ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 58 за 2021 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

Э.С. Клышинский, **В.А. Бунтякова,**
О.В. Карпик

Исследование грамматической неоднозначности наиболее частотных слов русского языка

Рекомендуемая форма библиографической ссылки: Клышинский Э.С., Бунтякова В.А., Карпик О.В. Исследование грамматической неоднозначности наиболее частотных слов русского языка // Препринты ИПМ им. М.В.Келдыша. 2021. № 58. 22 с. https://doi.org/10.20948/prepr-2021-58
https://library.keldysh.ru/preprint.asp?id=2021-58

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

Э.С.Клышинский, В.А.Бунтякова, О.В.Карпик

Исследование грамматической неоднозначности наиболее частотных слов русского языка

Москва — 2021

Клышинский Э.С., Бунтякова В.А., Карпик О.В.

Исследование грамматической неоднозначности наиболее частотных слов русского языка

В своих предыдущих исследованиях мы обнаружили, что в европейских языках грамматическая неоднозначность наиболее частотных слов ведет себя несколько иначе, чем в более редких словах. В данном исследовании мы более подробно анализируем причины этого явления, уделяя особое внимание первой тысяче наиболее частотных токенов. Исследование современных систем снятия омонимии и синтаксического анализа показало, что всплеск разнообразия частей речи, который наблюдается в наиболее частотных словах, приводит к увеличению числа ошибок на выходе этих систем.

Ключевые слова: грамматическая неоднозначность, квантитативный анализ, статистика распределения, русский язык

Eduard Klyshinsky, Valeria Buntyakova, Olesya Karpik

Investigation of Grammatical Ambiguity of Most Frequent words of the Russian Language

During our previous research, we found that the grammatical ambiguity of most frequent words of European languages has a different distribution in comparison with less frequent ones. In the current research, we investigate in more details the reasons of such a phenomenon; we pay a special attention to the first thousand of most frequent tokens. Our investigation of modern disambiguation systems demonstrated that the increase of language diversity, we had found for most frequent words, leads to increase of number of mistakes made by those systems.

Key words: grammatical ambiguity, quantitative analysis, distribution, statistics, the Russian language

Оглавление

Введение

Неоднозначность является одним из фундаментальных свойств всех языков, причем она может проявляться на разных уровнях анализа текстов. Выражение или высказывание называется неоднозначным, если оно может быть интерпретировано более чем одним способом [1]. В данной работе мы будем рассматривать грамматическую неоднозначность.

В целом лексическая неоднозначность делится на два типа: омонимию и полисемию. Лексема называется полисемичной, если у нее есть два и более связанных значений [1]. Полисемия широко распространена в языке и является закономерным результатом стремления к экономии. Например, в словаре С.А. Кузнецова для слова *падать* выделяется 10 значений: идти (об атмосферных осадках), уменьшаться, валиться вниз на землю под действием собственной тяжести и др.

Омонимия, напротив, — относительно редкое явление. Две лексемы называются полностью омонимичными, если они имеют несвязанные значения, но одинаковые грамматические свойства; две лексемы частично омонимичны, если они имеют несвязанные значения и совпадают в части форм [1]. Помимо этого, под омонимией также понимается совпадение словоформ одной лексемы, подобное явление также называется грамматической неоднозначностью. Например, *безопасности* может быть генитивом, дативом или локативом единственного числа лексемы *безопасность*, то есть эти словоформы омонимичны.

Грамматическая неоднозначность или омонимия мешает решению задач компьютерной обработки текста, поэтому необходимым этапом предобработки является снятие омонимии. Снятие (или разрешение) омонимии — этап анализа текста, на котором проводится выбор единственного варианта морфологического анализа для каждого токена [2].

Проблема снятия омонимии актуальна для множества задач: машинный перевод (например, в [3] показано влияние разрешения омонимии на точность перевода), автоматическое извлечение информации (см., например, [4]), автоматический анализ содержания (о влиянии снятия омонимии по лемме в [5]), обработка речи и текста.

Заметим, что в некоторых случаях возможно проведение анализа текста без снятия омонимии (см., например, [6]), однако подобный подход либо предполагает, что достаточно большой объем текстов выведет нужные леммы в наиболее часто встречающиеся за счет высокой встречаемости разных форм одной леммы, либо используемый метод нечувствителен к составу слов или добавлению новых.

Большинство исследований омонимии можно разделить на два больших направления: описание омонимии как языкового явления и снятие омонимии при автоматической обработке языка. Меньшая часть исследований посвящена квантитативному описанию омонимии.

Омонимия как языковое явление

Явление омонимии тесно связано с явлением полисемии, но, как отмечено в [7], в отличие от последнего “не приносит пользы никому кроме остряков и рифмоплетов”. Сложно приписать начало исследования омонимии конкретному человеку или времени: сам термин «омонимия» был введен еще Аристотелем. Тем не менее, начало разработки теории омонимии относится только к XX веку.

Начальные положения классификации омонимов были заложены в [8] и продолжены В.В. Виноградовым в серии публикаций (например, [9]). В этих статьях были сформулированы типы грамматической омонимии и критерии их определения. Целый ряд статей был направлен на совершенствование этой и других классификаций ([10, 11] и др.).

В связи с тем, что английские тексты имеют большое распространение, большое количество публикаций в области теоретической лингвистики было посвящено его исследованию. Ещё в [12] впервые было подсчитано количество английских омонимов; в [13] показано, что омонимия не должна считаться маргинальным языковым процессом; в [14] рассматриваются способы разрешения омонимии в языке; в [15] классификация типов омонимии также применяется к английскому языку.

В связи с автоматизацией процессов обработки текстов в настоящее время исследования омонимии сдвинулись из области теоретических описаний в сферу автоматической обработки языка и компьютерной лингвистики.

Снятие омонимии

Снятие омонимии подразумевает два процесса со своими особенностями и проблемами реализации: определение лексического значения слова определение части речи слова и его параметров.

Автоматическое снятие частеречной омонимии стало возможным после создания корпуса Брауна [16], первого большого корпуса английского языка, который можно было использовать для компьютерного анализа. В 1970-х была сделана автоматическая частеречная разметка [17], основанная на правилах следования частей речи в английском языке. Точность такой разметки составляла лишь около 70%, поэтому она проверялась вручную.

В материковой Европе был создан свой корпус английского языка — Lancaster-Oslo-Bergen Corpus of British English. В 1980-е он был размечен с использованием скрытых марковских моделей [18]. Этот метод основывался на анализе вероятности цепочек частей речи. Проблему составлял тот факт, что для определения вероятностей редких частей речи объема корпуса не хватало.

Проблема недостаточного объема корпусов была решена в 1988 с использованием методов динамического программирования, позволявших предсказать недостающие вероятности [19]. Помимо этого, такой метод был быстрее и эффективнее.

Все описанные выше методы требуют заранее размеченный корпус для обучения. Алгоритм, предложенный в [20], использует только словарь, а не

аннотированный корпус, что позволяет использовать его для языков без хорошо размеченных корпусов большого размера.

Снятие омонимии по параметрам может входить в процесс снятия омонимии по частям речи и использует те же методы. Однако если для морфологически бедных языков в приписываемый тег входит только часть речи, то для морфологически богатых в него входят еще и значения параметров. Например, в [21] для чешского языка используется 3127 тегов.

Из недавних исследований в области русского языка стоит отметить цикл статей, посвященных проблемам лексико-семантической разметки в НКРЯ в книге «Национальный корпус русского языка: 2003-2005. Результаты и перспективы» [22], а также серию публикаций на конференции «Диалог» на ту же тему [23-25], рассматривающих, среди прочего, использование семантических фильтров.

