Soviet), RuShiftEval-2 (Soviet:Post-Soviet), and RuShiftEval-3 (Pre-Soviet:Post-Soviet). The competition provided a test set of gold change scores for 99 Russian nouns corresponding to the transitions.

4.4 String Similarity Metrics in Model Testing

BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and BERT-F1 (Zhang* et al., 2020) metrics from the *evaluate* library (Hugging Face, 2023) were employed for the definitions generated using the test part of the MAS dataset. BLEU measures n-gram overlap between texts, ROUGE-L focuses on the longest common subsequence, and BERT-F1 leverages contextual embeddings for semantic similarity. The evaluation results[2] are presented in Table 3.

Table 3: Fine-tuning Results of FRED-T5-1.7B on the MAS Dataset

Metric	**Value**
BLEU	11.02
ROUGE-L	29.36
BERT-F1	75.22

Low BLEU and ROUGE-L scores indicate that the model generates definitions differently from the test set, although high BERT-F1 scores imply semantic similarity.

At this stage, self-referential errors were fixed by excluding tokens related to the target word from being sampled in the model's output.

4.5 Rushifteval Testing

The paraphrase-multilingual-mpnet-base-v2 model (Transformers, 2023), additionally fine-tuned on RuSemShift, a similar dataset (Rodina and Kutuzov, 2020), was used to vectorize definitions. The distances between the definitions were calculated using the cosine distance. Results were compared against approaches from the Rushifteval task, as shown in Table 4.

Table 4: Algorithm Results Compared to Rushifteval Teams

Team	**Average**	**Word Representation Type**	**Model Used**
DeepMistake (post-competition)	**0.850**	**Contextual Emb.**	**XLM-R**
Proposed Approach	**0.815**	**Generated Definitions**	**FRED-T5-1.7B**
GlossReader	0.802	Contextual Emb.	XLM-R
DeepMistake	0.791	Contextual Emb.	XLM-R
vanyatko	0.720	Contextual Emb.	RuBERT
Other 10 Teams	0.457-0.178	…	…

[1] A special denoiser token `<LM>`, dedicated to the task of text continuation, was utilized.

[2] Out of 100, higher is better.

The proposed approach outperforms most entries in the Rushifteval competition.

Table 5: Comparison with definition generation approaches

Method	RuShiftEval-1	RuShiftEval-2	RuShiftEval-3	Base Model
Proposed Approach without vectorizer fine-tuning	**0.722**	**0.763**	**0.749**	**FRED-T5-1.7B**
Fedorova et al. (2024)	0.488	0.462	0.504	MT0-XL

A shown in Table 5, the proposed approach significantly outperforms the results of Fedorova et al. (2024). The vectorizer fine-tuning step was omitted to ensure that the results are directly comparable.

It could be noted that Fedorova et al. (2024) appears to retain unhelpful definitions in the training data, unlike proposed approach in 4.2, possibly resulting in their model reproducing non-informative patterns and the lower performance of their approach.

4.6 Visualization

Generated word vectors were clustered using the DBSCAN algorithm. Each cluster is represented by a prototypical definition closest to its centroid. DBSCAN parameters (`eps` and `min_samples`) are manually tuned by incremental adjustment to ensure the formation of cohesive clusters. Then, the temporal distribution of these meanings is displayed using bar charts, as shown in Figure 1.

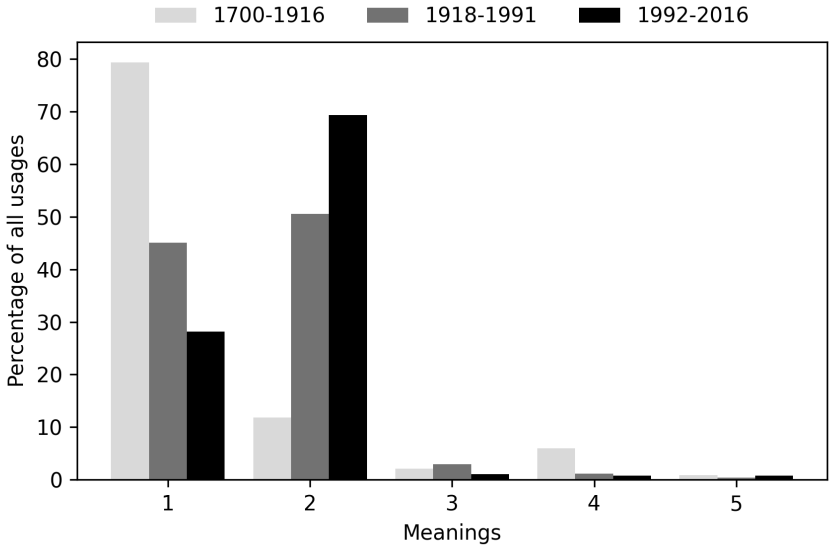

Figure 1: Semantic Shift of the Word *машина [machine/car]* (Parameters: eps=0.14, min_samples=5)

