Известия Российского государственного педагогического университета им. А. И. Герцена, 2026, № 219
Izvestia: Herzen University Journal of Humanities & Sciences
www.izvestiaherzen.ru

Филологические науки

УДК 81'33

EDN PPPOSQ
https://www.doi.org/10.33910/1992-6464-2026-219-269-279

Научная статья

Проблемы лемматизации художественного текста: нераспознанные слова в «Корпусе русского рассказа 1900–1930 гг.»

Т. Г. Скребцова ✉[1], А. О. Гребенников [1]

[1] Санкт-Петербургский государственный университет,
199034, Россия, г. Санкт- Петербург, Университетская наб., д. 11

Для цитирования: Скребцова, Т. Г., Гребенников, А. О. (2026) Проблемы лемматизации художественного текста: нераспознанные слова в «Корпусе русского рассказа 1900–1930 гг.». *Известия Российского государственного педагогического университета им. А. И. Герцена*, № 219, с. 269–279. https://www.doi.org/10.33910/1992-6464-2026-219-269-279 EDN PPPOSQ

Получена 31 июля 2025; прошла рецензирование 7 сентября 2025; принята 26 февраля 2026.

Финансирование: Исследование не имело финансовой поддержки.

Аннотация

Введение. Несмотря на то что современные морфоанализаторы способны не только обращаться к словарю словоформ, но и сегментировать незнакомые единицы и выводить гипотетические леммы, некоторые словоформы при лемматизации художественного текста все равно не удается идентифицировать.

Материалы и методы. В статье рассматривается проблема распознавания слов при автоматической обработке текстов на материале «Корпуса русского рассказа 1900–1930 гг.» (Корпуса) — представительного электронного корпуса, который содержит несколько тысяч текстов, написанных в России (а затем и в Советском Союзе) в первые три десятилетия XX века. Сопоставив данные частотного словаря для выборки из Корпуса с «Большим орфографическим словарем русского языка», мы получили список таких нераспознанных элементов для текстов русских рассказах начала XX века. Далее, попытавшись восстановить соответствующее исходное слово из текста и понять, почему оно не было распознано, мы идентифицировали типичные проблемы, с которыми сталкивается морфоанализатор при обработке литературных текстов.

Результаты исследования. Нераспознанные элементы подразделяются на несколько групп: аббревиатуры и сокращения, имена собственные, сложные слова, а также стилистически маркированные лексемы и слова, содержащие латинские буквы. Многие из них не зафиксированы в толковых словарях русского языка. В статье приводятся статистические данные по количеству нераспознанных единиц в каждой группе и динамике соответствующих изменений за указанные три десятилетия. Показано, что это количество неодинаково в разные периоды. Общая тенденция состоит в увеличении числа таких слов от первого, довоенного периода, ко второму (военно-революционному) и третьему (раннесоветскому). Наиболее заметный рост нераспознанных словоформ наблюдается в рассказах, созданных в советский период. Выдвигаются гипотезы, призванные объяснить существенные расхождения в количестве нераспознанных слов по отдельным периодам: прежде всего, мы апеллируем к экстралингвистическим факторам, а именно к изменениям в общественно-политической обстановке.

Заключение. Изменение окружающего мира неизбежно влечет за собой изменение языка, и именно рассказы, в силу их жанровой специфики, отражают эти изменения быстрее всего.

Ключевые слова: морфоанализатор, лемматизация, художественный текст, рассказ, Корпус русского рассказа

Lemmatization of fiction: Unidentified words in the Russian short story corpus 1900–1930

T. G. Skrebtsova ✉[1], A. O. Grebennikov [1]

[1] St. Petersburg State University, 11 Universitetskaya Emb., Saint Peterbsurg 199034, Russia

For citation: Skrebtsova, T. G., Grebennikov, A. O. (2026) Lemmatization of fiction: Unidentified words in the Russian short story corpus 1900–1930. *Izvestia: Herzen University Journal of Humanities & Sciences*, no. 219, pp. 269–279. https://www.doi.org/10.33910/1992-6464-2026-219-269-279 EDN PPPOSQ

Received 31 July 2025; reviewed 7 September 2025; accepted 26 February 2026.

Funding: The study did not receive any external funding.

Abstract

Introduction. Modern morphological analyzers are capable not only of referring to a dictionary of word forms, but also of segmenting unfamiliar units and deriving hypothetical lemmas. However, some word forms of literary texts still cannot be identified during lemmatization.

Materials and Methods. This study examines the problem of word recognition based on material from the Russian Short Story Corpus 1900–1930, a representative electronic corpus containing several thousand texts written in Russia and later the Soviet Union during the first three decades of the 20th century. By comparing frequency dictionary data of a sample from the Corpus with the Russian Orthographic Dictionary, we obtained a list of unrecognized word forms. We then attempted to restore the original words from the text and understand why they were not recognized, which allowed us to identify typical problems in the automatic lemmatization of literary texts.

Results. Unrecognized elements fall into several groups: abbreviations and acronyms, proper names, complex words, stylistically marked lexemes, and words containing Latin letters. Many of these are not found in Russian explanatory dictionaries. The article provides statistical data on the number of unrecognized units in each group and how these figures changed over the three decades. The general trend shows an increase in the number of such words from the first, pre-war period to the second (war and revolution) and third (early Soviet) periods. The most noticeable increase in unrecognized word forms is observed in stories from the Soviet period. We put forward hypotheses to explain the significant differences in the number of unrecognized words in different periods, primarily by appealing to extralinguistic factors, namely changes in the socio-political situation.

Conclusions. Changes in the socio-political environment inevitably lead to changes in language; due to their genre, short stories reflect these changes most rapidly.

