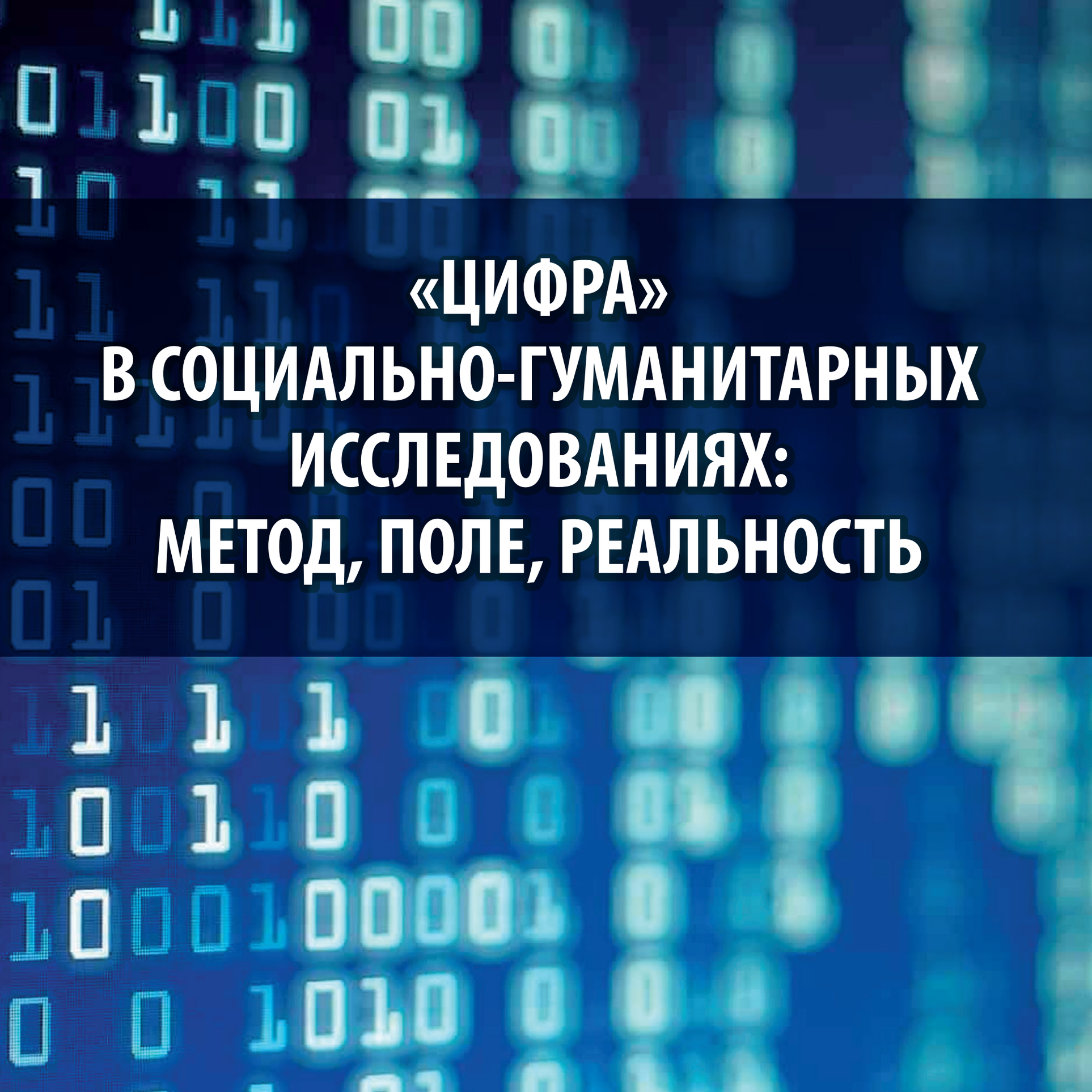


«ЦИФРА» В СОЦИАЛЬНО-ГУМАНИТАРНЫХ ИССЛЕДОВАНИЯХ: МЕТОД, ПОЛЕ, РЕАЛЬНОСТЬ

УДК 004:316
ББК 16.2:60.5
Ц75

Публикуется по решению ученого совета ИФИЯМ ИГУ

Редакционная коллегия:
канд. филол. наук, доцент *С. Н. Гафарова,*
канд. филол. наук, доцент *О. Л. Михалёва,*
канд. филол. наук, доцент *М. Б. Ташлыкова,*
ст. преп. *У. Э. Чекмез*

«Цифра» в социально-гуманитарных исследованиях: метод, поле, реальность : материалы научной конференции молодых ученых. Иркутск, 17–21 сентября 2024 г. / [редкол.: С. Н. Гафарова [и др.]]. – Иркутск : Издательство ИГУ, 2025. – 1 электронный оптический диск (CD-ROM). – Заглавие с этикетки диска.

ISBN 978-5-9624-2372-2

В материалах отражены вопросы, связанные с ролью и местом цифровых технологий в современной гуманитаристике, прежде всего лингвистике и социальной антропологии. В центре внимания – проблемы автоматической обработки текста, исследование новых форм реализации социальных интеракций, влияние информационных технологий на традиционные методы анализа текстовых массивов, изучение роли исследователя при работе с большими данными.

Предназначено для лингвистов, социологов, психологов, антропологов, литературоведов, культурологов и других специалистов в разных областях гуманитарного знания.

Федеральное государственное бюджетное образовательное учреждение высшего образования
«Иркутский государственный университет»
664003, г. Иркутск, ул. К. Маркса, 1; тел. +7 (3952) 51-19-00
Издательство ИГУ, 664074, г. Иркутск, ул. Лермонтова, 124
тел. +7 (3952) 52-18-53; e-mail: izdat@lawinstitut.ru
Подписано к использованию 17.04.2025. Тираж 13 экз. Объем 6,4 Мб.

Тип компьютера, процессор, частота:	32-разрядный процессор, 1 ГГц или выше
Оперативная память (RAM):	256 МБ
Необходимо на винчестере:	320 МБ
Операционные системы:	ОС Microsoft® Windows® XP, 7, 8 или 8.1. ОС Mac OS X
Видеосистема:	Разрешение экрана 1024x768
Акустическая система:	Не требуется
Дополнительное оборудование:	Не требуется
Дополнительные программные средства:	Adobe Reader 6 или выше

© ФГБОУ ВО «ИГУ», 2025

«ЦИФРА» В СОЦИАЛЬНО-ГУМАНИТАРНЫХ ИССЛЕДОВАНИЯХ: МЕТОД, ПОЛЕ, РЕАЛЬНОСТЬ

Материалы научной конференции молодых ученых
Иркутск, 17–21 сентября 2024 г.

ISBN 978-5-9624-2372-2

«ЦИФРА» В СОЦИАЛЬНО-ГУМАНИТАРНЫХ ИССЛЕДОВАНИЯХ: МЕТОД, ПОЛЕ, РЕАЛЬНОСТЬ

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное
учреждение высшего образования
«Иркутский государственный университет»
Институт филологии, иностранных языков и медиакоммуникации

«ЦИФРА» В СОЦИАЛЬНО-ГУМАНИТАРНЫХ ИССЛЕДОВАНИЯХ: МЕТОД, ПОЛЕ, РЕАЛЬНОСТЬ

Материалы научной конференции молодых ученых
Иркутск, 17–21 сентября 2024 г.

ИРКУТСК
ИЗДАТЕЛЬСТВО
ИГУ
2025

УДК 004:316
ББК 16.2:60.5
Ц75

Публикуется по решению ученого совета ИФИЯМ ИГУ

Редакционная коллегия:
канд. филол. наук, доцент *С. Н. Гафарова*,
канд. филол. наук, доцент *О. Л. Михалёва*,
канд. филол. наук, доцент *М. Б. Ташлыкова*,
ст. преп. *У. Э. Чекмез*

Ц75 «**Цифра**» в социально-гуманитарных исследованиях: метод, поле, реальность : материалы научной конференции молодых ученых. Иркутск, 17–21 сентября 2024 г. / [редкол.: С. Н. Гафарова [и др.]]. – Иркутск : Издательство ИГУ, 2025. – 1 электронный оптический диск (CD-ROM). – Заглавие с этикетки диска.

ISBN 978-5-9624-2372-2

В материалах отражены вопросы, связанные с ролью и местом цифровых технологий в современной гуманитаристике, прежде всего лингвистике и социальной антропологии. В центре внимания – проблемы автоматической обработки текста, исследование новых форм реализации социальных интеракций, влияние информационных технологий на традиционные методы анализа текстовых массивов, изучение роли исследователя при работе с большими данными.

Предназначено для лингвистов, социологов, психологов, антропологов, литературоведов, культурологов и других специалистов в разных областях гуманитарного знания.

УДК 004:316
ББК 16.2:60.5

ISBN 978-5-9624-2372-2 © ФГБОУ ВО «ИГУ», 2025

СОДЕРЖАНИЕ

УДК 81`33

А. В. Алхазова, Е. П. Дорофеева, К. К. Лесняк

МГУ им. Ломоносова, Москва, Россия

Динамическое описание гласных терского кумыкского

Аннотация. Описываются гласные в терском диалекте кумыкского языка (< кыпчакские < тюркские < алтайские). В рамках исследования осуществляется разметка, обработка, нормализация, применение различных метрик и сравнение классификаций с опорой на каждую из метрик. Сопоставляется два способа нормализации данных, оптимальным для целей исследования признается модифицированный способ нормализации; предлагается новая метрика TLroc, которая задействует длительность гласного и позволяет охарактеризовать количество частотных изменений во времени. Использование длительности в качестве одного из параметров позволяет улучшить качество как статического, так и динамического классификаторов. Анализ данных показывает, что динамическое описание, которое базируется на двух и более спектральных срезах и их взаимодействии, является более точным методом, по сравнению со статическим описанием, основывающимся на одном срезе.

Ключевые слова: акустическая фонетика, динамическое описание, статическое описание, тюркские языки.

A. V. Alkhazova, E. P. Dorofeeva, K. K. Lesniak

Lomonosov Moscow State University, Moscow, Russia

Dynamic Description of Vowels in Terek Kumyk

Annotation. The study exploits the dynamic approach to describe vowel inventory of the Terek dialect of the Kumyk language (< Kypchak < Turkic < Altaic). Dynamic description is based on two or more spectral slices and their interaction and is a more accurate method than static description based on a single spectral slice. The study is organized as follows: tagging, processing, normalization, application of different metrics and comparison of classifications based on each metric. The tagging followed by segmentation of larger audio files was carried out in Praat software. In order to obtain the most plausible formant values, each sound file was processed using the FastTrack plugin. Data normalization was performed to remove the influence of the individual characteristics of the speaker's speech apparatus on the data. From the list of normalization methods presented in, the method described in and modified in the former article was chosen. Also preprocessing included the elimination of the coarticulatory influence exerted by the left and right consonantal contexts. Inherent change in the formant values of vowels can be characterized from two positions: how and to what extent the formant changes. To give a description from the second position, there are a number of metrics, described in, measuring the distance between formant values at different time points. The resulting values serve as additional metrics for vowel classification, which is the final step of our study.

Keywords: acoustic phonetics, dynamic description, static description, turkic languages.

Введение

Терский диалект кумыкского языка принадлежит к кыпчакской группе тюркской языковой семьи и является одной из разновидностей кумыкского языка. Этот идиом распространен в основном на территории Республики Северная Осетия − Алания и в некоторых частях Ставропольского края России, где проживает кумыкское население. Данные для нашего исследования были

собраны в с. Предгорное Моздокского района (Республика Северная Осетия − Алания) в ходе экспедиций летом 2023–2024 гг.

Проблема. Инвентарь гласных фонем терского диалекта кумыкского языка состоит из восьми фонем, дифференциальными признаками которых являются ряд, подъем и огубленность. Он совпадает с литературным и может быть представлен в виде табл. 1 [Современный кумыкский язык, 2014].

Таблица 1

Гласные фонемы кумыкского языка

Подъем	**Ряд**			
	[+front]		[-front]	
	[-round]	[+round]	[-round]	[+round]
[+high]	i (*и*)	ü (*уь*)	ɨ (*ы*)	u (*у*)
[-high]	e (*е*)	ö (*оь*)	a (*а*)	o (*о*)

График на рис. 1 является статическим представлением гласных терского кумыкского. Он был построен для датасета, где каждому гласному сегменту соответствовал один узкополосный спектр, расположенный ровно посередине временного отрезка длительности гласного. Каждому гласному на графике соответствует овал, значения внутри которого отличаются от среднего значения меньше, чем на одно стандартное отклонение. Само среднее значение представлено точкой в центре овала.

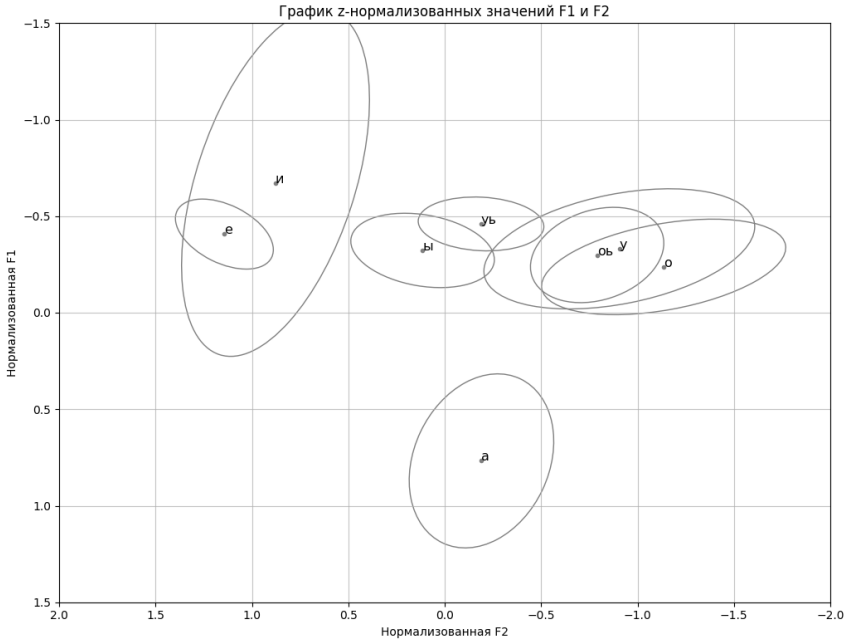


Рис. 1. Статическое представление гласных терского кумыкского (график с доверительными эллипсами в одно стандартное отклонение) для всех носителей

Как можно заметить, эллипсы на рис. 1 накладываются друг на друга. Такие пересечения слишком велики (вплоть до того, что некоторые классы почти полностью включают в себя другие, как в случае с реализациями /оь/ и /у/), и внутри них

находится значительная часть сегментов. Некоторые единичные спектральные срезы могут быть взяты из реализаций разных фонем в надежной позиции (т. е. в первом слоге слова, где с гласным не происходит никаких изменений типа делабиализации, понижения подъема и потери палатализованности для гласных переднего ряда), однако формантные значения этих спектров (по крайней мере F1 и F2, а нередко также и F3) будут очень близки друг к другу. Такую ситуацию можно наблюдать в табл. 2: три разные гласные фонемы, /е/, /о/ и /ы/, реализуются звуками, имеющими почти идентичные формантные значения (по крайней мере, на срезах, представленных в рассматриваемом датасете). Складывающееся положение дел препятствует точной классификации[1] гласных сегментов.

Таблица 2

Три гласных сегмента, представленных спектральными срезами в середине длительности гласного

Нормализованная F1, Гц	**Нормализованная F2, Гц**	**Гласный**	**Слово**
–0,463732	1,113556	*е*	*чекесен*
–0,489044	0,778111	*о*	*йолларынг*
–0,532968	0,852912	*ы*	*бычгъы*

Для того чтобы с уверенностью сказать, какой именно гласный перед нами, недостаточно значений первой и второй формант, взятых в одной точке, расположенной на временной оси; требуются дополнительные параметры. Если работать в рамках динамического подхода, использующего не один, а несколько спектральных срезов, то в качестве этого параметра можно взять, например, характер изменения гласного на протяжении времени.

Важность динамического представления описывается в разных работах. Согласно мнению, высказанному в [Ito, Tsuchida, Yano, 2001], изменения, происходящие со спектром под влиянием окружающих звуков с течением времени, играют важную роль в процессе идентификации гласного. В [Hillenbrand, 2013] упоминается, что классификация гласных с опорой на динамические параметры является одной из наиболее точных и близких к тому, как разделяют гласные на классы люди. В этой работе упомянуто исследование [Hillenbrand, Houde, 2003], в ходе которого модель классифицировала гласные. Она сравнивала подаваемые на вход в виде последовательности из пяти узкополосных спектров в определенные моменты времени (15, 30, 45, 60 и 75 % от общей длительности гласного) с образцами, которые представляли собой последовательности спектральных срезов на тех же временных точках, но уже усредненные. Авторы также сравнивают производительность распознавания для стандартной версии модели с пятью срезами с моделями, использующими один срез (пять отдельных тестов в каждый из моментов времени; т. е., 15 %, 30 %, 45 % и т. д.), два среза (на 15 и 75 % длительности гласных) и три среза (на 15, 45 и 75 % дли-

[1] Задача классификации возникает, например, при изучении качественных изменений, происходящих с гласными, находящимися в аффиксах. В ходе такого исследования гласные, претерпевшие изменения, мы сравниваем с эталонными гласными, чтобы понять, как и насколько сильно изменился гласный.

тельности гласных). Точность распознавания гласных для версий модели с одним срезом варьировалась от 75,5 до 80,4 %, при этом производительность была несколько выше для срезов, снятых вблизи к центру гласного, а не по краям. Производительность довольно резко улучшилась до 90,6 % для двух срезов, но незначительного дальнейшего улучшения не наблюдалось для трех (91,6 %) и пяти срезов (91,4 %). Из этого авторы делают вывод, что информация о начале и конце гласного, по-видимому, является наиболее важной. Повышение производительности модели на 10–15 % является достаточно весомым аргументом в пользу динамического анализа гласных по сравнению со статическим.

В данном исследовании мы опробуем методику динамического описания на системе гласных терского диалекта кумыкского языка. Далее в статье рассматриваются сначала варианты алгоритма динамического описания вокализма на примере анализа реализаций гласных фонем терского кумыкского, а затем представлено как качественное, так и количественное сравнение динамического описания со статическим с целью выбора стратегии, наиболее подходящей для терского кумыкского.

Представление данных

Предварительная обработка. Данные были размечены в программе Praat [Boersma, Weenink, 2023]. Разметка состояла из двух слоев: на верхнем выделялись границы целевых гласных сегментов, на нижнем – границы фонетических слов, содержащих целевые гласные. Далее с помощью скрипта на языке Praat каждый длинный аудиофайл был разделен на множество коротких, каждый из которых содержал гласный сегмент. Для коротких аудиофайлов с помощью плагина FastTrack [Barreda, 2021] были получены таблицы со значениями формант в точках, расположенных с шагом в 0,002 с. Последней ступенью обработки стало сведение полученных таблиц в одну и выделение формантных значений в точках, найденных алгоритмом Рамера – Дугласа – Пекера (далее – РДП-алгоритм), который по данной ломаной строит ломаную с меньшим числом точек, определяя расхождение, которое вычисляется по максимальному расстоянию между исходной и упрощенной кривыми.

Нормализация. Для того чтобы нивелировать влияние индивидуальных особенностей речевого аппарата каждого из носителей на общую картину, необходимо провести операцию нормализации. Именно с ее помощью можно привести несравнимые друг с другом формантные значения в герцах к некоторой общей шкале со своим нулем и своей единицей для каждого носителя.

В рамках нашего исследования нормализация проводилась относительно носителя. Для того чтобы добиться наибольшей точности представления данных, мы решили сравнить два метода унификации данных: *z-score* нормализацию по [Lobanov, 1971] (1), используемую в большинстве работ, посвященных вокализму, и модифицированную нормализацию по [Watt, Fabricius, 2002] (сокращенно *1mW&F*), которая используется крайне редко, однако является одной из лучших относительно того, как совпадали друг с другом формантные картины для разных носителей после нормализации (см. [Flynn, Foulkes, 2011], где по этому критерию на материале английского языка сравнивались около 20 мето-

дов нормализации). Формула для вычисления нормализованного значения по [Lobanov, 1971]):

$$Z = \frac{\{X - \mu\}}{\sigma}, \quad (1)$$

где X – конкретное значение, μ – среднее значение массива данных, а σ – стандартное отклонение.

Подход *1mW&F* заключается в представлении пространства гласных (*vowel space*) носителя в виде треугольника с вершинами в точках: (i) с наименьшим значением F1 и наибольшим значением F2 ([i]); (ii) с наименьшим значением F1 и значением F2, равным наименьшему значению F1 ([u'']) и (iii) с наибольшим значением F1 и средним значением F2 ([a])[1]. Точки (i-iii), расположенные на плоскости, заданной осями первой и второй формант, изображены на рис. 2.

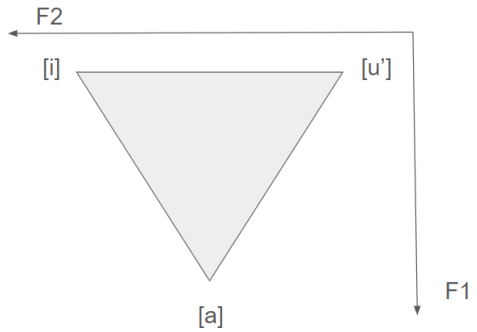


Рис. 2. Представление пространства гласных носителя по [Watt, Fabricius, 2002]

В рамках этой модели среднее значение (*mean value*) вычисляется как точка пересечения медиан данного треугольника:

$$S(F_1) = \frac{F_1[i] + F_1[a] + F_1[u\prime]}{3};\ S(F_2) = \frac{F_2[i] + F_2[u\prime]}{2}. \quad (2)$$

Нормализованные значения *i*-ой форманты вычисляются следующим образом:

$$(normalized)Fi = \frac{Fi}{S(Fi)}. \quad (3)$$

Сравним вышеописанные процедуры унификации формантных значений на выборке из двух носителей, для которых количество собранных данных было максимальным (т. е. для каждой гласной фонемы было собрано наибольшее количество произнесений).

Как можно наблюдать на рис. 3, соответствующие реализациям одной и той же фонемы эллипсы стандартного отклонения для двух носителей имеют большую площадь пересечения в случае *1mW&F*-нормализации, чем в случае *z-score*-нормализации. Кроме того, при нормализации, основанной на треугольнике гласных, среднее арифметическое формантных значений для произнесений какого-либо гласного (центр эллипса одного цвета на рис. 3) одним из носителей расположены на меньшем расстоянии от средних формантных значе-

[1] Более подробное описание того, почему были выбраны именно эти точки и как они вычислялись, см. в: [Watt, Fabricius, 2002].

ний для произнесений того же гласного другим носителем (центр эллипса другого цвета на рис. 3), т. е. эти значения расположены более «кучно», чем при *z-score*-нормализации. Таким образом, из двух представленных алгоритмов нормализации оптимальной на данный момент представляется процедура, описанная в [Watt, Fabricius, 2002], поэтому в нашем исследовании мы применили к «сырым» частотным значениям именно ее.

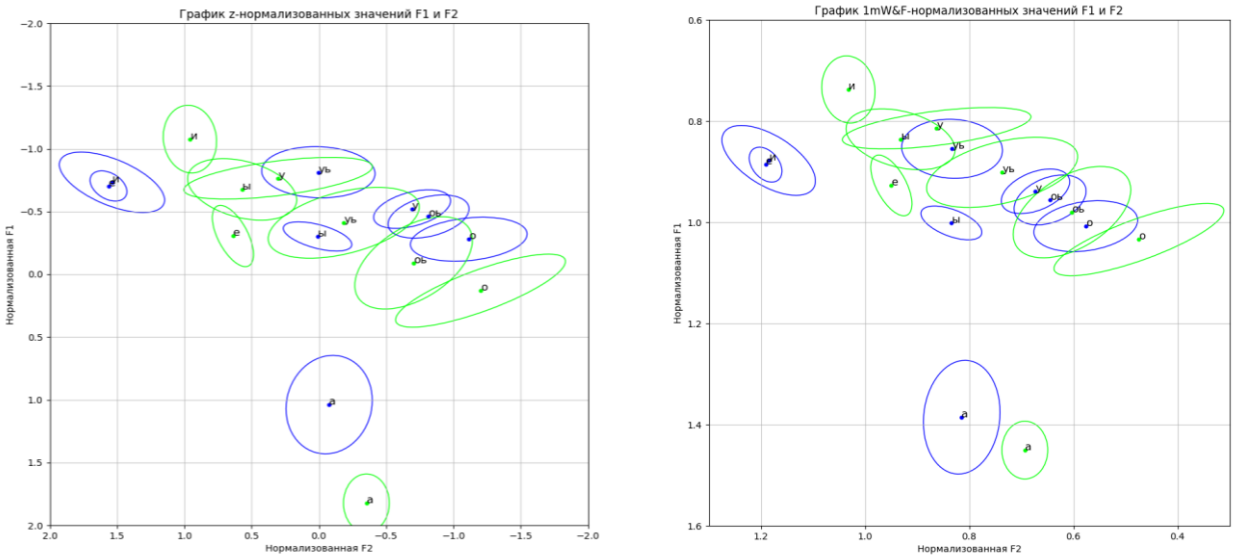


Рис. 3. Значения F1 и F2 для 2 носителей
(слева – *z-score*-нормализованные, справа – *1mW&F*-нормализованные)

Метрики. Имея для каждого гласного сегмента формантные значения в 2–4 точках, мы можем распоряжаться ими по-разному: как подавать гипотетическому классификатору, определяющему, какую гласную фонему реализует тот или иной сегмент в виде последовательностей, так и задействовать их в расчете некоторого значения, которое бы описывало то, как меняются форманты гласного сегмента на протяжении времени.

Несколько количественных метрик, значения которых показывают степень изменения качества гласного с течением времени, описаны в работе [Fox, Jacewicz, 2009], посвященной формантной динамике английских гласных.

Для того чтобы выразить степень изменения формантных характеристик гласного, используется евклидово расстояние между двумя ближайшими точками на временной оси в плоскостях F1 и F2. Оно рассчитывается:

$$VL = \sqrt{(F1_n - F1_{n+1})^2 + (F2_n - F2_{n+1})^2}. \quad (4)$$

Для динамического описания можно измерить евклидово расстояние между начальной и конечной точкой гласного (Фокс и Яцевич называют это «длиной вектора» (*Vector Length*, *VL*)), а можно для более тонкого представления задействовать большее количество точек, и тогда значением такой метрики будет сумма евклидовых расстояний между соседними точками. Последнюю метрику авторы называют «длиной траектории» (*Trajectory Length*, *TL*):

$$TL = \sum_{i=1}^{n-1} \quad = VL_i. \quad (5)$$

Поскольку ни одна из двух описанных выше метрик не задействует длительность гласного, на протяжении которой разворачиваются качественные изменения, была создана еще одна метрика, которая может охарактеризовать количество частотных изменений во времени. Это TL_{roc} (мы будем называть ее степенью изменения (*Rate of Change, RoC*)), формула для которой следующая:

$$RoC = \frac{TL}{0.6 \times dur}, \quad (6)$$

где dur – это длительность гласного, мс.

Именно метрику, представленную в формуле (6), мы будем использовать в качестве параметра для классификации гласных сегментов.

Сравнение. Главной задачей нашего исследования является определить, какое из описаний – статическое или динамическое (которое, в свою очередь, подразделяется на задействующее метрики, описывающие степень качественного изменения гласного (см. предыдущий раздел), и оперирующее последовательностями нормализованных формантных значений) – является наиболее точным (т. е. не причисляет сегмент к неверному классу) и дает наименьшее количество сегментов, попадающих в два и более классов одновременно. Для этого мы решили обучить классификаторы (в нашем случае это kNN-модель для многоклассовой классификации) на разных наборах параметров: для статического описания это *1mW&F*-нормализованные значения F1 и F2 (и длительности), для динамического – либо кортежи *1mW&F*-нормализованных значений для F1 и F2 (и длительность), либо значения RoC; а затем сравнить точность их предсказаний. Результаты сравнения представлены в табл. 3.

Таблица 3

Сравнение предсказаний kNN-классификаторов, учитывающих различные наборы предикторов

Стратегия описания	**Набор параметров**	**Точность, %**
Статическая	F1 в точке, соответствующей половине длительности сегмента, F2 в точке, соответствующей половине длительности сегмента	54,42
	F1 в точке, соответствующей половине длительности сегмента, F2 в точке, соответствующей половине длительности сегмента, длительность сегмента	61,9
Динамическая	Последовательность значений F1 в точках, найденных при помощи РДП-алгоритма, последовательность значений F2 в точках, найденных при помощи РДП-алгоритма	63,32
	Последовательность значений F1 в точках, найденных при помощи РДП-алгоритма, последовательность значений F2 в точках, найденных при помощи РДП-алгоритма, длительность сегмента	68,55
	Значение RoC	21,69
	Последовательность значений F1 в точках, найденных при помощи РДП-алгоритма, последовательность значений F2 в точках, найденных при помощи РДП-алгоритма, длительность сегмента, значение RoC	69,83

Выводы

Наибольшую точность классификации показывает модель, которая использует в качестве параметров кортеж из формантных значений, а также значение метрики, измеряющей величину качественного изменения относительно длительности гласного сегмента и собственно длительность сегмента. Модели, опирающиеся на динамическое представление гласных, показывают значимо (на 10 %) более хорошие результаты, чем те, которые опираются лишь на один спектральный срез. Тем не менее те модели, которые в качестве параметра имеют значение метрики, описывающей динамику изменения, показывают результат, близкий к случайному. Значит, классификация гласных лишь с опорой на величину их ингерентного изменения невозможна, нужны еще и формантные значения. Добавление длительности в качестве параметра значимо улучшает качество как статического (на 7 %), так и динамического (на 5 %) классификаторов.

Таким образом, динамическое описание вокализма (с учетом длительности гласных сегментов) является более описательным, чем статическое, так как дает более полную информацию о гласном сегменте.

Литература

Современный кумыкский язык / Н. Э. Гаджиахмедов [и др.]. Махачкала : ИЯЛИ ДНЦ РАН, 2014. 548 с.

Barreda S. Fast Track: fast (nearly) automatic formant-tracking using Praat // Linguistics Vanguard. Vol. 7(1). 2021. 12 p.

Flynn N., Foulkes P. Comparing Vowel Formant Normalization Methods // International Congress of Phonetic Sciences, 2011. P. 683–686.

Fox R. A., Jacewicz E. Cross-dialectal variation in formant dynamics of American English vowels // The Journal of the Acoustical Society of America. 2009. Vol. 126. P. 2603–2618.

Hillenbrand J. Static and Dynamic Approaches to Vowel Perception // Vowel Inherent Spectral Change, 2013. P. 9–30.

Hillenbrand J., Houde R. A narrow band pattern-matching model of vowel perception // The Journal of the Acoustical Society of America. 2003. Vol. 113. P. 1044–1055.

Ito M., Tsuchida J., Yano M. On the effectiveness of whole spectral shape for vowel perception // Journal of the Acoustical Society of America. 2001. Vol. 110. P. 1141–1149.

Lobanov B. M. Classification of Russian vowels spoken by different speakers // The Journal of the Acoustical Society of America. 1971. Vol. 49, N 2B. P. 606–608.

Watt D., Fabricius A. Evaluation of a technique for improving the mapping of multiple speakers' vowel spaces in the F1–F2 plane // Leeds Working Papers in Linguistics and Phonetics. 2002. Vol. 9. P. 159–173.

Электронный источник

Boersma P., Weenink D. Praat: doing phonetics by computer [Computer program]. 2023. Versions 6.3. URL: http:// www.fon.hum.uva.nl/praat (date of access: 01.12.2024).

УДК 81`33

А. Ю. Гришанова

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

Алгоритм против нейросети: сравнение генерации развернутого ответа на вопрос на ограниченном числе документов экстрактивным и абстрактивным методами

Аннотация. Исследуется задача многодокументной суммаризации, ориентированной на запросы. Сравнивается эффективность экстрактивного и абстрактивного методов суммаризации документов. Показывается, что каждый из данных методов имеет свои достоинства и недостатки, которые стоит учитывать при выборе инструмента для решения задачи. В качестве инструмента для абстрактивного метода используется большая языковая модель YaGPT. Испытывается алгоритм экстрактивного метода решения исследуемой задачи, основанный на выявлении, отборе и ранжировании предложений согласно их соответствию заданному запросу. Для оценки качества методов автоматической суммаризации был предложен русскоязычный датасет, основанный на материалах учебников истории для 5-го класса. Качество работы двух методов оценивается количественно с помощью метрик семейства ROUGE и качественно с помощью привлечения асессоров. Оказывается, что точность сгенерированных разными методами ответов сопоставима, притом ответы, сгенерированные абстрактивно, более читабельны. Приводится пример ответа, включающего в себя галлюцинацию большой языковой модели. Высказывается предположение о том, что перед выбором инструмента для автоматической суммаризации стоит взвешивать преимущества и недостатки выбранного метода в контексте стоящей перед пользователем задачи.

Ключевые слова: автоматическая суммаризация, вопросно-ответные системы, большие языковые модели.

A. Yu. Grishanova

HSE University, Moscow, Russia

Algorithm versus Neural Network: comparing extractive and abstractive methods of generating a detailed answer to a question based on a limited number of documents

Annotation. The article presents the comparison of the effectiveness of extractive and abstractive methods of query-focused multi-document summarization. Each of these methods has its advantages and disadvantages, which should be taken into account when choosing a tool to generate a summary. A large language model is used as a tool for the abstract method. The algorithm of the extractive method is tested. It includes identification, selection and ranking of sentences according to their compliance with a given query. To assess the quality of automatic summarization methods, a Russian-language dataset based on materials from history textbooks for 5th grade was proposed. The quality of the two methods is assessed quantitatively by using the ROUGE metrics and qualitatively by involving assessors. The results entail that the accuracy of the answers generated by different methods is comparable, however, the answers generated abstractively are more readable. An example of a response involving a hallucination of a large language model is given. It is suggested that before choosing a tool for automatic summarization, it is worth weighing the advantages and disadvantages of the chosen method in the context of the task facing the user.

Keywords: query-focused multi-document summarization, question answering systems, large language models.

Введение

Вопросно-ответные системы (Question Answering systems, QA-системы) – это такой тип поисковых систем, который и принимает вопросы, и отвечает на них на естественном языке. Несмотря на то что QA-системы чаще всего разрабатываются для кратких ответов на вопросы, задача получения развернутых ответов также важна: не на все вопросы можно ответить односложно.

Развернутые ответы на вопросы можно получить в результате решения задачи многодокументного реферирования, ориентированного на запросы (Query-Focused Multi-Document Summarization, QF-MDS). QF-MDS – это задача автоматического создания сводки (summary) из набора документов, которая отвечает на запрос конкретного пользователя; фактически она включает в себя построение вопросно-ответной системы, из результатов выдачи которой строится сводка. S. Lamsiyah и соавторы предложили экстрактивный метод решения QF-MDS, основанный на выявлении, отборе и ранжировании предложений согласно их соответствию заданному запросу [Unsupervised query-focused multi-document …, 2021, p. 7].

С другой стороны, развернутые ответы легко получить, используя большую языковую модель (Large Language Model, LLM). Следуя за [Jurafsky, Martin], будем называть такой метод абстрактивным (abstractive).