Meanings for *машина [machine/car]*:

1. A device or instrument for a specific task.
2. An automobile or vehicle.
3. An aircraft or helicopter.
4. A mechanically or thoughtlessly acting person.
5. A system of institutions or organizations.

4.7 Qualitative Analysis

For deeper examination, 20 words exhibiting semantic shifts from *Two Centuries in Twenty Words* (Dobrushina and Daniel', 2018) were selected: *знатный [noble]*, *кануть [to disappear]*, *классный [classy/cool]*, *мама [mom]*, *машина [machine/car]*, *молодец [young man/attaboy]*, *пакет*

[bag/package], *передовой [advanced]*, *пионер [pioneer]*, *пожалуй [perhaps]*, *пока [until/bye]*, *привет [hello]*, *пружина [spring]*, *публика [public]*, *свалка [landfill/fight]*, *сволочь [bastard]*, *стиль [style]*, *тётка [aunt]*, *тройка [three/a set of three]*, *червяк [worm]*. The usages were extracted from the diachronic sub-corpus of Russian National Corpus (Savchuk et al., 2024).

For each word, 300 instances were randomly sampled for each period of the corpus (pre-Soviet, Soviet, post-Soviet). The model generated definitions for each occurrence, followed by the creation of corresponding visualizations.

Next, the semantics of each word based on multiple dictionaries were described following Sternin and Rudakova (2017). To ensure comprehensive meaning descriptions, we synthesized information from 3 modern Russian dictionaries: *Big Explanatory Dictionary* (Kuznetsov, 1998), *Dmitriev's Explanatory Dictionary of the Russian Language* (Dmitriev, 2003), and *Ozhegov and Shvedova's Explanatory Dictionary*, in addition to *Two Centuries in Twenty Words*. Usage labels were omitted since the model wasn't trained to generate them.

The manually obtained semantic descriptions were compared with those in the visualization, and changes in their usage across periods for meanings corresponding to those in *Two Centuries in Twenty Words* were analyzed.

4.8 Qualitative Analysis of Generated Definitions

As a result of generalizing dictionary definitions, 121 meanings were compiled for 20 words. A total of 83 definitions were obtained using the proposed approach. Thus, excluding 5 incorrect definitions, 64.4% of the meanings were identified.

Table 6: Types of Definitions and Their Counts

Type of Definition	**Count**	**Percentage**
Correct	57	68.67%
Close	10	12.04%
Incorrect	5	6.02%
Insufficiently Specific	3	3.61%
Redundancy or Excessive Use of General Phrases	4	4.81%
Close, Redundancy or Excessive Use of General Phrases	1	1.20%
Overly Specific	3	3.61%
Self-reference	0	0.00%
Opposite Meaning	0	0.00%
Incorrect Part of Speech	0	0.00%

As shown in Table 6, the majority of definitions are correct without any errors or shortcomings (68.67%).

Common issues include close or incorrect meanings, such as defining *червяк [worm]* as an adult insect or describing *пожалуй [perhaps]* as a conjunction. Redundancy is present, exemplified by the repetitive "chaotic" in the definition of *свалка [landfill/fight]* ('Беспорядочная, беспорядочная схватка'), possibly due to the abundance of synonymous expressions in the training dataset, a common method in lexicology. Additionally, some definitions lack specificity, such as describing *мама [mom]* simply as 'a tender address to a woman.' These problems may arise from the model's limited world knowledge.

Another issue is insufficient context, leading to ambiguity in distinguishing meanings, as seen with *пионер [pioneer]* in *Pioneers listen to this and admire it [Пионеры слушают это и восхищаются]*.