Квантитативные описания языков (то есть описания не с качественной точки зрения, а связанные с расчетом числовых показателей) начинались с создания частотных словарей (для русского языка см. [27-30]), на основе которых уже проводились исследования. Частотный словарь, основанный на НКРЯ, был представлен в [31]. В [32] соединены данные частотных словарей, основанных на четырех корпусах русского языка, с целью создать наиболее полный частотный словарь для оценивания сложности текстов.

Новое направление исследований касалось подсчета распределения падежей и частей речи. В [33] сравниваются результаты Йоссельсона и Штейнфельдта, сделанные без больших корпусов, с данными, которые дают большие корпуса, и предлагается иерархия русских падежей по частотности. В [34] подсчитано распределение частей речи в зависимости от стиля текста. В классической книге [35] рассматриваются частоты слов и грамматических категорий для английского языка.

Квантитативный анализ грамматической неоднозначности слов распространен значительно меньше. В [21] рассматривается омонимия по частям речи, на основе которой предлагается классификация типов омонимии. В работе [36] приводятся такие данные, как количество неоднозначных токенов, среднее число словоформ на словоупотребление, посчитанные на материале венгерского и английского языков. В [37] приведены аналогичные данные для румынского языка. Работа [38] также использует классы неоднозначности, применяемые для исследования размеров и свойств этих классов в эстонском, литовском и латышском языках. В отличие от большинства предыдущих работ, в наших статьях [39, 40] было рассмотрено распределение типов омонимии для нескольких языков. Эти статьи показывают, что разным пластам лексики в зависимости от частотности свойственна разная омонимия, хотя снятие омонимии осуществляется одинаково для всех слов.

Текущая работа продолжает наши исследования в этой области. Её задача — во-первых, более подробно показать разницу между словоформами разной частотности с точки зрения типов омонимии, во-вторых, объяснить

эту разницу и, в-третьих, исправить некоторые недостатки метода и проверить полученные результаты на другом словаре.

Метод исследования и используемые данные

С целью повторить результаты, полученные в [40], в качестве исходных данных использовался тот же самый корпус — корпус новостей сайта Лента.ру 2005-2015 гг., размер корпуса — 32.4 миллиона словоупотреблений.

В отличие от [40], в данной работе мы рассчитывали частоты не лемм, а токенов. Подобный подход позволяет избежать неоднозначности интерпретации полученных результатов, связанной с тем, что для одного токена проводится увеличение частоты для всех лексем, входящих в его разбор. Таким образом, в предыдущем методе учитывалась частота не столько лексемы, сколько её склонность образовывать омонимы с другими лексемами. В текущей работе было выделено 20 000 самых частотных словоформ и 43 988 вариантов их грамматического разбора.

Для грамматического анализа использовался словарь OpenCorpora [41], реализованный в библиотеке pymorphy2. Тип омонимии определялся в соответствии с классификацией, представленной в [39, 40]. Кратко напомним суть данной классификации.

1. Однозначные: существует только один вариант анализа словоформы. (Пример: *Наташа*$_{\text{NOUN}}$.)

2. Неоднозначные по параметрам: часть речи и начальная форма во всех разборах совпадают, грамматические признаки — нет. (Пример: *Наташи*$_{\text{NOUN}}$ — генитив единственного или номинатив множественного.)

3. Неоднозначные по части речи: начальная форма во всех разборах совпадает, часть речи не совпадает. (Пример: *ученый*$_{\text{ADJF}}$ *кот* vs *ученый*$_{\text{NOUN}}$ *пришел*.)

4. Неоднозначные по лемме: часть речи совпадает, начальная форма - нет. (Пример: *Александра* — генитив от *Александр*$_{\text{NOUN}}$ или номинатив от *Александра*$_{\text{NOUN}}$.)

5. Неоднозначные по части речи и лемме: ни часть речи, ни начальная форма не совпадают. (Пример: *ясно*$_{\text{ADVB}}$ vs *ясно*$_{\text{ADJF}}$.)

6. Несловарные: словоформа отсутствует в словаре. (Пример: *natasha, 2020, путуньк*.)

Помимо типа, для каждой словоформы определяется частотность. Для 20 000 самых частотных словоформ также определяются все возможные частеречные теги и леммы. Pymorphy2 не использует контекст при морфологическом разборе, то есть выбор возможных тегов осуществляется только на основании формы токена.

Количественная оценка омонимии в русском языке в зависимости от частотности токена

Как это отмечалось выше, метод подсчета частотности слов с различными типами омонимии в статье [40] обладает одним серьезным недостатком. Для каждого токена берется множество всех возможных лемм, после чего частотность увеличивается для каждой из этих лемм. В итоге учитывается не только частотность лемм, но и склонность входящих в них токенов образовывать омонимичные группы. На практике при снятии омонимии более важным оказывается анализ самих токенов и их частоты, чем лемм. В связи с этим в данной работе мы проводим все расчеты не для лемм, а для отдельных токенов. Это несколько изменяет вид графиков, однако общие тенденции на них оказываются примерно теми же. Форма распределения отличается не только из-за разных методов подсчета, но и из-за использования разных морфологических анализаторов: pymorphy2 в текущей работе и Расширенная AOT.ru в предыдущих.

Итак, рассмотрим распределение типов омонимии по токенам в зависимости от их частотности. На рис. 1 показано распределение типов омонимии в первых 10 000 наиболее частотных токенов, разделенных на десять групп по тысяче токенов. Сравнение показывает, что разница распределений между разными группами не является такой резкой, как это было получено для распределения лемм. Однако и в текущей версии видно, что первая тысяча токенов в большей степени обладает омонимией по части речи, чем остальные частотные группы. Особенно это относится к словам, омонимичным только по части речи.

Далее мы углубили наш анализ, разделив первую тысячу наиболее частотных токенов на группы по 100 токенов. Полученный график распределения токенов по частотным группам приведен на рис. 2. Здесь видно, что различия принимают очень резкий характер. В первой сотне частеречная омонимия занимает почти половину распределения, тогда как доля неоднозначности по грамматическим параметрам снижается в несколько раз. Можно предположить, что форма распределения зависит от а) распределения частей речи в языке и б) особенностей этих частей речи. Исходя из этого, чтобы объяснить именно такую форму распределения для русского языка, необходимо, во-первых, понять, как части речи распределены по частоте, и, во-вторых, какие типы омонимии свойственны каким частям речи.

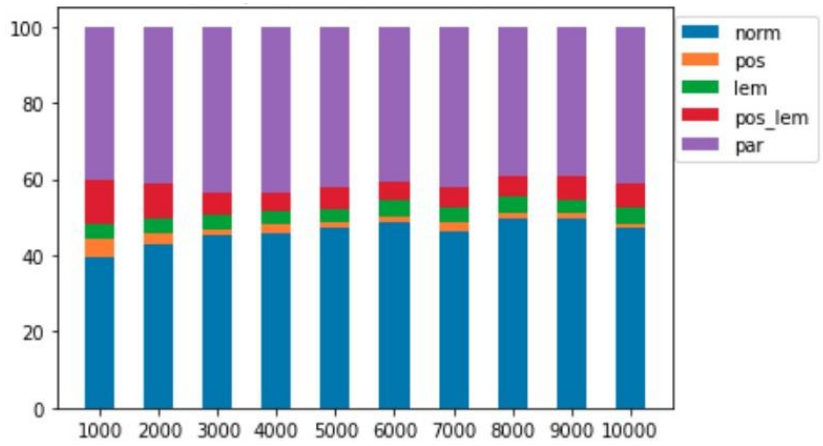

Рис. 1. Распределение типов неоднозначности в зависимости от частотности словоформ (для 10000 наиболее частотных токенов)