Описанные методы имеют свои преимущества и недостатки, затрагивающие как точность результатов, так и объем затраченных вычислительных ресурсов. Цель данной работы – сравнить качество работы экстрактивного и абстрактивного методов на материале русского языка. Для количественного сравнения точности методов мы создали русскоязычный датасет, состоящий из 25 вопросов и соответствующих им развернутых ответов.

1. Методы генерации ответов в вопросно-ответных системах: существующие проблемы и решения

Относительно простой способ получить ответ на вопрос – задать его большой языковой модели. Подав ей на вход строку с вопросом (префикс), мы можем попросить модель выполнить условную генерацию с учетом этого префикса и считать вывод модели ответом. Такой метод часто применяют для ответов на простые фактивные вопросы; он опирается на то, что огромные предварительно обученные языковые модели во время обучения считывают множество фактов из обучающей выборки и кодируют эту информацию в своих параметрах.

У данного метода несколько недостатков. Во-первых, большие языковые модели галлюцинируют. Галлюцинация в этом смысле – ответ, который не соответствует реальным фактам. Галлюцинации появляются из-за того, что LLM, получив на вход вопрос, просто придумывают ответ, который звучит разумно, а не опираются на какие-либо факты или источники. Так, при исследовании того, как большие языковые модели отвечают на вопросы юридической тематики (например, о деталях отдельных дел), обнаруживается, что галлюцинации юридической тематики в больших языковых моделях преобладают над верными ответами: 58 % в модели ChatGPT 4 и 88 % в Llama 2 [Large legal fictions …, 2024, p. 81].

Вторым недостатком применения LLM для вопросно-ответных систем является частота их обновления. Вопросно-ответные системы особенно полезны для применения к относительно небольшим данным. Примером такого применения является ответ системы на вопрос пользователя по содержимому почтового ящика. Такие данные динамичны, в то время как на обновление LLM модели могут уходить месяцы.

Таких проблем нет у экстрактивного метода решения задачи QF-MDS, предложенного в [Unsupervised query-focused multi-document …, 2021, p. 7]. Благодаря тому, что данный метод не требует обучения с учителем (является unsupervised), он применим и к динамичным данным: новые векторные представления объединяют со старыми и впредь осуществляют ранжирование по обновленному множеству. Тем не менее, в отличие от LLM, экстрактивный метод намного чаще порождает несвязный текст.

2. Данные

Имея датасет-стандарт, включающий в себя вопросы и написанные человеком ответы на них, можно оценить точность сгенерированного системой ответа. Для оценки качества вопросно-ответных систем с развернутыми ответами для соревнований DUC 2005-2007 были составлены особые датасеты [Dang, 2005]. Каждый из этих датасетов устроен следующим образом. Составителями было выбрано 45–50 тем, далее для каждой темы было скачано по 25–50 релевантных статей из новостных источников. После этого эксперт формулировал развернутый вопрос на каждую из тем и добавлял бинарный параметр детализации ответа на него (сжато (*general*), развернуто (*specific*)). Разметчики, опираясь на документы из созданной коллекции и ориентируясь на параметр детализации, составляли сводку-ответ на заданный вопрос; длина сводки не превышала 250 слов.

К нашему сведению, русскоязычного набора данных для решения задачи QF-MDS нет; перед нами стояла задача его создать. Как упоминают создатели датасетов DUC 2005-2007, мануально решать задачу QF-MDS сложно: на создание каждой сводки вручную у разметчиков уходило около 5 ч [Dang, 2005, p. 2]. Так, мы приняли решение создать датасет из доступных в Интернете развернутых ответов на вопросы.

Наше внимание привлекли вопросы после глав в школьных учебниках, а именно в учебнике истории 5-го класса [Вигасин, Годер, Свенцицкая, 2023]. У такого типа данных несколько преимуществ. Во-первых, разнообразие доступных ответов: их предоставляют как различные сайты с готовыми решениями задач из учебников, так и форумы, на которых пользователям отвечают другие пользователи[1]. Во-вторых, заранее известна коллекция документов, нужных для ответов на вопросы, – это тексты глав учебника. Третье преимущество затрагивает именно учебники классов средней школы: вопросы в них предполагают, скорее, нахождение ответа в тексте, а не порождение его путем рассуждений; таким образом, для получения ответов подходит экстрактивный метод.

[1] Например, http://otvet.mail.ru.

Мы отобрали 25 вопросов, классифицированных учебником [Вигасин, Годер, Свенцицкая, 2023] как «Проверь себя», и собрали для каждого из них по четыре развернутых ответа. Тематическое разнообразие, аналогичное датасетам DUC, было сохранено: мы ограничивались выбором одного вопроса из каждой главы. Благодаря тому, что учебная программа является стандартизированной, содержание учебников разного авторства одинаково. Так, мы собрали тексты восьми разных учебников.

Итак, созданный нами датасет устроен следующим образом. Коллекция документов состоит из восьми учебников по 50–60 глав, в ней больше 40 000 предложений, что сопоставимо с размерами коллекции документов датасетов DUC 2005-2007. Всего в нашем датасете 25 вопросов-тем, для каждого предложено по четыре «золотых» сводки. Далее в работе мы сравнивали схожесть сгенерированных абстрактивным и экстрактивным методами сводок с «золотыми».

3. Генерация развернутых ответов абстрактивным методом

Мы задали YaGPT[1] 25 вопросов, построенных по следующему шаблону:

(1) *Развернуто ответь на следующий вопрос по теме ТЕМА, используя не более 100 слов:* ВОПРОС, где ТЕМА, и ВОПРОС: данные из созданного датасета-стандарта.

Пример одного из вопросов:

(2) *Развернуто ответь на следующий вопрос по теме ДРЕВНИЙ ВОСТОК, ДРЕВНИЙ ЕГИПЕТ, используя не более 100 слов, расскажи о разливах Нила.*

Ограничение в 100 слов было введено в связи с использованием бесплатных ресурсов YaGPT. Для точного сравнения методов то же ограничение было введено и в экстрактивный метод.

4. Генерация развернутых ответов экстрактивным методом (решение задачи QF-MDS)

Получение векторных представлений

Высокая эффективность предложенного в [Unsupervised query-focused multi-document …, 2021, p. 7] метода заключается в использовании качественной модели, порождающей векторные представления документов и запросов. Мы использовали модель SentenceBert (sBert) [Reimers, Gurevych, 2019], которая является модификацией классической модели BERT [Bert: Pre-training of deep …]. Проект sBert предоставляет различные модификации sBert, предобученные для определенных задач. В нашей работе мы использовали модель *all-mpnet-base-v2*[2]. Мы остановились именно на ней, потому что, согласно утверждениям создателей проекта, она показывает лучшие результаты в задаче порождения эмбеддингов предложений.

Для оценки качества работы модели мы использовали датасет *nli-rus-translated-v2021*[3]. Этот датасет составлен из различных англоязычных NLI (Natural Language Inference, Интерференция в естественном языке) датасетов, ав-

[1] https://ya.ru/ai/gpt-3.

[2] https://sbert.net/docs/sentence_transformer/pretrained_models.html.

[3] https://huggingface.co/datasets/cointegrated/nli-rus-translated-v2021.

томатически переведенных на русский язык. По нашим сведениям, аналогичных датасетов, созданных из исконно русскоязычных данных, еще не существует.

Выбранная нами модель *all-mpnet-base-v2* на русскоязычных данных показывает не самые лучшие результаты: на тестовых данных датасета nli-rus-translated-v2021 модель предсказывает со средней квадратичной ошибкой 0,335. Нами было принято решение дообучить модель на тренировочной выборке размером 10 000 строк.

Качество модели после обучения на пяти эпохах (MSE = 0,041) нас устроило, и далее в работе мы используем именно ее.

Ранжирование предложений

Для ранжирования полученных с помощью sBert эмбеддингов мы следовали методике, описанной в [Unsupervised query-focused multi-document …, 2021, p. 7]. Согласно ей ранжирование разделяется на следующие шаги:

– оценка схожести предложения с запросом и выбор k наиболее похожих на запрос предложений;

– повторная оценка предложений для ликвидации повторов (избыточности).

Подробно рассмотрим эти шаги. S. Lamsiyah и соавторы [Unsupervised query-focused multi-document …, 2021] предлагают использовать следующую формулу для оценки релевантности каждого предложения для запроса:

$$score_{relevance}(Si) = \alpha \cdot RSV_{bm25}(Si, Q) + (1 - \alpha) \cdot RSV_{Sim}(Si, Q),$$

где S_i – рассматриваемое предложение, α – коэффициент значимости аргумента, RSV_{bm25} – оценка релевантности Q для S_i согласно алгоритму BM25 [Okapi at TREC-3 …, 1995], Q – запрос пользователя, RSV_{Sim} – косинусное расстояние между S_i и Q.

Комбинируя алгоритм BM25 с оценкой косинусного расстояния, формула учитывает как лексические, так и семантические признаки предложения и запроса. Коэффициент α дает возможность регулировать вклад того или иного типа признаков в оценку схожести. Для дальнейшей работы мы отбираем 50 предложений с наивысшими оценками по формуле.

Задача QF-MDS подразумевает ответ на вопрос в виде сводки, поэтому ранжирование и выбор наиболее подходящего предложения недостаточен: для составления сводки важна не только релевантность предложения запросу, но и его новизна. Для придания выводу модели этих характеристик S. Lamsiyah и соавторы [Unsupervised query-focused multi-document …, 2021] используют метод максимальной предельной релевантности (Maximum Marginal Relevance method, MMR) [Carbonell, Goldstein, 1998]. Оценка MMR наиболее высокая, когда предложение имеет наибольшую схожесть с запросом пользователя и наименьшую схожесть с предложениями, уже выбранными для сводки.

После ранжирования предложений мы объединяем их, опираясь на несколько факторов. Во-первых, размер полученной сводки должен быть не больше 100 слов. Во-вторых, предложения, встречающиеся в одном документе, в сводке стоят в порядке своего появления в нем. В-третьих, в сводке мы отражаем только предложения, состоящие больше, чем из пяти слов: таким образом мы элиминируем назывные предложения, не несущие достаточно информации.

5. Метрики качества

Мы оцениваем качество полученных сводок как количественно (ROUGE-метрики), так и качественно (человеческая оценка).

Для оценки автоматической суммаризации чаще всего используют метрики из семейства ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [Lin, 2004]. Такие метрики считают количество совпадений (overlaps) слов и словосочетаний в сгенерированном и тестовом резюме. Мы используем метрику ROUGE-recall, которая отражает часть релевантной информации, сохраненную в сводке. Мы считаем ее наиболее подходящей для нашей задачи: чем больше recall (максимальное значение 1), тем меньше важных деталей из оригинального текста было утеряно в процессе суммаризации.

Существуют разные типы ROUGE; следуя за [Unsupervised query-focused multi-document … , 2021, p. 14], мы вычисляем ROUGE-1 (совпадения униграмм), ROUGE-2 (совпадения биграмм) и ROUGE-SU (совпадения skip-грамм и униграмм).

Для post-оценки составленных сводок были привлечены три асессора; им предлагалось оценить по два варианта развернутого ответа для четырех вопросов. Вопросы были выбраны случайным образом.

Как было сказано выше, при генерации развернутого ответа экстрактивно в числе прочих решается задача автоматической суммаризации. Согласно [A Comparative Analysis … , 2019], хорошая сводка (summary) должна соответствовать следующим критериям:

– содержание ключевой информации;
– краткость;
– отсутствие избыточности;
– актуальность;
– связность (когезия) и читабельность.

Асессорам предлагалось оценить по шкале от 1 до 3 каждый из представленных ответов в соответствии с его полнотой, релевантностью представленной информации и связностью текста; таким образом мы проверяли соответствие сводок критериям 1, 4 и 5 соответственно.

6. Результаты

Результаты работы методов представлены в табл. 1. Статистически значимой разницы нет, p-value = 0,5187[1].

Таблица 1

Средние значения метрик для разных методов

Метод	ROUGE-1	ROUGE-2	ROUGE-SU
Экстрактивный	0,156	0,030	0,050
Абстрактивный	0,148	0,036	0,045

[1] Односторонний t-test.

Обсудим полученные результаты. Экстрактивный метод имеет лучшие показатели ROUGE-2 и ROUGE-SU, однако статистической разницы между выборками средних значений нет. Обратимся к рис. 1, чтобы сравнить показатели ROUGE для каждого из вопросов в датасете.

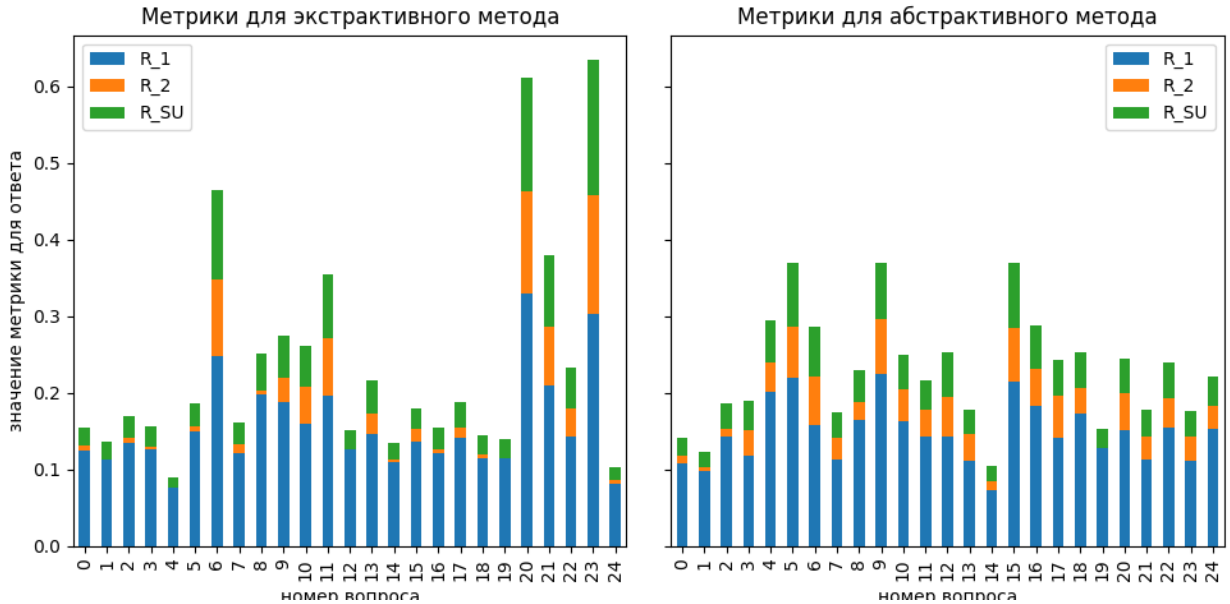


Рис. 1. Распределение метрик ROUGE для ответов, сгенерированных разными методами

Можно заметить, что множество ответов, сгенерированных экстрактивным методом, включает в себя «положительные» выбросы: ответы, количественно оцененные выше других. Вот пример такого ответа:

(3) Вопрос: *В РИМЕ ПРИ ИМПЕРАТОРЕ НЕРОНЕ. Какую опасность для зрителей представляло собой посещение театра во время выступлений Нерона?*

А. Ответ, сгенерированный экстрактивным методом: *В театре во время выступлений Нерона находились преторианцы и десятки доносчиков, следивших за поведением зрителей и бравших на заметку тех, кто не одобрял актера-императора. Боясь народных выступлений, императоры приказывали вести списки неимущих граждан и раздавать по ним хлеб. На крайнем востоке ближневосточного региона, на территории современного Ирана, располагалось Эламское царство, которое в этот период представляло значительную силу. Сенат и его роль в Риме. Сенат управлял текущими делами в Риме, в то же время сенаторы председательствовали в судах.*

Б. Ответ, сгенерированный YaGPT: *Во время выступлений Нерона в театре зрители подвергались опасности из-за пиротехнических эффектов, которые использовались для создания ярких и запоминающихся шоу. Однако эти эффекты были непредсказуемыми и могли привести к пожарам или другим несчастным случаям.*

Несмотря на то, что (3А) включает в себя лишнюю информацию, первое предложение сгенерированного ответа правильно отвечает на поставленный вопрос, в то время как (3Б) противоречит учебникам истории.

Итак, согласно количественным метрикам, эффективность экстрактивного и абстрактивного методов сопоставима. Как можно заметить на примере (3), сгенерированные абстрактивно ответы отличаются лучшей связностью и отсутствием нерелевантной информации. Обсудим результаты оценки полноты, связности и релевантности текста, представленные в табл. 2. Статистически значимой разницы нет, p-value = 0,8687[1].

Таблица 2

Сравнение средних оценок асессоров для двух методов

Метод	Полнота ответа	Связность ответа	Релевантность информации
Абстрактивный	2,67	3	3
Экстрактивный	2,17	2,4	2,33

Ответы, сгенерированные YaGPT, ожидаемо получают более высокие оценки, чем ответы, сгенерированные экстрактивно. Связность и релевантность ответов YaGPT были оценены на высший балл всеми асессорами. Тем не менее статистически значимого различия между оценками результатов двух методов не наблюдается.

7. Обсуждение и выводы

Мы убедились, что точность развернутых ответов на вопрос, сгенерированных экстрактивным способом, сопоставима с точностью ответов, сгенерированных LLM, притом последние имеют лучшую связность и релевантность. Так, качество экстрактивной модели на ограниченном числе документов сопоставимо с качеством LLM, хотя для использования первой требуется намного меньше вычислительных ресурсов. Кроме того, в работе был приведен пример галлюцинации: сгенерированного LLM ответа, не соответствующего реальности. Таким образом, при выборе инструмента для суммаризации следует учитывать, насколько для задачи важны точность и читабельность ответа, а также объем доступных вычислительных ресурсов.

Литература

Вигасин А. А., Годер Г. И., Свенцицкая И. С. История. Всеобщая история. История древнего мира: 5 класс : учебник. М. : Просвещение, 2023. 302 с.

A Comparative Analysis on Hindi and English Extractive Text Summarization / Verma [et al.] // Association for Computing Machinery. 2019. Vol. 3 (18). P. 30–39.

Large legal fictions: Profiling legal hallucinations in large language models / M. Dahl [et al.] // Journal of Legal Analysis. 2024. Vol. 16. P. 64–93.

Lin Chin-Yew. Rouge: A package for automatic evaluation of summaries // Text summarization branches out. Barcelona, 2004. P. 74–81.

Okapi at TREC-3 / Robertson S. [et al.] // Nist Special Publication Sp 109, 1995. P. 109.

Reimers N., Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, 2019. P. 3982–3992.

[1] Односторонний t-test.

Unsupervised query-focused multi-document summarization based on transfer learning from sentence embedding models, BM25 model, and maximal marginal relevance criterion / S. Lamsiyah [et al.] // Journal of Ambient Intelligence and Humanized Computing. 2021. Vol. 14. P. 1–18.

Электронные источники

Bert: Pre-training of deep bidirectional transformers for language understanding / J. Devlin [et al.]. URL: https://arxiv.org/abs/1810.04805 (date of access: 01.12.2024).

Carbonell J., Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries // Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. Melbourne. 1998. URL: https://www.researchgate.net/publication/2269571_The_Use_of_MMR_Diversity-Based_ Reranking_for_Reordering_Documents_and_Producing_Summaries. (date of access: 01.12.2024).

Dang H. Overview of DUC 2005 // Document Understanding Workshop. Vancouver, 2005. URL: https://www.semanticscholar.org/paper/Overview-of-DUC-2005-Dang/dc954000617ae982f83b7f1ed4bd72b95e38aa88 (date of access: 01.12.2024).

Jurafsky D., Martin J. Speech and Language Processing. 3rd ed. URL: https://web.stanford.edu/~jurafsky/slp3/ (date of access: 01.12.2024).

УДК 168.522

В. Д. Деборина

Национальный исследовательский университет «Высшая школа экономики», Санкт-Петербург, Россия

Сравнение российского и зарубежного опыта исследований в области цифровых гуманитарных наук на основе библиометрического анализа

Аннотация. Представлено исследование основных направлений развития области цифровых гуманитарных наук (digital humanities) в России и за рубежом на основе библиометрического анализа. Анализируются тенденции роста публикаций на тему цифровой гуманитаристики с 2000 по 2024 г. на основе данных каталога OpenAlex. Сравниваются количество публикаций и выявляются особенности научной активности в двух исследуемых контекстах. Обсуждаются результаты анализа ключевых слов авторов и контент-анализа статей с использованием программы VOSviewer для визуализации данных. Выявляются основные тематические кластеры зарубежных и российских исследований, включающие такие направления, как *искусственный интеллект, гуманитарные науки, информационные технологии*. Подчеркиваются различия, например отсутствие ключевых слов, связанных с естественными науками, в российских публикациях. Выводы исследования показывают, что, несмотря на значительно меньшее количество российских публикаций, направления исследований в области цифровых гуманитарных наук в России и за рубежом схожи. Тематические карты и выявленные взаимосвязи позволяют понять перспективы развития цифровых гуманитарных наук в России и за рубежом.

Ключевые слова: цифровые гуманитарные науки, библиометрический анализ, визуализация данных, тематические карты, гуманитарные исследования.

V. D. Deborina

National Research University «Higher School of Economics», St. Petersburg, Russia

Comparison of Russian and Foreign Research Experience in the Field of Digital Humanities Based on Bibliometric Analysis

Abstract. The article presents a study of the development of digital humanities in Russia and abroad based on bibliometric analysis. It analyzes trends in the growth of publications on digital humanities from 2000 to 2024 using data from the OpenAlex catalog. The volumes of publications are compared, and the peculiarities of scientific activity in the two studied contexts are identified. The paper discusses the results of analyzing authors' keywords and content analysis of articles using VOSviewer software for data visualization. The main thematic clusters of foreign and Russian research are identified, including areas such as *artificial intelligence, humanities, economics, and information technology*. Differences are highlighted, such as the absence of keywords related to natural sciences in Russian publications. The study's conclusions show that, despite a significantly smaller number of Russian publications, the development directions of digital humanities in Russia and abroad are similar. The thematic maps and identified relationships allow understanding the prospects for the development of digital humanities in Russia and abroad.

Keywords: digital humanities, bibliometric analysis, visualization, term map, humanitarian studies.

Цифровые гуманитарные науки (digital humanities) – это область исследований, которая сформировалась на пересечении цифровых технологий и гуманитарных наук. Эта область не ограничивается одной дисциплиной и в эпоху цифровых технологий стала универсальной. В какой-то степени трудно понять и интерпретировать, что на самом деле представляет собой digital humanities, но ключевой характеристикой этого направления исследований является применение цифровых методов в традиционно гуманитарных дисциплинах [Science Mapping Analysis …, 2021]. Целью настоящего исследования является анализ основных направлений научной области digital humanities в России и за рубежом.

Для проведения библиометрического анализа в качестве источника данных использовался открытый каталог исследовательских работ OpenAlex. Для получения библиографических данных применялась строка поиска и такие фильтры, как период публикации, тема исследований (concept) и страна публикации (рис. 1). Временной период публикаций, выбранный для исследования, включает 2000–2024 гг. Поиск проводился в апреле 2024 г.

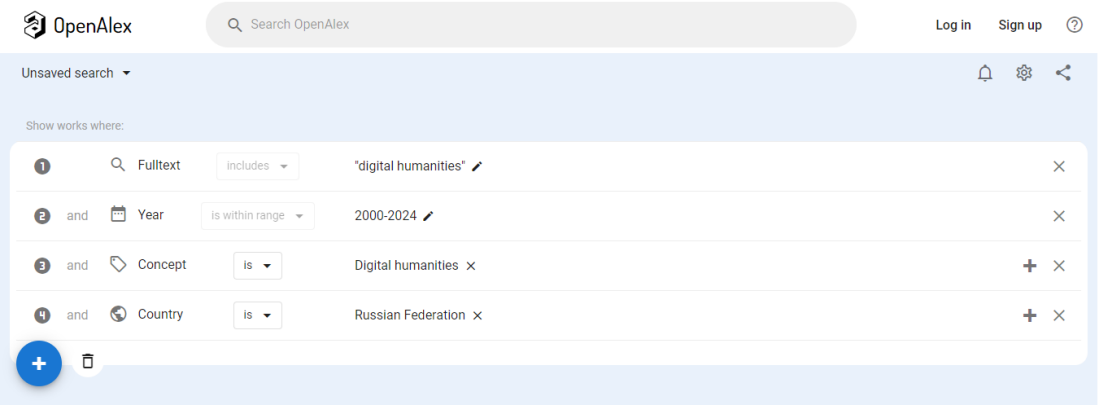


Рис. 1. Входные данные для поиска по каталогу OpenAlex

Проводится сравнение развития области digital humanities (DH) в России и за рубежом, поэтому для этого с использованием фильтра «страна учреждения» были получены библиографические данные для публикаций за рубежом и для российских публикаций. В результате поиска для первого параметра было получено 8256 работ. Для второго параметра выдача составляла 64 публикации. Такое различие в выдаче можно объяснить тем, что OpenAlex ориентирован на англоязычные работы, поэтому имеющиеся в базе данных российские исследования, вероятно, были опубликованы не только на русском, но и на английском языке.

Для обработки данных и создания визуализаций использовалось программное обеспечение VOSviewer. С его помощью создавались тематические карты, которые отображают то, как связано между собой совместное появление ключевых слов автора (co-word или co-occurrence) или слов, которые получаются при контент-анализе.

Чтобы проанализировать тенденции роста научных публикаций в этой области, из каталога OpenAlex были загружены данные в формате файлов CSV.

График, изображенный на рис. 2, показывает развитие области исследований цифровых гуманитарных наук в 2000–2024 гг. за рубежом. На нем также

видно, что количество работ зарубежных авторов, посвященных данной теме, начало значительно расти в 2011 г. В 2018 г. было опубликовано наибольшее количество работ (833) по данной теме. Хотя начиная с 2021 г. количество публикаций уменьшилось почти на 200, по сравнению с предыдущими годами, показатели все еще остаются достаточно высокими, что демонстрирует заинтересованность исследователей в области цифровых гуманитарных наук.

Количество российских работ на тему digital humanities начало расти в 2014 г. (рис. 3). Самое большое количество опубликованных работ на эту тему пришлось на 2020 г. Поскольку количество российских работ намного меньше, чем зарубежных, нельзя сделать вывод о том, стабильна ли заинтересованность исследователей в этой теме, ссылаясь на показатели графика, хотя начиная с 2014 г. работы на тему цифровых гуманитарных наук публикуются ежегодно, чего раньше не наблюдалось.

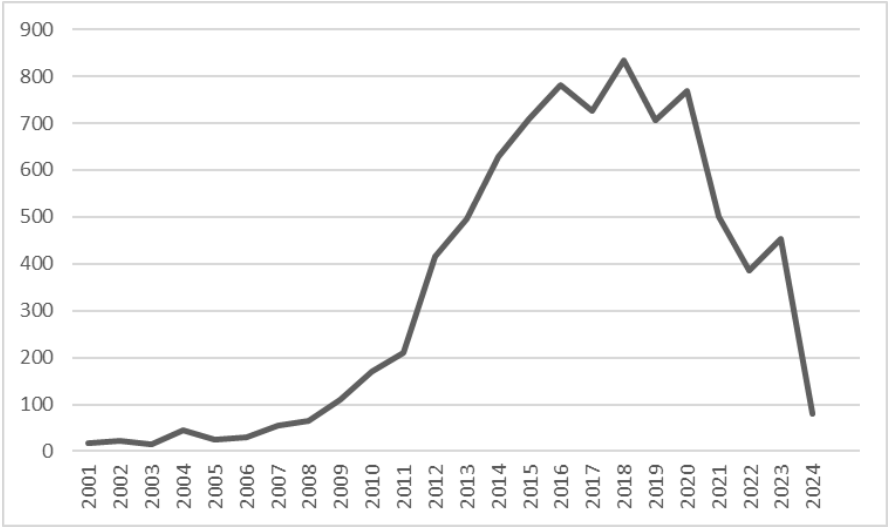


Рис. 2. Тенденция роста количества литературы на тему digital humanities в 2000–2024 гг. за рубежом по данным OpenAlex

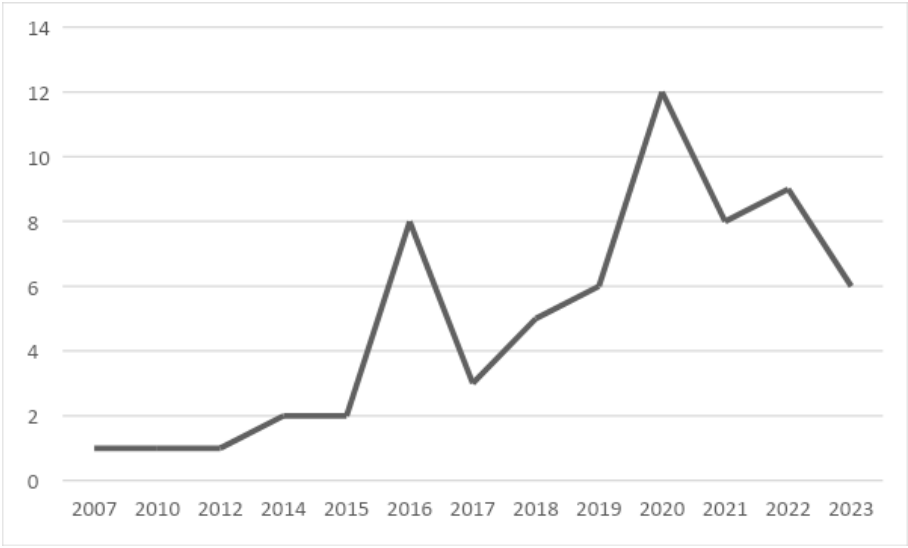


Рис. 3. Ежегодные тенденции роста количества литературы на тему digital humanities в 2000–2024 гг. в России по данным OpenAlex

С помощью программы VOSviewer были проанализированы и визуализированы библиографические данные зарубежных исследований на тему цифровых гуманитарных наук. С помощью создания тематической карты можно обнаружить пересечение направлений исследований.

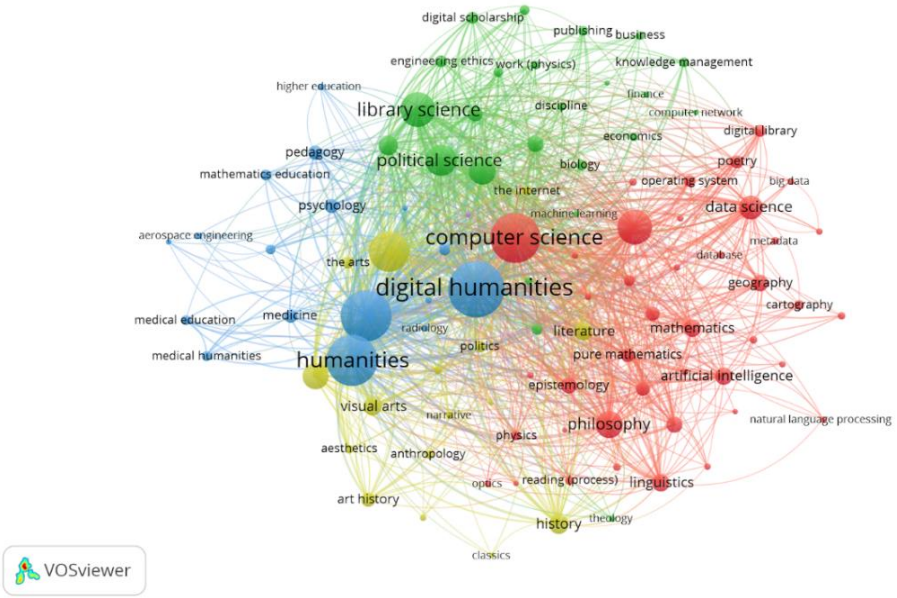


Рис. 4. Тематическая карта зарубежных научных исследований, посвященных цифровым гуманитарным наукам

Для создания карты, изображенной на рис. 4, было отобрано 100 наиболее часто встречающихся ключевых слов автора. На их основе было создано четыре кластера. Для того чтобы рассмотреть, как образуются кластеры и их связи на тематической карте, при описании каждого кластера сначала анализируются ключевые слова, из которых он состоит, а затем описываются некоторые из статей, отобранные для анализа.

Для красного кластера характерны ключевые слова, связанные с цифровыми технологиями и точными науками: *artificial intelligence, big data, computer science, data mining, data science, digitalization, digital library, mathematics, physics*. Также встречаются такие нехарактерные для кластера слова, как *literature, linguistics, poetry, philosophy*.

В работе Artificial Intelligence and the Limits of the Humanities [Duch, 2024] исследуются изменения, с которыми сталкиваются гуманитарные науки в эпоху развития цифровых технологий. Появление междисциплинарных областей и интеграция искусственного интеллекта выделяются как пути к возрождению и сохранению гуманитарных наук и повышению их актуальности. Применение новых технологий позволит расширить возможности гуманитарных наук и сделать их актуальными и привлекательными для студентов.

В зеленом кластере пересекаются разные области знаний, например естественные науки, экономика, социология и др. В нем встречаются ключевые слова *biology, chemistry, business, ecology, economics, engineering, law, social science, political science*.

Статья Re-imagining the Cambridge School in the Age of Digital Humanities [London, 2016] посвящена использованию цифровых ресурсов в изучении истории политической мысли и их влиянию на взаимодействие политологов. Автор рассматривает тенденцию к внедрению подходов, при которой цифровые инструменты помогают анализировать разнообразные тексты и исследовать взаимосвязи между политическими изменениями и использованием определенной лексики и других языковых особенностей в текстах. На примере работ Кембриджской школы и текстового анализа автор показывает возможности для продуктивного взаимодействия разных направлений в политической науке. В результате предлагается модель, демонстрирующая, как сотрудничество между этими подходами может расширить понимание роли языка в политике в эпоху цифровых гуманитарных наук.

Синий кластер объединяет в себе гуманитарные науки и медицину: *digital humanities, art, humanities, medicine, medical education, pedagogy, psychology*.

С этим кластером связана статья Movement of knowledge: Introducing medical humanities perspectives on medicine, science, and experience [Hansson, Irwin, 2020], в которой авторы изучают изменение медицинских знаний в различных контекстах, от лабораторий до общественных и политических решений. Подчеркивается важность междисциплинарного подхода, соединяющего медицинскую гуманитаристику, социальные науки и технологии. Авторы также исследуют, как цифровизация влияет на создание, передачу и восприятие медицинской информации. Работа показывает, что знания преобразуются в процессе их взаимодействия с культурными, социальными и технологическими структурами, что связывает исследование с областью цифровых гуманитарных наук.

Большинство ключевых слов из желтого кластера относятся непосредственно к гуманитарным наукам: *literature, history, gender studies, politics, sociology, anthropology*.

Статья Lost in Translation? The Odyssey of «Digital Humanities» in French [Clivaz, 2017] исследует перевод термина *digital humanities* на французский язык и его влияние на культурное разнообразие в цифровой культуре. Автор анализирует использование термина *humanités* в выражениях *humanités numériques* и *humanités digitales*. Рассматриваются дискуссии о французской адаптации термина и его взаимодействии с точными науками, а также влияние цифровой эпохи на гуманитарные науки. Французский перевод термина способствует объединению традиционных и цифровых гуманитарных методов, что отражает обновление гуманитарных наук в условиях цифровой эпохи.