4.9 Statistical Analysis of Semantic Shifts

For most of the words, the visualizations partially or fully align with the data from *Two Centuries in Twenty Words*, except for the word *пока [until/bye]*, where the visualization results contradict the study's

findings. Overall, main meaning changes consistent with the book's data were identified in 12 out of 20 words. Additionally, changes partially aligned in 4 other words.

One of the best visualizations was created for the word *пакет [bag/package]*. 7 definitions were identified correctly, 4 of which appear only in the post-Soviet period.

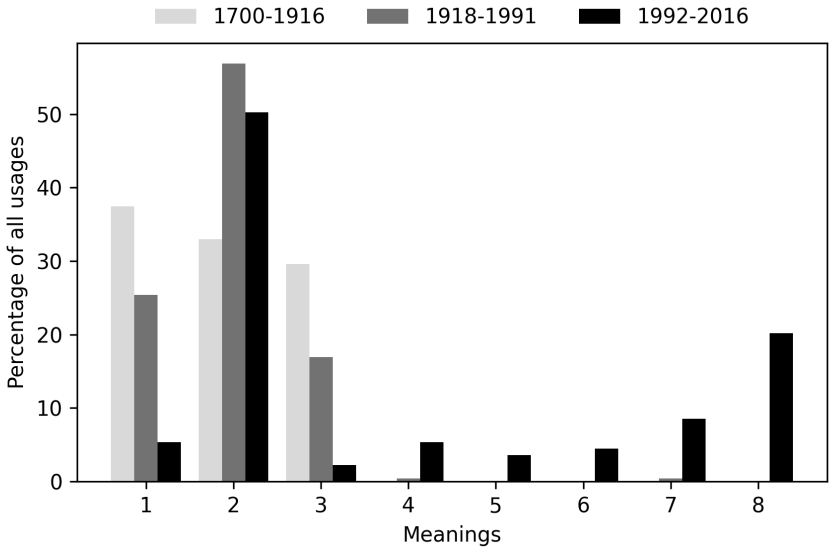

Figure 2: Semantic Shift of the Word *пакет [bag/package]* (Parameters: eps=0.11, min_samples=8)

Meanings for *пакет [bag/package]*:

1. A letter, parcel, etc., in such a form.
2. A paper or fabric pouch for storing, transporting, etc.
3. A letter, parcel, etc., sealed in such an envelope.
4. A collection of homogeneous, related objects, phenomena, etc.
5. A collection of software tools united by a certain criterion.
6. A part of something belonging to someone under certain conditions. (marked as incorrect)
7. A collection of homogeneous objects, documents, etc.
8. A collection of shares of a joint-stock company.

A comprehensive analysis is not feasible for *публика [public]* and *кануть [to disappear]*, because *Two Centuries in Twenty Words* does not provide sufficient usage frequency diagrams for their meanings.

Similarly, for *сволочь [bastard]*, only 2 out of 4 meanings were detected by the proposed approach (*употребляется как бранное слово [used as a swear word]* and *о подлом, гнусном человеке [referring to a vile, despicable person]*), both falling under 'Индивидуальное оскорбление [Individual insult]' in the book.

Conclusion

The study demonstrated the effectiveness of definition modeling in detecting and visualizing semantic shifts in the Russian language. A FRED-T5-1.7B model, fine-tuned on the MAS dictionary, was used to generate context-based word definitions. The model demonstrated high BERTScore similarity metrics on the test set, performed among the top solutions on the Rushifteval shared task and outperformed the results of Fedorova et al. (2024). A visualization algorithm was developed to represent semantic changes over time, allowing for reproducing a manual effort of studying semantic changes for a set of 20 words. Qualitative analysis of the results revealed that 68.67% of generated definitions were fully correct, with main meaning changes accurately detected in 12 out of 18 words available for analysis and partial alignment in 4 others. This shows that the approach could aid historical linguists and lexicographers in linguistic studies.

The findings can be applied to assess the extent of semantic shifts in lexemes, providing visualizations and definitions for each identified meaning.

Future research directions might include incorporating multiple dictionaries as training data or utilizing more advanced LLMs.