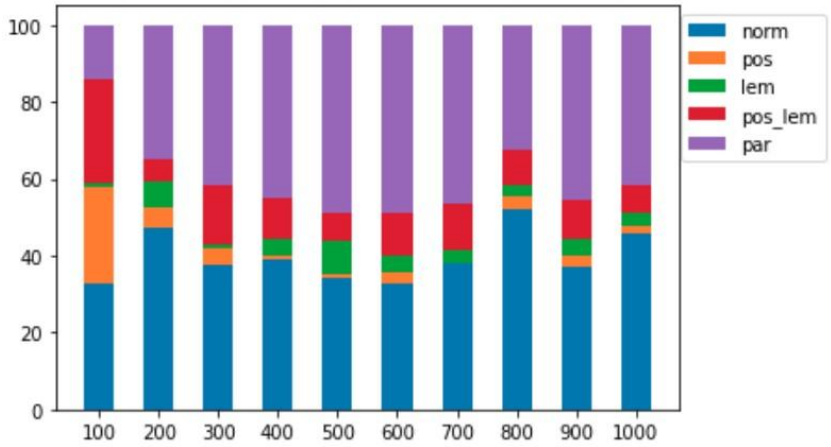

Рис. 2. Распределение типов неоднозначности в зависимости от частотности словоформ (для 1000 наиболее частотных токенов)

Для проведения такого анализа мы рассчитали распределение типов омонимии по частям речи для 10 000 самых частотных токенов. Список частей речи был взят из pymorphy2. При анализе брался наиболее вероятный разбор, возвращаемый первым в списке результатов. Например, словоформа *ясно*, неоднозначная по лемме и части речи, будет засчитана как наречие, т.к. это наиболее вероятный разбор. Результаты расчетов приведены в Табл. 1 в виде относительных значений. Сумма по строке здесь равна 100%. Таблица показывает, что наибольшее влияние на распределение типов омонимии оказывают существительные, полные прилагательные, глаголы и наречия, то есть части речи, самые частотные по корпусу в целом. Аналогичная информация в графическом виде приведена на рис. 3. На нем также дается разделение данных по частотным группам внутри первых 10 000 наиболее частотных токенов.

Таблица 1

Распределение частей речи по типам омонимии для 10 000 наиболее частотных токенов (относительные значения в процентах)

	NOUN	ADJF	ADJS	COMP	VERB	INFN	PRTF	PRTS	GRND	NUMR	ADVB	NPRO	PRED	PREP	CONJ	PRCL	INTJ
неоднозначно по лемме	89.89	6.65	0.00	0.14	2.77	0.00	0.00	0.00	0.00	0.00	0.00	0.55	0.00	0.00	0.00	0.00	0.00
неоднозначно по параметрам	59.97	31.78	0.00	0.00	2.17	0.74	4.21	0.27	0.04	0.60	0.00	0.22	0.00	0.00	0.00	0.00	0.00
неоднозначно по части речи	24.20	41.99	0.00	0.00	1.07	1.42	0.00	0.71	0.00	0.00	8.90	0.36	0.36	7.12	6.05	6.76	1.07
неоднозначно по части речи и лемме	25.09	32.74	3.64	0.37	3.64	0.37	4.85	2.99	0.28	0.65	19.40	1.49	0.47	1.49	1.96	0.47	0.09
однозначно	55.06	3.95	0.59	0.20	24.65	7.28	0.31	4.89	0.76	0.10	1.32	0.30	0.05	0.25	0.14	0.15	0.01

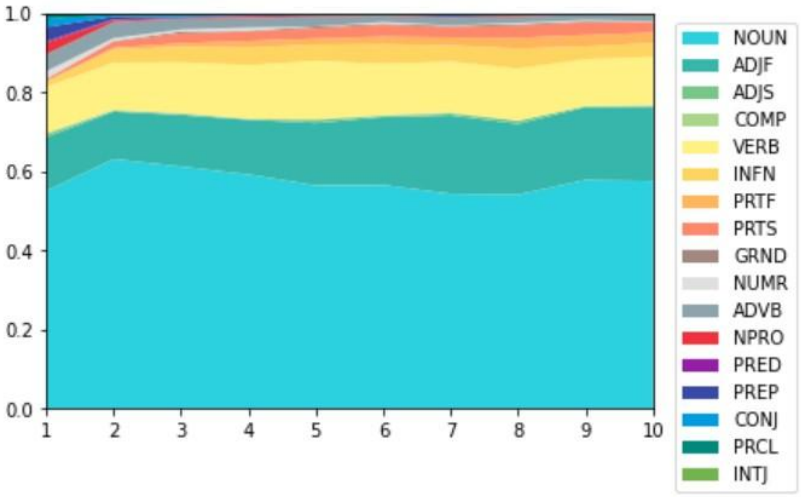

Рис. 3. График распределения частей речи по частоте (полный список частей речи)

Нетрудно заметить, что на рис. 3, где список частей речи является полным, отчетливо выделяются лишь существительные, полные прилагательные и глаголы. В связи с этим мы сгруппировали части речи, которые были объединены по наличию грамматических категорий: полные, краткие прилагательные, местоимения, полные и краткие деепричастия — представлены как прилагательные (изменяются по числу, роду, падежам (кроме кратких форм)); наречия, компаративы, деепричастия, инфинитивы и предикативы — сгруппированы в наречия (неизменяемые, самостоятельные); предлоги, частицы, союзы и междометия — в функциональные части речи (неизменяемые, несамостоятельные); личные глаголы, существительные, числительные — отдельно. Результаты показаны на рис. 4. На графике видно, что функциональные части речи участвуют в основном в первой тысяче наиболее частотных токенов, тогда как в остальном распределении их доля резко падает. Для более глубокого анализа был построен аналогичный график для первой тысячи наиболее частотных токенов, разделенных на сотни (см. рис. 5). Как видно из графика, распределение токенов по частям речи в зависимости от частоты встречаемости отличается и здесь.

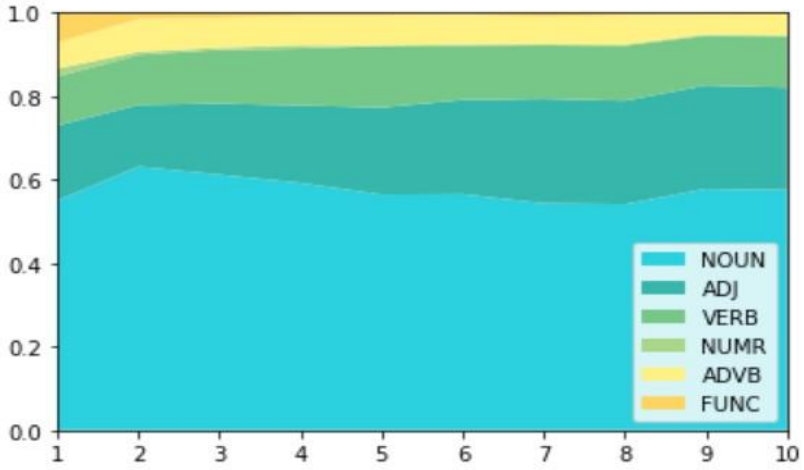

Рис. 4. График распределения частей речи по частоте (части речи с группировкой, первые 10 000 токенов)

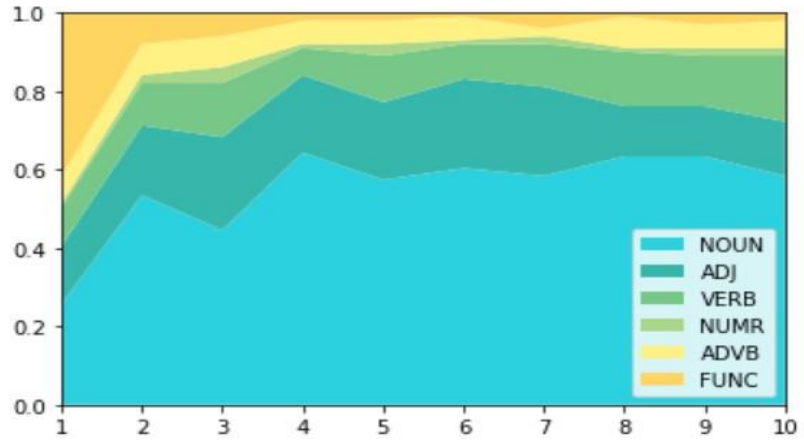

Рис. 5. График распределения частей речи по частоте (части речи с группировкой, первая 1 000 токенов)

Наконец, мы проанализировали совместную встречаемость частей речи в омонимичных словах. Заметим, что, если слово омонимично только по грамматическим параметрам, части речи будут совпадать. В Табл. 2 показана матрица таких совпадений частей речи. Все значения даны в процентах, сумма значений по ряду равна 100%. Расчет проводился по первым 10 000 наиболее частотных токенов.