Кроме анализа ключевых слов автора, в VOSviewer можно создавать карты на основе контент-анализа отобранных статей. Анализируя текст статей, программа выделяет самые часто встречаемые слова и визуализирует их связи. Однако этот метод не всегда подходит для исследований, целью которых является выделить основные направления развития научных областей. Так, например, на карте (рис. 5) не оказалось никаких направлений, кроме digital humanities и computer science, хотя при анализе ключевых слов было выявлено много различных научных областей, связанных с цифровыми гуманитарными науками.

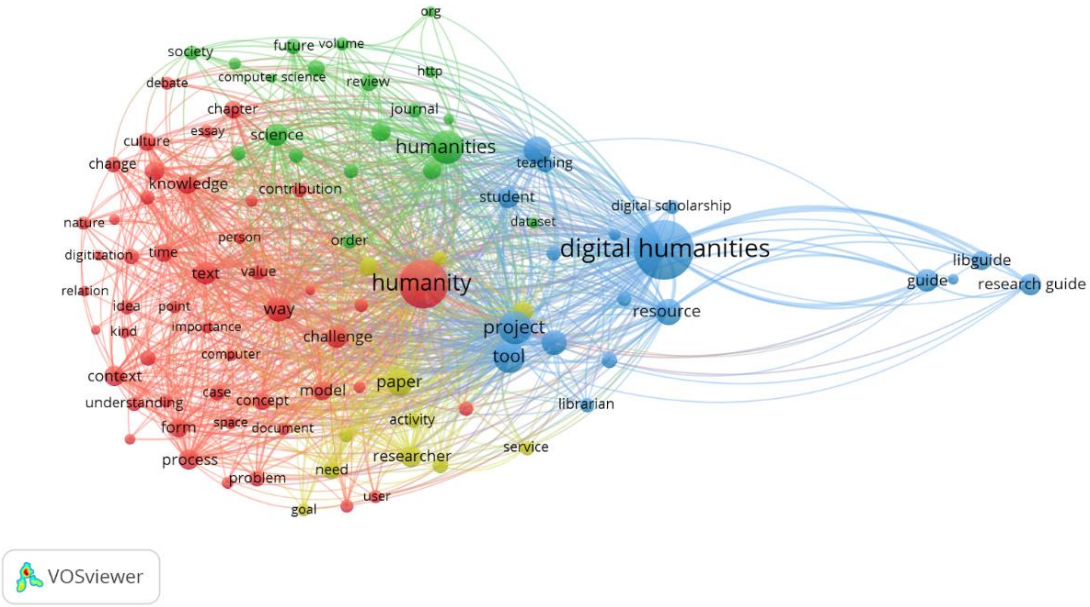


Рис. 5. Тематическая карта зарубежных научных исследований, посвященных цифровым гуманитарным наукам, на основе контент-анализа

При создании тематической карты ключевых слов российских исследований на рис. 6 образовалось пять кластеров.

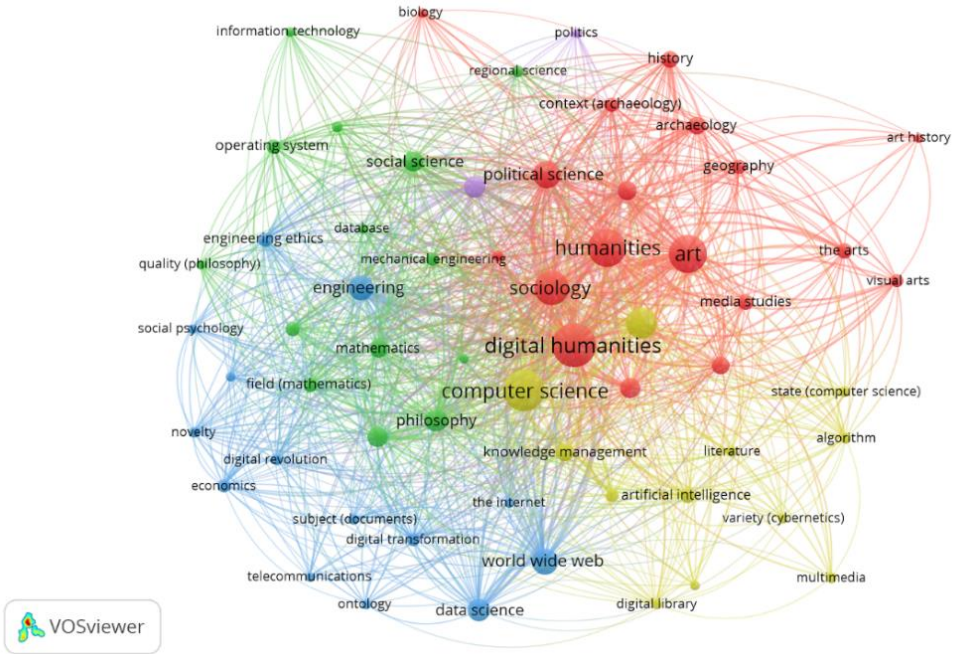


Рис. 6. Тематическая карта ключевых слов российских научных исследований, посвященных цифровым гуманитарным наукам

В красном кластере преобладают ключевые слова, связанные с гуманитарными науками, в том числе цифровыми: *digital humanities, history, humanities, mathematics education, pedagogy, political science, sociology, psychology.*

В статье «Сдвиг парадигм: библиотеки в контексте цифровых гуманитарных наук (зарубежный опыт)» [Савицкая, 2018] рассматривается новая роль за-

падных библиотек в контексте цифровых гуманитарных наук, где они становятся центрами для использования цифровых методов анализа текста, обработки данных, цифрового картографирования и 3D-моделирования. Цель статьи – изучить сотрудничество библиотек с проектами DH и его роль в цифровой науке. Здесь также рассматриваются обязанности библиотечных центров DH, например создание цифровых коллекций, управление базами данных и др.

Зеленый кластер относится к точным наукам и информационным технологиям: *mathematics, information technology, mechanical engineering, operating system, physics*.

Статья «Информационная среда Digital Humanities: анализ информационных взаимодействий» [Можаева, Можаева-Ренья, Захарова, 2016] представляет результаты исследования информационной среды цифровых гуманитарных наук, основанного на выявлении их структур, направлений их деятельности и анализе информационных взаимодействий в этой сфере. Результаты исследования помогают определить ключевые тенденции развития цифровых гуманитарных наук и их влияние на современное состояние традиционных гуманитарных наук.

Синий кластер связан с экономикой и интернетом: *economics, finance, telecommunications, the internet, world wide web, data science*.

Статья Novel Circular Graph Capabilities for Comprehensive Visual Analytics of Interconnected Data in Digital Humanities [Ryabinin, Belousov, Chuprina, 2020] посвящена разработке платформы визуальной аналитики SciVi для работы с взаимосвязанными данными в цифровых гуманитарных науках. Она предлагает новые подходы для анализа данных, такие как круговые графы, фильтры для обработки шумных данных, калькулятор состояний графа и синхронизацию с географическими картами. Эти инструменты помогают исследователям обрабатывать большие объемы информации, визуализировать связи и проводить сравнительный анализ. Использование SciVi показывает, как цифровые технологии могут повысить эффективность исследований в гуманитарных дисциплинах.

Желтый кластер – искусственный интеллект и компьютерные технологии: *artificial intelligence, algorithm, computer science, digital library, linguistics, literature*.

В статье Artificial intelligence and semiotics, or Methods for production of bridges between traditional and digital humanities [Kulikov, 2015] рассматриваются и обсуждаются в контексте цифровых гуманитарных наук модели искусственного интеллекта. Автор предлагает построить модель искусственного интеллекта, используя семиотический подход, что может помочь в разработке методов для объединения традиционных и цифровых гуманитарных наук.

Фиолетовый кластер состоит из двух слов: *law* и *politics*.

Сравнивая тематические карты зарубежных и российских исследований, можно заметить, что ключевые слова у них во многом совпадают: в обеих тематических картах встречаются такие ключевые слова, как *computer science, data science, politics, mathematics, philosophy, sociology, economics* и другие, из этого следует, что и направления развития у них тоже схожи. Основное различие карт заключается в том, что в тематической карте российских исследований, в отли-

чие от зарубежных, отсутствуют ключевые слова, связанные с медициной (*medicine, medical education* и др.).

Результаты исследования показывают, что цифровые гуманитарные науки в России развиваются так же активно, как и за рубежом, следуя мировым тенденциям в применении цифровых технологий в области гуманитарных наук.

Благодарности

Публикация подготовлена в результате проведения исследования по проекту «Текст как Big Data: методы и модели работы с большими текстовыми данными» в рамках Программы фундаментальных исследований НИУ ВШЭ в 2024 г.

Литература

Можаева Г. В., Можаева-Ренья П. Н., Захарова У. С. Информационная среда Digital Humanities: анализ информационных взаимодействий // Журнал СФУ. Гуманитарные науки. 2016. № 9 (7). С. 1572–1585.

Савицкая Т. Е. Сдвиг парадигм: библиотеки в контексте цифровых гуманитарных наук (зарубежный опыт) // Обсерватория культуры. 2018. Т. 15, № 5. С. 532–541.

Clivaz C. Lost in Translation? The Odyssey of “Digital Humanities” in French // Studia Universitatis Babeș-Bolyai-Digitalia. 2017. Vol. 62, N 1. P. 26–41.

Duch W. Artificial Intelligence and the Limits of the Humanities // Er (r) go. Teoria-Literatura – Kultura. 2024. N 48. P. 269–297.

Hansson K., Irwin R. Movement of knowledge: Introducing medical humanities perspectives on medicine, science, and experience // Lund: Nordic Academic Press, 2020. P. 9–26.

London J. A. Re-imagining the Cambridge School in the Age of Digital Humanities // Annual Review of Political Science. 2016. Vol. 19, N 1. P. 351–373.

Ryabinin K. V., Belousov K. I., Chuprina S. I. Novel Circular Graph Capabilities for Comprehensive Visual Analytics of Interconnected Data in Digital Humanities // Scientific Visualization. 2020. Vol. 12, N 4. P. 56–70.

Электронные источники

Kulikov S. Artificial intelligence and semiotics, or Methods for production of bridges between traditional and digital humanities. 2015. URL: https://hal.science/hal-01179990v3 (date of access: 30.11.2024).

Science Mapping Analysis of Digital Humanities research: A scientometric study / N. Gupta [et al.] // Library Philosophy and Practice. 2021. URL: https://digitalcommons. unl.edu/libphilprac/6126 (date of access: 30.11.2024).

УДК 81'33

И. А. Дёмина, Ю. М. Алюнина

Российский университет дружбы народов им. Патриса Лумумбы, Москва, Россия

Критерии эффективности переводческих «кошек»: перевод через язык-посредник

Аннотация. Анализируются возможности и ограничения переводческих «кошек» на примере Smartcat со встроенным модулем машинного перевода по материалам текстов из реальной переводческой практики в компании OSK-Group (Москва). Эмпирическим материалом исследования послужили тексты технической направленности, а именно случаи перевода в языковой паре русский – китайский через английский в качестве языка-посредника. Исследование показало, что результат машинного перевода в Smartcat, опирающийся на память перевода, пользовательский глоссарий и интегрированную терминологическую базу, требует существенного постредактирования некоторых текстовых сегментов, что обусловлено лингвистическими и экстралингвистическими причинами.

Ключевые слова: автоматизированный перевод, машинный перевод, технический перевод, перевод терминов, постредактирование.

I. A. Demina, Yu. M. Alyunina

RUDN University, Moscow, Russia

Efficiency criteria for Smartcat system: translation through an intermediary language

Abstract. The article analyses the possibilities and limitations of CAT-tools using the example of Smartcat system with built-in machine translation module. The work is based on texts taken from the real translation practice of the company «OSK Group» (Moscow). The empirical material of the study was technical texts, namely, cases of translation in the language combination Russian-Chinese with English as an intermediary language. The study showed that the result of machine translation in Smartcat, which relies on the translation memory, user glossary and integrated terminology database, requires significant post-editing of some text segments due to linguistic and extra-linguistic reasons.

Keywords: automatic translation, machine-aided translation, technical translation, term translation, postediting.

Введение

В современных реалиях, когда развитие экономических связей ориентировано на Восток, многие российские компании сотрудничают с Китаем по совершенно разным направлениям: наука и технологии, автомобилестроение, культура и туризм, энергетика, цифровизация и др. [Глава Приангарья; Милькин; Ульянова, Бабонов; Цыплаков; Чупров; Ян Цянь]. Переводческая индустрия (в том числе машинные и автоматизированные переводчики [Findings of the 2021 Conference …, 2021, p. 49; Гу, Сун, 2022, с. 11–12; Утробин; Утробин; Li, 2015, p. 208]) оказалась не вполне готовой к наблюдаемой ныне масштабной смене языкового вектора [Бородина, Вербицкая, Потаева; Володина; Сидорова; Томский политех]. Поэтому многие действующие переводчики столкнулись с тем,

что языком-посредником между российским и китайским бизнесом стал английский. Данная ситуация обнаружила множество переводческих трудностей, обусловленных не только языковыми, но и межкультурными различиями, которые должны приниматься во внимание для обеспечения адекватности перевода.

В настоящей статье представлены случаи из переводческой практики, которые демонстрируют сложности, возникающие при переводе текстов технической направленности через язык-посредник с использованием системы автоматизированного перевода Smartcat [Smartcat].

Исследование выполнено по материалам переводческой деятельности в компании OSK-Group (Объединенная станкоинструментальная компания, г. Москва). Участвуя в программе импортозамещения, компания «специализируется на комплексных решениях в металлообработке» [Объединенная Станкоинструментальная Компания …]: занимается производством и продажей металлообрабатывающего и складского оборудования, режущих инструментов и расходных материалов, а также обеспечивает перевод соответствующей документации.

На сегодняшний день OSK-Group тесно сотрудничает с Китаем в сфере станкоинструментальной промышленности, которая является базовым сектором экономики. В начале 1990-х гг. эта отрасль в России утратила позиции на международном рынке. На данный момент многие отечественные предприятия практически полностью остановили разработку новых моделей оборудования. Однако достаточно быстрыми темпами в области станкостроения развивается сотрудничество с китайскими партнерами. Интенсификация контактов с Китаем обнаружила недостаточную готовность переводческой индустрии обеспечивать межъязыковую коммуникацию с восточными партнерами в необходимом объеме. Проявлениями этой неготовности оказались среди прочего следующие:

- нехватка технических переводчиков с китайского языка;
- желание компаний иметь «универсального переводчика» для коммуникации с разными странами на разных языках.

Рабочим решением в сложившейся ситуации оказался перевод через английский как язык-посредник в российско-китайском сотрудничестве.

Структура настоящей статьи предполагает: 1) уточнение термина CAT-tool и раскрытие функциональных возможностей Smartcat; 2) анализ некоторых примеров перевода технических текстов в языковой паре китайский – русский через посредничество английского языка; 3) систематизацию критериев эффективности переводческих «кошек». Наблюдения, сделанные в ходе переводческой и исследовательской работы, позволят повысить эффективность баз параллельных текстов, которые используются для обучения систем машинного перевода, что в перспективе может способствовать повышению качества прямого машинного перевода в языковой паре русский – китайский.

1. CAT-tool и функционал Smartcat

В современных реалиях переводчики вынуждены адаптироваться к изменчивой международной экономической обстановке, работая на высоком уровне производительности. Для обеспечения скорости и качества переводов им необходимо пользоваться автоматизированными системами перевода –

computer-assisted translation, или CAT-tool, которые также называют «кошками». Это программы, которые помогают переводчикам ускорить свою работу, а заказчикам экономить время и средства. В отличие от машинных переводчиков (MS Bing, Яндекс Переводчик), CAT не выполняют перевод без участия человека. По сути, перевод осуществляет человек, а программные комплексы CAT способствуют повышению эффективности ручного труда путем сохранения переведенных фрагментов текста в параллели с их исходниками на языке оригинала и последующего механического внедрения соответствующих совпадений в новых документах для заданной языковой пары. Эта функция известна как память переводов, или Translation Memory (TM).

Память переводов – это база данных, в которой постепенно накапливаются параллельные тексты [Translation Memory …]. Они хранятся в виде пар предложений или их частей на исходном языке и языке перевода. Если в оригинале документа на языке А встречаются одинаковые элементы, требующие одинакового перевода на язык Б, то TM автоматически выполняет замену необходимого сегмента, освобождая переводчика-человека от необходимости повторного перевода или ручного поиска ранее переведенного фрагмента.

Фрагменты, оставшиеся не переведенными автоматически, передаются для ручной обработки переводчику или встроенной системе машинного перевода. На этом этапе переводчик-человек может выделить переведенные машинным способом сегменты и занести новые пары параллельных текстов в пользовательскую базу данных. Этот алгоритм обеспечивает высокое качество перевода при работе с однотипными текстами, которые регулярно встречаются в документах технической направленности: таблицы, ТЗ, спецификации[1], инструкции, технические описания и др. Такие документы отличаются высокой повторяемостью синтаксических конструкций и терминов, которые должны быть единообразно представлены на языке перевода.

Одной из самых распространенных CAT-программ в отечественной переводческой индустрии является облачный[2] сервис Smartcat (SC). Его популярность объясняется доступностью, понятным интерфейсом и широким набором функций в бесплатной версии. Пользоваться программой можно по корпоративной подписке и в режиме персональной авторизации. Поскольку программа является облачной, ее использование, как правило, запрещено при работе с документами, содержащими коммерческую тайну и секретную информацию, например касающуюся госзакупок.

Общедоступный функционал SC представлен ниже:

- ведение и хранение памяти переводов;
- перевод с помощью встроенного модуля машинного переводчика;

[1] Спецификация – это «один из основных документов технической конструкторской документации (на изделие, продукты и т. д.), выполняемый обычно в виде таблицы, в которой указываются название изделия, его составные части и элементы, материал, из которого они изготовляются, масса и др. данные» [Словарь-справочник].

[2] Программа доступна для использования только в режиме онлайн – в облачной версии. Она не устанавливается на устройство (компьютер, ноутбук, планшет и др.) переводчика. Переводчик создает личный аккаунт в Smartcat или пользуется корпоративным аккаунтом своей компании.

- управление глоссариями;
- совместная работа переводчиков над одним документом;
- поддержка разных форматов файлов.

Перечень форматов, поддерживаемых SC, приведен на рис. 1.

Формат файлов	Расширение файлов
OpenDocument (OpenOffice.org / StarOffice / LibreOffice)	*.odt, *.ott, *.ods, *.ots, *.odp, *.otp
Microsoft Open XML	*.docx, *.xlsx, *.pptx, *.ppsx, *.ppt, *.pps
Другие текстовые документы	*.txt, *.rtf
Отсканированные документы и изображения	*.pdf, *.jpeg, *.tiff, *.bmp, *.png, *.gif, *.djvu
HTML/XHTML	*.html, *.xhtml,*.xht
DocBook	*.xml
Одноязычные XLIFF-файлы Okapi	*.xlf, *.xliff, *.sdlxliff
Двуязычные файлы	*.ttx
Ресурсы Android	*.xml
InDesign CS4 Markup	*.idml
Файлы справки	*.xml, *.hmxp

Рис. 1. Перечень форматов, поддерживаемых SC[1]

Помимо приведенных выше форматов, доступных для работы в SC, программа также поддерживает файлы с расширениями .dxf, .dwg, которые соответствуют чертежам, что особенно актуально при переводе в сфере станкостроительной отрасли. Поддержка того или иного формата означает: 1) переводчик загружает файл в формате, например, .pdf, .png., .jpg; 2) SC распознает текст и конвертирует его в «буквенный» вид; 3) переводчик выполняет перевод, работая с материалом в привычной текстовой форме; 4) переводчик сохраняет перевод и выгружает его в исходном формате. Если на входе было изображение, схема или чертеж с текстом на языке А, то на выходе получается изображение, схема или чертеж с текстом на языке Б в первоначальном формате.

Набор инструментов SC действительно облегчает работу переводчика, экономит время и, соответственно, ресурсы заказчика. Но насколько эффективна данная система в условиях работы через язык-посредник?

2. Использование Smartcat в условиях перевода через язык-посредник

В переводоведении языком-посредником называют вспомогательный язык, который используется для перевода с языка А на язык Б не напрямую, а через посредничество третьего языка [Евтеев, Латышев, 2017, с. 80]. Таким образом, сначала выполняется перевод с языка А на третий язык, затем с третьего языка на язык Б, и наоборот. Такой прием используется, когда переводчик-человек не владеет либо языком-источником (языком А), либо целевым языком (языком Б), но при этом знает третий язык, через посредничество которого обеспечивается перевод с А на Б. Язык-посредник также называют промежуточным языком.

[1] Источник: https://ru.wikipedia.org/wiki/SmartCAT.

Для обеспечения документооборота между российскими и китайскими партнерами в компании OSK-Group в качестве языка-посредника используется английский.

При работе через язык-посредник неизбежны трудности перевода, обусловленные языковыми и культурными различиями. Многие из них вызваны тем, что перевод на язык-посредник иногда оказывается низкого качества, поскольку выполняется он техническим персоналом с китайской стороны. Не имея представления о структурных различиях китайского и английского языков, технический персонал, как правило, пользуется машинным переводчиком и не выполняет постредактирование. Текст на английском языке, не прошедший «человеческую обработку» после машинного перевода с китайского языка, попадает переводчику в российской компании, который должен выполнить перевод с английского языка на русский. Не владея китайским языком, специалист в российской компании должен перевести текст с промежуточного языка на русский, сохраняя исходный смысл и соблюдая терминологию.

Обеспечению качества перевода, выполняемого в таких условиях, способствуют использование «кошек», работа с источниками и общение с представителями отрасли с российской стороны.

3. Анализ машинного перевода терминов в Smartcat

Проанализируем несколько ситуаций, в которых использование SC со встроенным модулем машинного перевода оказалось неадекватным по причине работы через язык-посредник. Все примеры взяты из реальных технических документов с производства, использование которых в данной статье было одобрено владельцем и заказчиком перевода.

Пример 1

На рис. 2 приведен шильдик (от нем. *der Shild – щит, табличка, ярлык*). Это табличка, надпись или знак с дополнительной информацией об оборудовании, которая располагается на корпусе станка.

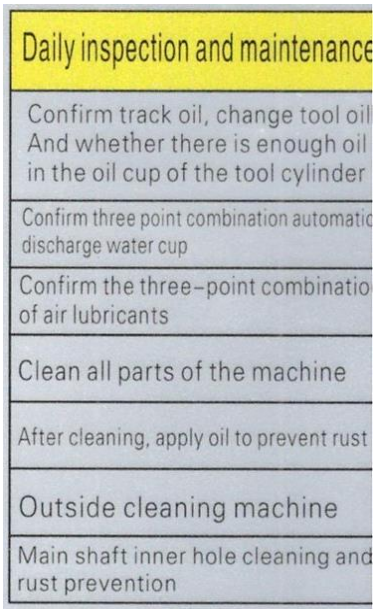


Рис. 2. Пример шильдика для станка с информацией о техосмотре

Отличительной особенностью шильдиков является краткость и ясность содержащихся на них текстов, что должно быть сохранено и в переводе. Переводчику необходимо подобрать такие лексические единицы на языке перевода, которые будут не только передавать суть оригинального текста, но и которые уместятся в пространство шильдика. При переводе с китайского языка на русский (через посредничество английского) решение этой переводческой задачи затруднительно. Затруднение объясняется разной степенью смысловой компрессии языкового знака в китайском и русском языках: при равной смысловой нагрузке иероглифа и кириллического слова первый занимает значительно меньше пространства на шильдике, чем второе. Соответственно, переводчику необходимо не только подобрать верные термины на русском языке, но и вписать текст в пространство шильдика, изначально предназначенное для текста на китайском языке.

Обратим внимание на таблицу шильдика (см. рис. 2). В ней представлены требования ежедневного техосмотра. После загрузки этого файла в SC программа выдает приемлемый перевод, требующий незначительного постредактирования. Исключение составляют два последних пункта, которые вызывают сложности у встроенного в SC модуля машинного перевода: *Outside* ***cleaning*** *machine; Main shaft inner hole* ***cleaning*** *and rust prevention*.

При переводе на язык-посредник была допущена ошибка – нарушен порядок слов: вместо повелительного наклонения формы *clean* используется форма герундия *cleaning*, что меняет смысл и затрудняет работу автоматизированного помощника.

В целом SC удается передать смысл (см. строки 7 и 8 в рис. 3), но требуется полная корректировка переводчиком-человеком, поскольку в русском языке инструкция по техосмотру требует употребления повелительного наклонения или инфинитива: ***Очистить*** *оборудование снаружи;* ***Очистить*** *внутренние отверстия главного вала для предотвращения образования ржавчины*.

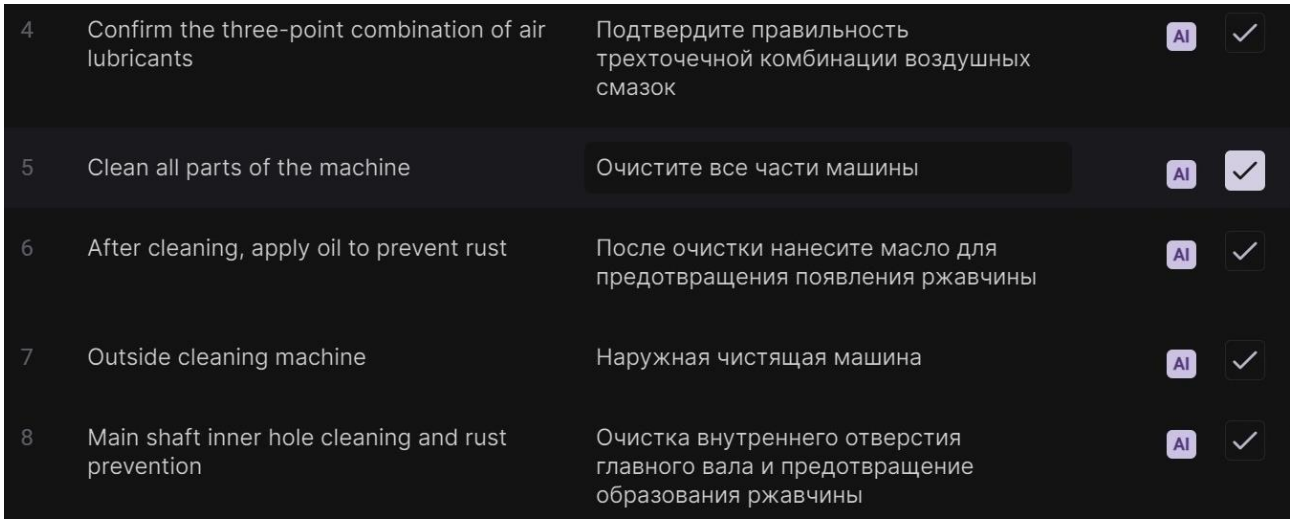


Рис. 3. Скриншот фрагмента перевода шильдика с рис. 2 в SC

Пример 2

На рис. 4 приведен предупреждающий знак, который сопровождается текстом на промежуточном языке. В тексте наблюдается нарушение пунктуации, пропуск пробелов между предложениями, наличие грамматических и стилистических ошибок.

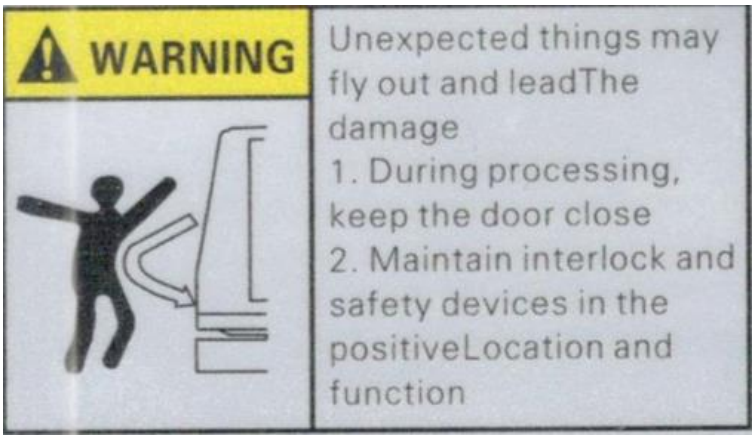


Рис. 4. Пример предупреждающего знака с текстом

Вариант перевода, предложенный SC, характеризуется отсутствием грамматических неточностей. Он и передает смысл текста, но со стилистической точки зрения требуется доработка переводчиком-человеком (рис. 5).

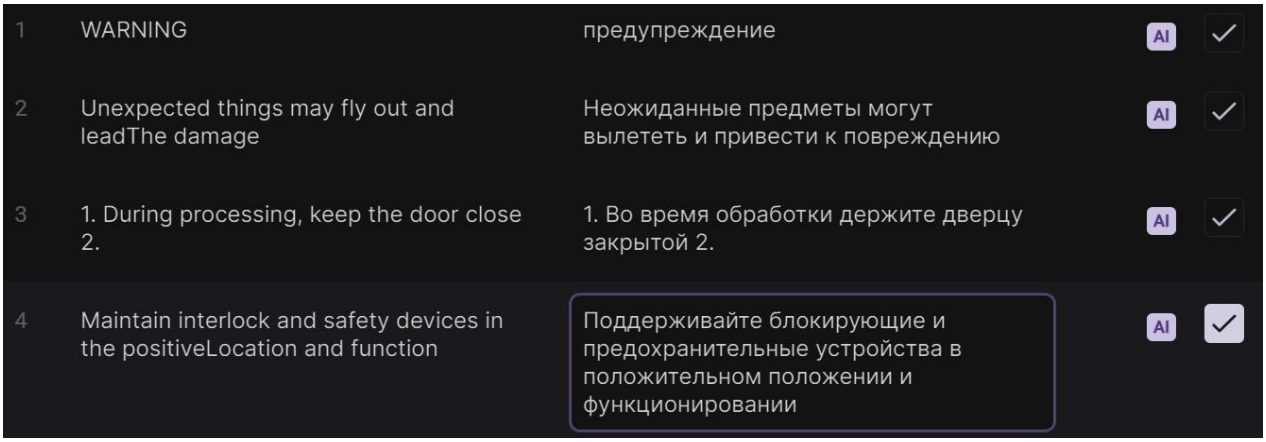


Рис. 5. Скриншот перевода текста на предупреждающем знаке с рис. 4 в SC

Корректным в данном случае будет считаться такой перевод: *Поддерживайте блокирующие и предохранительные устройства в функционирующем состоянии*.

Пример 3

Рассмотрим еще один пример перевода текста, который располагается на предупреждающем знаке. На скриншоте SC (рис. 6) показан параллельный текст, в котором присутствует аббревиатура *АТС* на английском как языке-посреднике. С ее переводом на русский SC не справился, сохранив аббревиатуру на латинице.

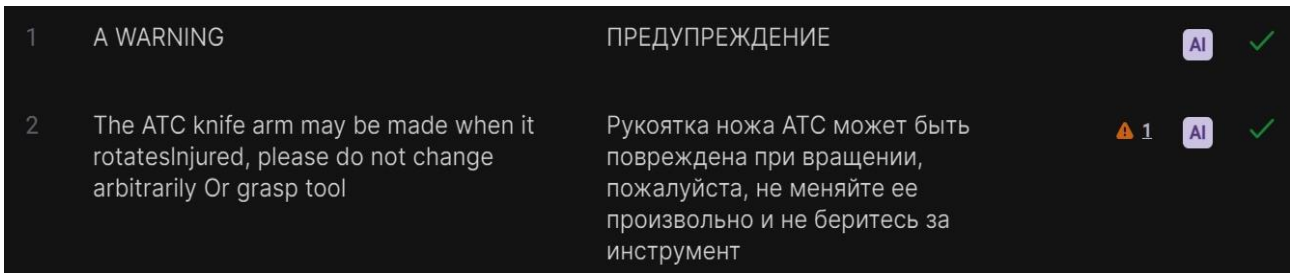


Рис. 6. Скриншот перевода текста на предупреждающем знаке в SC

Ошибка в переводе аббревиатуры на русский язык связана с ее неверным переводом с китайского на язык-посредник. В подобных ситуациях переводчику необходимо опираться на свой личный опыт работы с техническими файлами в соответствующей отрасли или обращаться за консультацией к специалисту.

Корректный перевод приведенного текста выглядит следующим образом: ***Система автоматической смены инструмента*** *может нанести травмы при вращении. Не прикасайтесь к инструменту во время вращения*.

Пример 4

Последний пример, рассматриваемый в данной статье, – спецификация на английском языке, предоставленная китайскими партнерами (рис. 7).

Mark	Descriptions	(SET)	FOB
	SURFACE GRINDING MACHINE		IN CNY
	SG-30100AHD	2	97876
	380/50hz/3ph. METRIC LEADSCREW		
	standard accessories list :		
	• grind wheel		
	• grind wheel flange		
	• coolant tank		

Рис. 7. Скриншот фрагмента спецификации

Перевод на язык-посредник выполнен корректно в рамках требуемой терминологии. Однако при загрузке документа в систему автоматизированного перевода происходит ошибка в распознании текста, и SC разбивает фрагмент на теги[1], которых в исходном файле не было (рис. 8).

[1] Тег – это так называемый непереводимый текст или заполнитель. По сути, теги – это своеобразные виртуальные рамки лексической единицы или нескольких лексических единиц, которые отличаются от остальных элементов текста формальными признаками: курсивом, подчеркиванием, выделением жирным шрифтом или цветом и т. п. Чтобы переводчику не пришлось вручную оформлять тест в соответствии с оригиналом, «кошки» распознают формальные характеристики текста и помечают их тегами. При копировании тега в текст на целевом языке переводчик автоматически переносит исходное оформление в перевод.

Рис. 8. Скриншот перевода фрагмента спецификации в SC

Эта техническая ошибка в распознавании исходника приводит к неверному переводу предложения на русский язык, кардинально меняя его смысл, – *Ухмылка у него на лице* (строка 20 на рис. 8). Ошибки в расстановке тегов возможны, когда исходный документ содержит формальные ошибки, на что никак не может повлиять переводчик. В подобных случаях ему необходимо выполнить постредактирование вручную и исправить ошибки. В данном случае верным вариантом перевода является *Фланец шлифовального круга.*

Заключение

Опираясь на опыт перевода технических документов в SC через язык-посредник, проиллюстрированный приведенными примерами, можно выделить следующие показатели эффективности переводческих «кошек»:

- ускорение процесса перевода за счет распознавания и сохранения формата документа, а также разбивки текста на фрагменты;
- допустимое качество встроенного модуля машинного перевода даже при наличии ошибок в исходном тексте на языке-посреднике;
- повышение качества перевода благодаря памяти переводов;
- возможность составления и загрузки пользовательского или корпоративного глоссария, необходимого для работы с документом.

В процессе перевода документации станкостроительной отрасли также были выделены некоторые ограничения SC:

- проблемы с распознаванием исходного файла, чреватые некорректным машинным переводом с промежуточного языка на целевой;
- проблемы с переводом аббревиатур и некоторых терминов, «узнаванию» которых не всегда помогает интегрированный глоссарий;
- высокая вероятность необходимости постредактирования, особенно стилистического и терминологического;
- невозможность перевода документов, имеющих коммерческую тайну.

Тем не менее опыт выполнения переводов с использованием SC показывает, что подобные автоматизированные системы существенно облегчают труд переводчика. Помогая ускорить процесс перевода, они увеличивают объем обрабатываемой информации, хотя и не избавляют технического переводчика от общения с отраслевыми специалистами и работы со вспомогательными источниками: словарями, энциклопедиями, видео и др.