The code for this project and the model are available on GitHub: `https://github.com/tatarinovst2/work-definition-modeling`

References

Nikolay Arefyev, Maksim Fedoseev, Vitaly Protasov, Alexander Panchenko, Daniil Homskiy, and Adis Davletov. 2021. Deepmistake: Which senses are hard to distinguish for a word-in-context model. // *Computational Linguistics and Intellectual Technologies*, volume 20, P 16–30, 06.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. // Jill Burstein, Christy Doran, and Thamar Solorio, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, P 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

D.V. Dmitriev. 2003. *Tolkovy slovar' russkogo yazyka: Ok. 2000 slovar. st., svyshe 12000 znacheniy [Explanatory Dictionary of the Russian Language: About 2000 Dictionary Entries, Over 12000 Meanings]*. Slovari Akademii Rossiyskoy. Astrel' [i dr.], Moskva. GUP IPK Ulyan. Dom pechati.

N.R. Dobrushina and M.A. Daniel'. 2018. *Dva veka v dvadtsati slovakh [Two Centuries in Twenty Words]*. Izdatel'skiy dom Vysshey shkoly ekonomiki, Moskva, 2 edition.

A.P. Evgenyeva. 1981-1984. *Slovar' russkogo yazyka: V 4-kh t. [Dictionary of the Russian Language: In 4 Volumes]*. Russkiy yazyk, Moskva, 4-e izd., ispr. i dop [4th ed., corrected and supplemented] edition. V 4-kh tomakh [In 4 volumes].

Mariia Fedorova, Andrey Kutuzov, and Yves Scherrer. 2024. Definition generation for lexical semantic change detection. // Lun-Wei Ku, Andre Martins, and Vivek Srikumar, *Findings of the Association for Computational Linguistics: ACL 2024*, P 5712–5724, Bangkok, Thailand, August. Association for Computational Linguistics.

Mario Giulianelli, Iris Luden, Raquel Fernandez, and Andrey Kutuzov. 2023. Interpretable word sense representations via definition generation: The case of semantic change analysis. // Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 3130–3148, Toronto, Canada, July. Association for Computational Linguistics.

Han Huang, Tomoyuki Kajiwara, and Yuki Arase. 2021. Definition modelling for appropriate specificity. // Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, P 2499–2509, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Hugging Face. 2023. Evaluate. `https://github.com/huggingface/evaluate`. Retrieved November 15, 2023.

Andrey Kutuzov and Lidia Pivovarova. 2021. Rushifteval: a shared task on semantic shift detection for russian. // *Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue*, P 533–545.

Andrei Kutuzov, V. Fomin, V. Mikhailov, and Julia Rodina. 2020. Shiftry: Web service for diachronic analysis of russian news. // *Computational Linguistics and Intellectual Technologies*, volume 19, P 500–516, 01.

S.A. Kuznetsov. 1998. *Bol'shoy tolkovy slovar' russkogo yazyka: A-Ya [Large Explanatory Dictionary of the Russian Language: A-Ya]*. Norint, SPb. RAN. Inst. lingv. issled. Sost., gl. red. kand. filol. nauk S.A. Kuznetsov.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. // *Text Summarization Branches Out*, P 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. // *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: learning to define word embeddings in natural language. // *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, P 3259–3266. AAAI Press.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. // Pierre Isabelle, Eugene Charniak, and Dekang Lin, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. // Marilyn Walker, Heng Ji, and Amanda Stent, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, P 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Maxim Rachinskiy and Nikolay Arefyev. 2021. Zero-shot cross-lingual transfer of a gloss language model for semantic change detection. // *Computational Linguistics and Intellectual Technologies*, volume 20, P 578–586, 06.

Julia Rodina and Andrey Kutuzov. 2020. RuSemShift: a dataset of historical lexical semantic change in Russian. // Donia Scott, Nuria Bel, and Chengqing Zong, *Proceedings of the 28th International Conference on Computational Linguistics*, P 1037–1047, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

S.O. Savchuk, T.A. Arkhangelskiy, A.A. Bonch-Osmolovskaya, O.V. Donina, Yu.N. Kuznetsova, O.N. Lyashevskaya, B.V. Orekhov, and M.V. Podryadchikova. 2024. Natsionalny korpus russkogo yazyka 2.0: novye vozmozhnosti i perspektivy razvitiya. *Voprosy Yazykoznaniya*, 2:7–34.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. // Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, P 1–23, Barcelona (online), December. International Committee for Computational Linguistics.

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. RussianSuperGLUE: A russian language understanding evaluation benchmark. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

I.A. Sternin and A.V. Rudakova. 2017. *Slovarnye definicii i semanticheskiy analiz [Dictionary Definitions and Semantic Analysis]*. Istoki, Voronezh.