Аналогичная информация с распределением по тысячам представлена на Рис. 6. На нем изображена доля выбранных пар частей речи, оказавшихся омонимичными. Так как общее число пар слишком велико, мы отобрали все, вошедшие в семь самых частотных пар для каждой тысячи, и отобразили только их. Заметим, что здесь первая тысяча вновь отличается от остальных, что объясняется другим набором частей речи в ней. На всём отрезке наиболее частая омонимия — прилагательных с существительными, вторая по частоте — наречий с краткими прилагательными, с третьей тысячи также значительно совпадение полных прилагательных и причастий.

Таблица 2

Матрица совпадений частей речи в омонимичных токенах

	ADJF	ADJS	ADVB	COMP	CONJ	GRND	INFN	INTJ	NOUN	NPRO	NUMR	PRCL	PRED	PREP	PRTF	PRTS	VERB
ADJF	80.39	0.14	0.14	0.09	0.19	0.14	0.00	0.00	11.65	1.12	0.00	0.19	0.00	0.00	5.71	0.00	0.23
ADJS	0.56	46.72	26.83	0.00	3.75	0.00	0.00	0.00	8.63	0.19	0.75	2.44	1.50	0.56	0.00	6.38	1.69
ADVB	0.49	23.44	49.51	0.66	6.23	0.00	0.00	0.16	6.89	0.66	1.48	4.10	0.49	3.93	0.00	0.33	1.64
COMP	5.26	0.00	10.53	57.89	5.26	0.00	0.00	0.00	10.53	0.00	0.00	5.26	0.00	5.26	0.00	0.00	0.00
CONJ	1.78	8.89	16.89	0.89	39.56	0.44	0.44	3.11	7.11	0.89	0.89	11.56	0.89	2.67	0.00	0.44	3.56
GRND	4.76	0.00	0.00	0.00	1.59	71.43	0.00	0.00	14.29	0.00	0.00	1.59	0.00	6.35	0.00	0.00	0.00
INFN	0.00	0.00	0.00	0.00	0.26	0.00	97.18	0.26	1.54	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.77
INTJ	0.00	0.00	2.63	0.00	18.42	0.00	2.63	36.84	7.89	0.00	0.00	18.42	2.63	7.89	0.00	0.00	2.63
NOUN	3.82	0.71	0.64	0.06	0.25	0.14	0.09	0.05	92.19	0.06	0.03	0.26	0.02	0.29	0.28	0.00	1.12
NPRO	22.22	0.93	3.70	0.00	1.85	0.00	0.00	0.00	3.70	63.89	0.00	3.70	0.00	0.00	0.00	0.00	0.00
NUMR	0.00	5.06	11.39	0.00	2.53	0.00	0.00	0.00	2.53	0.00	75.95	0.00	0.00	0.00	0.00	0.00	2.53
PRCL	2.22	7.22	13.89	1.11	14.44	0.56	0.00	3.89	9.44	2.22	0.00	40.56	1.11	2.22	0.00	0.00	1.11
PRED	0.00	25.81	9.68	0.00	6.45	0.00	0.00	3.23	3.23	0.00	0.00	6.45	41.94	3.23	0.00	0.00	0.00
PREP	0.00	2.22	17.78	1.48	4.44	2.96	0.00	2.22	14.07	0.00	0.00	2.96	0.74	50.37	0.00	0.00	0.74
PRTF	30.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.44	0.00	0.00	0.00	0.00	0.00	65.43	0.00	0.00
PRTS	0.00	11.60	0.68	0.00	0.34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	86.69	0.68
VERB	0.34	0.61	0.68	0.00	0.54	0.00	0.20	0.07	4.97	0.00	0.14	0.14	0.00	0.07	0.00	0.14	92.10

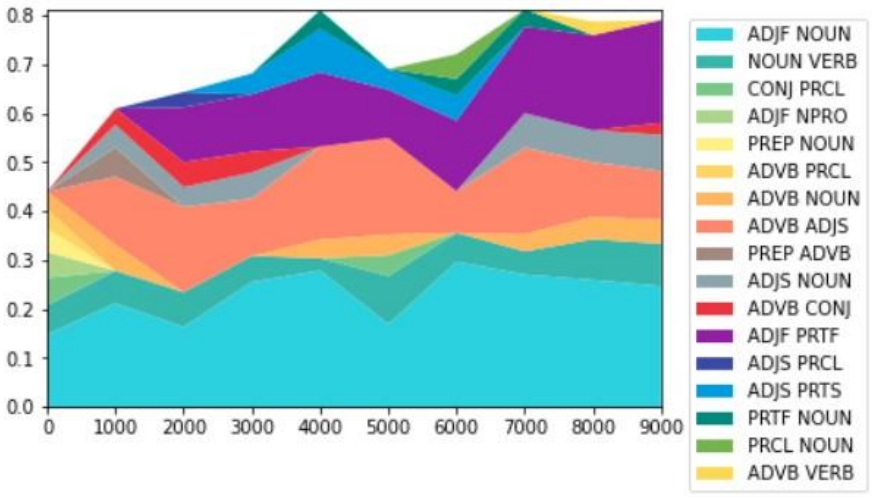

Рис. 6. Самые частотные омонимичные пары частей речи (внутри частотных групп токенов)

Для анализа неоднозначности по параметрам существительных была сделана аналогичная матрица. Распределение падежей по частотности было подсчитано, например, в [33], но распределение совпадений падежей в омонимах ранее не рассматривалось. Из полученных результатов следует, что омонимия по падежам лишь слегка зависит от частоты. На всех графиках явно выделяется треугольник номинатив-аккузатив-генитив, но в первой тысяче, в отличие от других, другие связи также значимы (например, датив-локатив). Получается, что состав токенов существительных в первой тысяче более разнообразен.

Таблица 3

Матрица совпадений падежей в омонимичных токенах

	loct	nomn	gen2	acc2	datv	accs	ablt	gent	voct	loc2
loct	2372	1022	0	0	1419	1057	619	1186	1	9
nomn	1022	5493	0	0	999	4181	600	2515	8	9
gen2	0	0	16	0	16	5	0	0	0	12
acc2	0	0	0	0	0	0	0	0	0	0
datv	1419	999	16	0	2049	1044	651	1111	1	38
accs	1057	4181	5	0	1044	5416	599	3171	12	16
ablt	619	600	0	0	651	599	1556	622	1	2
gent	1186	2515	0	0	1111	3171	622	4984	12	9
voct	1	8	0	0	1	12	1	12	12	0
loc2	9	9	12	0	38	16	2	9	0	38

Обсуждение результатов

Как это следует из полученной статистики, в первой сотне токенов, в отличие от всех остальных частотных групп, большую часть занимают неоднозначности по части речи и по части речи и лемме. Согласно Табл. 2, основной вклад вносят имена нарицательные и собственные, наречия и служебные части речи. Распределение имен и наречий почти неизменно (см. Рис. 5), тогда как служебные части речи встречаются почти исключительно в первой сотне. Получается, за уникальность распределения первой сотни токенов отвечают именно служебные части речи.