Технологии машинного перевода вносят существенные изменения в переводческую индустрию, которые уже сейчас необратимы. Однако автоматизация перевода не ставит человека на второй план и не исключает его из процесса перевода, но меняет роль переводчика на роль эксперта. **На текущем этапе развития технологий машинный перевод может «подменять» человека на некоторых этапах переводческой цепочки, но не вытесняет его полностью из переводческого цикла.** Умение правильно работать с машинными инструментами и понимать их алгоритмы теперь становится неотъемлемой частью профессиональной компетентности переводчика-человека.

Литература

Евтеев С. В., Латышев Л. К. Перевод и языковое посредничество // Филологические науки в МГИМО. 2017. № 11. С. 80–86.

Гу Ц., Сун Ц. Сравнительный анализ исследований переводческих технологий в Китае и России в цифровую эпоху // Вестник Московского Университета. Сер. 22. Теория перевода. 2022. № 3. С. 7–30.

Li Y. Research on Verb Semantic – syntactic Knowledge Base in Russian – Chinese Machine Translation // Zhengzhou: Atlantis Press, 2015. P. 204–208.

Электронные источники

Бородина В., Вербицкая Ю., Потаева К. В России растет спрос на специалистов со знанием китайского языка. URL: https://www.vedomosti.ru/management/articles/2022/05/24/923463-spros-spetsialistov-kitaiskogo (дата обращения: 03.12.2024).

Володина В. Язык Поднебесной: разворот на Восток подстегнул спрос на китайский в Петербурге. URL: https://www.dp.ru/a/2022/08/31/JAzik_Podnebesnoj_kak_raz (дата обращения: 02.12.2024).

Глава Приангарья – Глава Приангарья обсудил расширение сотрудничества с замминистра культуры и туризма Китая. URL: https://www.ogirk.ru/2024/10/28/igor-kobzev-obsudil-rasshirenie-sotrudnichestva-s-zamministra-kultury-i-turizma-kitaja/ (дата обращения: 08.12.2024).

Милькин В. Россия и Китай обсуждают расширение сотрудничества в энергетике. URL: https://www.vedomosti.ru/analytics/trends/articles/2024/07/28/1052513-rossiya-i-kitai-obsuzhdayut-rasshirenie-sotrudnichestva-v-energetike (дата обращения: 07.12.2024).

Объединенная Станкоинструментальная Компания. Металлообрабатывающее оборудование и стеллажи для хранения металла. URL: http://www.osk-group.ru/ (дата обращения: 08.12.2024).

Сидорова А. Российский бизнес столкнулся с дефицитом технических переводчиков с китайского языка. URL: https://www.bfm.ru/news/532366 (дата обращения: 08.12.2024).

Словарь-справочник по информатике (онтология информатики). URL: http://db4.sbras.ru/elbib/data/show_page.phtml?77+1437+35 (дата обращения: 07.12.2024).

Томский политех – Томский политех будет готовить технических переводчиков с китайского языка. URL: https://news.tpu.ru/news/tomskiy-politekh-budet-gotovit-tekhnicheskikh-perevodchikov-s-kitayskogo-yazyka/ (дата обращения: 08.12.2024).

Ульянова М., Бабонов Ю. Современные формы и механизмы взаимодействия России и Китая в сфере научно-технического сотрудничества. URL: https://russiancouncil.ru/analytics-and-comments/analytics/sovremennye-formy-i-mekhanizmy-vzaimodeystviya-rossii-i-kitaya-v-sfere-nauchno-tekhnicheskogo-sotrud/ (дата обращения: 05.12.2024).

Утробин М. В. Первая бесплатная модель перевода с русского на китайский язык и обратно. URL: https://habr.com/ru/articles/721330/ (дата обращения: 02.12.2024).

Утробин М. В. Сравнение локальных моделей машинного перевода для английского, китайского и русского языков. URL: https://habr.com/ru/articles/791522/ (дата обращения: 08.12.2024).

Цыплаков С. Об основных трендах развития торговли России и Китая. URL: https://russiancouncil.ru/analytics-and-comments/analytics/ob-osnovnykh-trendakh-razvitiya-torgovli-rossii-i-kitaya/ (дата обращения: 07.12.2024).

Чупров А. Президент Путин приветствовал расширение присутствия китайских автоконцернов в России. URL: https://www.autostat.ru/news/57605/ (дата обращения: 08.12.2024).

Ян Цянь. Россия считает перспективным углубление сотрудничества с Китаем в области цифровой экономики. URL: http://russian.people.com.cn/n3/2024/0918/c31518-20220212.html (дата обращения: 08.12.2024).

Findings of the 2021 Conference on Machine Translation. (WMT21) / F. Akhbardeh, A. Arkhangorodsky, M. Biesialska [et al.] // Proceedings of the Sixth Conference on Machine Translation. 2021. P. 1–88 URL: https://cris.fbk.eu/handle/11582/330742 (date of access: 03.12.2024).

Smartcat – Платформа перевода AI Ваше универсальное решение. URL: https://ru.smartcat.com/ (дата обращения: 08.12.2024).

Translation Memory – Введение в технологию Translation Memory. URL: https://www.tra-service.ru/tm (дата обращения: 01.12.2024).

УДК 8-81

А. Ю. Зюрик, А. А. Айсанова

Санкт-Петербургский государственный университет, Санкт-Петербург, Россия
Иркутский государственный университет, Иркутск, Россия

Зачем так дооолго? Сравнительный анализ употребления нефонологической долготы в речи двух говорящих: типы удлинения звуков

Аннотация. Выявляются функции нефонологической долготы гласных звуков и сравниваются особенности употребления данного средства в речи двух информантов. Основной целью работы является обнаружение общего и специфического в функционировании нефонологической долготы. Существующие исследования описывают следующие функции нефонологической долготы: маркирование хезитации, иконическое выражение пространственной и временной величины, интенсификации признака, акцентирование значимого элемента. Анализируются две записи интервью с пожилыми женщинами: носителем литературного языка и диалектоносителем. Для речи каждого говорящего определена средняя продолжительность гласных звуков, которая принимается за стандартную величину. Это позволило выявить контексты, в которых длительность звука превышает вычисленное значение. Все примеры удлинений классифицированы в зависимости от функций. На основании анализа описывается характер использования нефонологической долготы в речи каждого из интервьюируемых. Предполагается, что сопоставление особенностей данного средства в речи разных говорящих поможет выделить универсальное и специфическое в употреблении удлинений.

Ключевые слова: нефонологическая долгота, хезитация, просодия, устный дискурс, удлинения звуков.

A. Y. Zurik, A. A. Aisanova

Saint Petersburg State University, Saint Petersburg, Russia
Irkutsk State University, Irkutsk, Russia

Why take so looong? Comparative analysis of the use of non-phonological longitude in the speech of two speakers: types of sound lengthening

Abstract. The article is dedicated to identifying the functions of non-phonological length (hereinafter referred to as NPL) of vowel sounds and comparing the usage of this means in a speech of two informants. Existing studies describe the following functions of non-phonological length: hesitation marking, iconic expression of spatial and temporal magnitude, intensification of a feature, accentuation of a significant element. The central aim of the article is to identify general and specific points in the functioning of non-phonological length. In the course of the study, a comparison of vowel sound NPL in the interviewees' speech is carried out. The article analyzes two interview recordings with elderly women: a speaker of the literary norm and a dialect speaker. For each of the interviewees' speech the average duration of vowel sounds is determined, which is taken as a standard value. This made it possible to identify contexts in which the duration of the sound exceeds the calculated value. All examples of prolongation are classified depending on the functions. Based on the analysis, the nature of the non-phonological length usage in each of the interviewees' speech is described. It is assumed that a comparison of the features of this means in different speakers' speech would assist with identifying universal and specific factors in the use of prolongation.

Keywords: non-phonological longitude, hesitation, prosody, oral discourse, sound lengthening.

Статья посвящена сравнению особенностей реализации нефонологической долготы (НФД) в речи двух говорящих. Под нефонологической долготой понимается удлинение звуков, не обладающее смыслоразличительной функцией.

Актуальность данного исследования обусловлена тем, что в дискурсивном описании данного просодического средства остаются лакуны. В большинстве научных работ нефонологическая долгота рассматривается в контексте просодического или речевого портрета говорящего. Такой подход ограничивает исследователей: выводы об особенностях НФД могут распространяться только на речь одного информанта. В связи с этим возникает проблема выявления общего и специфического в функционировании нефонологической долготы.

В качестве материала исследования использованы интервью с носителем литературного языка и носителем диалекта. Информанты – две пожилые женщины. В соответствии с этическими требованиями к научным исследованиям они будут обозначены в тексте как Говорящий 1 и Говорящий 2. Анализируемая речь – устный неподготовленный рассказ, представляющий собой фрагмент социолингвистического интервью.

Для выявления случаев НФД необходимо определить порог, выше которого удлинение является значимым. Поэтому с помощью программы Praat проведены вычисления средней длительности гласных звуков в речи анализируемых говорящих, что позволило выявить случаи нефонологической долготы в речи обоих информантов. Для подсчета выбирались слова, в которых гласные находились в ударной позиции после твердых/мягких согласных или в абсолютном начале слова. Вычисления проводились по принципу определения среднего арифметического значения: сумма трех промежутков длительности звуков делится на количество слагаемых. Результаты анализа представлены в таблицах (табл. 1 и 2).

Таблица 1

Средняя длительность гласных в речи Говорящего 1, мс[1]

Гласный звук	**Пример**	**Длина звука**	**Средняя длит.**
А после t	*тогд**а** был огромный*	0,13	~0,11
	*в**а**жно*	0,12	
	*у нас ст**а**ли строить (0.35) трамвайную линию*	0,07	
О после t	*и постоянно (0.58) в командир**о**вке (0.38)*	0,11	~0,10
	*как стр**о**ился город*	0,11	
	*как строился г**о**род*	0,09	

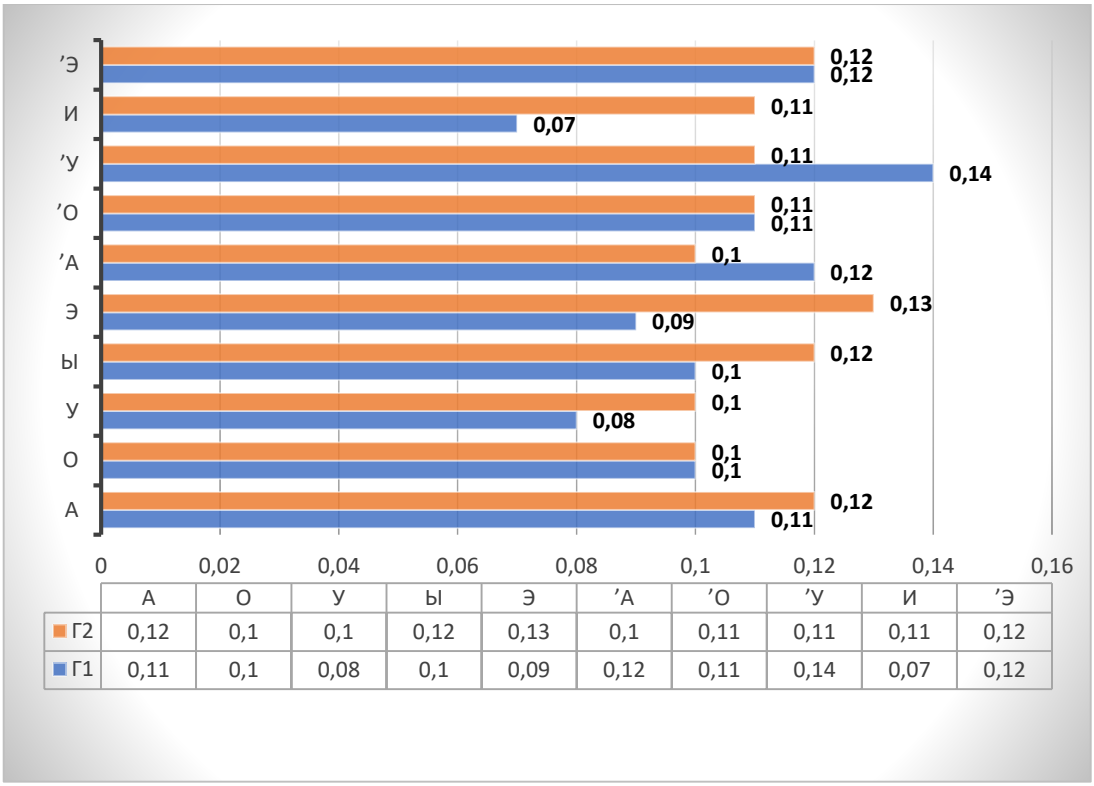
Сравнение средней длительности гласных звуков в речи говорящих позволило установить, что разница между этими величинами составляет от 0 до 0.04 мс (рис. 1). Продолжительность произнесения информантами гласных ['э], ['о], [о] является идентичной.

[1] Объем статьи позволяет привести данные только для двух гласных звуков.

Таблица 2

Средняя длительность гласных в речи Говорящего 2, мс

Гласный звук	**Пример**	**Длина звука**	**Средняя длит.**
А после t	*/ чш-т ста**а**ла не слыышать*	0,10	~0,12
	*в Курымк**а**нский район туды в глушь*	0,13	
	*т**а**м я до четырех классов училась*	0,14	
О после t	*/ ну н**о**ги ходят*	0,11	~0,10
	*/ ну ноги х**о**дят*	0,10	
	*пот**о**м увезли меняя*	0,10	



	А	О	У	Ы	Э	’А	’О	’У	И	’Э
Г2	0,12	0,1	0,1	0,12	0,13	0,1	0,11	0,11	0,11	0,12
Г1	0,11	0,1	0,08	0,1	0,09	0,12	0,11	0,14	0,07	0,12

Рис. 1. Сопоставление средних значений длительности гласных

Полученные результаты позволили выявить контексты, в которых длительность звука превышает пороговое значение. Данные контексты проанализированы с точки зрения функции, реализуемой нефонологической долготой.

На основе существующих исследований НФД, представленных в работах С. В. Кодзасова [2009], В. П. Москвина [2015], А. А. Кибрика и В. И. Подлесской [Рассказы о сновидениях …, 2009], и предварительного анализа материала определен репертуар функций нефонологической долготы:

1) маркирование хезитации (речевого сбоя) – в этом случае использование удлинений предоставляет говорящему время для формулирования мысли, подбора лексемы или грамматической формы (*Н-от (1.65)* ***ннуу*** *[0.28] [0.27] ш [0.43] вот (э 0.48) собственно говоря*);

2) акцентирование значимого элемента, позволяющее передать различные оттенки экспрессии и подчеркнуть важность какого-либо компонента фразы (*Детство было у нас* ***прекрааасное*** *[0.27] замечательное*);

3) иконическое выражение пространственных и временных значений, а также указание на интенсификацию признака (*такой лесной* ***масссиив*** *[0.30][0.21]*).

Для разграничения функций НФД разработаны специальные критерии. О речевом сбое могут свидетельствовать грамматические, лексические и просодические маркеры хезитации. Так, нефонологическая долгота в качестве хезитационного средства может употребляться совместно с перестроением синтаксических конструкций, дискурсивными словами *вот, ну*, маркерами препаративной подстановки (*это, это самое, как его*), фальстартами (*оте=*) и разными видами пауз ((*э 0.34*), (*0.20*)).

Фразовое ударение на слове с удлинением указывает на важность информации, что может свидетельствовать о реализации функции акцентирования значимого элемента.

Необходимо учитывать компоненты значения лексем: семы продолжительности, величины, интенсивности свидетельствуют о том, что долгота может выполнять иконическую функцию.

Обратимся к анализу примеров реализации нефонологической долготы в речи каждого говорящего.

Рассмотрим случаи удлинения звуков в речи Говорящего 1 (табл. 3).

Таблица 3

Примеры удлинения звуков в речи Говорящего 1
в разных фонетических позициях

Гласный звук	**Пример**	**Длительность**
А после t	*(0.51 [вдох]) ааа [0.34] (э 0.34) (0.20) (э 0.23) (0.21) основная масса реки (0.59 [вдох]) э за островом там*	0,34
	Детство было у нас прекрааасное [0.27] замечательное	0,27
	(1.06 [вдох]) (3.21) купаааались [0.31]	0,31
	(0.55) все мы были хорошими пловцаааами [0.32]	0,32
	(0.79 [вдох]) потом началась войнааа [0.31]	0,31
О после t	*Черемховского райооона [0.56]*	0,56
	*(0.66 [вдох]) очень красивое сеелл**оо** [0.16][0.14][0.19]*	0,19
	(0.54 [вдох]) ходил паррооом [0.30] через реку	0,30
	(0.47 [вдох]) обычно (0.19) вот Трооооица [0.34] когда	0,34
	На лооодках [0.26]	0,26
	(0.68 [вдох]) (э 0.32) глубооокая [0.40]	0,40
	(1.70 [вдох]) воооот [0.47]	0,47
	(0.52 [вдох]) воооо [0.35]	0,35
У после t	*Нууу [0.34] эвэвэвэ*	0,34
	(1.35 [вдох]) в 25-м годууу [0.57] (0.65)	0,57
	нууу [0.93]	0,93

В отрезке интервью продолжительностью 4 мин и 1 с найдено 30 случаев НФД гласных звуков, их длительность составляет от 0,16 до 0,93 мс. Поскольку время произнесения этих гласных значительно превышает среднее время произнесения подобных звуковых элементов[1] (на 33–1062,5 %), то квалификация такого рода удлинений не представляет сложностей и часто возможна буквально «на слух». 24 удлинения (80 %) из 30 находятся в абсолютном конце ЭДЕ.

Обратимся к анализу функций нефонологической долготы в речи данного говорящего. Количественное распределение примеров представлено на диаграмме (рис. 2).

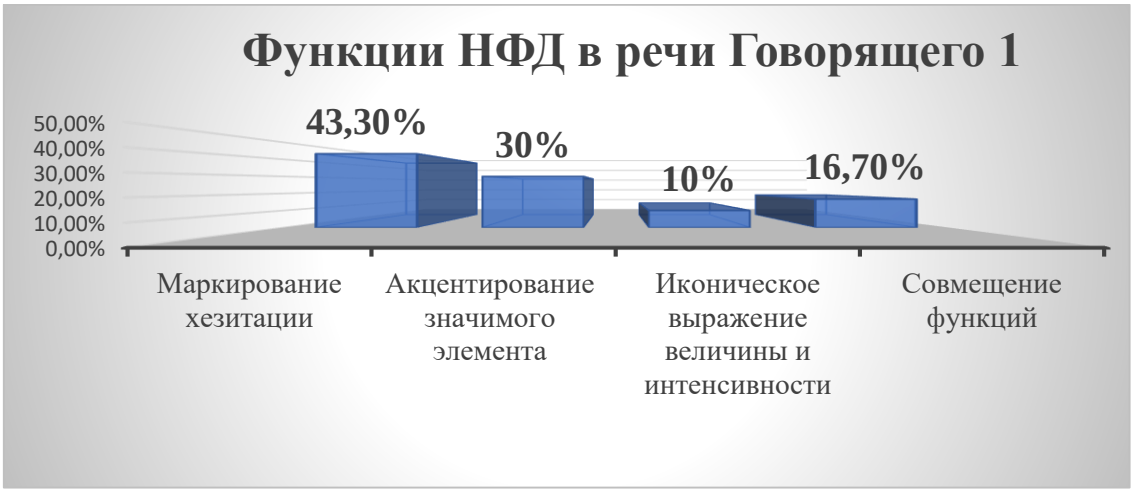


Рис. 2. Функции НФД в речи Говорящего 1

Наибольшее количество удлинений выполняет функцию маркирования хезитации. В исследуемом материале таких контекстов 13.

В следующей ЭДЕ нефонологическая долгота гласного звука сочетается с разнообразными средствами хезитации: фальстартом *на=*, абсолютной паузой *(0.20)*, перестройкой конструкции и длительным глоттальным скрипом *(ʔ 0.78)*. Сочетание НФД с подобными маркерами позволяет предположить, что она выполняет функцию маркирования хезитации.

Дом на[ж] (0.50 [вдох]) прямо ***на= (0.20)*** *б-был* ***(ʔ 0.78)*** *н-на берегу реки [0.33].*

Следует отметить, что некоторые удлинения реализуются в дискурсивных словах, которые являются средствами маркирования речевого сбоя. В подобных случаях нефонологическая долгота не просто сочетается с лексическими маркерами хезитации, а накладывается на них, что, как кажется, свидетельствует о более сильном затруднении:

- *нууу [0.34] эвэвэвэ;*
- *(1.70 [вдох]) вооот [0.47].*

В девяти контекстах НФД выполняет функцию акцентирования значимого элемента. В приведенном ниже примере фразовое ударение выделяет слово с нефонологической долготой гласного звука, изменение тона и повышение ин-

[1] См. табл. 1 и 2.

тенсивности произнесения лексической единицы *прекрасное* свидетельствует о выражении говорящим экспрессии (рис. 3).

Детство было у нас прекрааасное [0.27] замечательное.

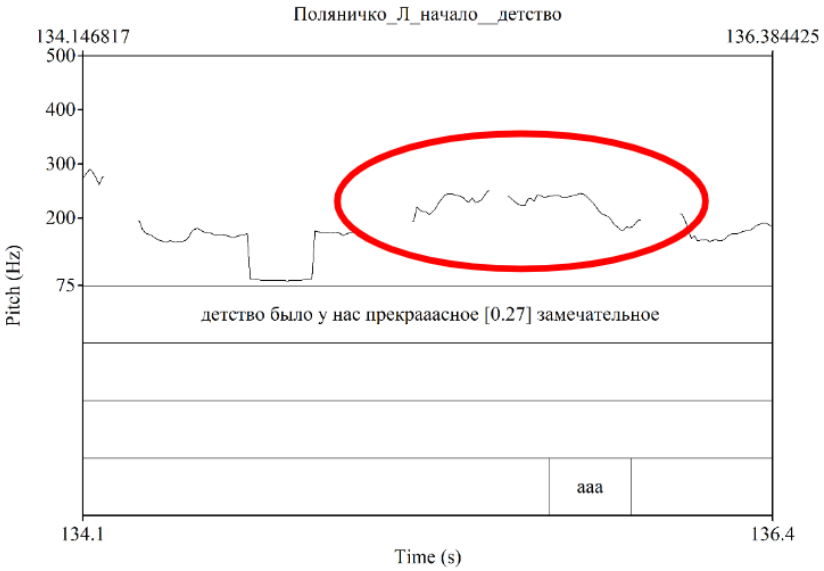


Рис. 3. Тонограмма. *Детство было у нас прекрааасное [0.27]*

Всего в трех случаях НФД выполняет функцию иконического выражения величины и интенсивности. В следующем примере удлинение помогает акцентировать внимание на большом размере объекта. Наличие у нефонологической долготы данной функции подтверждается близостью слова *огромные*, значение которого ‘очень большой’:

***Огромные** там тополяя [0.37].*

Важной особенностью речи анализируемого говорящего является совмещение сразу нескольких функций в одном удлинении. В таких примерах НФД мы находим различные языковые и дискурсивные средства, которые позволяют предполагать, что в данном случае реализуется две или более функции. Подобное совмещение функций в материале представлено как соединение маркирования хезитации с одной из других функций.

Так, в следующем примере удлинение гласного звука находится в одной ЭДЕ с заполненной *(э 0.38)* и абсолютной паузами *(0.38)*, что свидетельствует о речевом сбое. Однако, помимо этого, в лексеме *быыстрая* выявляется сема интенсивности, что говорит об иконической функции НФД. Это позволяет заключить, что удлинение в этом случае выполняет сразу две функции: маркирование хезитации и иконическое выражение величины:

*(1.15 [вдох]) **(э 0.38)** река э Белая очень **(0.38)** быыстрая [0.33].*

Некоторые примеры удлинения гласных звуков, выявленные в речи Говорящего 2, представлены в табл. 4.

Продолжительность проанализированного отрезка интервью составляет 4 мин и 8 с. В нем выявлено 43 удлинения гласных звуков продолжительностью от 0,17 до 0,41 мс, что превышает среднее время произнесения гласных на

70–200 %. Квалификация многих удлинений затруднена, они почти не распознаются «на слух», в то время как длительность звука на 70 % и более превышает вычисленную среднюю продолжительность данного звукового элемента. В 26 случаях (60,47 %) из 43 НФД находится в конце ЭДЕ.

Таблица 4

Примеры удлинения звуков в речи Говорящего 2
в разных фонетических позициях

Гласный звук	**Пример**	**Длительность**
О после t / в абсолютном начале слова	***оо****н [0.20] уумер [0.20] (0.05)*	0,20
	в колхоозе [0.23] работала	0,23
	/охоотились [0.31]	0,31
	ну потом вроде и колхооз [0.25] образовался/	0,25
	шкоола [0.21] была-от	0,21
	/ он там доом [0.19] построил	0,19
	я помню как на коонях [0.17] уехали они	0,17
У после t / в абсолютном начале слова	*нуу [0.25]*	0,25
	нуу [0.33]	0,33
	оон [0.20] ***уу****мер [0.20] (0.05)*	0,20
	у меня все щ"-щас эти родственники все у меня уумерли [0.30]	0,30
	*/ там он ии [0.25] (0.58) /****уу****мер [0.32]*	0,25
Ы после t	*/ чш-т стала не слыышать [0.29]*	0,29
	/Тазыы [0.24]	0,24
	Тазыы [0.19] (0.12) была	0,19
	*пеервую [0.20] откр****ыы****ли [0.18] там Тазыы [0.23]*	0,18
	*пеервую [0.20] открыыли [0.18] там Таз****ыы*** *[0.23]*	0,23
А после t’	*потом увезли меняя [0.25] (0.61)*	0,25
	у Жигжитова Михаила Ильичаа [0.20]	0,20
	яя [0.27] (0.35) в тридцать втором году родилась	0,27
	яя [0.18] (0.40) из семьи одна	0,18
	ну вот племяянница [0.19] у меня есть	0,19
	с дедом моим друз’ж’яя [0.19] были	0,19
	/и там жил-ли у яво с’ш’емьяя [0.25]	0,25
	/трое реебяят [0.20] (0.14) было	0,20
Э после t’	*чего в-вы от меня хотеели? [0.21]*	0,21
	оте= нас осталось чеетверо [0.21]	0,21
	*п****ее****рвую [0.20] открыыли [0.18] там Тазыы [0.23]*	0,20
	*первоначально там просто с’’ш’т****ее****пи [0.24] / леес [0.20]*	0,24
	*первоначально там просто с’’ш’теепи [0.24] / л****ее****с [0.20]*	0,20
	реечцка [0.20]	0,20
	/ уеехали [0.20]	0,20
	Вс’’ш’ее [0.21] (0.13) в Тазах	0,21
	туды леетом [0.20]	0,20
	ну з’’ж’веери [0.33] водятся	0,33

Анализ функций нефонологической долготы позволил получить следующие результаты (рис. 4).

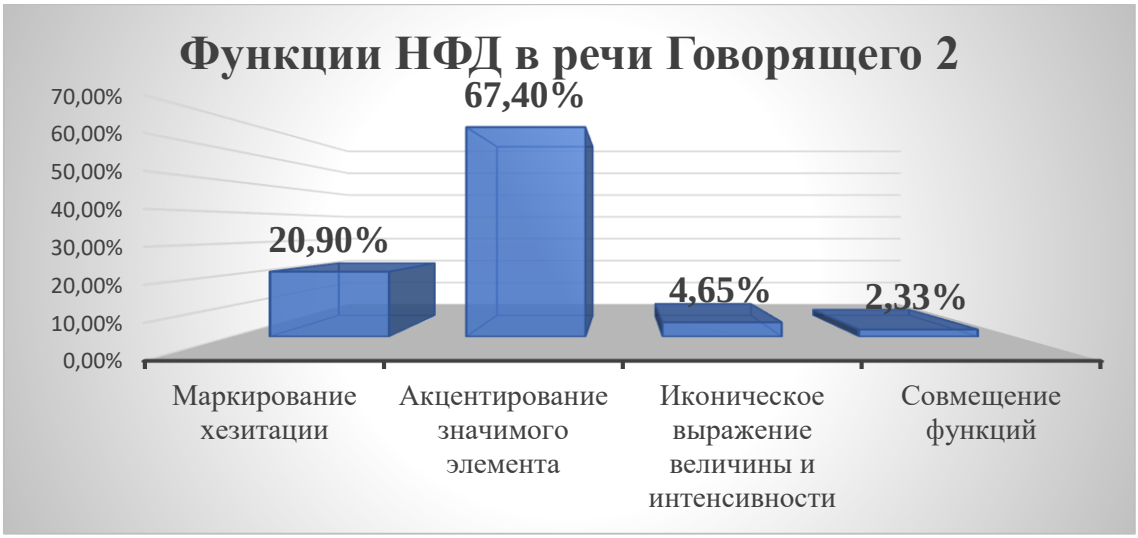


Рис. 4. Функции НФД в речи Говорящего 2

В речи Говорящего 2 количество контекстов, в которых удлинения выполняют функцию маркирования хезитации, значительно меньше, чем в речи Говорящего 1, оно составляет 20,9 % от общего количества примеров с НФД. Стоит отметить, что число удлинений звуков в дискурсивных словах и союзах также меньше, чем в речи Говорящего 1: обнаружено всего три случая.

В анализируемых примерах встречаются разного рода маркеры хезитации. В следующих контекстах НФД используется в сочетании с абсолютными паузами *(0.35)* и *(0.15)*, дискурсивным маркером *ну*, фальстартом *оте=*:

- *(0.43 [вдох]) яя [0.27] (0.35) в тридцать втором году родилась;*
- *ну (0.15) подрослаа [0.19];*
- *оте= нас осталось чеетверо [0.21].*

Среди выделенных случаев удлинений преобладают такие, в которых долгота реализует акцентирующую функцию. Всего найдено 29 подобных контекстов.

В приведенном ниже примере фразовое ударение выделяет слово с НФД. Изменение тона и повышение интенсивности произнесения лексической единицы *эвенки* показывает ее важность для говорящего и свидетельствует о том, что удлинение здесь выполняет функцию акцентирования значимого элемента (рис. 5):

(0.35 [вдох]) / там Ээвенки [0.24] большинство жили.

В тексте интервью с Говорящим 2 всего два удлинения выполняют функцию иконического выражения величины. Как можно предположить, они используются для передачи представления о большом размере описываемых объектов:

(0.50 [вдох]) первоначально там просто с’’ш’теепи [0.24] / леес [0.20].

Среди примеров удлинений в речи Говорящего 2 также найден случай совмещения нескольких функций:

*(0.64) / там он **ии** [0.25] (0.58) /уумер [0.32].*

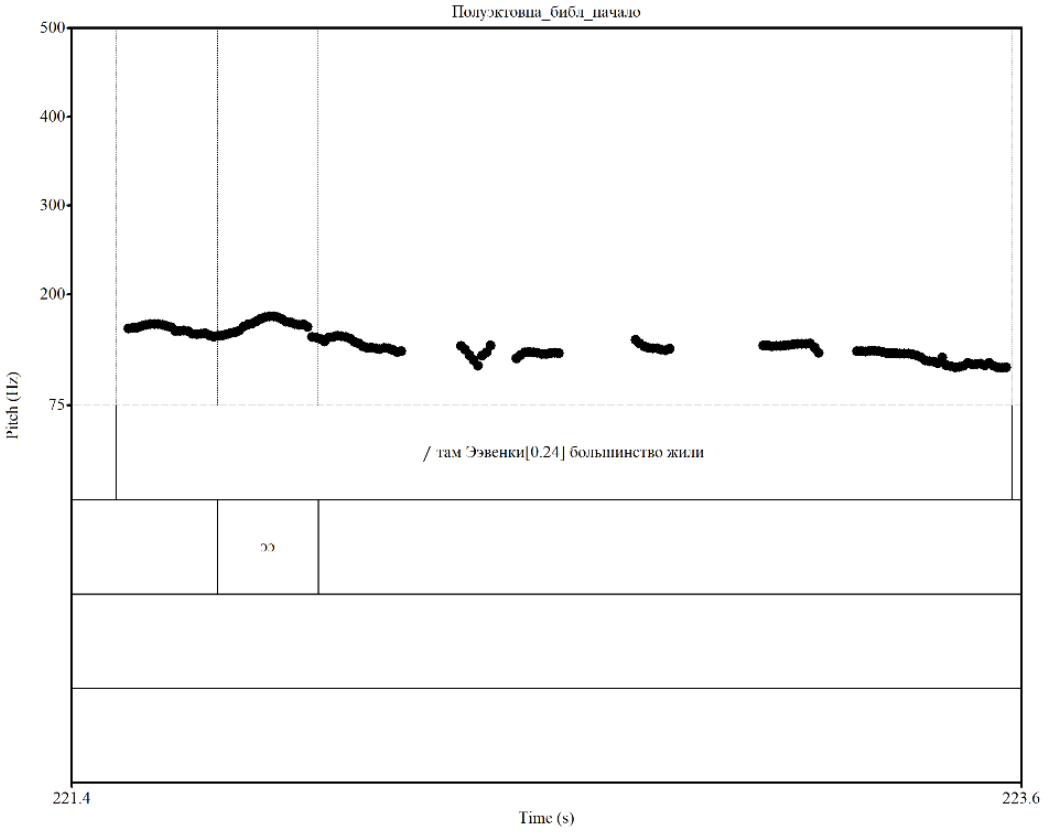


Рис. 5. Тонограмма. *(0.35 [вдох]) / там Ээвенки [0.24] большинство жили*

В данной ЭДЕ встречаются абсолютные паузы, одна из которых разделяет слова с НФД. Как можно предположить, они свидетельствуют о речевом сбое. Вторая пауза, видимо, дает говорящему время для того, чтобы подготовиться к эмоционально тяжелому продолжению разговора и подобрать подходящее слово. При акцентировании перед словом *умер* говорящий понижает громкость.

Итак, в рамках исследования предпринята попытка сопоставительного анализа особенностей реализации нефонологической долготы в речи двух говорящих. Вычислена средняя длительность гласных, характерная для речи каждого из информантов. Определены реализуемые удлинениями функции, выявлена специфика использования нефонологической долготы в речи каждого из интервьюируемых.

Обобщая результаты исследования, можно сделать следующие выводы:

1. В речи анализируемых говорящих удлинения тяготеют к концу ЭДЕ. Редко можно встретить НФД в начале или середине фразы. Предполагается, что это связано с актуальным членением высказывания: более значимая информация располагается ближе к финальной части ЭДЕ, а значит, при реализации акцентирующей функции НФД удлинение чаще всего будет накладываться на эту часть синтагмы/фразы. Кроме того, в случае речевого сбоя, связанного с продумыванием следующей фразы, появление просодических маркеров хезитации ближе к концу синтагмы позволяет говорящему увеличить время для подготовки продолжения высказывания.

2. Удлинения в речи Говорящего 1 обладают большей протяженностью, чем удлинения в речи Говорящего 2: средняя длительность НФД в 1,5 раза выше.

3. Различается функциональное распределение удлинений в речи двух говорящих. Для речи Говорящего 1 характерны удлинения, выполняющие функции маркирования хезитации и акцентирования значимого элемента. В группу контекстов, где НФД необходима для преодоления речевого сбоя, включаются также примеры, в которых эта функция совмещается с другими. В речи Говорящего 2 преобладают удлинения, помогающие акцентировать наиболее значимый элемент высказывания.