Sentence Transformers. 2023. paraphrase-multilingual-mpnet-base-v2. `https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2`. Retrieved April 19, 2024.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. // *International Conference on Learning Representations*.

Dmitry Zmitrovich, Aleksandr Abramov, Andrey Kalmykov, Vitaly Kadulin, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergei S. Markov, Vladislav Mikhailov, and Alena Fenogenova. 2024. A family of pretrained transformer language models for Russian. // Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, P 507–524, Torino, Italia, May. ELRA and ICCL.

Abstracts

RUSSIAN NATIONAL CORPUS 2.0: CORPUS PLATFORM, ANALYSIS TOOLS, NEURAL NETWORK MODELS OF DATA MARKUP (FULL VERSION)

Bonch-Osmolovskaya A. A., Vinogradov Russian Language Institute of the Russian Academy of Sciences, **Gladilin S. A.**, IITP (Kharkevich Institute), FRC CSC, **Kozerenko A. D.**, Vinogradov Russian Language Institute of the Russian Academy of Sciences, **Lyashevskaya O. N.**, HSE University, **Morozov D. A.**, NSU, **Kuznetzova Y. N.**, MSU, Institute of Linguistics of the Russian Academy of Sciences, **Makhova A. A.**, Vinogradov Russian Language Institute of the Russian Academy of Sciences, **Piskounova S. V.**, Independent Researcher, **Bujlova N. N.**, Lopukhin Federal Research And Clinical Center of Physical-chemical Medicine of Federal Medical Biological Agency, **Borodina D. G.**, St. Petersburg State University, **Vinogradova I. I.**, Prosveshchenie Publishers, **Sizov V. G.**, **Dyachenko P. V.**, **Kazennikov A. O.**, IITP (Kharkevich Institute), **Vlasova N. A.**, A.K. Ailamazyan Institute of Program Systems of the Russian Academy of Sciences, **Glazkova A. V.**, University of Tyumen, **Stolyarov S. S.**, **Garipov T. A.**, **Smal I. A.**, NSU, **Gubar'kova Ya. N.**, Yandex

The Russian National Corpus has existed for over 20 years and is a unique linguistic tool. However, the technical limitations of the software platform on which it was implemented significantly narrowed its development prospects. In 2020, work was launched on a comprehensive update of the RNC software platform, as a result of which the National Corpus switched to a new generation 2.0 platform. The implemented deep changes concerned both the development of functionality that meets modern approaches to corpus linguistics, and a fundamental restructuring of the platform architecture as a whole, from data preparation and indexing systems to the user interface. A separate area of development of the capabilities of the RNC was associated with the implementation of neural network models used for metadata tagging, disambiguation, word-formation markup, etc.

This article provides a short description of the new corpus platform as of 2024. The description includes key pa-rameters of changes in the architecture of the RNC platform and its user interface, descriptions of new corpus data analysis services and the specifics of their implementation, as well as a description of the experience of using neural network models for tasks related to corpus data markup.

The purpose of the article is to describe the technological layer of changes implemented in the National Corpus of the Russian Language as part of a large-scale update carried out in recent years.

LOGICAL STRESS AND GESTURE SYNONYMY IN THE CEPHALIC CHANNEL

Evdokimova A. A., Institute of Linguistics, Russian Academy of Sciences, Moscow, Russia

In this article, based on the material of the Russian-language reference subcorpus RUPEX and the Spanish-language monologue subcorpus CAFE (comunicación de los artistas flamencos españoles) annotated in ELAN, we analyze head gestures that mark logical stress and consider all the variations that occur in this position. When determining the position of coincidence of a gesture and stress, we took into account the principle described by Grishina about the anticipation or delay of gestures depending on the strategy chosen by the speaker. In 20% of cases, the gesture anticipates the logical stress, starting a little earlier, and in 13%, on the contrary, it occurs immediately after the stressed syllable, highlighted by the accent of the word. According to our research, some of these head gestures are synonymous with each other in other positions as well. Some gestures were chosen by the subjects according to the characteristic features of their cephalic portrait (for example, moving the head forward with a sideways tilt), while others turned out to be typical that highlight significant words in a monologue and are characteristic of both Russian and Spanish (for example, the pragmatic gesture Down). When compared with the data of the MUMIN corpus research group and the authors of the Spanish-language corpus collected from "spontaneous" speeches on talent shows, principles were developed for describing gestures in the cephalic channel in marked positions from the point of view of different functional approaches, taking into account the following factors: the influence of other kinetic channels, the presence/absence of a listener, and the superposition of gesture functions on top of each other. In the Russian language, in 73% of cases, the opposite movement is adjacent to or behind the gesture, which indicates the visualization of “emphatic tone curvature” or different types of “skid”. In the remaining cases, an intensification of the gesture is observed by anticipating it with the same one, but of smaller amplitude. Testing the annotation principles on two corpora showed their effectiveness as a basis for developing automatic head gesture annotation.