Keywords: morphological analyzer, lemmatization, fiction, short story, the Russian Short Story Corpus

Введение

Автоматический морфологический анализ текста на естественном языке на протяжении многих лет остается актуальной проблемой компьютерной лингвистики. Ключевым этапом морфоанализа является лемматизация — сведение словоформ, встретившихся в тексте, к леммам. Для языков с развитым словоизменением, таких как русский, лемматизация может быть сопряжена с трудностями. Современные морфоанализаторы (или, как их называют в последнее время, морфопроцессоры), помимо заложенного в них словаря словоформ, нередко включают в себя механизм морфологического предсказания (Большакова и др. 2017, 51–52; Кузнецов и др. 2019, 32–43). Если словоформа не распознана (например, в случае имен собственных), механизм предсказания пытается ее сегментировать так, чтобы вывести гипотетическую лемму. При этом, как правило, порождается не один, а несколько кандидатов в леммы.

В разных типах текстов проблемы с правильным выведением соответствующей леммы возникают по разным причинам. Так, академические тексты изобилуют специальными терминами, которые отсутствуют в общих словарях, а следовательно, и в парадигмах морфоанализатора. В официально-деловой литературе, помимо терминов, широко представлены номенклатурные обозначения и многоморфемные дериваты. Тексты средств массовой информации обычно включают в себя большое число имен собственных, относящихся к персоналиям, географии,

истории, культуре разных стран. При автоматической обработке литературных произведений также возникает ряд трудностей, характерных именно для этого типа текстов. В данной работе мы попытались создать предварительную типологию соответствующих случаев, опираясь на материалы «Корпуса русского рассказа 1900–1930 гг.».

Корпус русского рассказа 1900–1930 гг.

«Корпус русского рассказа 1900–1930 гг.» (далее — Корпус) — это представительный электронный ресурс, который содержит несколько тысяч текстов, написанных в России и Советском Союзе в первые три десятилетия XX века. Этот период был отмечен рядом драматических политических событий (Русско-японская война, Русская революция 1905 г., Первая мировая война, Февральская и Октябрьская революции, Гражданская война), которые кардинально изменили ход российской истории и оказали сильное воздействие на язык и литературу.

Основная идея Корпуса заключается в том, чтобы создать репрезентативное собрание рассказов, в котором будут широко представлены русские авторы, творившие в начале XX века — от всемирно известных (А. П. Чехов, Л. Н. Толстой, И. А. Бунин, М. Горький), до практически позабытых (например, Б. А. Верхоустинский, В. Я. Ленский, И. Ф. Колотовкин, Е. Н. Опочинин, О. П. Снегина). Весь временной отрезок 1900–1930 гг. делится на три периода в соответствии с ключевыми историческими вехами: 1) предвоенное время (1900–1913), 2) военное время, охватывающее Первую мировую войну, Февральскую и Октябрьскую революции, Гражданскую войну (1914–1922), и 3) послевоенное время, или ранний советский период (1923–1930).

На базе Корпуса была создана случайная выборка, содержащая 310 рассказов 300 писателей, приблизительно по 100 рассказов за каждый из трех периодов. Автор может быть представлен одним, случайно выбранным, рассказом за один период. Эта выборка служит своеобразным испытательным полигоном для лингвистов и литературоведов, позволяя им изучать жанровые особенности русского рассказа начала XX века и анализировать изменения, которые претерпевала его тематика, нарративные структуры, языковые и стилистические особенности при переходе от одного периода к другому (см. Sherstinova, Skrebtsova 2019; Skrebtsova 2021). Лингвостатистические методики используются с тем, чтобы получить количественные данные о лексике русских рассказов (Grebennikov et al. 2023; Sherstinova et al. 2020).

Материалы и методы

В результате лемматизации вышеупомянутой выборки было выделено 1 077 970 различных словоформ, которые были сведены к 124 081 лемме. В их числе оказались как реально существующие леммы, так и гипотетические (построенные посредством механизма предсказания). В свою очередь, среди гипотетических лемм высока доля фиктивных (не существующих в языке).

Заметим, что в целом отношения между словоформами и леммами можно охарактеризовать как многозначные. С одной стороны, к одной лемме может быть отнесено несколько словоформ. С другой стороны, словоформа может оказаться неоднозначной, допускающей сведение к нескольким различным леммам. Последнее часто случается и с нераспознанными словоформами, поскольку морфоанализатор пытается их альтернативно сегментировать и соотнести с имеющимися парадигмами. Таким образом, некоторые нераспознанные словоформы могут быть связаны с большим количеством гипотетических лемм, которые по большей части являются фиктивными.

Результаты лемматизации затем были наложены на список заголовочных слов, автоматически извлеченный из «Большого орфографического словаря русского языка» (далее — Орфографический словарь) (Бархударов и др. 2007). Те леммы, которые совпали с заголовочными словами Орфографического словаря, были исключены из дальнейшего рассмотрения, а оставшиеся (более 14 500 единиц суммарно за период 1900–1930 гг.) составили наш исследовательский материал. Примечательно, что их количество приблизительно одинаково в рассказах довоенного и военного времени (грубо говоря, по 5000 единиц), но резко возрастает в произведениях, созданных в Советской России, где оно достигает почти 7500 единиц (разумеется, списки лемм за разные периоды пересекаются). Следует особо подчеркнуть, что речь идет о количестве различных нераспознанных лемм в разные периоды, а не о частоте появления тех или иных единиц.

Было бы неверно утверждать, что рассматриваемый нами список состоит исключительно из фиктивных, не существующих в языке, лемм. Некоторые «правильные» леммы по тем или иным причинам не нашли отражения в Орфографическом словаре, в частности редупликации

типа *едва-едва, белый-белый*, сложные слова *старуха-мать, красно-белый*. Однако даже беглого взгляда на материал достаточно, чтобы понять, что фиктивных лемм в списке гораздо больше, чем правильных.