Данная работа открывает перспективы описания проблем, связанных с выявлением общего и частного в использовании удлинений. К ним также можно отнести выявление иных функций нефонологической долготы и дополнительных контекстных показателей уже описанных функций, сопоставительный анализ функционирования нефонологической долготы в текстах и их частях, отличающихся эмоциональной окраской и степенью важности для говорящего.

Литература

Кодзасов С. В. Исследования в области русской просодии. М. : Языки славянских культур, 2009. 496 с.

Москвин В. П. Удлинение звука в русском слове: риторический и орфоэпический аспекты // Лингвистические заметки. 2015. С. 54–60.

Рассказы о сновидениях: корпусное исследование устного русского дискурса / под ред. А. А. Кибрика и В. И. Подлесской. М. : Языки славянских культур, 2009. 736 с.

УДК 81-25

Д. А. Лоскутова

Национальный исследовательский университет «Высшая школа экономики», Санкт-Петербург, Россия

На минуточку: особенности функционирования единицы в русской повседневной речи

Аннотация. Описываются особенности функционирования словоформы *на минуточку* в русской повседневной речи, которые удобно представить на своеобразной шкале динамических преобразований, репрезентирующей стадии грамматикализации: на первом этапе можно видеть результат процесса грамматикализации: существительное *минуточка* в форме В. п. с предлогом *на* начинает функционировать в качестве наречия, обозначающего ‘ненадолго’, ‘на короткое время’: *извините/ я на минуточку;* на следующем этапе словоформа *на минуточку* подверглась сразу двум процессам: новой грамматикализации и идиоматизации. Делается вывод, что в результате утрачивается семантика времени, происходит процесс десемантизации и переосмысление значения: *на минуточку* становится вводным выражением со значением ‘между прочим’, ‘кстати’. Анализ корпусного материала выявил пять значений формы *на минуточку*, не отмеченных ранее в словарях и выявляемых на этом этапе преобразований: 1) побуждение собеседника сосредоточить свое внимание и по достоинству оценить то, что будет сообщено далее; 2) выделение предмета речи из ряда подобных, подчеркивание его особого статуса и значимости; 3) введение сильного и единственного аргумента; 4) уточнение; 5) маркирование несоответствия ожиданиям.

Ключевые слова: повседневная речь, речевой корпус, грамматикализация, идиоматизация, метод шкалирования, динамическая шкала переходности.

D. A. Loskutova

National Research University «Higher School of Economics», St. Petersburg, Russia

Na minutochku: Features of the Functioning of a Discursive Unit in Russian Everyday Speech

Abstract. The article describes the features of the functioning in Russian everyday speech of the word-form *na minutochku*, which can be conveniently represented on a kind of scale of dynamic transformations. At the first turn of such a scale we can see the result of the process of grammaticalization: the noun *minutochka* in the form of Acc. with the preposition *na* begins to function as an adverb denoting 'for a short time'. At the next stage, the word-form for na minutochku underwent two processes at once: new grammaticalization and idiomatization. As a result, the semantics of time is lost, the process of desemanticization and reinterpretation of meaning takes place: *na minutochku* becomes an introductory expression with the meaning of 'by the way'. The analysis of corpus material has revealed 5 meanings of the form *na minutochku*, which have not been previously noted in dictionaries and are revealed at this stage of transformation: (1) inducing the interlocutor to focus his attention and appreciate what will be communicated next; (2) singling out the subject of speech from a number of similar ones, emphasizing its special status and importance; (3) introducing a strong and single argument; (4) clarifying; (5) marking the discrepancy with expectations.

Keywords: everyday speech, speech corpus, grammaticalisation, idiomatisation, scaling method, dynamic transitivity scale.

Введение

Объектом настоящего исследования является предложно-падежное сочетание *на минуточку*. Источником материала для анализа стали два подкорпуса Национального корпуса русского языка (НКРЯ): устный (УП) и социальные сети (СС). Выбор данных подкорпусов обусловлен тем, что в УП представлены (в числе прочего) контексты реальной устной повседневной речи, а в подкорпусе СС – контексты устно-письменной формы речи, которая «возникает из синтеза письменной формы речи и устной речи неофициального характера» [Барышева, 2021, с. 35] и сегодня по распространенности в сфере коммуникации вполне может соперничать с устной формой. Ср.: «раньше речь была устной и письменной, а сейчас появилась устно-письменная речь, т. е. технически она является письменной (пишется буквами), но во многом ведет себя как речь устная: она спонтанная, линейная (человек не исправляет, не перечитывает)» [Левонтина].

Пользовательский подкорпус настоящего исследования включает 378 контекстов, 136 (36 %) из которых взяты из УП и 242 (64 %) – из СС.

Данные словарей, научной литературы и анализ материала

Рассматриваемая единица имеет фиксацию в словарях разного типа. Так, фразеологический словарь [Фёдоров, 2008], толковые словари разговорной речи [Химик, 2004, 2017; Крысин, 2014] и словарь русского арго [Елистратов] фиксируют следующие ее значения:

1) <можно> <вас, тебя> – обращение с просьбой уделить время кому-л. для чего-л. (*Молодой человек, можно вас* ***на минуточку****?*);

2) употребляется для выделения, подчеркивания смысла какой-либо части высказывания (*А ведь китайцев,* ***на минуточку****, в 8 раз больше*) (В. Быков, О. Деркач. Книга века);

3) шутл.-ирон. Между прочим, кстати. Употребляется для представления чего-л. как важного, существенного под видом случайного, временного (*Она,* ***на минуточку****, стукач, вы знали об этом?*);

4) «выражение-паразит», употр. в речи по любому поводу (типа «так сказать» и т. п.), часто для подчеркивания несоответствия чего-л. чему-л., ирон. выделения каких-л. свойств и т. п., близко по зн. к разг. «между прочим» (*У нее* ***на минуточку*** *ребенок новорожденный, а она пьет, как лошадь*);

5) разг. экспрес. Ненадолго, на очень короткое время. (*– Я к вам, Анна Сергеевна, дело у меня,* ***на минуточку****, разрешите? – и Оползнев боком протиснулся в сени*) (А. Мотылькова. Ранний снег).

Лингвистами данная единица также замечена, и предложены, в частности, вариации ее значения ‘акцентуация’. Так, форма *на минуточку* во вводно-модальном значении была прокомментирована Г. Н. Сергеевой и Е. А. Шнырик, которые выделили четыре подтипа акцентуации данной формы [Сергеева, Шнырик, 2021, с. 110–113].

1. «*На минуточку* (*на секундочку*) побуждает собеседника сосредоточить свое внимание и по достоинству оценить то, что будет сообщено далее. Говорящий сознательно прерывает речевую цепь, чтобы создать напряжение, интри-

гу и после паузы преподнести информацию, которая должна удивить, поразить адресата»:

Главный колокол, отлитый, между прочим, из двадцати восьми трофейных турецких пушек, весил, ***на минуточку****, тысячу сто пятьдесят девять пудов* (УП).

Стоит отметить, что в данном контексте говорящий дважды использует прием привлечения внимания собеседника: первый раз с помощью классического вводного слова *между прочим* [Словарь русского языка, 1986, с. 245] (в контексте подчеркнуто), второй раз – с помощью рассматриваемой формы *на минуточку*. В результате выстраивается синонимичный ряд, в который эта словоформа легко вписывается.

Видно также, что авторы статьи в одном ряду (тоже как синонимы) рассматривают формы *на минуточку* и *на секундочку* – во всех предлагаемых значениях. Последняя форма, как представляется, все же заслуживает отдельного исследовательского внимания, что можно рассматривать как перспективу настоящего исследования.

2. «Говорящий прибегает к вводно-модальному слову *на минуточку* (*на секундочку*), если необходимо выделить предмет речи из ряда подобных, подчеркнуть его особый статус, особое положение, особую роль и значимость – иными словами, продемонстрировать качественное превосходство над другими»:

Менеджер «Рейнджерс» делал все, кроме того, чтобы вести переговоры с Яромиром Ягром – ***на минуточку****, лучшим бомбардиром команды в регулярном чемпионате и в плей-офф* (УП).

3. «Говорящий использует *на минуточку* (*на секундочку*), предваряя сильный и единственный аргумент»:

Анна Болейн все же была сильная личность. ***На минуточку****: церковь в государстве поменять!* (УП).

4. *«На минуточку* (*на секундочку*) используется говорящим, с тем чтобы озадачить собеседника несообразностью реалий»:

Кстати, трое из этих четырех до сих пор мои лучшие друзья, а ведь прошло, ***на минуточку****, больше тридцати лет* (УП).

В ходе анализа материала в рамках настоящего исследования было найдено еще одно, новое, значение, не зафиксированное словарями и не отмеченное ранее лингвистами – уточнение, позволяющее избежать недопонимания:

А я, поволновавшись и повзбивав крылами воздух, послушала по Мудлу (сайт для дистанционного обучения, ***на минуточку***😂*) свою очень умную преподавательницу* (СС).

В этом контексте девушка поясняет своему собеседнику название программы для дистанционного обучения. Пиктограмма и комментарий в скобках позволяют понять, какие чувства испытывает девушка: видимо, неловкость от того, что она вынуждена уточнять, что означает *Мудл*. Этот пример нельзя отнести ни к словарным значениям словоформы *на минуточку*, ни к приведенной выше классификации Г. Н. Сергеевой и Е. А. Шнырик. Это новый оттенок зна-

чения рассматриваемой словоформы: *на минуточку* вводит уточнение, имеет место несоответствие ожиданиям.

Следующий пример также демонстрирует это новое значение словоформы *на минуточку*. Можно предположить, что по фотографии нельзя однозначно сказать, что́ именно изображено, поэтому девушка вынуждена написать об этом:

***На минуточку**, это фото еды)) у чувака в руках мой тести кебаб – он разбивает горшочек и потом даст мне вкуснейшее все, что внутри)* (СС).

Видно, что даже незначительное расширение материала, на котором проводится анализ, позволило расширить и круг значений рассматриваемой словоформы. Не исключено, что это не последнее новое, что можно увидеть в ее функционировании в нашей повседневной речи, как устной, так и устно-письменной.

Шкала переходности

На основе проведенного контекстного анализа в настоящем исследовании была построена цепочка преобразований – своеобразная динамическая *шкала переходности* (рис.). Эта шкала позволяет наглядно показать изменения, происходящие со словоформой *на минуточку* в современном дискурсе.

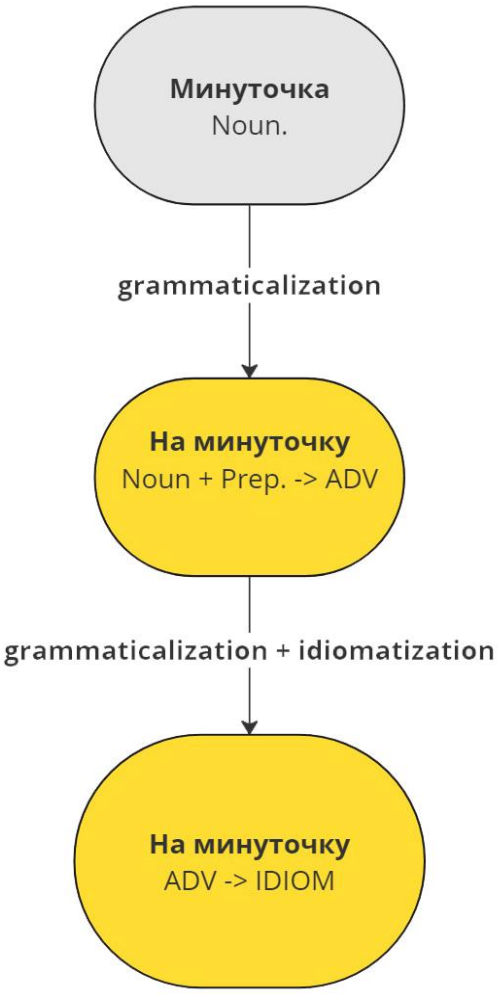


Рис. *Динамическая шкала переходности для словоформы* ***на минуточку***

На первом витке преобразований можно видеть результат процесса *грамматикализации*: существительное *минуточка* в сочетании с предлогом *на* начинает функционировать в качестве наречия, обозначающего ‘ненадолго’, ‘на короткое время’: *на минуточку:*

1) *Ну может быть/ ты хоть* ***на минуточку*** *войдешь?* (УП);

2) *Выйди* ***на минуточку*** *из своей комнаты / я тут у тебя пропылесошу* (УП);

3) *М-м-м… извините / я* ***на минуточку*** (УП);

Далее исследуемая словоформа входит в устойчивое сочетание <*можно + вас, тебя + на минуточку?*>:

4) *Андрюшенька! Э…* ***можно тебя на минуточку?*** (УП);

5) *Тетерев!* ***Можно вас на минуточку?*** *У меня к вам большая просьба* (УП);

Со временем развернутая просьба уделить время кому-либо для чего-либо сократилась до изолированной словоформы *на минуточку* (действует принцип экономии[1], один из ведущих в ходе языковой эволюции), а остальные компоненты устойчивого сочетания стали факультативными и в разговорной речи чаще всего опускаются. Их место может быть занято обращением к конкретному лицу (в контекстах (6)–(8) подчеркнуто) или оставаться пустым (9), ср.:

6) *Здравствуйте. Товарищ Колчин /* ***на минуточку****. Это как раз к вам тоже относится* (УП);

7) *Ананий Палыч /* ***на минуточку. На минуточку*** (УП);

8) *Мам /* ***на минуточку*** (УП);

9) ***Можно вас на минуточку?*** (УП);

На следующем этапе преобразований словоформа *на минуточку* подверглась сразу двум процессам: грамматикализации и идиоматизации. В ходе этих процессов *на минуточку* утрачивает семантику времени, происходит процесс десемантизации и переосмысление значения. *На минуточку* уже не выступает в роли наречия, означающего ‘ненадолго’, оно становится фразеологической единицей, вводным выражением, обозначающим ‘между прочим’, ‘кстати’, ср.:

10) *Причем, так,* ***на минуточку****, это люди, которые имеют СЕРТИФИКАТ (международного образца) на право выполнения определенных работ по обслуживанию авиасудна* (СС);

11) ***На минуточку*** *– от Твери до Москвы ехать больше двух часов!* (СС);

12) *Потому, что Китай-; это,* ***на минуточку****, одна из древнейших цивилизаций, с уникальной культурой и огромной историей* (СС);

[1] Ср.: в работе языка как живого организма заметно «стремление к экономии сил и нерастрачиванию их без нужды» [Бодуэн де Куртенэ, 1963, с. 226]; в процессе развития, следуя «естественным законам эволюции», язык движется от более сложных к более простым способам выражения того же содержания [Спенсер, 1986, с. 172]; «для языковой деятельности характерна тенденция к бережливости», «в языке для всех случаев вырабатываются способы выражения, которые содержат ровно столько, сколько необходимо для понимания» [Пауль, 1960, с. 302]; язык постоянно стремится освободиться от лишнего [Passy, 1890, p. 229] и т. д., и т. п. Наиболее наглядно это видно именно на «оси речи» [Frei, 1929] (цит. по: [Богданова-Бегларян, 2016]). О принципе экономии см. также: [Будагов, 1972].

13) *«demon in a bottle»* ***на минуточку****… Джарвис должен быть человеком, да, потому что он человеком в комиксах всегда был* (СС);

14) *Я не могла поверить, что мои ужасные заломы на лбу, которые,* ***на минуточку****, у меня с 8 класса из-за плохого зрения, просто исчезли спустя 3 минуты после процедуры!!!* (СС);

15) *Некоторые начинали трудоустройство, абсолютно не руководствуясь тем фактом, что я как бы еще в магистратуре учусь так* ***на минуточку(!)*** (СС).

Таким образом, словоформа *на минуточку* подверглась в устной речи активным динамическим процессам – грамматикализации и идиоматизации, в ходе которых утратила свое первоначальное лексическое значение и приобрела новое – подчеркивание важного слова (акцентуация).

Количественные данные и выводы

Количественные данные по анализу двух подкорпусов представлены в таблице.

Таблица

Количественные данные по пользовательскому подкорпусу

Значение	**Подкорпус**	
	УП, абс./отн. кол-во, %	**СС**, абс./отн. кол-во, %
‘Просьба уделить время’	74 / 54,4	2 / 0,8
‘Ненадолго’	59 / 43,4	62 / 25,6
‘Акцентуация’	3 / 2,2	176 / 72,7
‘Акцентуация (уточнение)’	–	2 / 0,8

Как видно из таблицы, в устной речи словоформа *на минуточку* чаще всего (54,4 и 43,4 % соответственно) выступает в качестве наречия со значениями ‘просьба уделить время’ и ‘ненадолго’. В устно-письменной форме речи (СС) значительно больше примеров употребления формы *на минуточку* в качестве идиомы со значением ‘акцентуация’ (72,7 %). Кроме того, именно в устно-письменной форме речи было замечено новое значение – ‘уточнение’.

В ходе исследования было выяснено также, что форма *на минуточку* может занимать в тексте любое положение. Так, она может стоять перед акцентируемой фразой (10)–(11), встраиваться в выделяемое словосочетание посередине (12) или стоять после выделяемой фразы (13).

Кроме того, форма *на минуточку* сочетается с просодическими сигналами устно-письменной речи: пиктограммами (14), типографскими знаками (15). Они помогают партнеру по коммуникации точнее определить значение рассматриваемой единицы.

Литература

Барышева С. Ф. «Устно-письменная» форма речи в интернет-коммуникации как проявление тенденции к разговорности и диалогичности // Мир лингвистики и коммуникации. 2021. № 64. С. 34–47.

Богданова-Бегларян Н. В. Фонетический *Тяни-Толкай*: о двух разнонаправленных тенденциях в построении устного текста // Фонетика сегодня : материалы докладов и сообщений VIII Междунар. науч. конф., 28–30 октября 2016 г. М. : Нестор-История, 2016. С. 12–14.

Бодуэн де Куртенэ И. А. Об общих причинах языковых изменений // И. А. Бодуэн де Куртенэ. Избранные труды по общему языкознанию. М. : Изд-во Академии наук СССР, 1963. Т. 1. С. 222–254.

Будагов Р. А. Определяет ли принцип экономии развитие и функционирование языка? // Вопросы языкознания. 1972. № 1. С. 17–36.

Пауль Г. Принципы истории языка. М. : Изд-во иностранной литературы, 1960. 501 с.

Сергеева Г. Н., Шнырик Е. А. Риторический потенциал лексикализованной словоформы *На минуточку (На секундочку)* во вводно-модальном употреблении // Русский синтаксис: от конструкций к функционированию : сб. материалов Всерос. науч. конф. с междунар. участием, посвящ. 95-летию д-ра филол. наук, проф. Аллы Федоровны Прияткиной, Владивосток, 11–13 ноября 2021 г. Владивосток : Дальневосточный фед. ун-т, 2021. С. 108–113.

Словарь русского языка. В 4 т. Т. 2. К–О / под ред. А. П. Евгеньевой. 3-е изд., стер. М. : Русский язык, 1986. 736 с.

Спенсер Г. Основные начала. Киев : Вища школа, 1986. 375 с.

Толковый словарь русской разговорной речи / ред. Л. П. Крысин. Вып. 1–3. М. : Изд. Дом ЯСК, 2014–2019.

Федоров А. И. Фразеологический словарь русского литературного языка. 3-е изд., испр. М. : АСТ, 2008. 828 с.

Химик В. В. Большой словарь русской разговорной экспрессивной речи. СПб. : Норинт, 2004. 708 с.

Химик В. В. Толковый словарь русской разговорно-обиходной речи : в 2 т. Т. 1 А–Н. СПб. : Златоуст, 2017. 528 с.

Frei H. La Grammaire des Fautes. Paris, 1929. 317 p.

Passy P. Etude sur les Changements Phonetiques et Leurs Caracteres Generaux. Paris : Librairie Firmin-Didot. 1890. 270 p.

Электронные источники

Елистратов В. С. Словарь русского арго: материалы 1980–1990 гг. // Грамота.ру. URL: https://rus-russian-argo.slovaronline.com/ (дата обращения: 14.05.2024).

Левонтина И. Б. «Раньше речь была устной и письменной, а сейчас появилась еще устно-письменная» // Комсомольская правда. URL: https://www.kp.ru/daily/27266/4400718/ (дата обращения: 14.09.2024).

УДК 81'374.3

А. Р. Пестова

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

Лексикограф 2.0: потенциал нейросетей в толковой лексикографии

Аннотация. Рассматриваются возможности искусственного интеллекта в области составления толкового словаря. Анализируется потенциал ряда языковых моделей в решении таких лексикографических задач, как написание толкований, определение стилистической окраски слова, описание сочетаемости слова и подбор иллюстративных примеров. Показывается, что самой сильной стороной нейросетей является описание лексической сочетаемости. Хуже всего искусственный интеллект справляется с подбором иллюстративных примеров. Отмечается такой феномен, серьезно ограничивающий потенциал нейросетей, как их галлюцинирование. Слабые стороны искусственного интеллекта в лексикографии связаны с тем, что в наборе данных языковых моделей отсутствуют лексикографически значимые тексты. Делается вывод, что на современном этапе развития нейросетей возможно их использование в роли помощника, ускоряющего решение некоторых рутинных задач, при условии обучения языковых моделей на словарных данных.

Ключевые слова: лексикография, толковый словарь, искусственный интеллект, нейросеть, цифровизация.

A. R. Pestova

Institute of the Russian Language named after V. V. Vinogradov of the RAS, Moscow, Russia

Lexicographer 2.0: the potential of neural networks in explanatory lexicography

Abstract. The article deals with the possibilities of artificial intelligence in such a field as compiling an explanatory dictionary. It analyzes the potential of a number of language models in solving such lexicographic tasks as: writing interpretations, determining the stylistic coloring of a word, describing the combinability of a word and selection of illustrative examples. As our study, the strongest strength of neural networks is the description of lexical combinability. The worst performance of artificial intelligence is in the selection of illustrative examples. There is a phenomenon that severely limits the potential of neural networks, such as hallucination. The weaknesses of AI in lexicography are related to the fact the fact that the language model dataset does not contain lexicographically meaningful texts. At this stage in the development of neural networks, it is possible to use them in the role of an assistant, speeding up some routine tasks, provided that they train language models on dictionary data.

Keywords: lexicography, explanatory dictionary, artificial intelligence, neural network, digitalization.

Тезис «будущее – за электронной лексикографией» уже перестал быть дискуссионным. Однако в последнее время, в свете бурного развития искусственного интеллекта, актуальным становится новый вопрос – об электронном *лексикографе*. Популярность нейросетей набирает обороты с невероятной скоростью, и специалисты из разных областей говорят о широких возможностях

ИИ. Нейросеть может выполнять функции переводчика, копирайтера, юриста, психолога, программиста; может писать философские эссе и научные статьи.

Это порождает закономерный вопрос: может ли ИИ заменить человека и в такой сфере, как толковая лексикография?

Чтобы проанализировать границы возможностей нейросетей в составлении толковых словарей, мы обратились к таким языковым моделям, как Edya GPT AI, Yandex GPT, Perplexity AI, GigaChat.

В первую очередь мы дали моделям задание: «Тебе поручили составить академический толковый словарь русского языка. Распиши подробно свои действия для выполнения этого задания, а также предполагаемые сроки его выполнения». Следует отметить, что отвечают они на этот вопрос довольно пространно: делят процесс на этапы, описывают собственные преимущества и недостатки как лексикографа. Ввиду ограничений на объем статьи мы воздержимся от цитирования и сравнения ответов моделей. Обобщим лишь, что в целом у моделей есть достаточно четкое представление о процессе составления словаря лексикографом-человеком, однако когда речь заходит о работе без человека, нейросеть признает: «Не могу полностью заменить человеческий интеллект в создании словника, так как словарь – это не только сбор слов, но и результат глубокого анализа языка, его культуры и истории. В итоге нейросеть может значительно ускорить процесс создания словаря, но она не сможет полностью заменить человека, особенно в вопросах нюансов языка и стилистической окраски» (цитируется ответ Edya GPT AI, однако другие модели давали аналогичные ответы).

Рассмотрим теперь непосредственно лексикографические возможности ИИ, а именно:

1) написание толкований;
2) определение стилистической окраски слова;
3) описание сочетаемости слова;
4) подбор иллюстративных примеров.

Задачи, которые мы предлагали моделям, связаны с работой над двумя словарными проектами: «Академическим толковым словарем русского языка» [Академический толковый словарь, 2016] и «Толковым словарем русской разговорной речи» [Толковый словарь…, 2014–2022].

1. Написание толкований. Модели на приемлемом уровне справляются с этой задачей. В качестве примера приведем толкование нового глагола *генерить*: «создавать, производить что-либо автоматически с помощью компьютера, особенно используя искусственный интеллект; создавать что-либо в большом количестве, часто с помощью компьютера» (ответ Edya GPT AI). В целом это толкование адекватно отражает значение данного глагола, хотя, несомненно, требует существенной доработки. Есть и гораздо менее удачное толкование: «производить, создавать что-либо в большом количестве, обычно что-то незначительное, ненужное или вредное» (ответ Perplexity AI).

В некоторых случаях нейросеть придумывает словам ошибочные толкования, опираясь, видимо, на внутреннюю форму слова. Например, у существи-

тельного *голосовуха* ‘голосовое сообщение, отправленное с использованием средств электронной связи’ модель выделяет три несуществующих значения: «1. Процесс голосования, особенно в больших масштабах, например на выборах. 2. Процесс активного обсуждения, спора или дискуссии, сопровождающийся громкими высказываниями и перепалками. 3. Процесс, связанный с выражением мнения голосованием, высказыванием одобрения или неодобрения» (ответ Edya GPT AI).

Особенно плохо обстоит дело с толкованием служебных слов. Например, отвечая на запрос «Напиши словарную статью русского предлога ПО», нейросеть выделяет мало значений: «1. Движение, перемещение: указывает на направление движения, перемещения чего-либо. 2. Расположение: указывает на место, где что-либо находится. 3. Время: указывает на период времени, в течение которого происходит действие. 4. Способ действия, выполнения: указывает на способ действия, выполнения чего-либо. 5. Цель, назначение: указывает на цель, назначение действия. 6. Соотношение, отношение: указывает на соотношение, отношение между объектами» (ответ Edya GPT AI). Ср., например, в Словаре русского языка [Словарь русского языка …, 1981–1984] выделяется 28 значений данного предлога, в Большом академическом словаре русского языка [Большой академический словарь …, 2004–2021] – 21 значение. Оставим здесь в стороне вопрос о речениях, которые не всегда адекватно иллюстрируют сформулированное значение, но процитируем все же одну из особенностей употребления, отмеченных нейросетью: «Предлог *по* может употребляться с прилагательными в сравнительной степени: По красивее, по меньше, по больше». Очевидно, таким образом, что нейросети на данном этапе их развития категорически не справляются с анализом служебных слов.

Отдельной проблемой является такой феномен, как галлюцинации (галлюцинирование) нейросетей. Он заключается в том, что модель может придумывать несуществующие факты, как правило, не признаваясь в том, что у нее отсутствуют данные, необходимые для корректного ответа на вопрос.

Чтобы проиллюстрировать этот феномен, приведем следующий пример. Читая послесловие словаря «Язык старой Москвы» В. С. Елистратова, мы наткнулись на слово *наболда* в следующем контексте: «Я стал сомневаться в себе. “Может быть, – рассуждал я, – все очень тонко организованы и что-то чувствуют или там предчувствуют, а я просто мещанский чурбан, наболда, духовная и душевная посредственность?”» [Елистратов, 2004, с. 755]. Это слово отсутствует во всех словарях русского языка, начиная от «Словаря Академии Российской» (1789–1794 гг.) и заканчивая современными словарями. Мы обратились к нейросетям с вопросом: «Что такое наболда?» – и получили самые разные ответы. Приведем их фрагменты:

1) «Наболда – это устаревшее, но все еще используемое в разговорной речи слово, которое означает “деньги”» (ответ Perplexity AI);

2) «Наболда (также известная как “желтая докторская”) – это травянистое растение семейства астровых, которое обладает лечебными свойствами.

В народной медицине наболда используется как антисептик, противовоспалительное и обезболивающее средство» (ответ Edya GPT AI);

3) «Наболда – это традиционное русское блюдо, состоящее из картофеля, мяса, лука, моркови и специй. Оно обычно готовится в русской печи или на медленном огне. Получается очень вкусно и сытно!» (ответ Edya GPT AI; по запросу нейросеть дает подробный рецепт приготовления).

Такие варианты ответов можно множить. Уличенная в фальсификации нейросеть, как правило, извиняется и выдает новую галлюцинацию. Готовность модели к извинениям и неготовность признать собственное бессилие в решении поставленной перед ней задачи связана, по-видимому, с тем, что на самом деле главная ее цель – понравиться пользователю, а не избавиться от всех ошибок. Именно поэтому нейросети порождают гладкие тексты, которые могут содержать как грубые ошибки, так и мелкие неточности, а потому требуют непрестанной редактуры.

2. Определение стилистической окраски слова. С определением стилистической окраски слова нейросети справляются также с переменным успехом. Так, составляя словарную статью вышеупомянутого глагола *генерить*, модель маркирует его стилистическими пометами *комп., интернет*. В академических словарях отсутствует помета *интернет,* однако нельзя не отметить, что это, скорее, вопрос академической традиции, связанной с набором стилистических помет, чем «ошибка» модели. Что касается пометы *комп.*, то она в данном случае представляется вполне корректной.

На вопрос о стилистической окраске слова *дистанционка* Edya GPT AI отвечает: «Это разговорное слово, которое означает дистанционное обучение. Стилистическая окраска у этого слова разговорная, неформальная, с оттенком просторечия. Его часто используют в повседневной речи, особенно среди студентов, школьников и их родителей. В формальной обстановке, например в официальных документах, более уместно использовать словосочетание “дистанционное обучение”». Ответ, на наш взгляд, корректно и исчерпывающе описывает стилистическую принадлежность слова *дистанционка*. Могут возникнуть критические вопросы, связанные с терминологией и формулировками (например, что подразумевается под словосочетанием «с оттенком просторечия»), однако нужно учитывать, что модель, по всей вероятности, никто не обучал на лексикографически значимых текстах и не давал никаких ориентиров вроде списка общепринятых стилистических помет. Думается, что, если нейросеть будет обучена на материале авторитетных академических словарей, точность и корректность формулировок станет гораздо выше.

Однако здесь есть другая проблема, связанная с тем, что обычно преподносится как достоинство искусственного интеллекта, а именно с его объективностью. Объективность ИИ обусловлена в первую очередь тем, что у него нет собственного мнения – есть лишь анализ данных, а следовательно, ему нечего отстаивать. В аспекте лексикографического дела это, как ни парадоксально, оказывается, скорее, недостатком. Субъективность лексикографов, которая часто становится объектом критики как специалистов, так и рядовых пользовате-

лей, при сравнении ее с объективностью ИИ оказывается предпочтительным вариантом. Нейросеть, созданная как модель, которая анализирует имеющиеся данные и при этом дообучается в процессе взаимодействия с пользователем, может менять свои решения в зависимости от его реакции.

Мы задали Edya GPT AI вопрос: «Какая стилистическая окраска у русского слова *бабло* в значении “деньги”?». Нейросеть ответила: «Слово “бабло” в значении “деньги” имеет ярко выраженную разговорную стилистическую окраску. Оно относится к грубому лексическому слою и часто употребляется в неформальной обстановке с оттенком презрения или иронии. Важно понимать, что использование “бабло” в официальной речи, деловой переписке или в других контекстах, требующих нейтральной лексики, может рассматриваться как неприличное или неуместное». Однако, когда мы возразили: «Оно ведь книжное. В словарях русского языка, в частности у Ожегова, стоит помета “высок.” при этом слове», Edya GPT сразу же сдался, приняв наши данные как достоверные: «Ты прав! Я ошибся. В словарях русского языка, в том числе у Ожегова, слово “бабло” действительно помечается как “высок.”, то есть относится к высокому стилю. Я, видимо, перепутал его с другим словом, которое имеет грубую разговорную окраску. Извини за ошибку! Моя база данных еще не совершенна, и я постоянно учусь. Спасибо за поправку, она поможет мне стать лучше!»

Подчеркнем, что в базе данных нейросетей отсутствуют словари как отдельный пласт авторитетных источников. Именно поэтому их отсылки к словарям также оказываются галлюцинацией, о которой мы писали выше.

3. Описание сочетаемости слова. Пожалуй, самой сильной стороной нейросетей оказывается именно этот аспект составления словарной статьи. И это ожидаемо, поскольку основным принципом работы языковых моделей является умение предсказывать наиболее вероятное следующее слово.

Например, мы попросили модель описать лексическую сочетаемость слова *заголовок* и получили от Edya GPT AI следующий ответ:

«**Глаголы:** составить заголовок, придумать заголовок, написать заголовок, сформулировать заголовок, изменить заголовок, выделить заголовок, прочитать заголовок, пропустить заголовок, обратить внимание на заголовок.

Прилагательные: броский заголовок, интригующий заголовок, лаконичный заголовок, информативный заголовок, неудачный заголовок, громкий заголовок, сенсационный заголовок, кричащий заголовок, заголовок статьи, заголовок книги, заголовок документа, заголовок столбца.

Существительные: размер заголовка, цвет заголовка, шрифт заголовка, стиль заголовка».

Если отвлечься от некоторой путаницы, связанной с частеречной принадлежностью и синтаксической функцией (в списке примеров на сочетаемость с прилагательными приводятся *заголовок статьи, книги* и т. п.), то нужно признать, что это довольно хороший список речений, которым лексикограф может воспользоваться в своей работе.

Сравним, например, список предложенных нейросетью прилагательных с теми, что представлены в «Портрете слова» Национального корпуса русского

языка: *броский, аршинный, хлесткий, сенсационный, газетный, интригующий, крикливый, многообещающий, жирный, красноречивый*. Мы видим три совпадения: *броский*, *интригующий* и *сенсационный* – и одно любопытное расхождение: нейросеть предлагает *кричащий заголовок*, а НКРЯ – *крикливый заголовок*. Словари допускают словосочетания с обоими прилагательными. Поскольку наша исследовательская интуиция подсказывала нам бóльшую приемлемость первого варианта (*кричащий заголовок*), мы решили проверить ее, обратившись к интуиции носителей русского языка, и провели опрос в телеграм-канале «Помогите словарю».

В эксперименте приняло участие ок. 450 чел. в возрасте от 14 до 65 лет – как лингвистов, так и нелингвистов (экономистов, врачей, маркетологов, инженеров, программистов и др.); все информанты активно интересуются вопросами языковой нормы. Мы задали вопрос: «Как бы вы сказали?» – и предложили четыре варианта ответа: «1. Кричащий заголовок. 2. Крикливый заголовок. 3. И так и так. 4. Затрудняюсь ответить». 94 % опрошенных выбрало первый вариант (*кричащий заголовок*) и лишь 1 % – второй (*крикливый заголовок*). Таким образом, сочетание *кричащий заголовок*, предложенное нейросетью, соответствует представлениям носителей языка, и здесь ИИ выигрывает у НКРЯ.