BACKTRANSLATION INVARIANCE BOOSTS EFFECTIVENESS OF NON-ENGLISH PROMPTS

Kurtukova A.[1, 2], **Kozachenko A.**[1], [1]NTR Labs, Tomsk, Russia; [2]Tomsk State University of Control Systems and Radioelectronics, Tomsk, Russia

We present an approach to improving non-English prompts based on backtranslation invariance (the semantics of the prompt should not change after automatic translation to English and back). It improves prompts in non-English languages for a variety of Large Language Models (LLMs), including GPT-4-o, Llama-3.1, and Mixtral8x7B. We evaluate the approach for Russian and Finnish languages. In the benchmark of removing commas from a sentence, the proposed approach achieved an accuracy increase of 42% for Russian and 54% for Finnish compared to non-invariant prompts (LLaMA). In the benchmark of counting commas, accuracy increase of 19% for Russian and 11% for Finnish (GPT).

THE METHODOLOGY OF MULTI-CRITERIA EVALUATION OF TEXT MARKUP MODELS BASED ON INCONSISTENT EXPERT MARKUP

Levikin A.[1], **Khabutdinov I.**[2], **Grabovoy A.**[2], **Vorontsov K.**[3], [1]MSU, Moscow, Russia; [2]Antiplagiat Company, Moscow, Russia; [3]MSU Institute for Artificial Intelligence, Moscow, Russia

A wide class of natural language processing tasks is solved using markup. At the moment, the vast majority of models and datasets rely on a simple markup structure containing only fragments and labels. Moreover, simple classification metrics such as F1, Precision, Recall are used to evaluate the model’s accuracy. The problem with such metrics is that they do not take into account all aspects of the markup structure and that they are applicable only under the assumption of the existence of an ideal markup. This paper proposes a more general and universal markup structure that allows solving complex problems and builds a methodology for multi-criteria evaluation of text markup models based on inconsistent expert markup. After that, the application of the constructed

method is considered to assess the quality of the model obtained within the winning algorithm of the "READ//ABLE" competition, which focused on building an effective essay markup system. The results demonstrate that the new markup structure and evaluation approach provides a more comprehensive and accurate assessment of model performance, addressing the limitations of traditional metrics by accounting for complex markup scenarios and expert inconsistencies.

SYSTEM OF INTERJECTION REACTIONS TO MAINTAIN COMMUNICATION BY A COMPANION ROBOT

Luria A. S., Lobachevsky State University of Nizhni Novgorod, Russia, **Kotov A. A.**, Kurchatov Institute National Research Center, Russian State University for the Humanities, Moscow, Russia

The article presents a classification of interjections developed to describe a wide range of interjection reactions in a multimodal corpus, as well as aimed to be applied by companion robots that communicate through speech and gestures. As part of the study, an experiment was conducted to test part of the developed classification. In the experiment, the F-2 robot used interjections (and accompanying gestures) from the developed classification, as well as, for comparison, automatically synthesized interjections. The accuracy of human recognition of the illocutionary force of interjections was assessed. The conducted experiment showed that the level of understanding of the interjection reactions developed within this work is higher than the level of understanding of automated reactions. The results of the experiment confirmed the effectiveness of the developed classification of interjections in the framework of communication between a robot and a user.