Двигаясь от материала и анализируя каждую из единиц списка, мы пытались мысленно восстановить соответствующее исходное слово из текста и понять, почему оно не было распознано морфоанализатором. Обобщая этот опыт, мы пришли к выделению типичных проблем, с которыми сталкивается морфоанализатор при обработке литературных текстов. Ниже представлен обзор соответствующих групп слов, снабженный примерами из нашего материала. Заметим, что некоторые группы пересекаются, т. е. единицы могут присутствовать более, чем в одной группе. В рамках каждой группы мы также прослеживаем количественную динамику по периодам, выявляем соответствующие тенденции и пытаемся предложить им объяснения.

Заранее следует оговорить тот факт, что мы не выделяем в отдельную группу опечатки, а также написание слов, не соответствующее современным орфографическим нормам (некоторые рассказы полузабытых авторов не переиздавались в советское время). Эти случаи носят сугубо формальный характер.

Основные категории нераспознанных слов

1. Слова, написанные латиницей

Очевидно, что процедура распознавания слов не работает, если слово написано с использованием букв другого алфавита. Как правило, в тексте рассказов имеет место вкрапление целых иноязычных словоформ, гораздо реже — отдельных компонентов (ср. *ex-поэт, п-ский*).

В довоенных рассказах встречаются в основном французские слова (*mademoiselle, chef, ingenieur, journal, chose, merci, laissez-passer* и т. д.), а также широко известные латинские слова и выражения (*modus vivendi, volensnolens, ergo, disputandum, gaudeamus*). Русские фамилии могут быть написаны на французский манер (*Arbousoff*).

В следующий период, в связи с началом Первой мировой войны, в текстах появляются также немецкие (*donnerwetter, noch, wohin, gewaltiger, Zeitung*), итальянские (*attenti, fermato, giorno*) и польские (*powracam, wrzystko, wygoi, rosbici, przez*) словоформы.

В советское время присутствие в рассказах иностранных слов заметно снижается: от 203 лемм в довоенный период к 124 в военное время и далее вплоть до 75 в советский период, что объясняется, по-видимому, изменившейся общественно-политической обстановкой.

2. Аббревиатуры

Аббревиатуры принято считать отличительной чертой русского языка советского времени (ср., Панов 1968; Фесенко, Фесенко 1955); неудивительно, что они широко представлены в нашем материале.

Хотя отдельные сокращения были в употреблении еще до 1914 г. (Карцевский 1923, 28), важным толчком к усилению этого словообразовательного механизма стала Первая мировая война, которая привела к появлению новых реалий (прежде всего военных учреждений и должностей) и, следовательно, новых номинаций. В обиходной речи эти, нередко длинные и громоздкие, словосочетания заменялись аббревиатурами (Мазон 2013, 206–20). Февральская, а затем Октябрьская революции 1917 г., существенно расширили сферу применения сложносокращенных слов: «Октябрьская революция распространила на весь правительственный, административный и общественный аппарат, который она пыталась создать, условную терминологию, построенную по принципу слоговой аббревиации» (Мазон 2013, 206–20). Анализируя язык этого периода, С. И. Карцевский отмечает: «Среди языковых новшеств периода 1905–1922 гг., сокращения являются самыми “революционными” и заслуживают внимательного рассмотрения» (Карцевский 1923, 4). В трудах лингвистов — современников революций и языковых изменений — можно найти подробную морфологическую классификацию бытовавших в то время аббревиатур, подкрепленную богатым фактическим материалом (Карцевский 1923, 45–52; Мазон 2013, 205–209; Селищев 2008, 157–17). Исследователи сходятся в том, что для 1920-х гг. наиболее характерными были различные варианты слоговых сокращений.

С точки зрения современного читателя, многие аббревиатуры, встречающиеся в русских рассказах раннесоветского периода, представляют сложность с точки зрения членения на компоненты и, следовательно, понимания (ср. *волъячейка, земгусар, замнаркомпуть, пролеткульт, губнаробраз, предревком, предбумтрест, упродкомиссар, церабкоон)*. Большинство из них не зафиксировано в толковых словарях русского языка ни целиком, ни по частям. Чем больше времени отделяет нас от того периода, тем менее понятным становится описываемый в рассказах

 https://www.doi.org/10.33910/1992-6464-2026-219-269-279

предметный мир (Маркасова 2011; Скребцова 2021).

Дело дополнительно осложняется тем, что, входя в язык, многие аббревиатуры быстро ассимилировались, развивали словоизменительные и производные формы (ср. *исполкомовец, женделегатка, губнаробразовский, горсоветский*). Излишне говорить, что эти слова также (за редким исключением) не получили словарной фиксации.

По материалам рассматриваемой выборки, дореволюционные аббревиатуры представлены такими конвенциональными единицами, как *д-р, г-н, г-жа*. Впоследствии большинство из них вышло из употребления, а вместо них появились новые, относящиеся к советским реалиям. Общее количество сокращений последовательно увеличивается с 12 в довоенных рассказах до 45 и до 139 в рассказах, созданных в Советской России. Этот рост весьма показателен в свете отмеченных выше тенденций.