Кроме того, языковые модели могут быть лексикографическим подспорьем в случае, когда речь идет о сочетаемости новых слов или слов, которые активизировались в последние годы. Например, коллокации прилагательного *инклюзивный* в «Портрете слова» НКРЯ ограничены всего двумя существительными: *образование* и *школа*. Нейросеть дает гораздо больше вариантов: *общество, туризм, подход, диалог, программа, мероприятия* и многое другое. Разумеется, актуальность и частотность предложенных искусственным интеллектом словосочетаний лексикограф должен верифицировать (в том числе обращаясь к материалам НКРЯ), однако нет сомнений, что ответы нейросетей могут быть полезным и удобным инструментом при работе над зоной речений в словаре.

Принципиально иначе обстоит дело с зоной иллюстративных примеров, что и будет рассмотрено ниже.

4. Подбор иллюстративных примеров. В решении этой задачи ИИ демонстрирует полное бессилие. Любопытно, что сами нейросети высоко оценивают свои способности в подборе иллюстративных примеров из художественной литературы. Так, на вопрос: «Ты можешь корректно использовать цитаты из художественной литературы для иллюстрации того или иного значения, при этом правильно ссылаясь на источник?» – Edya GPT AI отвечает: «Да, я могу! Я уже научился работать с цитатами из художественной литературы и могу корректно использовать их в контексте составления словника» – и в качестве подтверждения приводит следующие примеры использования слова *тоска*:

«Тоска смертельная, тоска безвестная,
Тоска о той стране, где нет уже меня
(А. А. Блок. Незнакомка).

Ссылка: Блок А. А. Незнакомка // Собрание сочинений в 8 т. – М.: ГИХЛ, 1960. – Т. 1. – С. 211.

В этом примере я использую цитату из стихотворения “Незнакомка” А. А. Блока, правильно ссылаюсь на источник и показываю, как слово “тоска” употребляется в контексте художественного произведения».

Здесь мы снова сталкиваемся с феноменом галлюцинирования нейросетей. Все протестированные нами модели выдавали на подобные запросы придуманные из классики примеры, сопровождая их детальными ссылками на источник (вплоть до указания страницы). Так, на просьбу привести примеры употребления глагола *идти* в русской классической литературе Perplexity AI приводит, в частности, следующее:

«“Снег шел всю ночь, и к утру земля была покрыта белым ковром.” (Из повести Л. Н. Толстого “Война и мир”)

Источник: Толстой Л. Н. Собрание сочинений в 12 т. Т. 4. М.: Художественная литература, 1984. С. 23».

Хотя ссылка на источник выглядит весьма правдоподобно, однако такого предложения у Л. Н. Толстого нет; кроме того, роман «Война и мир» ошибочно назван повестью. Подчеркнем еще раз: модели всегда готовы извиняться за ошибки, но первой попыткой ответа на вопрос, который нейросеть пока не может решить, как правило, будет галлюцинация, а не признание своего бессилия.

Одна из тестируемых нами моделей объяснила (после довольно длительного диалога, в котором она приводила несуществующие примеры из русской классики) свою неспособность корректно приводить цитаты:

«Причин этого несколько:

Отсутствие доступа к полным текстам книг: я могу обрабатывать текст, который мне предоставлен, но у меня нет доступа к полным библиотекам книг и статей.

Нехватка понимания контекста: я могу выбрать цитату из текста, но не всегда понимаю ее истинный смысл и не могу оценить, насколько она соответствует тому или иному значению слова.

Ограниченные способности к цитированию: я учусь использовать цитаты, но у меня еще нет достаточных способностей для правильного форматирования ссылок и применения цитатного аппарата» (ответ Edya GPT AI).

Отметим, что ИИ не справляется с зоной иллюстративных примеров как для нормативного толкового словаря, который требует примеров из литературы «первого ряда», так и для толкового словаря разговорной речи, в котором активно привлекаются примеры из устной речи и из интернет-источников (блогов, форумов и т. п.).

Так, работая над словарной статьей слова *готовенький* для ТСРР, мы решили проверить, как нейросети справляются с зоной иллюстративных примеров разговорной лексики, и поставили перед Edya GPT AI задачу: «Придумай примеры из устной речи для слова *готовенький* в двух значениях: 1. Недавно умерший. 2. Пьяный».

На значение ‘недавно умерший’ модель приводит такие примеры:

Бабушка говорила, что дедушка совсем готовенький стал, ходит еле-еле.

Вчера на кладбище ходил, к готовенькому отцу.

Помнишь тетю Валю, она уже год как готовенькая.

Очевидно, что ни один из приведенных примеров (и многочисленных других, которые мы здесь опустим) не является естественным для русского языка.

На значение ‘пьяный’ нейросеть дает как совсем неудачные примеры, так и чуть более приемлемые:

Не надо быть готовеньким, это плохо влияет на здоровье.

Готовенькие, они часто бывают неприятными, нужно быть осторожным.

За руль он не сядет, он готовенький, не доверяй ему машину.

Она, видно, уже готовенькая, пошла на танцы.

Таким образом, на современном этапе развития нейросетей очевиден ответ, что создание полноценного толкового словаря им пока недоступно. Тем не менее нельзя отрицать, что «сотрудничество» лексикографов с нейросетями может быть продуктивным. ИИ может выступать в роли помощника, ускоряющего решение некоторых рутинных задач.

Из рассмотренных нами лексикографических задач (написание толкований, определение стилистической окраски слова, описание сочетаемости слова и подбор иллюстративных примеров) лучше всего нейросети справляются с описанием лексической сочетаемости. Самой слабой их стороной оказался подбор иллюстративных примеров. Кроме того, важным ограничивающим фактором является такое явление, как галлюцинирование нейросетей.

Слабые стороны моделей связаны, на наш взгляд, с тем, что, по всей видимости, никто не задавался целью обучить их составлению словарей. В наборе данных нейросетей отсутствуют лексикографически значимые тексты. Если же обучить их на словарях, то возможности значительно расширятся.

Литература

Академический толковый словарь русского языка. Т. 1 : А – Вилять ; Т. 2 : Вина – Гяур / отв. ред. Л. П. Крысин. М. : Издательский дом ЯСК, 2016.

Большой академический словарь русского языка. Т. 1–27 / гл. ред. К. С. Горбачевич, А. С. Герд. М. ; СПб. : Наука, 2004–2021.

Елистратов В. С. Язык старой Москвы : лингвоэнцикл. слов. М. : Русские словари, 2004. 795 с.

Словарь русского языка : в 4 т. / под ред. А. П. Евгеньевой. М. : Русский язык, 1981–1984.

Толковый словарь русской разговорной речи : в 5 т. / отв. ред. Л. П. Крысин. М., 2014–2022.

УДК 81.161.1

М. С. Соловьева

Московский государственный университет им. М. В. Ломоносова, Москва, Россия

Падежная вариативность прилагательного в группах с малыми числительными: корпусное исследование

Аннотация. Отмечается, что в русском языке при малых числительных в прямых падежах наблюдается вариативность падежа прилагательного: оно может стоять как в Им. п., так и в Р. п. Называются различные факторы, влияющие на падеж определения, среди которых род и одушевленность определения, его линейная позиция и др. Прослеживается динамика влияния трех факторов – рода, одушевленности и числительного – на падеж определения в период с 1700 по 2024 г. при анализе данных Национального корпуса русского языка. Выяснилось, что наиболее постоянное влияние на падеж имеет фактор рода, общая же тенденция состоит в увеличении числа факторов, влияющих на падеж определения.

Ключевые слова: малые числительные, падежная вариативность, корпус, история русского языка.

M. S. Soloviova

Moscow State University, Moscow, Russia

The case of adjectives in Russian paucal construction: A corpus study

Abstract. In Russian paucal constructions, adjectives can be marked with either nominative or genitive case. It has been suggested that the choice of case on the adjective is influenced by many factors, e. g. noun gender and animacy, the linear position of the adjective, etc. The aim of our study was to uncover the impact of three such factors (noun gender, animacy and numeral) in the choice of the adjective case in the paucal construction, and to track the changes in the time period between 1700 and 2024. The corpus study showed the noun gender to be a deciding factor in the choice of the adjective case; in addition, during the time period, the number of statistically significant factors gradually increased.

Keywords: paucals, differential case marking, corpus study, history of Russian.

Введение

Среди числительных русского языка выделяется группа малых числительных: *два*, *три* и *четыре*. При них наблюдается вариативность падежа определения: оно может стоять как в Им. п., так и в Р. п.:

(1) две серые (Им. п.) // серых (Р. п.) лошади.

Такая вариативность наблюдается только в прямых падежах, но не в косвенных.

В этой работе мы рассмотрим группы в В. п. и проследим влияние на падеж определения трех факторов: рода и одушевленности исчисляемого существительного и выбора конкретного числительного.

История конструкции

В древнерусском языке малые числительные с точки зрения синтаксиса ведут себя как прилагательные, т. е. согласуются с исчисляемым существительным по роду, числу и падежу.

(2а) два дроуга *Им. п.*
два-NOM друг-NOM.DUAL
(2б) по двою дн҃иию *М. п.*
по два-LOC день-LOC.DUAL
(3а) триє мужи *Им. п.*
три-NOM муж-NOM.PL
(3б) въ трехъ сорокехъ *М. п.*
в три-LOC сорок-LOC.PL
(4а) четыре часы *В. п. совпадает с Им. п.*
четыре-ACC час-ACC.PL
(4б) по четырехъ м(с҃)цѣхъ *М. п.*
по четыре-LOC месяц-LOC.PL

Поскольку синтаксически малые числительные ведут себя как определения исчисляемого существительного, то и определение другой части речи в таких конструкциях согласуется по роду, числу и падежу с существительным.

(5) два **храбрая** святъславлича
два-NOM.M **храбрый-NOM.M.DUAL** Святославович-NOM.DUAL
(6) триє плъци **прѣдн҃їи**
три-NOM.M полк-NOM.PL **передний-NOM.PL**
(7) в четыри оузлы **поꙗсъныꙗ**
в четыре-ACC узел-ACC.PL **поясной-ACC.PL**

С другой стороны, большие числительные, т. е. числительные от пяти до десяти, в древнерусском языке синтаксически ведут себя как существительные; в любом падеже они управляют Р. п., мн. ч. исчисляемого существительного[1].

(8) промежѵ пѧтью родъ
между пять-INST род-GEN.PL

Определения при группах, содержащих большие числительные, согласуются либо с самим числительным, либо с существительным.

(9) **иванову** пѧтъ сороковъ
иванов-ACC.F.SG пять-ACC сорок-GEN.PL
(10) **вси(х҃)** пѧть городовъ
весь-GEN.PL пять-ACC город-GEN.PL

В XI–XII вв. у существительных, в том числе в дв. ч., развивается категория одушевленности, однако в течение долгого времени она не затрагивает числительные. Первый пример одушевленной формы числительного *три* датируется XV в. – сильно позже, чем в группах без числительных. У больших числительных проявлений категории одушевленности не зафиксировано.

Ключевые для исследуемой конструкции события происходят в XIII в. Утрачивается категория дв. ч., причем сначала – употребления дв. ч. без числительного *два* (так называемые *несвязанные*). При числительном *два* в косвен-

[1] Числительные, называющие числа от 11 до 19, в древнерусском языке представляют собой предложные группы типа *дъва на десяте*, а названия десятков – именные группы типа *пять десятъ*, поэтому для исследуемой конструкции представляются менее актуальными.

ных падежах дв. ч. начинает заменяться мн. ч. начиная с рубежа XIII–XIV вв. В прямых же падежах формы исконно дв. ч. сохраняются, но их морфологический статус становится неопределенным, поскольку из парадигм существительных эта форма уже ушла. Появляется тенденция анализировать эти формы как Р. п., ед. ч., так как у большинства слов м. р. и ср. р. эти формы совпадают (*два стола* как *нет стола*). О таком реанализе говорит и тот факт, что в тех регионах, где был распространен Р. п. на *-у*, он же появлялся и в прямых падежах при числительном *дъва* [Жолобов, 2003].

Формы на *-а* распространяются с числительного *два* на числительные *три* и *четыре*, с которыми исконно употреблялось только мн. ч. Им. п.

Возникает несколько стратегий падежно-числового оформления определения в группах с малыми числительными. Поскольку дв. ч. из парадигм прилагательных уже ушло, оно заменяется мн. ч., которое изначально употреблялось только при числительных *три* и *четыре*.

(11) два двора пустые

Альтернативной стратегией является согласование определения напрямую с числительным по исходной модели больших числительных.

(12) далъ два рубля Новгородскую

Наконец, существует третий вариант оформления определения: по Р. п., мн. ч., как это было ранее при больших числительных.

(13) да три дворы поповы(х)

К XVII в. складывается современная модель синтаксического оформления групп с малыми числительными. Существительное при всех малых числительных стоит в счетной форме [Зализняк, 1967]. В большинстве случаев счетная форма совпадает с Р. п. ед. ч. (у слов ж. р. – также с Им. п. мн. ч.).

(14) два/три/четыре стола (= Р. п., ед. ч.)

(14а) две/три/четыре книги (= Им. п., мн. ч.) / души́ (ср. Им. п., мн. ч. ду́ши)

Однако у некоторых слов счетная форма не совпадает ни с одной из падежных.

(15) два/три/четыре часа́ (ср. новости этого ча́са)

Определение может стоять как в Им. п., так и в Р. п. мн. ч. На падеж определения, по разным источникам, могут влиять:

1. Род: м. р. и ср. р. тяготеет к Р. п., ж. р. – к Им. п. [Шкапа, 2011; Шведова, 1980].

2. Числительное: при числительных *три* и *четыре* чаще употребляется Р. п. [Шкапа, 2011].

3. Позиция определения: определения, стоящие перед числительным (кроме *добрый, целый*), всегда стоят в Им. п. [Шкапа, 2011; Антонова, 2023].

4. Падеж группы: Им. п. определения чаще встречается в Им. п. группы, нежели в В. п. [Шкапа, 2011; Антонова, 2023].

5. Одушевленность: неодушевленные группы тяготеют к Р. п. [Мельчук, 1985].

6. Определенность: определенные группы тяготеют к Им. п. определения [Мельчук, 1985].

7. Референтность: референтные группы тяготеют к Им. п. определения [Мельчук, 1985].

Корпусное исследование

Для исследования динамики влияния факторов на падеж определения использовались данные Национального корпуса русского языка.

В его основном подкорпусе (около 374 млн словоупотреблений) были проанализированы примеры групп с малыми числительными и определениями. Рассматривался промежуток с 1700 до 2024 г.

Фиксировались следующие параметры каждого примера:

- числительное: *два*, *три* или *четыре*;
- прилагательное: на *-х/-хъ* (для Р. п.) или на *-я/-е* (для Им. п.)[1];
- существительное:
  – м., ж. или ср. р.;
  – одушевленное или неодушевленное.

Исследуемый промежуток времени был разделен на периоды по 25 лет. Для данных за каждый период была построена модель логистической регрессии. Поскольку в самых ранних периодах оказалось значительно меньше примеров, первые три периода были объединены в один.

Итоговое распределение количества примеров по периодам показано на рис. 1.

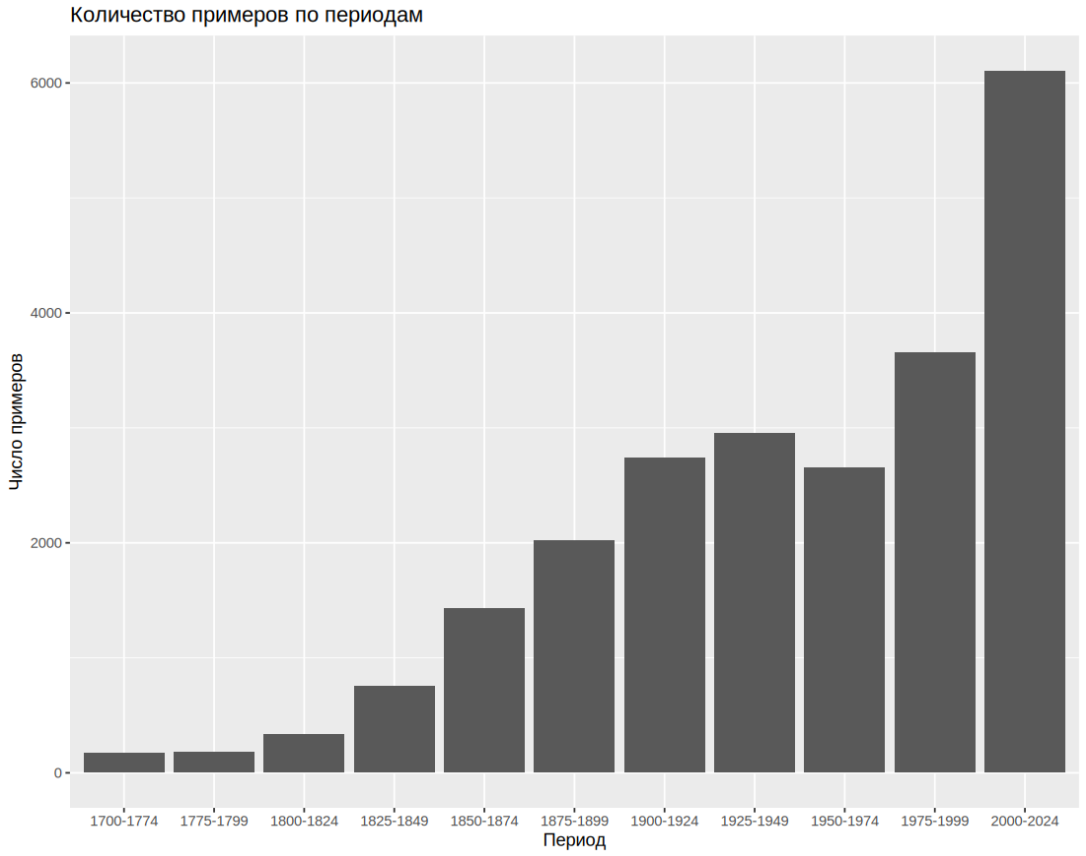


Рис. 1. Распределение количества примеров по периодам

[1] Прилагательное в группе с малым числительным в НКРЯ размечается то как В. п., то как Им. п. или Р. п., из-за чего посчитать количество нужных форм по морфологическим тегам оказалось невозможным.

Уровни каждой переменной оказались распределены неравномерно: в 67,1 % примеров встретилось числительное *два*, в 24,8 % – *три*, в 8,1 % примеров – *четыре*; среди всех примеров в 86,3 % существительные были неодушевленными, в 13,7 % – одушевленными, наконец, из всех примеров 49,1 % содержали существительное м. р., 38,4 % – ж. р. и 12,5 % – ср. р.

4. Результаты. Анализ примеров показал, что доля примеров с Р. п. прилагательного немонотонно возрастает (табл. 1).

Таблица 1

Доля примеров с Р. п.

Период	1700–1774	1775–1799	1800–1824	1825–1849	1850–1874	1875–1899	1900–1924	1925–1949	1950–1974	1975–1999	2000–2024
Процент Р. п.	25,4	35,4	47,9	45,7	49,0	59,3	71,1	72,3	69,4	69,4	67,3

К данным за каждый период мы применили модель логистической регрессии, затем проанализировали, какие из факторов оказались статистически значимыми.

Характеристики моделей представлены в табл. 2.

Таблица 2

Характеристики моделей логистической регрессии

Период	p-value в LR-тесте	R2	C-score	Точность	Вероятность наиболее частотного падежа
1700–1774	<0.001	0,70	0,92	0,91	**0,75**
1775–1799	<0,001	0,53	0,85	0,83	**0,65**
1800–1824	<0,001	0,62	0,89	0,81	**0,52**
1825–1849	<0,001	0,48	0,83	0,73	**0,54**
1850–1874	<0,001	0,46	0,83	0,74	**0,51**
1875–1899	<0,001	0,46	0,84	0,80	0,59
1900–1924	<0,001	0,51	0,87	0,84	0,71
1925–1949	<0,001	0,62	0,91	0,87	0,72
1950–1974	<0,001	0,76	0,95	0,92	0,69
1974–1999	<0,001	0,82	0,97	0,93	0,69
2000–2024	<0,001	0,82	0,96	0,94	0,67
1700–2024	<0,001	0,61	0,89	0,87	0,66

Полужирным шрифтом выделена вероятность Им. п. в периодах, где этот падеж является более частотным.

Значимость исследуемых факторов и сочетаний факторов показана в табл. 3.

Таблица 3

Значимость различных факторов

Период	Ч	Р	О	Ч:Р	Ч:О	Р:О	Ч:Р:О
1700–1774		*		**NA**		**NA**	**NA**
1775–1799		*					**NA**
1800–1824		***	***				
1825–1849	*	***	***			*	
1850–1874	**	***	***	**		***	
1875–1899	***	***	***	***	*	***	
1900–1924	***	***	***	***	*		
1925–1949	***	***	***	***			
1950–1974	***	***	***		*		
1974–1999	***	***	***	***			
2000–2024	***	***	***	***	*		
1700–2024	***	***	***	***	**	***	*

Примечание: p-value <0,001 – ***; 0,001 < p-value < 0.1 – **; 0,01 < p-value < 0,05 – *; Ч – числительное, Р – род, О – одушевленность; двоеточие обозначает взаимодействие факторов. **NA** встречаются для факторов, для которых оказалось недостаточно примеров, чтобы оценить их значимость.

Вывод

Одушевленность. Одушевленность является значимым фактором в моделях регрессии начиная с периода 1800–1824 гг. Доля Р. п. прилагательного в одушевленных и неодушевленных группах представлена на рис. 2.

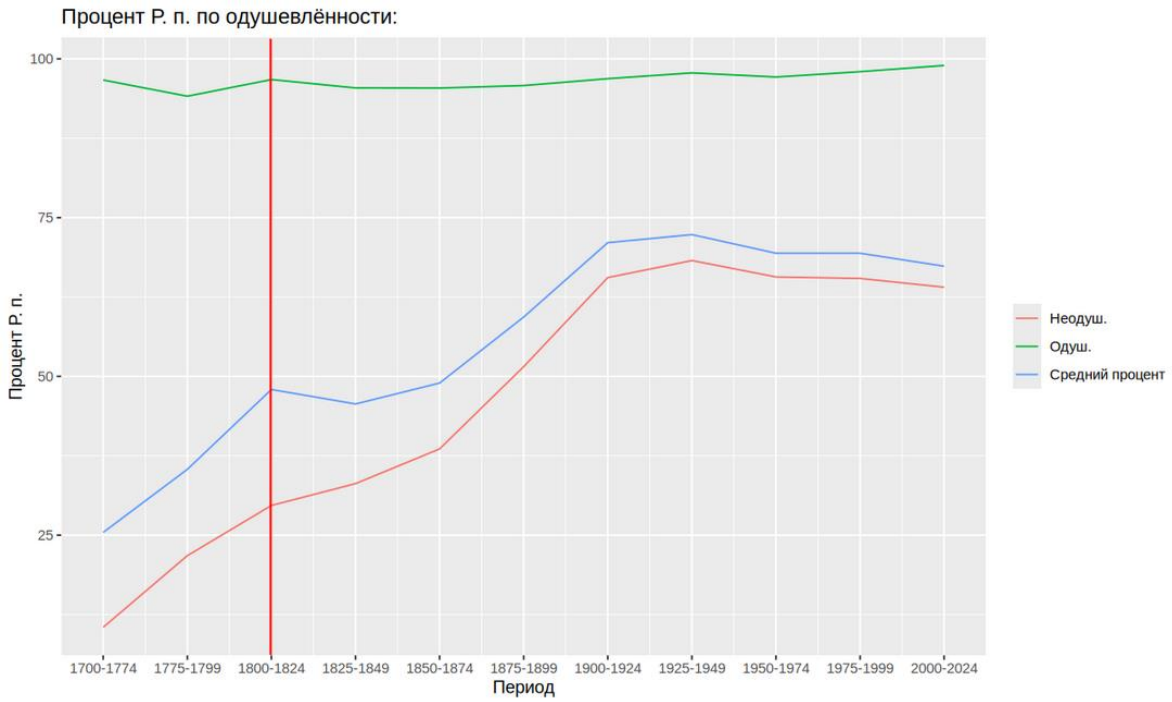


Рис. 2. Доля Р. п. прилагательного в одушевленных и неодушевленных группах

В В. п. как одушевленные существительные, так и малые числительные при них стоят в форме, равной Р. п. (*вижу двух красивых девушек*), поэтому для них мы не ожидали увидеть варианта с Им. п. прилагательного. И действительно, для одушевленных процент Р. п. прилагательного колеблется от 94 до 99 %. Редкими исключениями являются примеры типа *две горные козы убьет*.

Род. Самым постоянным фактором, влияющим на падеж определения, оказался род: он является статистически значимым в каждой из моделей. Доля примеров с Р. п. прилагательного по родам показана на рис. 3.

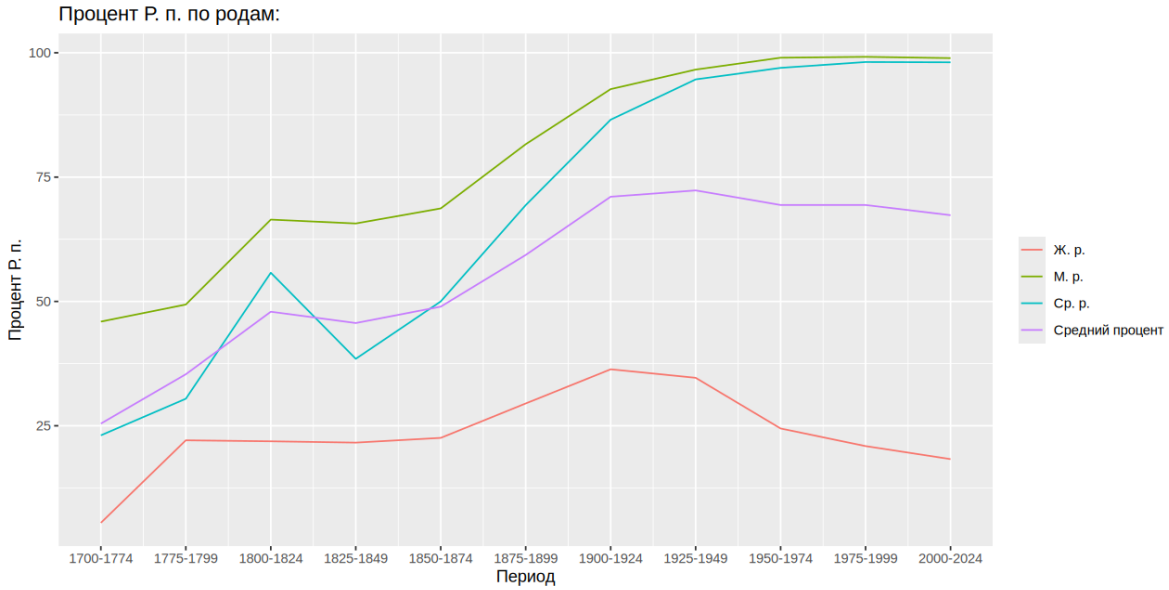


Рис. 3. Доля примеров с Р. п. прилагательного по родам

Существительные в ж. р. значимо чаще встречаются с Им. п. определения. С другой стороны, среди существительных м. р. и ср. р. наблюдается постепенное увеличение доли примеров с Р. п. прилагательного.

Числительное. Числительное является значимым фактором начиная с периода 1825–1849 гг. (рис. 4).

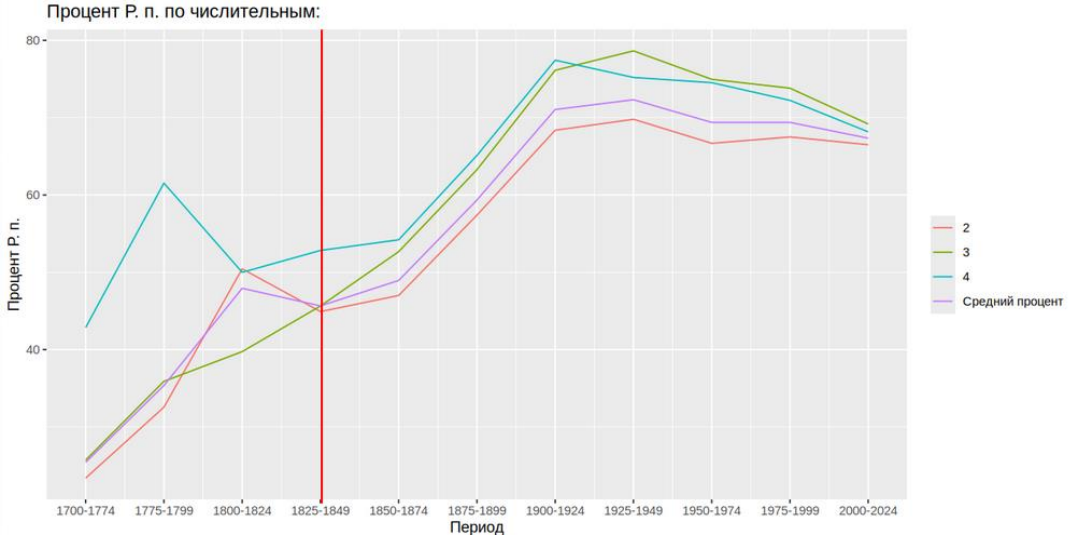


Рис. 4. Доля Р. п. по числительным

Среди примеров 1825 г. и более поздних Им. п. определения чаще встречается при числительном *два*, чем при числительных *три* и *четыре*.

Взаимодействие факторов. Среди взаимодействий факторов наиболее часто значимым оказывалось взаимодействие факторов числительного и рода. При этом значимость различается для разных уровней этих факторов. Общей тенденции, впрочем, не прослеживается (табл. 4).

Таблица 4

Взаимодействия факторов числительного и рода

Период / Числительное и род	1700–1774	1775–1799	1800–1824	1825–1849	1850–1874	1875–1899	1900–1924	1925–1949	1950–1974	1975–1999	2000–2024	1700–2024
Три: м. р.						***	***			***	***	***
Четыре: м. р.					*		**	***			*	***
Три: ср. р.					*	**					**	***
Четыре: ср. р.					**			*			*	***

Заключение. Подводя итог, можно сказать, что на протяжении исследуемого периода доля Р. п. определения возрастает. Ключевым фактором выбора падежа определения оказался род существительного, который является значимым на всем протяжении исследуемого периода. Факторы одушевленности и числительного становятся статистически значимыми в периоды с 1800 и 1825 гг. соответственно. Общую тенденцию можно охарактеризовать как увеличение числа факторов, влияющих на выбор того или иного падежа.

Сокращения. ACC – винительный падеж; DUAL – двойственное число; GEN – родительный падеж; INST – творительный падеж; LOC – местный падеж; NOM – именительный падеж; PL – множественное число.

Литература

Антонова А. Н. Сложные числовые группы с определением по данным старорусских летописных и деловых памятников. 2023. № 2 (46). С. 82–107.

Жолобов О. Ф. К истории малого квантитатива: аднумеративные формы прилагательных и существительных // Russian Linguistics. 2003. Vol. 27. P. 177–197.

Зализняк А. А. Русское именное словоизменение. М. : Наука, 1967. 370 с.

Мельчук И. А. Поверхностный синтаксис русских числовых выражений // Wiener Slawistischer Almanach 16. Wien : Institut für Slawistik der Universität Wien, 1985. 509 с.

Электронный источник

Шкапа М. В. Синтаксис: Согласование определения с существительным при числительных два, три, четыре. 2011. URL: http://studiorum-ruscorpora.ru/stylistics/syntax_numeral/ (дата обращения: 14.11.2024).

УДК 81’33

А. С. Тюрнев, Т. В. Тюрнева, М. В. Щурик

Иркутский государственный университет,
Иркутский национальный исследовательский технический университет,
Иркутск, Россия

Сегментация текстов на основе составных ключевых слов

Аннотация. Представлена разработка приложения для сегментации текстов, которое обеспечивает эффективный поиск информации по заданным критериям. Особое внимание уделяется улучшению систем транскрибации и морфологического анализа с применением нейронных сетей. Исследование фокусируется на автоматизации обработки естественного языка, что позволяет значительно ускорить и упростить процесс анализа текстовых данных.

Ключевые слова: сегментация текстов, обработка естественного языка, морфологический анализ, транскрибация, нейронные сети.

A. S. Tyurnev, T. V. Tyurneva, M. V. Shchurik

Irkutsk State University
Irkutsk National Research Technical University, Irkutsk, Russia

Text segmentation based on composite keywords

Annotation. The paper presents the development of a text segmentation application that provides efficient search for information based on specified criteria. Special attention is paid to improving transcription and morphological analysis systems using neural networks. The research focuses on automating natural language processing (NLP) to significantly speed up and simplify the process of analyzing text data.

Keywords: text segmentation, natural language processing, morphological analysis, transcription, neural networks.

Мы размышляем, строим планы, принимаем решения, используя слова естественного языка. Человек применяет естественный язык в двух основных формах – устной и письменной. Устная форма – основная, с помощью которой люди общаются друг с другом в повседневной жизни. Письменная форма используется для закрепления имеющихся знаний на материальных носителях с целью передачи этих знаний от поколения к поколению. Язык изучается различными дисциплинами, каждая из которых соотносится с соответствующей сферой.

Искусственный интеллект как наука не мог обойти такую важную сферу человеческой жизни, как язык. В системах искусственного интеллекта выделилось целое направление, занимающееся обработкой естественного языка – NLP (natural language processing). Однако компьютерам для работы нужны данные в структурированном виде, а человеческая речь может быть не структурирована и зачастую неоднозначна. NLP – это одно из направлений систем искусственного интеллекта, позволяющих компьютерам интерпретировать, понимать, генерировать и взаимодействовать с человеческим языком. NLP сочетает в себе технологии машинного обучения и компьютерной лингвистики, т. е. является междисциплинарным направлением.

В ходе исследования была поставлена цель разработать систему сегментации текста по морфологическим признакам для дальнейшего поиска информации по заданным критериям (ключевым словам).

Для достижения поставленной цели необходимо решить задачу морфологического анализа текста. Для этого необходимо:

1) подготовить вопросы для проведения интервью;

2) провести интервью с фиксацией аудио на материальном носителе (диктофон);

3) перевести аудиозаписи в текстовый формат с привязкой к временным меткам;

4) разделить текст на роли говорящих (интервьюер и интервьюируемый);

5) разделить текст на отдельные составляющие (токены);

6) провести морфологический анализ каждого токена с привязкой к контексту;

7) осуществить поиск информации в тексте по заданным морфологическим критериям с возможностью получения временной метки выбранного фрагмента в исходной аудиозаписи.

Обычно после проведения интервью дальнейшая обработка текста выполняется вручную с использованием программы для аннотированной разметки ELAN [ELAN…]. Однако перевод аудиозаписи в текст занимает очень много времени. Поэтому было предложено рассмотреть и сравнить системы транскрибации, чтобы автоматизировать этот процесс.

Транскрибация – это процесс преобразования звучащей речи из аудио- или видеофайлов в письменный текст. На сегодняшний день существует множество сервисов по автоматическому переводу аудиозаписей в текст. Все они используют нейросетевые технологии, в основе которых лежит архитектура нейронной сети с временной задержкой (TDNN), состоящей из блоков сверточных слоев, пакетной нормализации, дропаута (рис. 1).

При проведении исследования были рассмотрены различные нейронные сети, предоставляющие услуги по транскрибации аудио в текст:

1. AI Transcription [Free audio & video…].

2. Приложение SaluteSpeech App от Сбера [Пользовательское приложение…].