PRECEDENT TEXTS OF THE CORPUS "ONE SPEECH DAY" AND COMIC PASSE-PARTOUT OF EVERYDAY COMMUNICATION

Peresypkina X. A., **Bogdanova-Beglarian N. V.**, SPbSU, Saint Petersburg, Russia

One of the phenomena of modern communication should be recognized as "comic passe-partouts" (CP)—special linguacultural units present in the mental lexicon of native speakers, actively functioning in colloquial speech and illustrating both the frequency of use of ready-made units and all kinds of constructions peculiar to the speakers and the regularity of spontaneous speech creation. CPs arise on the basis of precedent texts, reflect their inherent ability to modifications, have, as a rule, a structure of construction and are realized in speech (oral and oral-written) in many variants, cf.: *"Slovo patsana. Krov' na asfal'te": Slovo brevna. Shchepki na asfal'te ulitsy Rusa; Slovo kotana. Sherst' na divane; "Vostok – delo tonkoe": Dizajn — delo tonkoe, dazhe esli rech' pro prostoj ulichnyj znak, vyvesku ili nadpis'; Mda… zagorat' na plyazhe s napil'nikom v ruke — delo tonkoe.* The phenomenon of CP is an example of the "unlocking" of the semantics of the precedent text and the functioning of the "construction" created on its basis in a conditionally infinite number of contexts. The existence of this new speech phenomenon cannot be ignored in any description of modern communication, especially in various applied aspects of linguistics, as well as in the creation of automatic systems for processing natural speech or artificial intelligence.

READABILITY ASSESSMENT OF WRITTEN ADYGHE USING A BASELINE APPROACH

Petrunina U.[1], **Zdorova N.**[1, 2], [1]Center for Language and Brain, HSE University, Moscow, Russia; [2]Institute of Linguistics RAS, Moscow, Russia

The study introduces a cross-linguistic approach extending the English-based Flesch Reading Ease formula for the assessment of Adyghe texts' readability level. The method relies on the corpus-based analysis of Adyghe shallow linguistic features, i.e. syllable length, word count, and sentence length. It allows to adjust the Flesch formula in accordance with these features by means of natural language processing (NLP) and corpus data analysis. Preliminary results showed that the adapted formula could overall adequately differentiate texts according to their complexity levels although it lacked precision in distinguishing between texts belonging to the same complexity range. The approach can be easily extended to other typologically different minority languages subject to their corpora size and availability.

TOWARDS THE TASK OF FACTUALITY ASSESSMENT

Suleymanova E. A., **Momot S. R.**, **Vlasova N. A.**, **Vozdvizhensky I. N.**, A. K. Aylamazyan Program Systems Institute of RAS, Pereslavl-Zalessky, Russia

Factuality concerns the extent to which the propositional content of a sentence conforms with the real world, according to the speaker. Factuality assessment is of practical interest for those NLP applications that have to do with textual information analysis: information extraction, information retrieval, text summarization, question-answering systems.

Existing practical approaches address the problem of factuality assessment using the "quantitative" measures of certainty and probability and thus have limited possibilities for annotating real texts. There is a need for a more appropriate model of factuality that could serve as the basis of the annotation scheme.

We suggest a model of factuality that makes use of speaker's cognitive attitudes as one of the parameters for discriminating between degrees of factuality. We also present a dataset that was manually annotated with factuality values in terms of the model.

INTERPRETABLE APPROACH TO DETECTING SEMANTIC CHANGES BASED ON GENERATED DEFINITIONS

Tatarinov M., **Demidovsky A.**, HSE University, Nizhny Novgorod, Russia

This paper investigates definition modeling as an approach to semantic change detection, which offers the advantage of providing human-readable explanations, unlike traditional embedding-based approaches that lack interpretability. Definition modeling leverages large language models to generate dictionary-like definitions based on target words and their contextual usages. Despite its potential, practical evaluations of this method remain scarce. In this study, FRED-T5 was fine-tuned using the Small Academic Dictionary for the task of definition modeling. Both quantitative and qualitative assessments of definition modeling's effectiveness in detecting semantic shifts within the Russian language were conducted. The approach achieved a Spearman's rank correlation coefficient of 0.815 on the Rushifteval task, demonstrating strong alignment with expert annotations and ranking among the leading solutions. For interpretability, a visualization algorithm was proposed that displays semantic changes over time. In the qualitative evaluation, our system successfully replicated manual linguistic analysis of 20 Russian words that had undergone semantic shifts. Analysis of the generated meanings and their temporal frequencies showed that this approach could be valuable for historical linguists and lexicographers.

Авторский указатель

Author Index

Научное издание

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной
международной конференции «Диалог»

Выпуск 23, 2025
Дополнительный том

Ответственный за выпуск **А. В. Ульянова**
Вёрстка **К. А. Климентовский**