3. Стилистически маркированные слова

В нашем исследовании под стилистически маркированными словами понимается широкий и весьма разнородный круг единиц, обычно не включаемых в толковые словари русского языка. Это региональные и диалектные слова (*шо, бульба, шашмура, кулын, баско, цукор, шанюшка, челдон, хуварак, шалага*), эмоционально окрашенные дериваты (*Россиюшка, французик, бумаженка, чаишко, стервюга, стакашка, фуражонка, студентишко, худобенький, цыганщинка*), устаревшие (*днесь, клегтать, человецех, шематон*) и просторечные слова и словоформы (*пущать, убег, графья, ихний, опосля, фершал, откудова, наскрозь, сурьезный, струмент, испужать, пачпорт, пымать, рупь, лягешь, могем, делов*), жаргонизмы (*целкач, мозляк*), звукоподражательные междометия (*гы-гы-гы, ур-ра-а, э-эх, у-у, а-ха-ха, бряц-бряц-бряц*), экспрессивные редупликации (*крепко-крепко, большой-большой, погоревал-погоревал, совсем-совсем*), транслитерированные иностранные слова (*вурст, катцен-яммер, югенд, аттанде-с*). Сюда же мы относим графически искаженные слова (в том числе неологизмы и аббревиатуры), призванные отразить специфику их артикуляции персонажем (*ерой, емназистка, хрещеный, машкарад, дилектор, камунист, тернационал, електрофикация, дикдатура, сафнарком, РЫ-СЫ-ФЫ-СЫ-РЫ, РЕ-ЕС-ДЕ-РЕ-ПЕ, хле-еб, ш-ш-ш-шесть, да-ай*). Общим для всех таких единиц является выраженный коннотативный компонент значения, который может включать в себя множественные аспекты: эмоциональный, оценочный, стилистический, социальный, территориальный, исторический, профессиональный и т. д. (ср. Стернин 1979, 68–99).

Употребление стилистически маркированных слов в художественном произведении в принципе может вытекать из особенностей личности автора (места проживания, социокультурного происхождения и пр.). Однако обычно такие единицы используются намеренно, для создания задуманного эффекта. Вкладывая стилистически окрашенную лексику в уста литературного персонажа, писатель выдает ему своеобразный «языковой паспорт», способный характеризовать его с разнообразных точек зрения, включающих национальность, социальную и профессиональную принадлежность, возраст, физическое и эмоциональное состояние, уровень культуры и степень образованности, отношение к собеседнику или окружающей обстановке и т. п. (Стернин 1979, 92–94).

Как и аббревиатуры, стилистически маркированные элементы вносят значительный вклад в количественное расхождение нераспознанных лемм по периодам. И также, как аббревиатуры, они демонстрируют многократный рост от первого периода к третьему, ср. 210:891:1798. Аналогичным образом, этот рост принято объяснять внешними причинами. Так, анализируя «язык революционной эпохи», А. М. Селищев отмечает «крепкие словечки и выразительные сочетания языка деревни, фабрики, низших слоев населения города» (Селищев 2008, 69), фамильярные и грубые выражения, циничную ругань и мат, воровской жаргон (Толстая 2020, 68–85). Подводя итог наблюдениям, автор констатирует: «Эмоционально-экспрессивная функция речи имела огромное значение в революционные годы <…>. Элементы, служащие для выражения эмоциональности речи, становятся в течение времени модными и часто употребляются в среде советских граждан» (Селищев 2008, 121–122). В другом известном исследовании, посвященном влиянию Первой мировой войны и революции на русский язык, находим утверждение о «ходовом употреблении в литературном языке таких слов и выражений, которые совершенно не допускались в прежние годы литературными приличиями» (Карцевский 1923, 70).

Применительно к художественной литературе военно-революционного и раннесоветского периода, по-видимому, можно говорить о воздействии нескольких факторов (все из которых носят социальный характер). Во-первых, после революции кардинальным образом обновился литераторский состав: многие профессиональные авторы были вынуждены эмигрировать,

а на их месте появились так называемые пролетарские писатели, язык и стиль которых был совершенно иным. Среди новых авторов, к тому же, были выходцы из провинции, изображавшие жизнь далеких регионов, которые прежде нечасто становились местом действия литературных произведений: Средней Азии, Сибири, Крайнего Севера. Соответственно, региональные слова попадали в текст (наряду с региональными именами собственными, см. ниже пункт 4).

Во-вторых, радикальные преобразования, вызванные Октябрьской революцией, привели к существенным изменениям в персонажах, обстановке, темах и сюжетах повествования. В новом советском государстве вымышленные персонажи, как правило, имели иную (крестьянскую или пролетарскую) биографию, действовали в иных обстоятельствах, и говорили непривычно по сравнению с героями классической литературы.

В-третьих, другим стал массовый адресат литературных произведений. Круг читателей расширился и социально изменился. Подводя итог, можно сказать, что новые (пролетарские) писатели писали для новых (советских) читателей о новом (послереволюционном) мироустройстве.

Но и это еще, по-видимому, не всё. В более широком смысле, совершенный партией большевиков переворот в политическом строе как будто требовал переворота и в языке, так как старый казался не годным для передачи новых смыслов и ценностей. При этом базовая денотативно-ориентированная лексика сохранилась; изменения отразились больше на выражении чувств, установок, оценок. Отсюда — существенно возросшая доля стилистически окрашенной, эмоционально-экспрессивной лексики.

4. Имена собственные

Имена собственные, по определению, не включаются в толковые словари, и, следовательно, при морфологической обработке попадают в число нераспознанных элементов. Две основные категории литературной ономастики — это имена лиц (антропонимы) и названия мест (топонимы), хотя в художественном тексте могут встречаться и другие виды «именованных сущностей»: клички животных, названия кораблей, заводов, музеев, и т. д.