В служебные части речи входят предлоги, союзы, частицы и междометия. Согласно Табл. 2, предлоги омонимичны наречиям и существительным; союзы — существительным, кратким прилагательным, наречиям и частицам; частицы — существительным, союзам, наречиям; междометия — существительным.

Особенности этого типа неоднозначности можно объяснить несколькими факторами. Во-первых, pymorphy2 часто разбирает короткие слова, которыми и является большинство лексем служебных частей речи, как аббревиатуры, то есть существительные: например, разбор для союза *и* включает в себя не только непосредственно союз (а также частицу и междометие), но и аббревиатуру *и* (например, *И. А. Тургенев*) во всех возможных формах. Необходимо отметить, что это происходит не со всеми однобуквенными словами. Так, разбор для *а* включает только союз, частицу и междометие. Непонятно, чем обусловлена эта непоследовательность. Во-вторых, многие служебные части речи являются производными от имен и наречий. Так, омонимичны существительным: *путём*$_{\text{PREP}}$, *правда*$_{\text{CONJ}}$, *марш*$_{\text{INTJ}}$; прилагательным: *очевидно*$_{\text{CONJ}}$; наречиям: *прежде*$_{\text{PREP}}$, *безусловно*$_{\text{CONJ}}$, *там*$_{\text{PRCL}}$. В-третьих, омонимия частиц и союзов (*Наташа ушла в отпуск, я же*$_{\text{CONJ}}$ *осталась работать. Вот же*$_{\text{PRCL}}$ *она!*), скорее всего,

объясняется нетривиальными историческими процессами, обсуждение которых лежит вне границ этой работы. Наконец, из-за того, что служебные слова обычно короткие, в них больше чем в среднем вероятность случайного совпадения словоформ, чем объясняется совпадение частиц и существительных.

Как и в предыдущем пункте, неоднозначность по части речи и лемме объясняется случайными совпадениями из-за малой длины слова. По графику зависимости доли однозначных и неоднозначных словоформ от их длины (Рис. 7) видно, что короткие словоформы (<=6 знаков) более неоднозначны, чем словоформы средней длины (7-10 знаков). Для более длинных словоформ неоднозначность вновь возрастает.

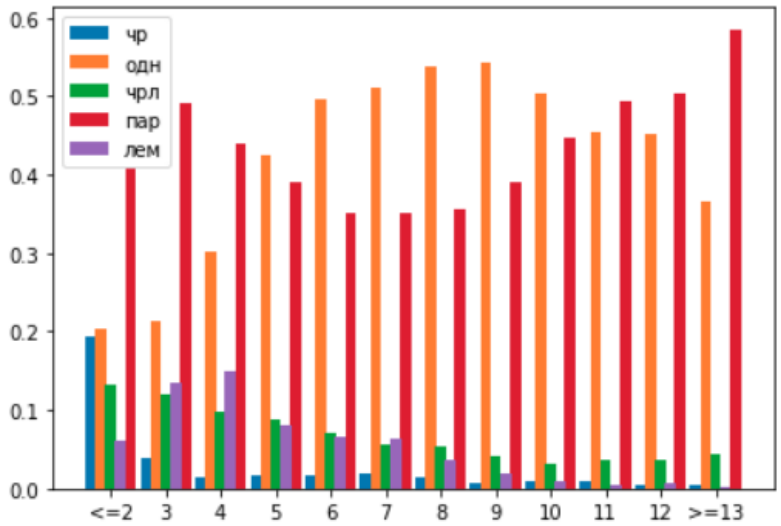

Рис. 7. Зависимость доли словоформ определенного типа в зависимости от их длины

Для иллюстрации приведем наиболее частотные слова, однозначные по лемме и/или части речи и неоднозначные по лемме и/или части речи. Даже по ним видно, что неоднозначные словоформы короче, чем однозначные. Помимо этого, данный список явно показывает, что лексика новостная.

Однозначные: *не, года, за, от, он, также, сообщает, до, году, россии, был, будет, время, сша, долларов, словам, во, заявил, ранее, который, которые, человек, более, они, тысяч, была, компании, процентов, рублей, кроме*.

Неоднозначные: *в, и, на, по, что, с, из, о, как, к, а, для, его, этом, об, после, при, у, со, это, были, того, однако, было, уже, том, их, но, еще, все.*

Вне первой сотни самых частотных токенов неоднозначность по параметрам возникает, когда в парадигме встречаются омонимичные показатели. Этот тип неоднозначности вызван именами: 64% — существительные, 28% — полные прилагательные. В этом нет ничего неожиданного: в русском именном склонении нередко совпадают формы одного слова за счет совпадения грамматических показателей.

У существительных в зависимости от одушевленности аккузатив в обоих числах совпадает либо с номинативом, либо с генитивом. Помимо этого, существуют и другие совпадения. В качестве примера для совпадающих форм в Табл. 4 показано склонение слова *игра* (тип 1d по классификации Зализняка). Совпадающие формы отмечены отдельными цветами, здесь их 5 из 12. Нераспространенные падежи, форм для которых нет у большинства лексем, (второй родительный, второй винительный, второй предложный и вокатив) в этой и следующей таблице не рассматриваются. Количество словоформ, в которых совпадают конкретные падежи (без разделения по числам), показано в Табл. 3.

Таблица 4

Парадигма слова *игра*

	единственное число	*множественное число*
nom	игр-а	игр-ы
gen	игр-ы	игр-0
dat	игр-е	игр-ам
acc	игр-у	игр-ы
abl	игр-ой	игр-ами
loc	игр-е	игр-ах

В парадигме прилагательных еще больше совпадающих форм. В Табл. 5 форм слова свободный каждое совпадение отмечено одним цветом, и незакрашенных ячеек, то есть однозначных форм, всего 3 из 24.

Несмотря на то что у прилагательных совпадающих форм намного больше, чем у существительных, существительные в этом типе встречаются в два раза чаще. Это объясняется преобладанием существительных над прилагательными во всём корпусе (см. Рис. 4). Неоднозначность по лемме возникает, когда у двух слов одной части речи совпадают формы (необязательно в одних и тех же ячейках парадигмы), то есть неоднозначность по лемме в большинстве случаев включает в себя неоднозначность по параметрам. Из-за омонимичности показателей подавляющую часть словоформ этого типа занимают существительные — 88%.

Лексемы, к которым относятся словоформы этой группы, отличаются попарно только родом или числом и, соответственно, типом склонения. Словоформы этого типа – совпадающие формы в двух разных парадигмах. В Табл. 6 рассмотрены совпадения на примере слов *франций* и *Франция*.

Таблица 5

Парадигма слова *свободный*

	единственное число			*множественное число*
	м.р.	*ж.р.*	*ср.р.*	
nom	свободн-ый	свободн-ая	свободн-ое	свободн-ые
gen	свободн-ого	свободн-ой	свободн-ого	свободн-ых
dat	свободн-ому	свободн-ой	свободн-ому	свободн-ым
acc	свободн-ый	свободн-ую	свободн-ое	свободн-ые
abl	свободн-ым	свободн-ой	свободн-ым	свободн-ыми
loc	свободн-ом	свободн-ой	свободн-ом	свободн-ых

Таблица 6

Парадигма слов *франций* и *Франция*

	sg	*sg*
nom	франций	Франция
gen	франция	Франции
dat	францию	Франции
acc	франций	Францию
abl	францием	Францией
loc	франции	Франции

Другие примеры: совпадение с именем собственным (*марта* - *март*$_{\text{NOUN}}$ и *Марта*$_{\text{NOUN}}$; *газа* - *газ*$_{\text{NOUN}}$ и *Газа*$_{\text{NOUN}}$); одна из лексем имеет оба числа, другая — только одно (*выборы* - *выбор* и *выборы, час* и *часы*); лексемы различаются только родом (*банк* и *банка*, *политик* и *политика*).