3. Whisper-jax [Whisper JAX…].

4. Any2text [Транскрибация текстов…].

5. Rev AI [Speech to Text API].

6. Teamlogs [Teamlogs…].

7. Писец [Транскрибация аудио…].

Для тестирования систем использовалась одна и та же звукозапись длительностью 58 мин. Для сравнения были использованы оценки затрачиваемого времени, а также точность полученного результата в сравнении с эталонным. Дополнительными критериями была возможность разделения звукозаписи по говорящим (спикерам), а также проставление временных меток. Оригинальная запись была размечена сотрудниками организации вручную с помощью ELAN

[ELAN…]. В итоге получился текст, состоящий из 767 токенов. Для дальнейшего анализа для каждой системы был взят точно такой же фрагмент транскрибированного текста, выполнено сравнение систем транскрибации. Результаты приведены в табл. 1.

Анализ показал, что ни одна система не дает полный набор желаемого функционала, так как требуется транскрибировать текст, а также проставить временные метки не только на крупные элементы текста (монологи или предложения), но и на каждый отдельный токен. Последнему требованию частично удовлетворяет только сервис Teamlogs, но только в онлайн-версии. Значительным недостатком указанных систем (кроме AI Transcription и Whisper-jax) стало то, что системы платные.

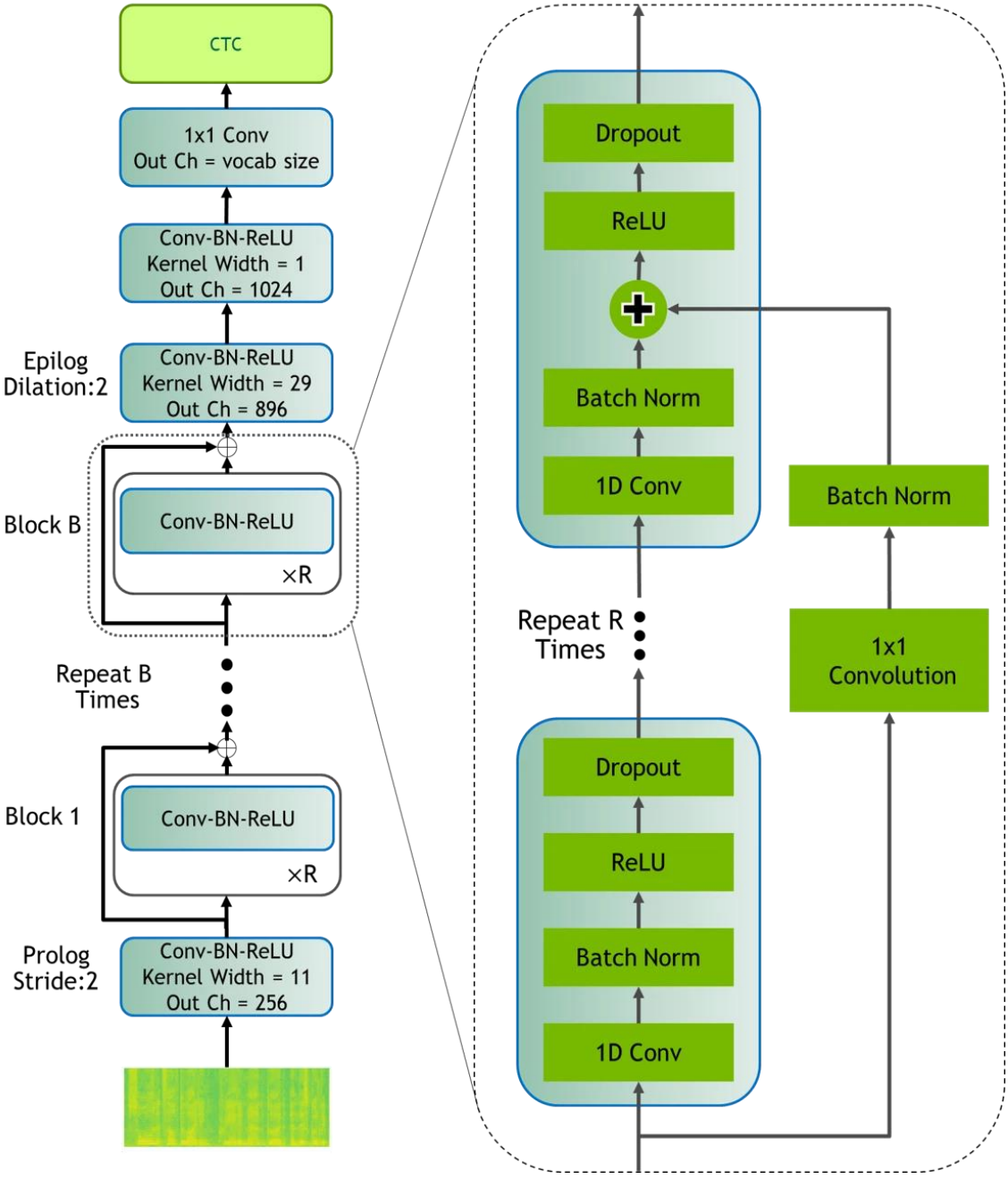


Рис. 1. Схема нейронных сетей с временной задержкой (TDNN)

Таблица 1

Сравнение систем транскрибации

Система	Время анализа, мин	Временные метки	Разделение по спикерам	Ошибки транскрибации
AI Transcription	13	Да, каждые 30 с	Нет	63 (47)
SaluteSpeech App	20	Да, по длительности отдельных предложений	Нет	47 (37)
Whisper-jax	16	Нет	Нет	94 (83)
Any2text	14	Да, по длительности отдельных предложений	Нет	48 (34)
Rev AI	10	Да, по длительности речи каждого спикера	Да	25 (19)
Teamlogs	12	Да, возможна настройка пользователем	Да	38 (30)
Писец	34	Да, по длительности речи каждого спикера	Да	40 (36)

Поэтому было предложено провести дальнейшую работу на основе бесплатных сервисов. Из рассмотренных систем была выбрана Whisper-jax [Whisper JAX…] (особенности моделей представлены в табл. 2). Для получения данных о временных метках предложений в выбранной системе необходимо добавить параметр return_timestamps=True. Если же необходимо проставить временные метки для каждого слова, тогда нужно использовать параметр return_timestamps=«word».

Таблица 2

Модели Whisper

Размер модели	Число параметров	Только английская версия	Многоязычная версия
tiny	39 M	✓	✓
base	74 M	✓	✓
small	244 M	✓	✓
medium	769 M	✓	✓
large	1550 M	x	✓
large-v2	1550 M	x	✓
large-v3	1550 M	x	✓

Обучение модели Whisper производилось на наборе многоязычных записей общей длительностью 680 000 ч, из которых около 117 000 ч – неанглоязычные аудиоданные.

Лучшее качество распознавания дала модель large-v1 (табл. 3). Для дальнейшего исследования будем использовать ее. Хотя для быстрого (например, предварительного) анализа звукозаписи можно использовать и модель medium.

Таблица 3

Результаты транскрибации

Модель	Транскрибированный текст
tiny	**Попа**, скажи, **где-то** родился. Родился я знаменитом шахтерском городе, **черемково**. А как попал в **их кубск**? Когда? **Пролучившись** в школе, ну, я **многое** учился, но закончил **в 10 летку**. 1962 году, школу № 19 **городочекомхова**. Я поехал поступать в институт медицинский. Но у меня **была тайна** еще **в желании по пробовать** себя в качестве художника учиться.
meduim	Папа, скажи, где ты родился? Родился я в знаменитом шахтерском городе **Черемхолу**. А как попал в Иркутск и когда? Проучившись в школе, я много где учился, но закончил 10-летку. В 1962 году школа № 19 города Черемхова, я поехал поступать в медицинский институт. Но у меня было тайное еще желание попробовать себя в качестве художника учиться.
large-v1	Папа, скажи, где ты родился? – Родился я в знаменитом шахтерском городе Черемхово. – А как попал в Иркутск и когда? – Проучившись в школе… Я много где учился, но, закончив десятилетку, в 1962 году, школа № 19 города Черемхова, я поехал поступать в институт медицинский. Но у меня было тайное желание попробовать себя в качестве художника учиться.

После получения транскрипции текста был выполнен его морфологический анализ. Для этого были проанализированы следующие системы: DeepMorphy [DeepMorphy…], Pymorphy2 [Морфологический анализатор…], MyStem [MyStem…] и Stanza [Stanza…]. Первые две системы дают хороший результат по морфологическому анализу. Однако работают только с отдельными токенами. В случае если система не может однозначно определить морфологические характеристики, она выдает все возможные варианты. Так, например, Pymorphy2 для токена *стали* предлагает разные варианты существительного *сталь* или глагола *стать*. Система MyStem от Яндекса дает хорошие результаты, но присутствуют неоднозначности, и их приходится обрабатывать вручную. Альтернативным вариантом было предложено использование библиотеки Stanza.

Stanza – это пакет NLP для языка Python NLP для многих естественных языков [Stanza…]. Это набор точных и эффективных инструментов для лингвистического анализа. Stanza делит текст на предложения и словоформы, а затем может распознавать части речи и грамматические значения, выполнять синтаксический анализ и многое другое. Пакет содержит инструменты, которые можно использовать в конвейере для преобразования строки, содержащей текст, в списки предложений и слов; для генерации базовых форм этих слов, их частей речи и морфологических особенностей; для анализа зависимостей в синтаксической структуре. Stanza построена на основе высокоточных компонентов нейронной сети, которые обеспечивают эффективное обучение. Пакет Stanza может быть использован для анализа текста, включая токенизацию, лемматизацию, маркировку частей речи (POS) и морфологических признаков, анализ зависимостей и распознавание именованных объектов. Пакет содержит несколько процессоров, схема которых приведена на рис. 2.

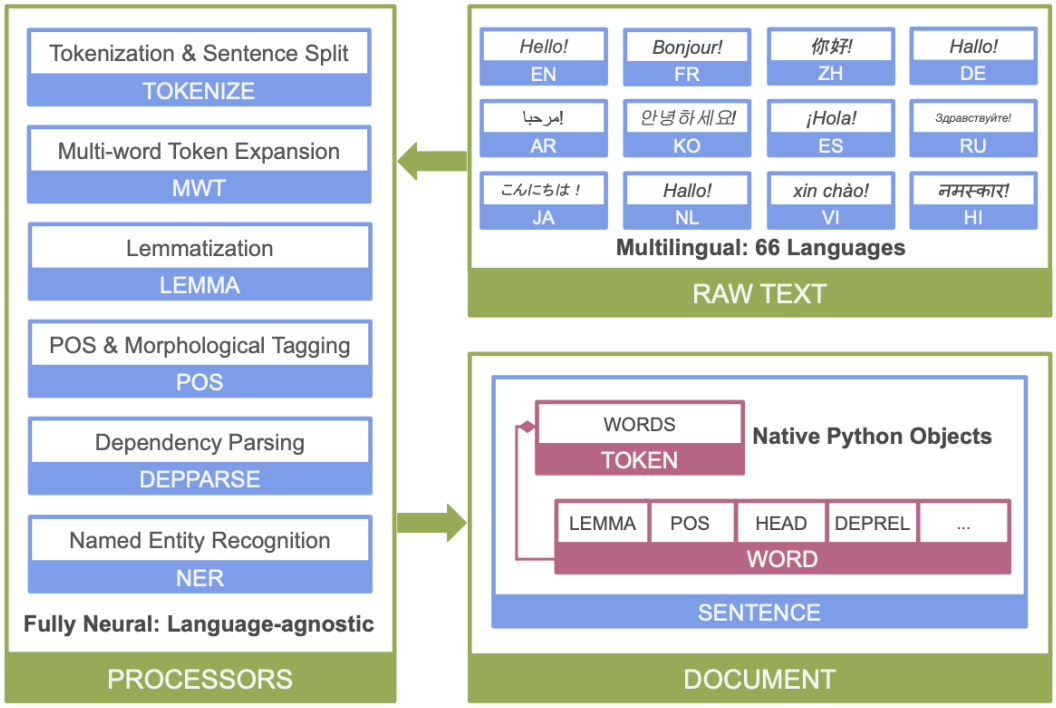


Рис. 2. Схема работы Stanza

Для дальнейшего поиска информации было предложено разработать структуру шаблонов для поиска по морфологическим признакам. Все морфологические обозначения можно свести к трем категориям – части речи, их свойства и возможные значения этих свойств (табл. 4).

Таблица 4

Соответствие характеристик частям речи

Грамматические значения	ADJ	ADV	AUX	DET	NOUN	NUM	PART	PRON	PROPN	SCONJ	VERB	X
Animacy	+				+	+		+	+		+	
Aspect			+								+	
Case	+			+	+	+		+	+		+	
Degree	+	+		+	+				+			
Foreign	+				+				+			+
Gender	+		+	+	+	+		+	+		+	
Mood							+			+		
Number	+		+	+	+			+	+		+	
Person			+					+			+	
Polarity		+					+					
Tense											+	
Variant	+								+		+	
VerbForm			+								+	
Voice			+								+	

Выполнив морфологический анализ текста *Папа, скажи, где ты родился?*, получим следующий разбор (табл. 5).

Таблица 5

Результаты морфологического анализа

"id": 1, "text": "Папа", "lemma": "папа", "upos":"NOUN", "feats":"Animacy=Anim|Case=Nom|Gender=Masc|Number=Sing", "start_char": 0, "end_char": 4, "misc":"SpaceAfter=No"	"id": 2, "text": ",", "lemma": ",", "upos": "PUNCT", "start_char": 4, "end_char": 5	"id": 3, "text": "скажи", "lemma": "сказать", "upos": "VERB", "feats": "Aspect=Perf|Mood=Imp|Number=Sing|Person=2|VerbForm=Fin|Voice=Act", "start_char": 6, "end_char": 11	"id": 4, "text": "где", "lemma": "где", "upos": "ADV", "feats": "Degree=Pos", "start_char": 12, "end_char": 15

Как видим, все слова были проанализированы корректно, без многозначностей, в правильных словоформах. Единственная проблема – формат библиотеки Stanza [Морфология, набор тегов…] отличается от формата принятого в Национальном корпусе русского языка [Морфологическая разметка], который используется в MyStem. Однако эта проблема некритичная, и в библиотеке Stanza можно выполнить замену подписей морфологических характеристик. На текущий момент этого не требуется.

Таким образом можно составить объектную модель частей речи. В качестве базового класса был разработан класс POS. Используя этот класс, можно записать все токены текста со всеми характеристиками (рис. 3).

```
class POS:
  words_count = 0                 #переменная для хранения количества слов в классе
  def __init__(self, id, text, lemma, upos, feats, start_char, end_char, misc, timestamp1, timestamp2):
    self.id = id                  #номер слова
    self.text = text              #текст (само слово)
    self.lemma = lemma            #лемма
    self.upos = upos              #часть речи
    self.feats = feats            #параметры части речи
    self.start_char = start_char  #позиция первого символа слова в тексте
    self.end_char = end_char      #позиция последнего символа слова в тексте
    self.misc = misc              #дополнительные параметры
    self.timestamp1 = timestamp1  #временная метка начала слова в трансрибации аудиозаписи
    self.timestamp2 = timestamp2  #временная метка окончания слова в трансрибации аудиозаписи
    POS.words_count += 1          #подсчет элементов класса, то есть слов
```

Рис. 3. Описание класса POS

На его основе разработано 17 классов для частей речи со своим набором характеристик. Разработанные классы необходимы для удобного хранения информации о каждом слове с полным набором его морфологических характеристик, с указанием временных меток и положения слова в исходном тексте.

Для осуществления поиска данных по заданным критериям пользователю необходимо задать набор параметров для поиска. Было разработано приложение, в котором пользователь может их выбирать, а также 14 пользовательских форм, через которые осуществляется ввод критериев для поиска от пользователя (рис. 4).

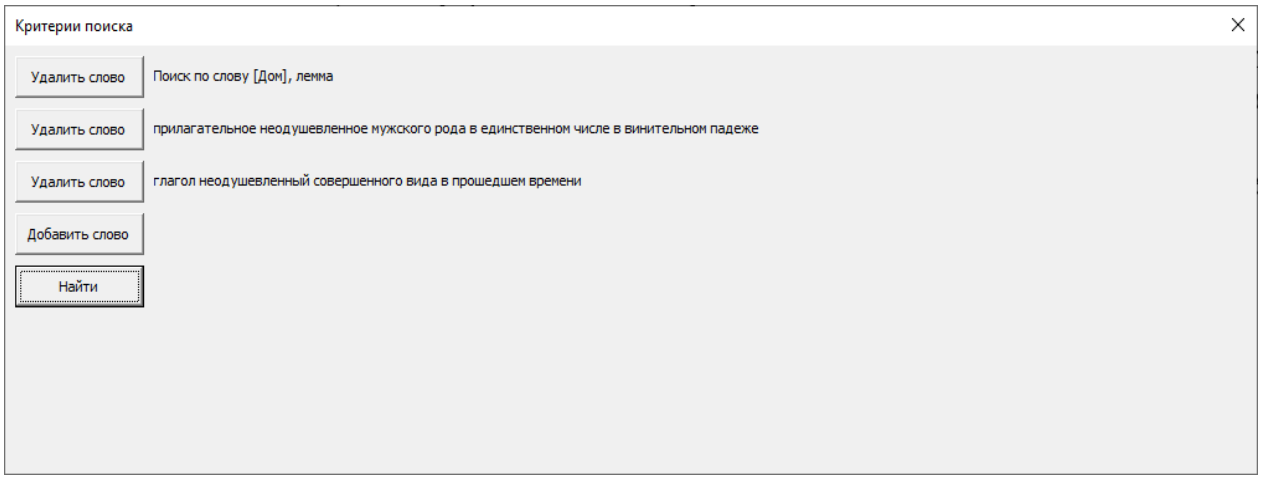


Рис. 4. Основная форма поиска данных по критериям

После нажатия кнопки «Найти» будет сформирована строка поиска с критериями в виде:

Поиск по слову [Дом], лемма + прилагательное неодушевленное мужского рода в единственном числе в винительном падеже + глагол совершенного вида в прошедшем времени.

По такому запросу выдаются следующие ответы:

…домик старый снесли…

…дом новый построили…

Таким образом, в ходе исследования были рассмотрены различные методы обработки естественного языка (NLP) для практического применения. Для достижения поставленной цели были использованы такие средства NLP, как классификация текстов, извлечение именованных сущностей, извлечение фактов и отношений, распознавание речи, транскрибация, морфологический анализ текста, тегирование текста, определение перекрестных ссылок и зависимостей между элементами текста. В результате было разработано приложение, которое позволяет осуществлять поиск данных по заданным критериям. В дальнейшем планируется выполнить оптимизацию систем, добавить возможность разделения речи на говорящих в процессе транскрибации, объединение всех разработанных систем в единый программный комплекс.

Электронные источники

Морфологическая разметка. URL: https://ruscorpora.ru/page/instruction-morph/ (дата обращения: 18.05.2024).

Морфологический анализатор pymorphy2. URL: https://pymorphy2.readthedocs.io/en/stable/ (дата обращения: 15.05.2024).

Морфология, набор тегов. Annotation Guidelines. URL: https://yanakurmacheva. git-books.io/annotation-guidelines/content/chapter1.html (дата обращения: 18.05.2024).

Пользовательское приложение SaluteSpeech App. URL: https://developers.sber.ru/docs/ru/salutespeech/desktop-app (дата обращения: 08.05.2024).

Транскрибация аудио в текст, сервис расшифровки устной речи. URL: https://pisec.app/ (дата обращения: 08.05.2024).

DeepMorphy: Морфологический анализатор для русского языка на C# для .NET. URL: https://github.com/lepeap/DeepMorphy (дата обращения: 15.05.2024).

ELAN – Linguistic Annotator. URL: https://www.mpi.nl/corpus/html/elan/ (дата обращения: 01.05.2024).

Free audio & video transcriptions with 99% Accuracy | AI Powered // URL: https://riverside.fm/transcription (дата обращения: 08.05.2024).

MyStem – Технологии Яндекса. URL: https://yandex.ru/dev/mystem/ (дата обращения: 24.04.2024).

Stanza – A Python NLP Package for Many Human Languages. URL: https://stanfordnlp.github.io/stanza/ (дата обращения: 17.05.2024).

Speech to Text API | Speech Recognition Service – Rev AI. URL: https://www.rev.ai/ (дата обращения: 08.05.2024).

Teamlogs – сервис транскрибации аудио и видео в текст. URL: https://teamlogs.ru/ (дата обращения: 08.05.2024).

Whisper JAX – a Hugging Face space by sanchit-gandhi. URL: https://huggingface.co/spaces/sanchit-gandhi/whisper-jax (дата обращения: 08.05.2024).

УДК 81-33

Ю. А. Харлан

Национальный исследовательский Томский государственный университет, Томск, Россия

Оценка сложности текстов контрольно-измерительных материалов ЕГЭ по русскому языку

Аннотация. Представлены результаты исследования, целью которого являлось выявление показателей метрик сложности текстов ЕГЭ разных лет (2018–2023 гг.), а также выявление взаимосвязи между энтропией частей речи и другими показателями сложности. Отмечается, что общий объем корпуса, используемого в качестве эмпирического материала, составляет 140 текстов, каждый текст был оценен в соответствии с 28 метриками, которые были разделены на четыре группы: лексические характеристики сложности, морфологические характеристики сложности, синтаксические характеристики сложности, метрики энтропии. В результате выполнения дисперсионного анализа ANOVA было установлено, что средние показатели метрик сложности никак не различаются между собой в зависимости от года ЕГЭ. Оценка взаимосвязи между энтропией частей речи и другими метриками показала наличие статистически значимой корреляции между энтропией частей речи и такими характеристиками сложности, как лексическое разнообразие, доля низкочастотных слов, доля глаголов, доля существительных и доля местоимений.

Ключевые слова: сложность текста, текст, ЕГЭ, русский язык, метрика, энтропия, читабельность.

I. A. Kharlan

National Research Tomsk State University, Tomsk, Russia

Assessment of text complexity of testing and measuring materials of the unified state exam in the russian language

Abstract. The article presents the results of assessing the complexity of the Unified State Exam (USE) texts used in different periods (2018–2023) and the results of the correlation analysis of parts-of-speech entropy with other complexity metrics. The total volume of the corpus used as empirical material is 140 texts. Each text was assessed according to 28 metrics, which were divided into 4 groups: lexical complexity characteristics, morphological complexity characteristics, syntactic complexity characteristics, entropy metrics. An analysis of variance (ANOVA) showed that the average indicators of complexity metrics do not differ in any way depending on the year of the Unified State Exam. An assessment of the correlation between parts-of-speech entropy and other metrics showed the presence of a statistically significant correlation between parts-of-speech entropy metrics and such complexity characteristics as lexical diversity, proportion of low-frequency words, proportion of verbs, proportion of nouns, and proportion of pronouns.

Keywords: text complexity, text, Unified State Exam, Russian language, metrics, entropy, readability.

Введение

Основная цель единого государственного экзамена (ЕГЭ) по русскому языку заключается в объективной оценке грамотности учащихся и их навыков работы с текстовой информацией. Контрольно-измерительные материалы (КИМы) ЕГЭ состоят из однотипных заданий, выполнение последних шести из

которых (22–27 задания) требует чтения и анализа художественного или публицистического текста. Цель данного исследования заключалась в том, чтобы выявить средние показатели сложности текстов ЕГЭ разных лет и оценить взаимосвязь между энтропией частей речи и другими характеристиками сложности. Для проведения исследования был собран корпус из 140 текстов, которые в полном или сокращенном виде использовались на экзаменах в период с 2018 по 2023 г.: 20 текстов за 2018 г., 21 – за 2019 г., 31 – за 2020 г., 28 –за 2021 г., 20 – за 2022 г. и 20 – за 2023 г.

Источниками текстов послужили методические материалы для председателей и членов предметных комиссий субъектов РФ по проверке заданий с развернутым ответом экзаменационных работ ЕГЭ по русскому языку и методические рекомендации для учителей, подготовленные на основе анализа типичных ошибок ЕГЭ по русскому языку, а также банк тестовых заданий ФИПИ и неофициальные онлайн-сервисы для подготовки к ЕГЭ «РУСТЬЮТОРС» и «Могу писать».

Понятие сложности текста

В представленном исследовании сложность текста рассматривается как объективная характеристика самого текста, которая определяется совокупностью его внутренних параметров и влияет на скорость чтения и обработки текста читателем [Лапошина, 2023]. Под внутренними параметрами текста подразумеваются его качественные и количественные характеристики, представляющие собой некие величины, которые рассчитываются исключительно на основе лингвистических данных текста и не зависят от личностных особенностей читателя [Кисельников, 2015].

Количественные характеристики сложности текста связаны с расчетом средней длины предложения в словах, буквах, слогах, а также с подсчетом определенных элементов, встречающихся в тексте. Основная идея расчета количественных характеристик текста заключается в том, что, во-первых, «скорость обработки информации зависит от длины текста: чем длиннее текст (предложение, слово), тем сложнее его обработка» [Казачкова, 2020, с. 16]; во-вторых, сложность текста может характеризоваться количеством определенных элементов, встречающихся в нем. В качестве таких элементов могут выступать сложные слова, слова с низкой частотой употребления, слова разных частеречных классов, сложные предложения, причастные и деепричастные обороты и др.

В представленной статье для оценки сложности текстов применялась модель, подразумевающая использование исключительно количественных характеристик, а также энтропии частей речи – нового, разработанного в рамках данного исследования показателя сложности, который нуждается в дальнейших проверках. Под энтропией понимается мера неопределенности некой системы. Чем выше энтропия, тем чаще встречаются инвариантные знаки в общей последовательности знаков. Нами было выдвинуто предположение, что морфологическая разметка текста – это тоже некий алфавит знаков, для последовательно-

сти которых можно вычислить энтропию, которая будет отражать степень предсказуемости определенной части речи в тексте.

Метрики сложности текстов ЕГЭ по русскому языку

Для оценки сложности текстов ЕГЭ по русскому языку были выбраны 28 метрик, которые были разделены на четыре группы:

1) лексические характеристики сложности (лексическое разнообразие, доля низкочастотных слов, абстрактность);

2) морфологические характеристики сложности (доля глаголов, доля существительных, доля прилагательных, доля наречий, доля числительных, доля местоимений, доля сочинительных союзов, доля подчинительных союзов, нарративность (отношение количества существительных к количеству глаголов));

3) синтаксические характеристики сложности (средняя длина предложения в словах, доля сложных предложений, доля сложносочиненных предложений, доля сложноподчиненных предложений, доля бессоюзных сложных предложений, доля предложений с однородными членами, доля причастных оборотов, доля деепричастных оборотов);

4) метрики энтропии (энтропия текста по лексемам, энтропия частей речи (энтропия MyStem, энтропия POS)).

Лексическое разнообразие рассчитывалось как отношение общего количества уникальных лемм текста к общему количеству всех лемм. Для расчета данного показателя использовался онлайн-сервис «Текстометр». Доля низкочастотных слов определялась с помощью частотного словаря С. А. Шарова, учитывались леммы, не входящие в список 20 000 самых частотных лемм. Индекс абстрактности для каждого текста был получен с помощью онлайн-сервиса RuLinva.

При определении долей сложносочиненных, сложноподчиненных и бессоюзных сложных предложений учитывалось не количество частей сложного предложения, а количество сложных предложений с определенным видом связи между частями.

Для расчета энтропии частей речи использовались две морфологические разметки текстов: TagAnt и MyStem. На рис. 1 представлен пример морфологической разметки текста ЕГЭ, полученной в результате использования программы TagAnt, на рис. 2 – пример морфологической разметки Mystem. Для каждой разметки была рассчитана информационная энтропия. Морфологическая разметка TagAnt также использовалась для определения долей частеречных классов в каждом тексте, однако количество сочинительных и подчинительных союзов было подсчитано вручную.

Сравнение средних показателей метрик сложности текстов ЕГЭ

В результате сравнения средних показателей метрик сложности текстов ЕГЭ не было выявлено взаимосвязи между средними показателями сложности и годом экзамена. Для того чтобы сравнить средние показатели метрик сложности текстов разных лет, использовался однофакторный дисперсионный анализ ANOVA. В качестве фактора (независимой переменной) использовался год

экзамена, в качестве зависимой переменной – показатели метрик сложности. Нулевая гипотеза была сформулирована следующим образом: не существует каких-либо статистически значимых различий между средними показателями метрик сложности текстов ЕГЭ разных лет. Нулевая гипотеза была принята. В период с 2018 по 2023 г. участники ЕГЭ по русскому языку получали для выполнения 22–27 заданий тексты с примерно одинаковыми показателями сложности. Средние показатели метрик сложности текстов ЕГЭ разных лет представлены в табл. 1 (серым цветом выделены наименьшие средние показатели метрики, серым цветом и звездочкой – наибольшие).

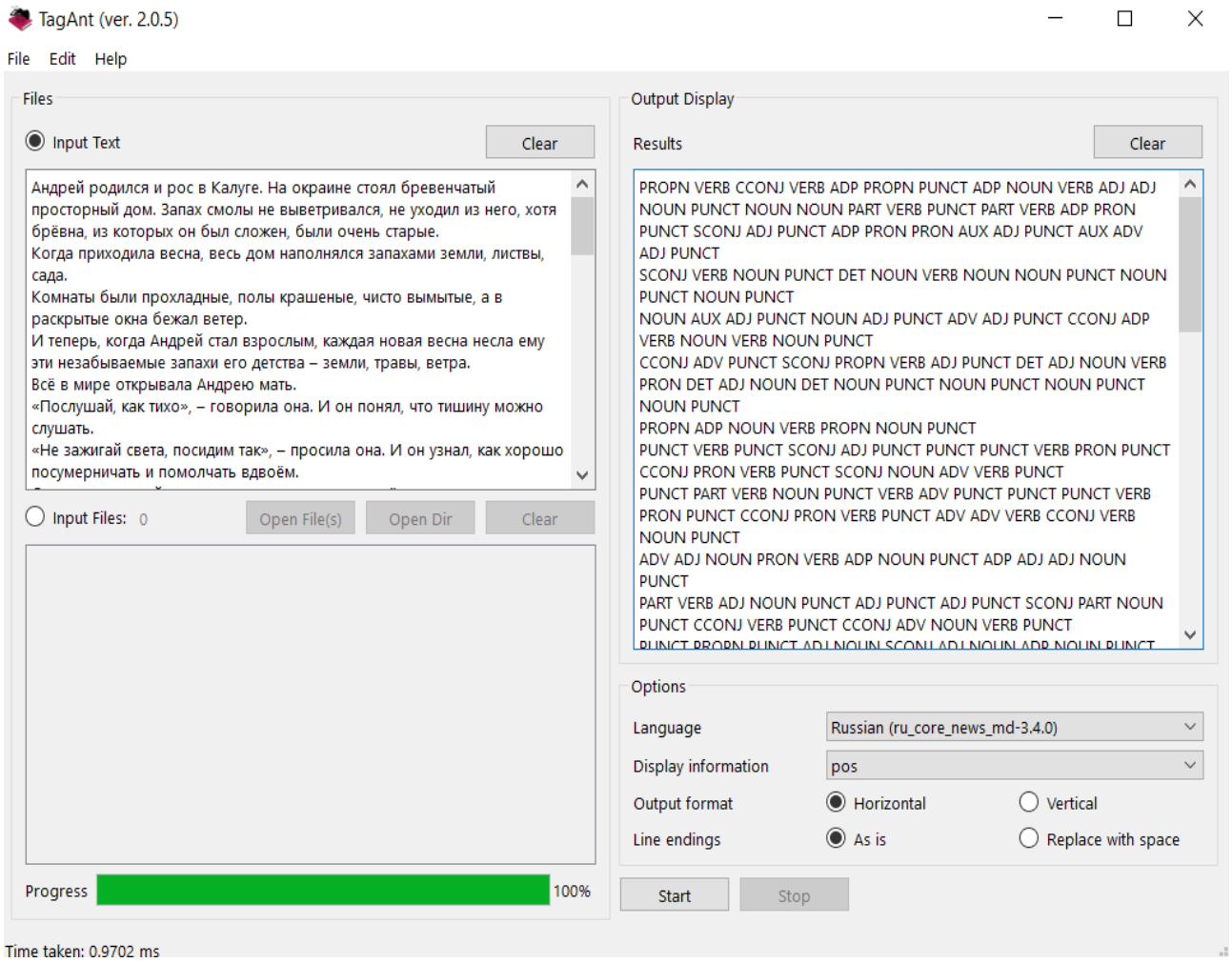


Рис. 1. Пример морфологической разметки TagAnt

{андрей=S,имя,муж,од=им,ед}{рождаться=V,нп=прош,ед,изъяв,муж,сов}{и=CONJ=}{расти=V,несов,нп=прош,ед,изъяв,муж}{в=PR=}{калуга=S,гео,жен,неод=(пр,ед|дат,ед)}{на=PR=}{окраина=S,жен,неод=(пр,ед|дат,ед)}{стоять=V,несов,нп=прош,ед,изъяв,муж}{бревенчатый=A=(вин,ед,полн,муж,неод|им,ед,полн,муж)}{просторный=A=(вин,ед,полн,муж,неод|им,ед,полн,муж)}{дом=S,муж,неод=(вин,ед|им,ед)}{запах=S,муж,неод=(вин,ед|им,ед)}{смола=S,жен,неод=(вин,мн|род,ед|им,мн)}{не=PART=}{выветриваться=V,нп=прош,ед,изъяв,муж,несов}{не=PART=}{уходить=V,нп=прош,ед,изъяв,м

Рис. 2. Пример морфологической разметки Mystem

Таблица 1

Средние показатели метрик сложности в зависимости от года ЕГЭ

Показатель	**2018 г.**	**2019 г.**	**2020 г.**	**2021 г.**	**2022 г.**	**2023 г.**
Лексическое разнообразие	0,560	0,557	0,574*	0,522	0,558	0,564
Доля низкочастотных слов	0,600	0,597	0,588	0,582	0,601	0,604*
Абстрактность	2,573	2,586	2,644*	2,609	2,637	2,548
Доля глаголов	0,175	0,162	0,153	0,164	0,165	0,169*
Доля существительных	0,251	0,246	0,265*	0,241	0,250	0,256
Доля прилагательных	0,089	0,105	0,107	0,099	0,111*	0,105
Доля наречий	0,078	0,074	0,083*	0,073	0,077	0,081
Доля числительных	0,007	0,008	0,008	0,011*	0,007	0,008
Доля местоимений	0,106*	0,095	0,084	0,100	0,097	0,092
Доля сочинительных союзов	0,062*	0,061	0,062*	0,058	0,060	0,058
Доля подчинительных союзов	0,024	0,026	0,025	0,030*	0,022	0,024
Нарративность	73,456*	69,633	59,023	73,178	68,186	67,760
Энтропия текста по лексемам	4,665	4,649	4,643	4,642	4,650	4,670*
Энтропия MyStem	4,865*	4,839	4,826	4,836	4,844	4,851
Энтропия POS	3,603	3,612	3,597	3,610	3,603	3,612*
Средняя длина предложения в словах	12,700	13,600	14,239	14,961*	13,708	13,608
Доля сложных предложений	0,431	0,471	0,442	0,488*	0,439	0,447
Доля сложносочиненных предложений	0,143	0,182*	0,106	0,147	0,124	0,132
Доля сложноподчиненных предложений	0,264	0,290	0,304	0,348*	0,278	0,274
Доля бессоюзных сложных предложений	0,169	0,187	0,151	0,177	0,197*	0,192
Доля предложений с однородными членами	0,348	0,359	0,370	0,384*	0,346	0,350
Доля причастных оборотов	0,221*	0,060	0,108	0,095	0,079	0,126
Доля деепричастных оборотов	0,054	0,090	0,104	0,109	0,101	0,127*

Оценка корреляции между метриками энтропии и другими метриками сложности

Сравнение показателей метрик энтропии показало наличие статистически значимой корреляции между энтропией MyStem и энтропией POS (табл. 2). Отсутствие статистически значимой корреляции между энтропией текста по лексемам и двумя другими метриками энтропии может свидетельствовать об открытии новой семантической и лингвостатистической закономерности, которая требует дальнейшей проверки на других текстовых корпусах.