Имена лиц — это, как правило, русские антропонимы, то есть имя, отчество и фамилия, которые могут использоваться как по отдельности, так и в различных сочетаниях. В рассказах встречаются также прозвища и иностранные имена. В принципе, частотные антропонимы можно было бы добавить в словарь морфоанализатора и таким образом автоматически распознавать. Проблема, однако, заключается в том, что русские личные имена обычно имеют большое число дериватов, которые с трудом поддаются учету. К примеру, для встретившихся в текстах производных единиц от имени *Вася/Василий* морфоанализатор выдал список из 20 лемм: *Васен, Васенька, Васи, Василие, Василить, Василич, Василишка, Васильe, Васильев, Васильевна, Васильевный, Васильевой, Васильевский, Васильич, Васина, Васить, Васиша, Васишка, Вась, Васютка*, — большинство из которых, как нетрудно заметить, фиктивные. В целом, фиктивные леммы, образованные от личных имен, составляют заметную часть общего списка нераспознанных элементов.

Помимо имен вымышленных персонажей, в русских рассказах начала XX века можно встретить упоминание известных поэтов и писателей (*Шекспир, Мильтон, Буссенар, Боккаччо, Бодлер, Пушкин, Достоевский, Некрасов, Тургенев, Надсон, Маяковский*), композиторов (*Шопен, Шуберт, Моцарт, Бетховен, Чайковский, Мопассан*), философов (*Ницше, Шопенгауэр, Энгельс, Маркс*), монархов и государственных деятелей (*Габсбурги, Цезарь, Ллойд Джордж*), ученых (*Эвклид, Эйнштейн, Фрейд*), бизнесменов (*Ротшильд, Нобель*) и прочих известных личностей (*Иисус, Колумб, Спартак, Шаляпин, Ливингстон, Клеопатра*). Встречаются также отсылки к литературным и мифологическим персонажам, таким как *Раскольников, Дон Кихот, Гамлет, Джульетта, Фальстаф, Мефисто, Немезида, Артемида*.

Схожим образом, названия мест в художественном тексте могут быть как реальными, так и вымышленными. Первые обычно отсылают к хорошо известным объектам — странам, регионам, крупным городам, океанам и морям, большим рекам, высоким горам и т. д. Упоминание в тексте подобных мест, помимо номинативной функции, выполняет важную прагматическую задачу, активизируя читательские фоновые знания и тем самым способствуя ассоциативному приращению смысла, ср. (Воронова 2000). Что касается названий небольших населенных пунктов, рек, озер и т. п., они в художественной литературе обычно вымышленные.

Помимо географических объектов нашей страны, в русских рассказах начала XX века упоминается множество мест по всему миру, включая *Сан-Франциско, Нью-Йорк, Шотландия, Монблан, Монпелье, Лодзь, Клондайк, Капри, Греция, Гольфстрим, Генуя, Висбаден, Версаль, Бельгия, Австрия, Япония, Чикаго, Флоренция,*

 https://www.doi.org/10.33910/1992-6464-2026-219-269-279

Палестина, Иерусалим, Вавилон, Бразилия, Брюссель, Биарриц, Афины, Александрия, Цусима, Харбин, Сербия, Персия, Палермо, Ницца, Лондон, Африка, Дарданеллы, Афон, Париж, Монако, Европа и др.

Вне зависимости от того, реальное название или вымышленное, его лемма далеко не всегда правильно восстанавливается анализатором, ср. *Японие, Шотландий, Могилева, Филадельфии, Гибралтарый* и пр.

Общее количество имен собственных в рассказах также последовательно растет от периода к периоду, причем существенный скачок снова происходит в советское время, ср. 1160:1210:1795. Этот факт сам по себе примечателен и не вполне понятен. Трудно вообразить, чтобы в 1920-е гг. советские люди много путешествовали по свету. Остается предположить, что прирост происходит за счет большего разнообразия в месте действия (в пределах СССР) и в личных именах (прежде всего, за счет увеличения числа эмоционально окрашенных дериватов).

5. Сложные и составные слова

В данном разделе речь пойдет о большой и весьма неоднородной группе элементов, которую мы выделяем по сугубо формальному признаку, а именно наличию двух и более корней в составе одной графематической единицы. Присутствие дефиса, разделяющего корни (или целые слова) внутри графематической единицы, является частотным, но необязательным признаком. В своем практическом исследовании мы не считаем нужным вдаваться в тонкости неустоявшейся терминологии и классификации (ср. *сложные, составные, сложносоставные слова, аппозитивные сочетания, композиты, слова-гибриды* и пр.), а будем рассматривать соответствующий материал исключительно с точки зрения тех трудностей, которые возникают при лемматизации.

В формальном аспекте рассматриваемые единицы представлены несколькими структурными моделями, такими как «существительное + существительное» (*девушки-сестры*), «прилагательное (основа) + прилагательное (полное)» (*бело-голубой*), «наречие + наречие» (*верх-вниз*), «числительное + числительное» (*два-три*), «числительное + существительное» (*пятикурсник*), «числительное + прилагательное» (*двухглавый*), «прилагательное + прилагательное» (*белый-белый*).

С содержательной точки зрения, рассматриваемые единицы могут относиться к разным аспектам картины мира: люди (*сын-гимназист, студент-технолог*), время (*час-два, годок-другой*), пространство (*далекий-далекий, вверх-вниз*), количество (*пять-шесть, рупь-тридцать*), цвет (*серо-лиловый, светло-сиреневый*), религиозные и культурные явления и пр. В последнем случае обе составные части обычно пишутся с большой буквы, например, *Царь-Колокол*, *Иуда-Предатель*, *Иоанн-Креститель*. В номинациях людей могут сочетаться разные аспекты идентичности, например, имя и профессия (*Сологуб-поэт*), профессия и национальность (*доктор-немец*), национальность и оценка (*полька-дура*) и др. (ср. *буржуй-жених, красавица-жена, Ирка-пионерка, Степушка-покойник, гений-поэт, враги-соседи, гости-мужчины, подруга-насмешница*).