Несмотря на то что в полных прилагательных совпадений в парадигме еще больше, чем в существительных, этот тип неоднозначности среди них не распространен из-за того, что у прилагательного нет лексикализованного рода или числа, то есть причина, по которой возникает неоднозначность в существительных, для них просто невозможна.

Однозначные слова представлены в основном существительными (55%) и глаголами (24%). В именном склонении много показателей, уникальных в парадигме. Так, в Табл. 4, из 12 клеток парадигмы 7 определяются единственным образом.

В отличие от именной, в глагольной парадигме неоднозначностей почти нет: 85,9% глагольных словоформ однозначны. Еще 7,5% неоднозначны по части речи, лемме или части речи и лемме, что объясняется случайными совпадениями (например, *времени* — это либо *время*$_{\text{NOUN}}$, либо *временить*$_{\text{VERB}}$). Оставшиеся 6,6% словоформ неоднозначны по параметрам. В них входят двувидовые глаголы (*констатировали*), глаголы, у которых не различаются переходные и непереходные формы (*выстрелить*), глаголы в первом лице множественного числа, совпадающие с гортативом (*пойдем*), а также пары глаголов, у которых совпадают леммы и формы настоящего времени несовершенного вида и будущего времени совершенного вида (например, *находиться* ‘располагаться где-то’ и *находиться* ‘провести много времени, ходя’).

Описание неоднозначности по части речи осложняется тем, что, собственно, нельзя просто сказать, какие части речи ее образуют. По Табл. 2, в которой учитывается самый вероятный разбор, большую часть составляют существительные, полные прилагательные, наречия и служебные части речи, которые уже были рассмотрены выше.

Существительные оказываются омонимичны полным прилагательным. Существительные, омонимичные полным прилагательным, занимают чуть меньше четырех процентов среди всех существительных, но из-за того, что существительные занимают такой большой пласт омонимии по частям речи, омонимия их с прилагательными значительна. Эта омонимия объясняется наличием субстантивов (*ванная*$_{\text{NOUN}}$ vs *ванная*$_{\text{ADJF}}$) и фамилий, произошедших от притяжательных прилагательных (*Антонов*$_{\text{NOUN}}$ vs *антонов*$_{\text{ADJF}}$).

Полные прилагательные омонимичны существительным и полным причастиям. Омонимия с существительными рассмотрена выше, а омонимия с причастиями объясняется тем, что причастия и отглагольные прилагательные образуются от глаголов одними и теми же способами, изменяются по одной и той же схеме, а отличаются друг от друга только сочетаемостью (*бывший*$_{\text{ADJF}}$ *сотрудник* vs *бывший*$_{\text{PRTF}}$ *в гостях*).

Наречия омонимичны кратким прилагательным в форме среднего рода (*поле широко*$_{\text{ADJS}}$ vs *широко*$_{\text{ADVB}}$ *в плечах*) и служебным частям речи, производным от них (*точно*$_{\text{ADVB}}$ vs *точно*$_{\text{PRCL}}$).

Слова, неоднозначные по части речи и лемме, сравнительно мало распространены вне самых частотных словоформ. Этот тип неоднозначности свойственен существительным (21%), полным прилагательным (30%) и наречиям (21%) (а также служебными частями речи, рассмотренными выше). Такое распределение объясняется рассмотренными выше факторами, а именно: наличием субстантивов (например, *ведущего* это *ведущий*$_{\text{NOUN}}$ *передачи* или *ведущий*$_{\text{ADJF}}$ *производитель*), обеспечивающих частую омонимию между существительными и полными прилагательными; совпадением наречий и кратких прилагательных (*выглядеть официально*$_{\text{ADVB}}$ (лемма *официально*) vs *это платье слишком официально*$_{\text{ADJS}}$ (лемма *официальный*)).

Оценка точности снятия омонимии в современных системах анализа текстов

Для того чтобы оценить влияние подобного разброса на результаты работы, мы проанализировали ошибки, которые выдают две библиотеки синтаксического анализа: UDPipe и spaCy. Библиотеки синтаксического анализа были выбраны в связи с тем, что в их состав входят модули снятия омонимии. Первая из библиотек работает в версии 2019 г. и с тех пор не обновлялась, очередная версия второй библиотеки была выпущена в 2021 г.

Для проверки мы использовали материалы, использовавшиеся в соревновании GramEval'2020, проводившемся в рамках конференции «Диалог» [42]. Данные материалы основаны на стандарте разметки Universal Dependencies [43], который также использовался для обучения языковых моделей выбранных библиотек. В связи с этим мы были гарантированы от ошибок, связанных с конвертацией частеречных систем разного формата.

Для получения оценки мы искали несовпадения в разметке библиотек с использованным «золотым стандартом». Ошибки были сгруппированы по частотным интервалам в соответствии с частотами токенов, рассчитанными на первом этапе исследования. Мы сформировали 20 интервалов: сотни из первой тысячи самых частотных слов, тысячи самых частотных слов с частотой от 1 000 до 10 000, слова с рангом больше 10 000. Мы проанализировали три вида ошибок: ошибка в части речи, в лемме, одновременно в лемме и части речи. Помимо абсолютного числа ошибок мы рассчитали долю ошибочных словоупотреблений во всех словоупотреблениях данного частотного интервала, а также долю ошибок данного частотного интервала во всех ошибках, совершенных библиотекой.

Итоговая статистика для UDPipe и spaCy показана на Рис. 8 и 9 соответственно. Как видно из рисунков, библиотека spaCy совершает больше ошибок в грамматической разметке слов. Однако как в одной, так и в другой библиотеке на первую сотню самых частотных токенов приходится самое большое количество ошибок, если исключить из анализа все токены с рангом больше 10 000. Заметим, что в проанализированных текстах встретилось более миллиона различных токенов, то есть на токены с рангом больше 10 000 приходится 99% интервала, но не больше 40% словоупотреблений. Доля ошибок, приходящаяся на некоторые интервалы, приведена в Табл. 7.

Исходя из полученных данных, мы можем утверждать, что синтаксические анализаторы UDPipe и spaCy совершают больше ошибок на первой сотне самых частотных токенов. Для этого есть две причины. Во-первых, это большая доля слов, которая по закону Ципфа [44] приходится на самый частотный интервал. То есть при сохранении доли ошибок само число ошибок в более частотном интервале будет больше. Вторая причина для увеличения числа ошибок — это «видовое разнообразие» характеристик наиболее частотных слов. Как мы видели это в предыдущих разделах, первая сотня слов обладает большей долей функциональных частей речи, а сам частеречный состав слов значительно отличается от остальных интервалов.