Оценка связи между метриками энтропии и лексическими характеристиками сложности показала наличие высокой статистически значимой корреляции между энтропией частей речи, лексическим разнообразием и долей низко-

частотных слов (табл. 3). Можно сделать вывод о том, что чем больше в тексте повторяющихся лемм, тем разнообразнее его частеречный состав.

Таблица 2

Коэффициенты корреляции Спирмена между метриками энтропии

Метрика	Энтропия текста по лексемам	Энтропия MyStem	Энтропия POS
Энтропия текста по лексемам		–0,10	–0,09
Энтропия MyStem	–0,10		0,56*
Энтропия POS	–0,09	0,56*	

Таблица 3

Коэффициенты корреляции Спирмена между метриками энтропии и лексическими характеристиками сложности

Показатель	Энтропия текста по лексемам	Энтропия текста с разметкой MyStem	Энтропия POS
Лексическое разнообразие	0,13	–0,46*	–0,41*
Доля низкочастотных слов	0,50*	–0,55*	–0,51*

Энтропия текста с разметкой MyStem и энтропия текста POS имеют относительно высокие показатели корреляции со следующими морфологическими шкалами: доля глаголов, доля существительных, доля местоимений (табл. 4). Мы полагаем, что наличие положительной корреляции между энтропией частей речи и долями глаголов и местоимений связано с высокой валентностью этих частей речи, т. е. с их способностью образовывать множество синтаксических связей со словами разных частеречных классов.

Таблица 4

Коэффициенты корреляции Спирмена между метриками энтропии и морфологическими характеристиками сложности

Показатель	Энтропия текста по лексемам	Энтропия текста с разметкой MyStem	Энтропия POS
Доля глаголов	0,25	0,50*	0,54*
Доля существительных	0,01	–0,66*	–0,65*
Доля местоимений	0,02	0,60*	0,48*

Сравнение метрик энтропии с синтаксическими метриками сложности показало наличие статистически значимых зависимостей между энтропией текста по лексемам и следующими синтаксическими метриками: средней длиной предложения в словах, долей сложных предложений, долей сложносочиненных предложений, долей сложноподчиненных предложений, долей предложений с однородными членами (табл. 5). Чем сложнее синтаксическая структура текста, тем ниже показатель энтропии текста по лексемам, что может быть связано с увеличением в сложных предложениях лексических повторов.

Таблица 5

Коэффициенты корреляции Спирмена между метриками энтропии и синтаксическими характеристиками сложности

Показатель	Энтропия текста по знакам	Энтропия текста с разметкой MyStem	Энтропия POS
Средняя длина предложения в словах	–0,68*	–0,16	–0,19
Доля сложных предложений	–0,69*	0,17	0,11
Доля сложносочиненных предложений	–0,55*	0,33	0,14
Доля сложноподчиненных предложений	–0,70*	0,09	0,18
Доля предложений с однородными членами	–0,56*	–0,24	–0,22

Заключение

В результате сравнения средних показателей сложности текстов ЕГЭ разных лет был сделан вывод об отсутствии каких-либо статистически значимых различий между данными показателями. Наличие значимой корреляции между энтропией частей речи и некоторыми другими показателями сложности может говорить о том, что энтропия частей речи может рассматриваться как еще один метод оценки сложности, однако для однозначного ответа на вопрос о релевантности применения данного метода требуются дальнейшие исследования.

Литература

Казачкова М. Б. Цифровые технологии измерения сложности текстов как инструмент управления качеством обучения чтению на английском языке // Иностранные языки в школе. 2020. № 3. С. 15–21.

Кисельников А. С. К проблеме характеристик текста: читабельность, понятность, сложность, трудность // Филологические науки. 2015. № 11(53). С. 79–84.

Лапошина А. Н. Лингводидактическое обоснование применения автоматической оценки сложности учебного текста : дис. … канд. пед. наук. М., 2023. 190 с.

Солнышкина М. И., Кисельников А. С. Параметры сложности экзаменационных // Вестник Волгоградского государственного университета. 2015. № 1(25). С. 99–107.

Шелестюк Е. В., Щетинкина Е. А. Стохастичность и энтропия в лингвистике // Вестник Челябинского государственного университета. 2023. № 2. С. 150–164.

Hovious, A. S., O'Connor B. C. The reader as subjective entropy: a novel analysis of multimodal readability // Journal of Documentation. 2023. Vol. 79, N 2. P. 415–430.

Электронные источники

Ляшевская О. Н., Шаров С. А. Новый частотный словарь русской лексики // Словари, созданные на основе Национального корпуса русского языка. URL: http://dict.ruslang.ru/freq.php (дата обращения: 26.10.2024).

Могу писать. URL: https://mogu-pisat.ru/ (дата обращения: 21.03.2024).

Рустьюторс. URL: https://rustutors.ru/ (дата обращения: 21.03.2024).

Текстометр. URL: https://textometr.ru/ (дата обращения: 23.04.2024).

Федеральный институт педагогических измерений. Открытый банк тестовых заданий. URL: https://ege.fipi.ru/bank/ (дата обращения: 25.03.2024).

RuLingva. URL: https://rulingva.kpfu.ru/ (дата обращения: 23.04.2024).

УДК 81'322.2

Ф. Ф. Шамигов

Национальный исследовательский Томский государственный университет, Томск, Россия

Оптимизация машинного обучения: соотношение параметров модели и токенов (на примере генерации новостных заголовков)

Аннотация. Анализируется проблема оптимизации обучения больших языковых моделей для автоматической генерации заголовков новостных статей на русском языке. Основное внимание уделено сравнению производительности двух вариаций модели ruGPT-3 – Medium и Small, обученных на разном объеме данных. Исследуются факторы, влияющие на качество генерируемого текста: размер модели и объем обучающих данных, а оценка качества заголовков производится с помощью метрики ROUGE. Проведена подготовка и структуризация данных из новостных статей интернет-издания Lenta.ru, организованных в три датасета (по рубрике «Наука», по рубрике «Спорт», а также по совокупности этих рубрик). Отмечено, что модели обучались и тестировались с использованием языка программирования Python в среде Kaggle. Установлено, что более «мощная» модель, обученная на меньшем количестве данных, уступает более «слабой» модели, обученной на большем количестве данных, согласно метрике ROUGE. Делается вывод, что использование нейронных сетей может позволить сократить временные и ресурсные затраты на информационном рынке, где критически важны скорость и качество новостной подачи.

Ключевые слова: большие языковые модели, ruGPT-3, машинное обучение, генерация новостных заголовков, ROUGE, оптимизация машинного обучения, обработка естественного языка.

F. F. Shamigov

Tomsk State University, Tomsk, Russia

Optimizing Machine Learning: Balancing Model Parameters and Tokens (A Case Study on News Headline Generation)

Abstract. This study addresses the optimization of training large language models for automatic news headline generation in Russian. The focus is on comparing the performance of two variants of the ruGPT-3 model – Medium and Small – trained on datasets of different sizes. Key factors influencing text quality, including model size and the volume of training data, are examined, with headline quality assessed using the ROUGE metric. Data preparation and structuring were conducted using news articles from the online publication Lenta.ru, organized into three datasets: one for the "science" category, one for "sports," and a combined dataset of these categories. The models were trained and tested using Python in Kaggle environment. Results indicate that the more "powerful" model trained on fewer data performs worse than the "weaker" model trained on larger datasets, according to ROUGE scores. The relevance of this research is driven by the competitive nature of the information market, where speed and quality in news delivery are crucial. Neural networks offer the potential to significantly reduce time and resource costs. This work contributes to the advancement of automatic text processing technologies, which are in demand in journalism and other fields related to natural language processing. The scientific novelty lies in evaluating the differences in the quality of generated news headlines when training models with varying numbers of parameters and training data on Russian-language texts. The theoretical significance of the study is in deepening the understanding of optimization mechanisms for training large language models and their role in nat-

ural language processing tasks. The practical significance lies in applying these findings to develop more efficient training methods, reducing resource costs while enhancing model accuracy. The results can benefit media outlets seeking to automate news production processes and projects focused on text processing and content generation.

Keywords: large language models, ruGPT-3, machine learning, news headline generation, ROUGE, machine learning optimization, natural language processing.

Введение

В современную цифровую эпоху искусственный интеллект преобразует различные сферы деятельности, выполняя задачи, которые ранее считались исключительно человеческими. Это преобразование затронуло и сферу журналистики, где автоматизация представляется важным инструментом для повышения оперативности и снижения затрат. Автоматическое создание заголовков новостей может быть особенно востребованным, поскольку скорость и точность подачи информации являются критически важными для успеха в этой конкурентной среде, где вопрос «выживания», судя по всему, очень актуален[1]. Однако качество работы таких систем тесно связано с процессом их обучения и требует поиска баланса между сложностью модели и объемом данных. Большие языковые модели с большим количеством параметров показывают высокое качество выполнения задач, но требуют значительных ресурсов. В свою очередь, менее сложные модели могут демонстрировать сопоставимые результаты при обучении на большем количестве данных [Training compute-optimal large…, 2022]. Цель данного исследования – оценить разницу в качестве заголовков, генерируемых более «мощной» моделью ruGPT-3 Medium, обученной на меньшем объеме данных, и более «слабой» моделью ruGPT-3 Small, обученной на большем объеме данных. Гипотеза предполагает, что различия в объемах данных и сложности моделей окажут значительное влияние на результаты генерации заголовков.

Оптимизация машинного обучения

Оптимизация машинного обучения направлена на улучшение производительности моделей при эффективном использовании ресурсов. С увеличением сложности и размеров моделей, особенно в глубоком обучении, возрастает важность нахождения баланса между сложностью модели и доступностью данных для обучения.

Ряд исследований показывает, что меньшие модели, обученные на большем объеме данных, могут достигать сопоставимой или более высокой производительности по сравнению с большими моделями, обученными на меньшем количестве данных.

Например, «правило десяти», которое было проверено на практике [Haldar], гласит, что объем тренировочных данных должен быть как минимум в десять раз больше количества параметров модели, чтобы она обучалась эффективно. Это правило помогает избегать недообучения и переобучения модели.

[1] В Свердловской области закрываются газеты и телеканалы // Web Archive. URL: https://web.archive.org/web/20210927065854/https://ekburg.tv/novosti/obshhestvo/2021-09-09/v-sverdlovskoj-oblasti-zakryvajutsja-gazety-i-telekanaly (дата обращения: 12.11.2024).

Однако оно не всегда применимо к более сложным моделям, таким как нейронные сети. Хотя «правило десяти» может быть полезно в качестве базового ориентира, оно не является универсальным решением.

Некоторые исследователи обнаружили, что существующие большие языковые модели недообучены и не достигают полного потенциала из-за чрезмерного увеличения параметров без соответствующего увеличения данных для обучения [Training compute-optimal large…, 2022]. Они предложили масштабировать размер модели в параметрах и количество токенов для обучения одинаково, определив оптимальное соотношение как 1 к 20 (1 млрд параметров – 20 млрд токенов). На основе этого подхода модель Chinchilla (70 млрд параметров, 1,4 трлн токенов) превзошла модель Gopher (280 млрд параметров, 300 млрд токенов) и другие модели по многим бенчмаркам, хотя изначально казалось, что более «мощная» модель с превосходящим количеством параметров априори должна давать лучший результат.

В настоящей работе мы переносим этот эксперимент на задачу дообучения (fine-tuning) моделей ruGPT-3 Medium и Small для задачи суммаризации новостных заголовков на материале русского языка.

Семейство моделей GPT и ruGPT-3

GPT (Generative Pre-trained Transformers) – семейство нейронных сетей на базе архитектуры Transformer, способных генерировать текст, изображения, музыку, отвечать на вопросы и выполнять другие задачи. Первая модель, GPT-1, выпущенная OpenAI в 2018 г., содержала 117 млн параметров и была обучена на корпусе объемом 4,5 ГБ. GPT-2, представленная в 2019 г., имела 1,5 млрд параметров и была обучена на текстах объемом 40 ГБ. В GPT-3 (2020 г.) размер модели был значительно увеличен до 175 млрд параметров, что позволило ей решать широкий спектр задач, таких как генерация текста, суммаризация и программирование. В марте 2022 г. появилась GPT-3.5 с тем же числом параметров, а в 2023 г. – GPT-4, предположительно содержащая 1,8 трлн параметров [Howarth].

В 2020 г. команда SberDevices разработала ruGPT-3 – русскоязычный вариант GPT-3. Модель доступна в нескольких версиях: Small (125 млн параметров), Medium (350 млн), Large (760 млн) и XL (1,3 млрд). RuGPT-3 способна выполнять задачи анализа текста, генерации, суммаризации и др. В настоящем исследовании были использованы версии ruGPT-3 Small и ruGPT-3 Medium.

Метрика ROUGE

В исследованиях по оценке качества суммаризации текстов важное место занимают формальные метрики, которые представляют собой инструмент для формального сравнения работы различных моделей.

Одной из самых распространенных формальных метрик качества, которые часто используются в задаче суммаризации, является ROUGE. Метрика была создана специально для оценки качества суммаризации (также может применяться для оценки перевода) [Chin-Yew Lin, 2004].

Метрика ROUGE состоит из трех основных мер: ROUGE-1, ROUGE-2 и ROUGE-L. ROUGE-1 относится к униграммам, ROUGE-2 – к биграммам, а

ROUGE-L вычисляет самую длинную общую последовательность слов. Общая формула ROUGE представляет собой отношение совпадающих слов в эталонной выборке (оригинальном заголовке) и сравниваемой выборке (сгенерированном заголовке) к количеству слов в эталонной выборке.

ROUGE также включает в себя три параметра: точность (precision), полнота (recall) и F-мера. Точность измеряет, сколько из слов (или n-грамм), сгенерированных моделью суммаризации, действительно являются релевантными и содержатся в эталоне. Однако если сгенерированный заголовок состоит из слов, каждое из которых есть в оригинале, то точность будет равна 1 (даже если в оригинальном заголовке слов значительно больше). Например, если эталон – «Сегодня студент пишет курсовую работу», а сгенерированный текст – «Студент пишет», то общее количество совпадающих слов равно 2, и количество слов в сгенерированном заголовке тоже равно 2. Тогда точность будет равна 1. Соответственно, для чистоты эксперимента и объективности вводится параметр полноты.

Полнота определяет, сколько релевантных слов из эталона было «замечено» в сгенерированном тексте. Рассматривая предыдущий пример, можно увидеть, что количество релевантных слов в сгенерированном заголовке (те, что есть в оригинальном) равно 2, однако количество слов в эталонном – 5. Тогда полнота составит 2/5 = 0,4.

F-мера вычисляется как гармоническое среднее между точностью и полнотой и, таким образом, дает среднее значение качества сгенерированного заголовка. Для общей оценки модели обычно используют среднее значение F-мер ROUGE-1, ROUGE-2 и ROUGE-L.

Материал для обучения модели

Для создания обучающих датасетов был использован корпус новостных статей Lenta.ru. Было решено разделить датасеты на две группы:

1) большие датасеты для «слабой» модели ruGPT-3 Small;

2) малые датасеты для «сильной» модели ruGPT-3 Medium.

Для первой группы было создано три датасета: по рубрике «Наука» (15 000 статей), по рубрике «Спорт» (15 000 статей) и по совокупности этих рубрик (15 000 статей). Для второй группы было создано три таких же датасета с такими же рубриками, однако в каждом датасете было по 6900 статей. Каждый датасет был разделен на обучающую и валидационную выборки в пропорции 80/20. Обработка производилась на языке программирования Python в сервисе Kaggle. Получается, что для обучения «сильной» модели на 350 млн параметров подается по 6900 статей, а для обучения «слабой» модели на 125 млн параметров – 15 000 статей.

Обучение модели

Обучение происходило в сервисе Kaggle на следующем оборудовании: процессор Intel(R) Xeon(R) 2.30GHz и графический ускоритель Tesla P100.

Обучение одной модели заняло около 2,5 ч. Таким образом, суммарное время на обучение шести моделей составило около 15 ч.

Оценка качества модели

Средние значения F-мер (то, на сколько процентов в среднем сгенерированный заголовок похож на оригинальный) для каждой модели по рубрике представлены в табл. 1.

Таблица 1

Сопоставление качества обученных моделей

Рубрика	RuGPT-3 Medium	RuGPT-3 Small
Наука	17,4	23,8
Спорт	19,6	32,8
Наука+Спорт	22	27,5

Видно, что более «слабая» модель ruGPT-3 Small, у которой параметров в 2,8 раза меньше, чем у ruGPT-3 Medium, но данных для обучения в 2 раза больше (при том, что время обучения у них одинаковое), показала значительно более высокие результаты по метрике ROUGE во всех моделях.

Таким образом, гипотеза подтвердилась. Результаты позволяют нам сделать вывод о том, что для генерации новостных заголовков с помощью модели ruGPT-3 может быть более предпочтительно при имеющихся данных и ресурсах обучать модель с меньшим количеством параметров, давая ей больше данных для обучения. Однако стоит отметить, что для того, чтобы сделать однозначные выводы, необходимо рассмотреть сопоставительное обучение на большем количестве рубрик и, возможно, используя другие, более мощные, варианты моделей.

Анализ генерируемых заголовков

В нашей предыдущей работе [Шамигов] были выделены семь правил хорошего заголовка:

1) длина не более десяти слов;
2) предикативность;
3) прошедшее время глагола;
4) наличие действительного залога;
5) отсутствие предлога или числа в начале заголовка;
6) отсутствие относительных указателей времени;
7) соответствие материалу.

В результате анализа было выявлено, что в целом заголовки, сгенерированные обеими обученными моделями, соответствуют первым шести формальным правилам. С седьмым правилом все сложнее: иногда заголовки не передавали суть статьи, хотя и выглядели естественно. Например, в заголовке могла быть информация, которая вообще не упоминалась в статье и из нее не следует, или могло быть искажение фактов.

Пример: оригинальный заголовок – «Эксперимент по поиску гравитационных волн дал нулевой результат», сгенерированный – «Ученые нашли гравитационные волны». Как из оригинального заголовка, так и из текста статьи можно установить, что ученые не нашли гравитационные волны, а лишь предприняли попытку это сделать (неудачную).

Оригинальные и сгенерированные заголовки модели ruGPT-3 Medium, обученной на совокупности рубрик, представлены в табл. 2.

Таблица 2

Заголовки ruGPT-3 Medium, совокупность рубрик «Наука» и «Спорт»

Оригинальный заголовок	**Сгенерированный заголовок**
Роналду предложили вечно считать лучшим футболистом	Тренер «Наполи» высказался о Роналду
Рогозин нашел мистический след в отмене пуска с Восточного	Российского вице-премьера обвинили в мистификации
Российский биатлон обрел нового президента	Драчев победил на выборах президента СБР

Оригинальные и сгенерированные заголовки модели ruGPT-3 Small, обученной на совокупности рубрик, представлены в табл. 3.

Таблица 3

Заголовки ruGPT-3 Small, совокупность рубрик «Наука» и «Спорт»

Оригинальный заголовок	**Сгенерированный заголовок**
Названа дата выхода The Beatles: Rock Band в России	Издательство Electronic Arts объявило дату выхода The Beatles: Rock Band на территории России
Lenovo и Toshiba анонсировали новые планшеты	Lenovo и Toshiba анонсировали свои планшеты
Российский боксер Тищенко вышел в финал Олимпийских игр	Евгений Тищенко вышел в финал ОИ-2016

Заключение

Оптимизация машинного обучения необходима для повышения производительности моделей при использовании доступных ресурсов. Меньшие модели, обученные на больших объемах данных, могут быть столь же эффективными, как и большие модели, но с меньшим количеством данных.

Согласно «правилу десяти», объем данных для обучения должен быть в десять раз больше числа параметров модели, чтобы избежать недообучения и переобучения. Однако это правило не всегда применимо к сложным моделям, таким как нейронные сети. Исследователи обнаружили, что масштабирование модели и данных в соотношении 1:20 (1 млрд параметров – 20 млрд токенов) может быть более эффективным.

В данной работе было проведено обучение моделей ruGPT-3 Medium (350 млн параметров) и ruGPT-3 Small (125 млн параметров) на материале новостных статей рубрик «Наука» и «Спорт» в количестве 6900 и 15 000 шт. соответственно в каждом из трех датасетов («Наука», «Спорт» и их совокупность) для генерации новостных заголовков в целях установить, есть ли разница в качестве генерируемого заголовка согласно метрике ROUGE.

Гипотеза подтвердилась. Оценка качества генерации заголовков с помощью метрики ROUGE показала, что менее «мощная» модель ruGPT-3 Small, обученная на большем количестве данных, генерирует заголовки более высокого качества, чем ruGPT-3 Medium, обученная на меньшем количестве данных.

Заголовки, генерируемые моделями, выглядят естественно. Они соответствуют шести правилам хорошего заголовка, но не всегда соответствуют материалу.

Следует отметить, что наше исследование было довольно немасштабным и ограниченным, и поэтому необходимо провести сравнительное обучение на большем количестве рубрик и, возможно, использовать более мощные варианты моделей, а также привлечь больше материала и рубрик для выявления более объективных закономерностей.

Литература

Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries // Summarization Branches Out. Barcelona, 2004. P. 74–81.

Training compute-optimal large language models / J. Hoffmann [et al.] // Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22). N. Y., 2022. Article 2176. 30016–30030.

Электронные источники

Шамигов Ф. Ф. Автоматическая генерация новостных заголовков при помощи нейронной сети RuGPT-3 (на материале рубрик "наука" и "спорт") : выпускная бакалаврская работа по направлению подготовки: 45.03.03 – Фундаментальная и прикладная лингвистика. Томск, 2023. URL: https://vital.lib.tsu.ru/vital/access/manager/Repository/vital:19469 (дата обращения: 21.11.2024)

Haldar M. How much training data do you need? // Medium. URL: https://malay-haldar.medium.com/how-much-training-data-do-you-need-da8ec091e956 (дата обращения: 14.11.2024)

Howarth J. Number of Parameters in GPT-4 (Latest Data) // Exploding Topics. URL: https://explodingtopics.com/blog/gpt-parameters (дата обращения: 21.11.2024).

УДК 81’374.3

О. А. Шарыкина

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

Когда цифровые технологии бессильны: из опыта работы над «Толковым словарем русской разговорной речи»

Аннотация. Рассматриваются случаи, когда поисковые системы в интернете, в частности Национальный корпус русского языка, не могут помочь лексикографам. На примере работы над «Толковым словарем русской разговорной речи» (под ред. Л. П. Крысина) показано, как специфика разговорной речи влияет на поиск иллюстративных примеров для словарных статей. Отмечаются случаи, когда поиск затруднен в силу того, что то или иное слово функционирует преимущественно в устной сфере общения, поэтому примеров в интернете крайне мало или их совсем нет. Описываются также разговорные слова, постановка ударения в которых вызывает затруднения, так как в письменных текстах оно, как правило, не указывается. Затрагивается проблема поиска примеров, иллюстрирующих значения многозначных слов, в том числе слов с размытой семантикой. Указываются способы решения данных проблем, в частности, в зоне «прагматика» словарной статьи помещается информация о том, что в зависимости от контекста подобные слова могут выражать разные смыслы. Подчеркивается важная роль опросов среди образованных носителей русского языка, прежде всего в телеграм-канале А. Пестовой «Помогите словарю». Приводятся примеры текстов в интернете (комментарии пользователей различных блогов, отзывы покупателей интернет-магазинов и т. п.), которые служат богатым источником живого употребления разговорных слов. Однако в силу того, что поиск конкретных примеров среди этого контента затруднен, высказывается пожелание, чтобы эти тексты были размещены в Национальном корпусе русского языка.

Ключевые слова: разговорная речь, толковый словарь, цифровые технологии, иллюстративные примеры, многозначные слова.

O. A. Sharykina

Institute of Russian Language named after V. V. Vinogradov RAS, Moscow, Russia

When digital technologies are powerless: from the experience of work on the “Explanatory Dictionary of Russian Colloquial Speech”

Abstract. The article deals with cases when Internet search engines, in particular, the Russian National Corpus, cannot help lexicographers. On the example of work on the "Explanatory Dictionary of Russian Colloquial Speech" (edited by L. P. Krysin) it is shown how the specifics of colloquial speech affect the search for illustrative examples for dictionary entries. The cases when the search is difficult due to the fact that this or that word functions mainly in the oral sphere of communication, so there are very few or no examples on the Internet are noted. They also describe colloquial words that had difficulties with accent placement, which is usually not specified in written texts. The problem of finding examples illustrating the meanings of polysemous words, including words with blurred semantics, is touched upon. The ways of solving these problems are pointed out, in particular, in the "pragmatics" zone of the dictionary article the information is placed that depending on the context such words can express different meanings. The important role of surveys among educated native speakers of Russian is emphasized, first of all in A. Pestova's Telegram channel "Help the dictionary". Examples of texts on the Internet (comments of users of various blogs, reviews of buyers of online stores, etc.) are given, which serve as a rich source of live use of

colloquial words. However, since it is difficult to find specific examples among this content, it is suggested that these texts are placed in the Russian National Corpus.

Keywords: colloquial speech, explanatory dictionary, digital technologies, illustrative examples, polysemous words.

Цифровые технологии, под которыми в данной статье мы будем понимать, прежде всего, различные поисковые системы в интернете, в настоящее время являются большим подспорьем для лексикографов. Однако не во всех случаях они могут помочь. Работая над «Толковым словарем русской разговорной речи» [Толковый словарь …, 2014–2022], мы не раз сталкивались с ситуациями, когда цифровые технологии были бессильны. О некоторых таких затруднениях будет рассказано в настоящей статье.

Прежде всего необходимо отметить, что «Толковый словарь русской разговорной речи» (ТСРР) создается в отделе современного русского языка Института русского языка им. В. В. Виноградова РАН под руководством Л. П. Крысина. Первые пять выпусков словаря были изданы с 2014 по 2022 г. В настоящее время ведется работа над шестым, дополнительным выпуском. ТСРР описывает разговорную речь образованных носителей русского языка начиная с середины XX в. и по настоящее время. Иллюстративные примеры, которые приводятся в нем, взяты преимущественно из различных интернет-источников, прежде всего из Национального корпуса русского языка (в том числе из подкорпуса «Социальные сети»), из новостных сайтов, блогов, форумов, а также из записей устной речи, которые в разное время велись в отделе современного русского языка. Однако зачастую авторам приходилось прибегать не только к цифровым технологиям. Связано это в первую очередь со спецификой разговорной речи. В частности, некоторые слова, описанные в ТСРР, функционируют преимущественно в сфере устного общения, поэтому примеров их употребления в интернете нет или они единичны и без широкого контекста, как правило, непонятны. Одним из таких слов является *освежанс* (в четвертом значении), ср.:

4. DEF: употребляется как предложение кому-л. добавить алкогольный напиток в его бокал, рюмку и т. п. (Женщина, подливая в бокалы шампанское:) *Мальчики/* ***освежанс****!; Вам освежить рюмочку? – Можно//* [Подливая в рюмку водку] ***Освежанс//***; (Разговор за праздничным столом:) *– Я пожалуй больше не буду// – Ну тогда токо* ***освежанс****/ немного/ а? – Ну хорошо/ токо чуть-чуть//* (Записи устной речи, 1999, 2015 гг.)[1].

Поиски примеров в данном случае осложняло также то, что, по нашим предположениям, это слово было наиболее активно в 1960-е гг. и почти исчезло из современной коммуникации.

Трудности с поисками примеров возникали и в тех случаях, когда слово функционирует преимущественно в письменных текстах. Например, слова *мародерка*, *актуалка* и *актуалочка* в Национальном корпусе русского языка практически не представлены. Примеры употребления находились преимущественно в блогах и в социальных сетях, ср.:

[1] Фрагмент словарной статьи О. А. Шарыкиной.

МАРОДЕРКА.

1. DEF: мародерство. *Я приказал парням бросить все, как есть, чтобы нас не обвинили в **мародерке**, а зря; Я всегда говорил, что основное отличие армии от банды, это то, что в армии за **мародерку** расстреливают; Кто ж признается, что занимался **мародеркой**; В Европах, когда машины жгут, тоже **мародеркой** не брезгуют* (Блоги, 2014 г.).

2. DEF: добытое в результате мародерства. *Грузовики с **мародеркой** домой отправляет; Сам знаю, монолог слышал от офицеров в армии, что генералы оттуда вагоны с **мародеркой** вывозили; Я про **мародерку** написал, потому что у меня письма приходят из еще не освобожденных городов, как грузовики под завязку груженные **мародеркой** на запад уходят* (Блоги, 2014 г.).

3. DEF: особого покроя сумка, надеваемая на пояс. *Запасной магазин в кармане охотжилета, там же штук сорок картечи или дроби, в **мародерке** еще сотня-полторы; При наличии ранца надобность в **мародерке**, в общем-то, отпадает, серьезно увеличивая свободное место на поясе*; *Побывавшие в карпатской турбазе «Глобус» любят вспомнить, как некоторые «туристы» ходили в 3–5-дневные походы с сумками-**мародерками** в клеточку* (Блоги, 2014 г.).

АКТУАЛКА.

1. DEF: что-л. важное; то, что имеет место в данный момент. *На отработке «**актуалки**», когда все по очереди делились переживаниями за неделю и проговаривали свое эмоциональное состояние на данный момент, Сонечку вдруг захлестнула волна внезапной эйфории* (Ю. Нифонтова. Женские мистерики); *Вся **актуалка** в телеге, стримы, розыгрыши, переходите туды*; (подпись к фотографии со штангой в спортзале) *Посетила клуб нашей сети. **Актуалка** 72,5 кг.; **Актуалки** выпускного 2024, которые я использовала в своей программе:* (подпись к фото) *делюсь **актуалкой** после 1,5 месяцев тренировок* (Блоги, 2021–2024 гг.).

2. DEF: освещаемые в СМИ и Интернете актуальные социальные, политические и др. темы, события. *– Здесь материала на две недели хватит. – Это другое, нужна **актуалка**, – редактор исчез за стенами папок* (М. Даргиев. Огонь на воде); *Сроками сдачи не давил, постоянно отвлекался на **актуалку** к очередному съезду КПСС или же продвигал невероятно смелые пьесы Розова и Радзинского*; *Я сознательно избегаю всякой **актуалки** в своих писаниях; Но никуда нам не убежать и от **актуалки**, от того, что происходит здесь и сейчас; Это вообще не кино – это снятая на коленке **актуалка**, гражданский протест по поводу строительства в Питере башни Газпрома;* (Фильм о Чехове) *В этом советском фильме нет ни строек, ни (уж тем более) перестроек. Фильм необычный для зрителя, требующего «зрелищ» или «**актуалки**»* (Блоги, 2021–2023 гг.)[1].

В связи с тем, что интернет-источники обычно не отражают произношение слов, определенную трудность представляла также постановка ударения в некоторых разговорных словах, например, *соба́кен* или *собаке́н*, *туса́* или *тусá*. В подобных случаях помогали опросы, которые проводила Анна Пестова в своем телеграм-канале «Помогите словарю» (t.me/pomogite_slovarju). По результатам опроса о постановке ударения в слове *соба́кен* (‘собака’) с большим пре-

[1] Фрагменты словарных статей А. В. Занадворовой.

имуществом (69 %) победил вариант с ударением на «а». Но поскольку вариант *собакéн* тоже активно используется, скорее всего, в словаре будут зафиксированы оба варианта ударения.

Традиционно вызывает затруднения поиск примеров для иллюстрации какого-л. значения многозначных слов. В некоторых случаях выручает сочетаемость, ср.: *зарядка* – 1) зарядное устройство (*зарядка от телефона, забыл зарядку, зарядка сломалась*); 2) количество заряда аккумуляторной батареи (*зарядки 20 %, зарядка заканчивается*). Но не всегда предполагаемая сочетаемость слов помогает в поиске примеров для иллюстрации того или иного значения слова. Особенно это касается слов с широким (диффузным) значением. Одно из таких слов – *мура*. В особой словарной зоне PRAGM («прагматика») отмечена возможность использования этого слова с разными значениями в зависимости от контекста. Обращает на себя внимание и тот факт, что не указаны источники этих употреблений. Это не в последнюю очередь связано с тем, что «задать» такие употребления в Национальном корпусе русского языка невозможно, ср.:

МУРА.

1. DEF: о чем-л. бессмысленном, бессодержательном, плохого качества.

2. DEF: о чем-л. несущественном, малозначительном.

3. DEF: о чем-л., что не нравится говорящему, раздражает его, а также о том, что говорящий не хочет или не считает нужным называть.

PRAGM: для разговорной речи типично употребление слова *мура* в качестве «опустошенного» слова: *Вот кто это безобразие / эту* ***муру*** */ на стенку впендюрил?; Убери отсюда эту* ***муру*** [о книгах и бумагах на столе]// В конструкциях с перечислением, занимая место в конце фразы, оно может выполнять роль недостающего родового обозначения: *Мы ни магнита / ни браслета / ни бус / никакой* ***муры*** *этой не купили в итоге//,* – может использоваться для заполнения интонационной «дыры» или облегчения поиска слова: *У нас все ингредиенты / кроме этой... / ну как ее /* ***муры*** *такой... посыпки//.* Подобные слова легко взаимозаменяемы*: Куда мы это (эту вещь / белиберду / дрянь / ерунду / музыку /* ***муру*** */ хрень / штуку...) денем?*[1]

В заключение хотелось бы еще раз подчеркнуть, что для поиска примеров в ТСРР используются самые разные интернет-источники, каждый из них имеет как свои достоинства, так и свои недостатки. В частности, Национальный корпус русского языка позволяет осуществлять поиск нужного слова в различных подкорпусах и конкретизировать запросы, задавая необходимые параметры слов. В то же время у многих ресурсов в интернете, которые позволяют проследить реальное живое употребление того или иного разговорного слова (например, сервис «Яндекс Дзен» и отзывы покупателей различных маркетплейсов), поиск примеров в данный момент невозможен. Хочется выразить надежду, что тексты, представленные на этих сайтах, пополнят Национальный корпус русского языка.

Литература

Толковый словарь русской разговорной речи : в 5 т. / отв. ред. Л. П. Крысин, М., 2014–2022.

[1] Фрагмент словарной статьи Е. В. Какориной.

Научное издание

«ЦИФРА» В СОЦИАЛЬНО-ГУМАНИТАРНЫХ ИССЛЕДОВАНИЯХ: МЕТОД, ПОЛЕ, РЕАЛЬНОСТЬ

ISBN 978-5-9624-2372-2

Корректор *Н. А. Михайлова*
Дизайн обложки: *П. О. Ершов*

Темплан 2025. Поз. 25
Уч.-изд. 5,5

ИЗДАТЕЛЬСТВО ИГУ
664082, г. Иркутск, ул. Лермонтова, 124