Семантические отношения между составными частями разнообразны и недостаточно изучены (так, в основательном исследовании (Толстая 2020) акцент сделан на синтаксической классификации сложных слов, а характеристика смысловых связей ей подчинена). Однако даже при поверхностном просмотре материала можно заметить некоторые распространенные виды связей.

В модели «существительное + существительное» компоненты могут иметь взаимодополняющие значения, представляя один и тот же объект с разных сторон, как в примере *мальчик-учитель*, где одна часть передает идею юного возраста, а другая указывает на профессию. Оба компонента денотативно ориентированы и связаны между собой отношением сочинения. В других примерах компоненты имеют неравный статус. Так, в *земля-матушка* первая часть выражает денотативное значение, а вторая — коннотацию (эмоциональное отношение), причем линейное расположение компонентов может быть и обратным, ср. *добряк-муж*. В последних двух примерах составные части связаны между собой тем, что можно назвать отношением квалификации. Встречается также отношение спецификации, где компоненты связаны между собой родо-видовыми отношениями, ср. *орел-кондор, рабочий-декоратор, шапка-ушанка.*

Отмеченные отношения, разумеется, не исчерпывают всех типов семантических связей в модели «существительное + существительное» (подробнее см. в (Перцова 2000)). Мы останавливаемся на них потому, что именно случаи такого рода вызывают наибольшие проблемы при лемматизации. Во-первых, будучи открытым классом, соответствующие единицы в принципе не могут быть исчислены. Во-вторых, они пишутся через дефис, причем в косвенных формах словоизменение обычно затрагивает обе части. Однако лемматизатор устроен так,

что он пытается вывести лемму только для последнего компонента, оставляя начальный нетронутым; в итоге лемма выводится неверно (ср. *сороки-воровка, торговцев-китаец, шорохи-вздох, товарищами-студент*). Напротив, другие виды семантических отношений между компонентами рассматриваемой модели обычно встречаются в сложных словах со слитным написанием. Такие слова, как правило, содержатся в словарях, но даже если нет, лемма может быть восстановлена правильно, так как изменению подвергается лишь финальная часть (ср. *киносъемщик, радиотрубка, народоармеец, крутосклон*).

Модель «прилагательное (основа) + прилагательное (полное)» охватывает написанные через дефис единицы; в более подробном виде их структура описывается формулой «основа прилагательного + интерфикс + прилагательное». Эта модель представляется наиболее частотной в нашем материале (ср. *сумеречно-таинственный, спокойно-радостный, скипидарно-смолистый*). Семантические отношения между компонентами многообразны и не были тщательно изучены и описаны. Если говорить, в частности, о названиях цветов, то можно предложить предварительную типологию, включающую такие отношения, как сочетание (*бело-зеленый*), уточнение (*темно-красный*), сравнение (*молочно-голубой*), метафору (*сурово-черный*). В то же время неясно, насколько предложенные виды отношений релевантны для всей группы в целом.

Модель «числительное + числительное» обычно передает идею приблизительной оценки, как, например, в *пять-шесть*. Модель «прилагательное + прилагательное» проявляется в редупликации (*белый-белый, маленький-маленький*), делая обозначаемый признак более выраженным, акцентированным. Схожая семантика присуща многим реализациям модели «наречие + наречие» (*быстро-быстро, скоро-скоро*).

Если рассматривать подобный материал с точки зрения лемматизации, очевидно, что шансы на успех у единиц, построенных по разным моделям, различен. Правильные леммы будут с большей вероятностью получены для слитно написанных сложных слов, а также для экземпляров моделей «наречие + наречие» и «прилагательное (основа) + прилагательное (полное)», поскольку их начальный компонент не затрагивается словоизменением. Для косвенных словоформ, построенных по другим моделям, лемма может быть корректно выведена только для конечного компонента. Начальный компонент остается не проанализированным, и слово целиком не распознается.

В статистическом аспекте, использование сложных и составных слов (вместе взятых, без какой-либо дифференциации) выглядит довольно стабильным в первый и второй период. В советское время оно становится более частым, ср. 567:570:767. Таким образом, эта группа лексики также вносит вклад в общий рост числа нераспознанных единиц в последний период.

Возвращаясь к цветообозначениям, можно наблюдать, хоть и в меньшем масштабе, схожую статистическую картину, ср. 45:49:68. Примечательно, что в довоенный период самой частой нераспознанной единицей является лексема *бледно-голубой* (19 случаев). В рассказах военного времени ее частота падает до единицы, а лидером становится *ярко-красный* (пять случаев). В рассказах послевоенной эпохи *серо-зеленый* и *темно-синий* возглавляют список (по пять случаев), а *бледно-голубой* и *ярко-красный* упоминаются лишь по одному разу. Если рассматривать эти изменения в историческом контексте, то они могут показаться символически значимыми. Так мирный, романтичный и нежный *бледно-голубой* уступил место интенсивному *ярко-красному* цвету революционных знамен и крови, чтобы в свою очередь смениться более сдержанными и размытыми оттенками *темно-синего* и *серо-зеленого*.

Выводы

Выше мы рассмотрели наиболее часто встречающиеся проблемы, связанные с лемматизацией художественного текста. Если единица отсутствует в словаре словоформ, морфоанализатор пытается ее сегментировать, с тем чтобы прийти к гипотетической лемме (или леммам), которая может оказаться реально существующей или фиктивной. Общий список лемм (реальных и фиктивных), полученных по словоформам из выборки русских рассказов, стал отправной точкой настоящего исследования.