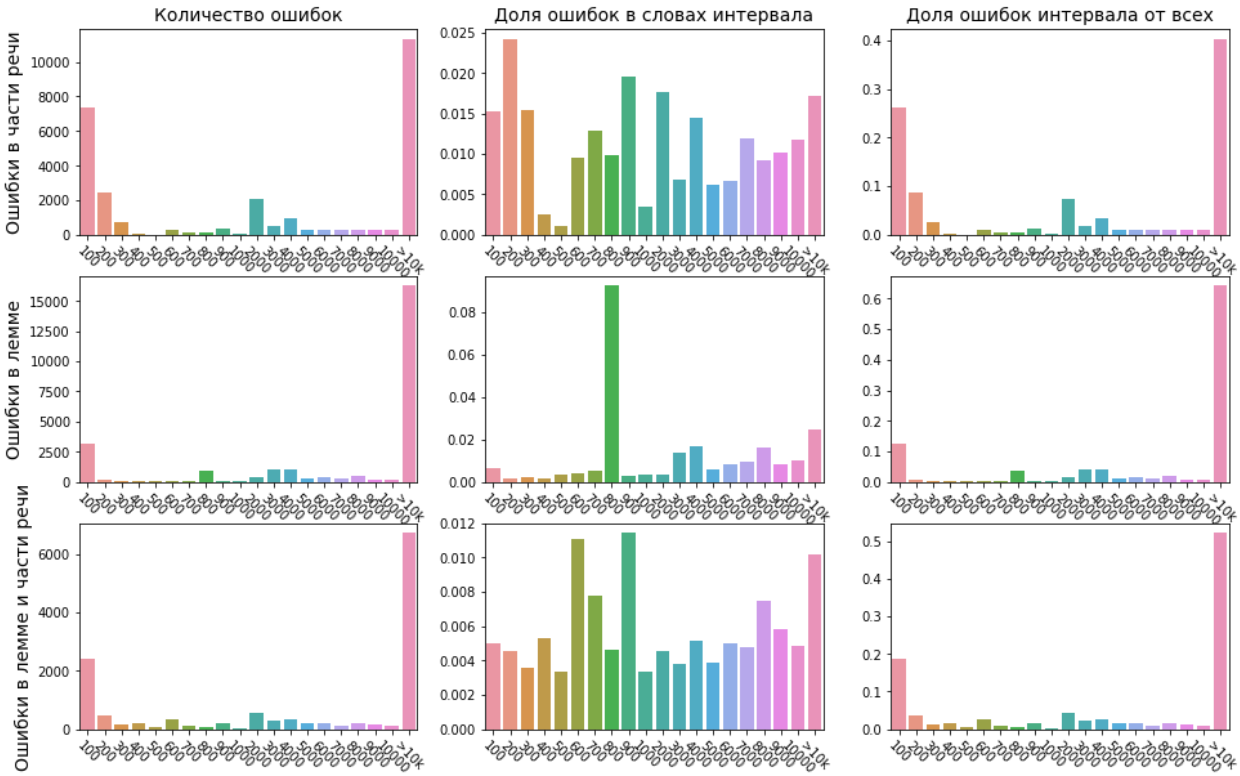

Рис. 8. Распределение ошибок по частотным интервалам (UDPipe)

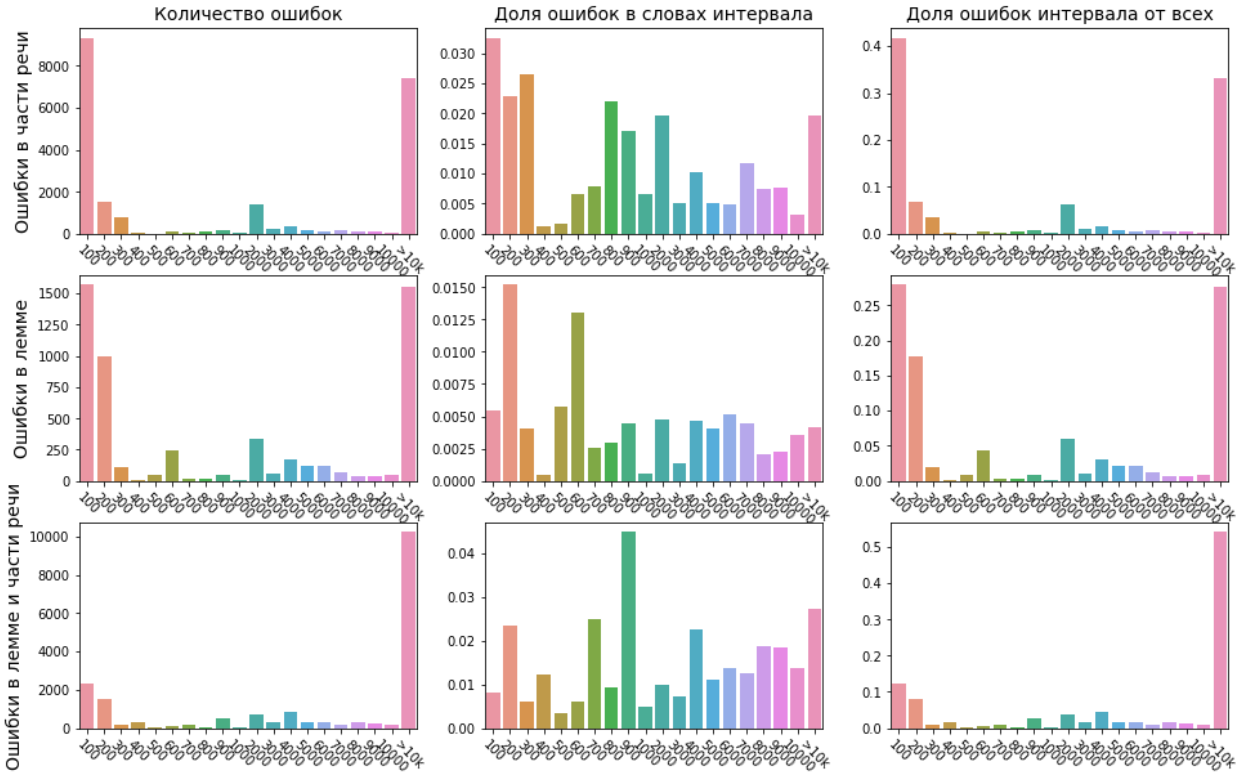

Рис. 9. Распределение ошибок по частотным интервалам (spaCy)

Таблица 7

Доля ошибок в зависимости от частоты

Библиотека	Тип омонимии	[0; 100]	[0; 1000]	[1000; 10000]	>10000
UDPipe	часть речи	0.261	0.485	0.111	0.403
	лемма	0.123	0.204	0.154	0.642
	часть речи и лемма	0.186	0.353	0.125	0.523
Spacy	часть речи	0.424	0.617	0.055	0.328
	лемма	0.217	0.516	0.116	0.369
	часть речи и лемма	0.233	0.415	0.173	0.412

Заключение

Как это видно из проведенного исследования, на распределение типов омонимии влияют свойства самых частотных частей речи: в русском языке (в отличие от, например, английского: см. [Клышинский, 2020]) однозначные слова и слова, неоднозначные по параметрам, сильно перевешивают слова всех остальных типов. Первое объясняется омонимичными формами в парадигмах существительных и прилагательных, а второе — неомонимичными формами в парадигмах существительных и глаголов.

Во-первых, из этого следует, что методы снятия частеречной омонимии не могут быть полностью языконезависимыми: в разных языках разные части речи будут вносить разные вклады в неоднозначность (см., например, [26]). Во-вторых, несмотря на то, что язык, как мы показали, разнообразен, омонимия, как правило, разрешается одними и теми же методами. Это приводит к тому, что большое количество ошибок сосредотачивается в более частотной области. Наша гипотеза состоит в том, что обучение отдельных классификаторов для разных частотных диапазонов может привести к увеличению точности разбора. При этом в случае русского языка, следует рассчитывать именно точность работы системы снятия омонимии, не беря в расчет однозначные слова, доля которых велика.