Этот список был сопоставлен со статьями Орфографического словаря. В результате он претерпел существенное сокращение за счет удаления совпадающих элементов. Остались леммы, которые не были обнаружены в Орфографическом словаре: в основном, они были фиктивными. Внимательное изучение позволило реконструировать правильные леммы и выделить основные группы случаев, вызывающих проблемы при лемматизации художественного текста. Помимо устаревших написаний и опечаток, правильное распознавание леммы

 https://www.doi.org/10.33910/1992-6464-2026-219-269-279

затруднено, если словоформа является аббревиатурой, именем собственным, сложным или составным словом, стилистически маркированной лексемой или содержит латинские буквы.

Статистические данные свидетельствуют, что количество нераспознанных единиц не одинаково в разные периоды. Общая тенденция состоит в увеличении числа таких слов от первого, довоенного периода, ко второму (военно-революционному) и третьему (раннесоветскому). Исключение составляют слова, написанные латиницей, но их процент невелик во все периоды, поэтому они не оказывают существенное влияние на общую динамику. Наиболее заметный рост нераспознанных словоформ наблюдается в рассказах, созданных в советский период. Пытаясь объяснить данные статистики, мы апеллируем к экстралингвистическим факторам, а именно к изменениям в общественно-политической обстановке. Меняется мир — меняется и язык, а рассказ, в силу своей жанровой специфики (короткого объема, публикации в периодических изданиях), достаточно быстро фиксирует эти изменения. Рассказы 1920-х гг. служат тому наглядным подтверждением.

Устный доклад по теме статьи был представлен на V Международной научной конференции по инженерной и прикладной лингвистике «Пиотровские чтения 2024» (Санкт-Петербург, 22 ноября 2024).

Конфликт интересов

Авторы заявляют об отсутствии потенциального или явного конфликта интересов.

Conflict of Interest

The authors declare that there is no conflict of interest, either existing or potential.

Вклад авторов

Авторы внесли равный вклад в подготовку статьи.

Author Contributions

The authors participated in writing the article equally.

Список литературы

Бархударов, С. Г., Протченко, И. Ф., Скворцов, Л. И. (2007) *Большой орфографический словарь русского языка: более 106000 слов*. Москва: Оникс; Мир и Образование, 1160 с.

Большакова, Е. И., Воронцов, К. В., Ефремова, Н. Э. и др. (2017) *Автоматическая обработка текстов на естественном языке и анализ данных*. Москва: Изд-во НИУ ВШЭ, 269 с.

Воронова, И. Б. (2000) *Текстообразующая функция литературных имен собственных (на материале эпических произведений XIX–XX вв.). Автореферат диссертации на соискание степени кандидата филологических наук*. Волгоград, Волгоградский государственный педагогический университет, 24 с.

Карцевский, С. И. (1923) *Язык, война и революция*. Берлин: Русское универсальное издательство, 72 с.

Кузнецов, С. А., Скребцова, Т. Г., Суворов, С. Г., Клементьева, А. В. (2019) *Лингвистический анализатор: преобразование текста в метаязыковую структуру данных*. Санкт-Петербург: Изд-во СПбГУ, 238 с.

Мазон, А. (2013) Лексика войны и революции в России (1914-1918). Введение. Аббревиация. *Политическая лингвистика*, № 1, с. 203–210.

Маркасова, Е. В. (2011) Проблемы поиска и лексикографического описания советизмов 1920-30 гг. *Russian Language Journal*, т. 61, с. 94–118.

Панов, М. В. (1968) *Русский язык и советское общество: Социолого-лингвистическое исследование. Словообразование современного русского языка*. Москва: Наука, 300 с.

Перцова, Н. Н. (2000) Некоторые проблемы семантики словосложения. В кн.: А. С. Нариньяни (ред.). *Труды международного семинара Диалог'2000 по компьютерной лингвистике и ее приложениям в двух томах. Том 1. Теоретические проблемы*. Протвино: [б. и.], с. 246–247.

Селищев, А. М. (2008) *Язык революционной эпохи: из наблюдений над русским языком последних лет (1917–1926)*. Москва: УРСС, 248 с.

Скребцова, Т. Г. (2021) Новые реалии общественно-политической жизни 1920-х гг. и их отражение в русской литературе и лексикографии (на примере сложносокращенных слов). *Политическая лингвистика*, № 2 (86), с. 146–154.

Стернин, И. А. (1979) *Проблемы анализа структуры значения слова*. Воронеж: Изд-во Воронежского государственного университета, 156 с.

Толстая, С. М. (2020) Сложные слова и словосочетания: синтаксис и семантика. *Rocznik Slawisticzny*, т. LXIX, с. 167–180. https://www.doi.org/10.24425/rslaw.2020.134712

Фесенко, А. В., Фесенко, Т. (1955) *Русский язык при Советах*. Нью- Йорк: Rausen Bros., 222 с.
Grebennikov, A. O., Marusenko, N. M., Skrebtsova, T. G. (2023) Mapping word frequencies in fiction on sociopolitical context: the case of early 20th century Russian short stories. *Terra Lingüística*, vol. 14, no. 1, pp. 21–30. https://doi.org/10.18721/JHSS.14103
Sherstinova, T., Grebennikov, A., Skrebtsova, T. et al. (2020) Frequency word lists and their variability (the case of Russian fiction in 1900-1930). In: *Proceeding of the 27th Conference of Fruct Association*. Helsinki: FRUCT Oy Publ., no. 27, pp. 366– 373.
Sherstinova, T., Skrebtsova, T. (2019) Russian literature around the October revolution: A quantitative exploratory study of literary themes and narrative structure in Russian short stories of 1900-1930. In: *Ceur Workshop Proceedings. International Conference "Internet and Modern Society"*. Aachen: RWTH Aachen University Publ., vol. 2813, pp. 117– 128.
Skrebtsova, T. G. (2021) Thematic tagging of literary fiction: the case of early 20th century Russian short stories. In: *Ceur Workshop Proceedings. International Conference "Internet and Modern Society"*. Aachen: RWTH Aachen University Publ., vol. 2813, pp. 265–276.

References

Barkhudarov, S. G., Protchenko, I. F., Skvortsov, L. I. (2007) *The Large Spelling Dictionary of the Russian Language. Over 106 000 words*. Moscow: Oniks Publ.; Mir i Obrazovaniye Publ., 1160 p. (In Russian)
Bolshakova, Ye. I., Vorontsov, K. V., Yefremova, N. E. et al. (2017) *Automatic processing of natural language texts and data analysis*. Moscow: HSE University Publ., 269 p. (In Russian)
Fesenko, A. V., Fesenko, T. (1955) *The Russian language in the Soviet Era*. New York: Rausen Bros. Publ., 222 p. (In Russian)
Grebennikov, A. O., Marusenko, N. M., Skrebtsova, T. G. (2023) Mapping word frequencies in fiction on sociopolitical context: the case of early 20th century Russian short stories. *Terra Lingüística*, vol. 14, no. 1, pp. 21–30. https://doi.org/10.18721/JHSS.14103 (In English)
Kartsevskiy, S. I. (1923) *Language, War and Revolution*. Berlin: Russian Universal Publ., 72 p. (In Russian)
Kuznetsov, S. A., Skrebtsova, T. G., Suvorov, S. G., Klementyeva, A. V. (2019) *Linguistic analyser: converting text into a meta-language data structure*. Saint Peterburg: St. Petersburg State University Publ., 238 p. (In Russian)
Markasova, E. V. (2011) Issues of Identification and Lexicographic Description of Sovietisms of 1920-30s. *Russian Language Journal*, vol. 61, pp. 94–114. (In Russian)
Mazon, A. (2013) Lexis of war and revolution in Russia (1914—1918). Introduction. Abbreviation. *Political Linguitsics*, № 1, pp. 203–210. (In Russian)
Panov, M. V. (1968) *Russian Language and Soviet Society: A Sociological and Linguistic Study. Word formation of the modern Russian language*. Moscow: Nauka Publ., 300 p. (In Russian)
Pertsova, N. N. (2000) Some aspects of semantics of compound words. In: A. S. Narinyani (ed.). *Proceedings of the Internationl Seminar on Computaional Linguistics "Dialog" 2000. Vol. 1*. Protvino: [s. n.], pp. 246–247. (In Russian)
Selishchev, A. M. (2008) *The Language of The Revolutionary Epoch: Ovserving The Russian Language of Recent Years (1917–1926)*. Moscow: URSS Publ., 248 p. (In Russian)
Sherstinova, T., Grebennikov, A., Skrebtsova, T. et al. (2020) Frequency word lists and their variability (the case of Russian fiction in 1900-1930). In: *Proceeding of the 27th Conference of Fruct Association*. Helsinki: FRUCT Oy Publ., no. 27, pp. 366– 373. (In English)
Sherstinova, T., Skrebtsova, T. (2019) Russian literature around the October revolution: A quantitative exploratory study of literary themes and narrative structure in Russian short stories of 1900-1930. In: *Ceur Workshop Proceedings. International Conference "Internet and Modern Society"*. Aachen: RWTH Aachen University Publ., vol. 2813, pp. 117–128. (In English)
Skrebtsova, T. G. (2021) New sociopolitical realities of the 1920s as reflected in Russian literature and lexicography (on the basis of syllabic acronyms). *Political Linguitsics*, № 2 (86), pp. 146–154. (In Russian)
Skrebtsova, T. G. (2021) Thematic tagging of literary fiction: the case of early 20th century Russian short stories. In: *Ceur Workshop Proceedings. International Conference "Internet and Modern Society"*. Aachen: RWTH Aachen University Publ., vol. 2813, pp. 265–276. (In English)
Sternin, I. A. (1979) Issues of analysing the structure of word meaning. Voronezh: Voronezh State University Publ., 156 p. (In Russian)
Tolstaya, S. M. (2020) Complex words and phrases: syntax and semantics. *Rocznik Slawisticzny*, vol. LXIX, pp. 167–180. https://www.doi.org/10.24425/rslaw.2020.134712 (In Russian)
Voronova, I. B. (2000) *The text-forming function of literary proper names (based on epic works of the XIX–XX centuries). Extended abstract of the PhD dissertation (Philology)*. Volgograd, Volgograd Sate Pedagogical University, 24 p. (In Russian)

Сведения об авторах

Скребцова Татьяна Георгиевна, кандидат филологических наук, доцент, доцент кафедры математической лингвистики, Санкт-Петербургский государственный университет.
SPIN-код: 1500-6308, Scopus AuthorID: 57217422722, ORCID: 0000-0002-7825-1120, e-mail: t.skrebtsova@spbu.ru

Гребенников Александр Олегович, кандидат филологических наук, доцент, доцент кафедры математической лингвистики, Санкт-Петербургский государственный университет.
SPIN-код: 8561-9537, Scopus AuthorID: 57447286700, ORCID: 0000-0003-2856-5049, e-mail: a.grebennikov@spbu.ru

Authors

Tatyana G. Skrebtsova, PhD (Linguistics), Associate Professor, Associate Professor at the Department of Mathematical Linguistics, Saint Petersburg State University.
SPIN: 1500-6308, Scopus AuthorID: 57217422722, ORCID: 0000-0002-7825-1120, e-mail: t.skrebtsova@spbu.ru

Alexander O. Grebennikov, PhD (Linguistics), Associate Professor, Associate Professor at the Department of Mathematical Linguistics, Saint Petersburg State University.
SPIN: 8561-9537, Scopus AuthorID: 57447286700, ORCID: 0000-0003-2856-5049, e-mail: a.grebennikov@spbu.ru