Библиографический список

1. Löbner S. Understanding Semantics (2nd ed.) // Routledge. 2013. 392 с.
2. Большакова Е.И., Пескова О.В., Клышинский Э.С., Носков А.А., Ландэ Д.В., Ягунова Е.В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика. М.: МИЭМ. 2011. 272 с.
3. Fabricz K. Particle homonymy and machine translation // IIn Proc. of International Conference on Computational Linguistics. 1986. С. 59-61.
4. Krovetz R. Homonymy and polysemy in information retrieval. In 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics. 1997. С. 72-79.
5. Litowski Kenneth C. Desiderata for tagging with WordNet synsets or MCAA categories. ACL-SIGLEX Workshop “Tagging Text with Lexical Semantics: Why, What, and How?” April 4-5, 1997, Washington, D.C., USA. 1997. С. 12-17.
6. Баданина Н.Д., Судаков В.А. Модели машинного обучения для классификации отзывов о банках // Препринты ИПМ им. М.В.Келдыша. 2021. No 50. 14 с. https://doi.org/10.20948/prepr-2021-50
7. Ullmann S. Semantics. An Introduction to the Science of Meaning. Oxford University Press. 1964. 278 с.
8. Булаховский Л.А. Из жизни омонимов // Русская речь. 1928. 3. С. 47-60.
9. Виноградов В.В. (1960). Об омонимии и смежных явлениях // Вопросы языкознания, 1960. 1. С. 295-312.
10. Колесников Н.П. Словарь омонимов русского языка. Изд-во Тбил. Ун-та. 1978. 631 с.
11. Щепин А.Г. Характеристика омонимов в плане фонетического и лексико-грамматического членения речи // Учен. зап. Читинск. гос. пед. ин-та: Общественные и гуманитарные науки. 1963. Вып. IX. С. 154–170.
12. Bridges A.S. On English homophones. Clarendon Press. 1919. 70 с.
13. Menner R.J. The conflict of homonyms in English // Language. 1936. 12(4). С. 229-244.
14. Jespersen O. Essentials of English grammar. London: Allen & Unwin. 1933. 333 с.
15. Тышлер И.С. О классификации омонимов в современном английском языке // Вопросы романо-германского языкознания, 1975, (4), С. 3-24.
16. Kucera H., Francis W.N. Computational analysis of present-day American English. University Press of New England. 1967. 424 с.
17. Greene B.B., Rubin G.M. Automatic grammatical tagging of English // Department of Linguistics, Brown University. 1971.
18. Garside R. The CLAWS word-tagging system. The Computational analysis of English: A corpus-based approach. London: Longman. 1987. С. 30-41.
19. DeRose S.J. Grammatical category disambiguation by statistical optimization // Computational linguistics, 1988, 14(1). С. 31-39.

20. Brill E., Pop M. Unsupervised learning of disambiguation rules for part-of-speech tagging // In Natural language processing using very large corpora. Springer, Dordrecht. 1999. С. 27-42.
21. Hajič J., Vidová-Hladká B. Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. // In Proc. of the COLING-ACL Conference. 1998. С. 483 – 490.
22. Плунгян В.А. Национальный корпус русского языка: 2003-2005. Результаты и перспективы. М.: Индрик, 2005, 344 с.
23. Толдова С.Ю., Кустова Г.И., Ляшевская О.Н. Семантические фильтры для разрешения многозначности в национальном корпусе русского языка: глаголы // Труды конференции «Диалог». 2008. С. 522-529.
24. Рахилина Е.В., Кобрицов Б.П., Кустова Г.И., Ляшевская О.Н., Шеманаева О.Ю. Многозначность как прикладная проблема: лексико-семантическая разметка в национальном корпусе русского языка // Компьютерная лингвистика и интеллектуальные технологии. 2006. С. 445-451.
25. Шеманаева О.Ю., Кустова Г.И., Ляшевская О.Н., Рахилина Е.В. (2007). Семантические фильтры для разрешения многозначности в Национальном корпусе русского языка: прилагательные // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог2007». 2007. С. 582–587.
26. Protopopova E.V., Bocharov V.V. Unsupervised learning of part-of-speech disambiguation rules // In Proc. of Computational Linguistics and Intellectual Technologies (Dialog-2013). 2013. С. 655-675.
27. Josselson H.H. Подсчет ходовых слов русского литературного языка. Detroit (MI). 1953. 274 с.
28. Šteinfeldt E. Russian Word Count. M.: Прогресс. 1963. 228 с.
29. Засорина Л.Н. (ред.) Частотный словарь русского языка. М.: Русский язык. 1977. 935 с.
30. Лённгрен Л. Частотный словарь современного русского языка. Uppsala. 1993. 188 с.
31. Ляшевская О.Н., Шаров С.А. Частотный словарь национального корпуса русского языка: концепция и технология создания. Режим доступа: http://www.dialog-21.ru/digests/dialog2008/materials/html/53.htm.
32. Blinova O.V., Tarasov N.A., Modina V.V., Blekanov I.S. Modeling Lemma Frequency Bands for Lexical Complexity Assessment of Russian Texts // In Proc. Of Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2020”. 2020. С. 76-92.
33. Копотев М.В. К построению частотной грамматики русского языка: падежная система по корпусным данным // Slavica Helsingiensia, 34. С. 136-151.
34. Браславский П.И. Морфологический строй функциональных стилей // Известия Уральского государственного университета. 2001. № 21. С. 9-17.
35. Francis W. Nelson, Kucera H., Mackie A. W. Frequency analysis of English usage: lexicon and grammar // Boston (Mass.): Houghton Mifflin. 1982. 561 с.

36. Oravecz C., Dienes P. Efficient Stochastic Part-of-Speech Tagging for Hungarian // In Proc. of Third Int. Conf. on Language Resources and Evaluation (LREC'02). 2002. С. 710-717.
37. Tufiş D. Using a large set of EAGLES-compliant morpho-syntactic descriptors as a tagset for probabilistic tagging // In Proceedings of Second International Conference on Language Resources and Evaluation. Athens. 2000.
38. Pinnis M., Goba K. Maximum Entropy Model for Disambiguation of Rich Morphological Tags // In Proc. of Systems and Frameworks for Computational Morphology - Second International Workshop (SFCM 2011). 2011. С. 14-22.
39. Клышинский Э.С., Логачева В.К., Мансурова О.Ю., Максимов В.Ю., Карпик О.В., Зиязтдинов И.Б., Макеенко П.А. Исследование неоднозначности употребления слов в европейских языках // Препринты ИПМ им. М.В.Келдыша. 2015. No 4. 31 с. URL: http://library.keldysh.ru/preprint.asp?id=2015-4
40. Клышинский Э.С., Логачева В.К., Карпик О.В., Бондаренко А.В. Количественная оценка грамматической неоднозначности некоторых европейских языков // Вестник Новосибирского государственного университета. Серия: Лингвистика и межкультурная коммуникация. 2020. 18(1). С. 5-21.
41. Bocharov V.V., Alexeeva S.V., Granovsky D.V., Protopopova E.V., Stepanova M.E., Surikov A.V. Crowdsourcing morphological annotation // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 12 (19). М.: РГГУ, 2013. С. 109-114.
42. Lyashevskaya O., Shavrina T., Trofimov I., Vlasova N. GRAMEVAL 2020 Shared Task: Russian Full Morphology and Universal Dependencies Parsing // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2020”. 2020. 19 (26). С. 553–569.
43. Nivre J., de Marneffe M.-C., Ginter F. et al. Universal Dependencies v1: A Multilingual Treebank Collection // In Proc. of LREC-2016. 2016. С. 1659-1666.
44. Кочеткова Н.А., Клышинский Э.С., Ермаков П.Д. Подчиняются ли составные конструкции закону Ципфа? // Системный администратор. № 11. 2016. С. 89-95