УДК 81'322.2:81'373.423+519.257
DOI 10.25205/1818-7935-2020-18-1-5-21

Количественная оценка грамматической неоднозначности некоторых европейских языков

Э. С. Клышинский [1], **В. К. Логачева** [2], **О. В. Карпик** [3], **А. В. Бондаренко** [4]

[1] *Национальный исследовательский университет «Высшая школа экономики»*
Москва, Россия

[2] *Сколковский институт науки и технологий*
Москва, Россия

[3] *Институт прикладной математики им. М. В. Келдыша РАН*
Москва, Россия

[4] *Государственный научно-исследовательский институт авиационных систем*
Москва, Россия

Аннотация

Неоднозначность слов по их грамматическим категориям является хорошо исследованной областью, однако существующие методы ее оценки в текстах на различных естественных языках являются скорее количественными, чем качественными. В данной статье предлагается разделение всех слов на несколько классов неоднозначности. Подобное разделение позволяет ввести количественный метод оценки, основанный на расчете статистики употребления слов. В статье проводится исследование неоднозначности для таких языков, как английский, немецкий, шведский, испанский, каталанский, французский, итальянский, португальский, русский, польский, словенский, турецкий. Нами было численно показано, что распределение слов по классам неоднозначности зависит от выбранного корпуса или системы морфологического анализа, однако остается уникальным для заданного языка. Так, славянские языки, а также французский и итальянский, обладают самой низкой частотой слов, неоднозначных по части речи. Наибольшей неоднозначностью по собственно грамматическим параметрам обладают славянские языки, немецкий и шведский. Кроме того, была обнаружена зависимость неоднозначности от частотности слова. В статье показывается, что наибольшей степенью неоднозначности обладают слова из первой тысячи самых частотных слов. Для большинства исследованных языков при снижении частоты слов также падает и процент слов, неоднозначных по части речи.

Учет разных классов неоднозначности позволяет более корректно проводить оценку систем снятия грамматической неоднозначности, применяемых для разных языков. Обычно сравнение проводится на всем тексте, тогда как мы предлагаем сравнивать результаты только на неоднозначных словах, поскольку их процент существенно отличается от языка к языку. Наши эксперименты, не вошедшие в данную статью, показали, что учет класса неоднозначности позволяет несколько улучшить работу системы автоматического снятия неоднозначности.

Ключевые слова

автоматическая обработка текстов, грамматическая неоднозначность, статистика употребления

Для цитирования

Клышинский Э. С., *Логачева В. К.*, *Карпик О. В.*, *Бондаренко А. В.* Количественная оценка грамматической неоднозначности некоторых европейских языков // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2020. Т. 18, № 1. С. 5–21. DOI 10.25205/1818-7935-2020-18-1-5-21

Quantitative Estimation of Grammatical Ambiguity: Case of European Languages

Eduard S. Klyshinsky [1], **Varvara K. Logacheva** [2], **Olesya V. Karpik** [3]
Alexander V. Bondarenko [4]

[1] *National Research University Higher School of Economics*
Moscow, Russian Federation

[2] *Skolkovo Institute of Science and Technology*
Moscow, Russian Federation

[3] *Keldysh Institute of Applied Mathematics RAS*
Moscow, Russian Federation

[4] *State Research Institute of Aviation Systems*
Moscow, Russian Federation

Abstract

The grammatical ambiguity (multiple sets of grammatical features for one word form or coinciding surface forms of different words) can be of different types. We distinguish six classes of grammatical ambiguity: unambiguous, ambiguous by grammatical features, by part of speech, by lemma, by lemma and part of speech, and out-of-vocabulary words. These classes are found in all languages, but word distribution may vary significantly. We calculated and analysed the statistics of these six ambiguity classes for a number of European languages. We found that the distribution of ambiguous words among these classes depends primarily on basic linguistic features of a language determining its typology class. Although it is influenced by text style and the considered vocabulary, the distinctive shape of the distribution is preserved under different conditions and differs significantly from distributions for other languages. The fact that the shape is primarily defined by linguistic properties is corroborated by the fact that closely related languages demonstrated in our research similar properties as far as their ambiguous words are concerned. We established that Slavic languages feature a low rate of part-of-speech ambiguous words and a high rate of words which are ambiguous by grammatical features. The former is also true for French and Italian, while the latter holds for German and Swedish, whereas the combination of these traits is characteristic of Slavic languages alone.

The experiments showed that reduction of the grammatical feature set does not change the shape of distribution and therefore does not reflect similarity among languages. On the other hand, we found that the top 1000 most frequent words in all the languages considered have different distribution in ambiguity classes unlike in the rest of the words. At the same time, for the majority of considered languages, less frequent words are less unambiguous by part of speech. In Romance and Germanic languages, the ambiguity is reduced for less frequent words. We also investigated the differences in statistics for texts of different genres in the Russian language. We found out that fiction texts are more ambiguous by part of speech than newswire, which are in turn more ambiguous by grammatical features.

Our results suggest that the quality of multilingual morphological taggers should be measured relying only on ambiguous words as opposed to all words of the processed text. Such an approach can help get a more objective linguistic picture and enhance the performance of linguistic tools.

Keywords

natural written language processing, grammatical ambiguity, statistics of occurrence

For citation

Klyshinsky, Eduard S., Logacheva, Varvara K., Karpik, Olesya V., Bondarenko, Alexander V. Quantitative Estimation of Grammatical Ambiguity: Case of European Languages. *Vestnik NSU. Series: Linguistics and Intercultural Communication*, 2020, vol. 18, no. 1, p. 5–21. DOI 10.25205/1818-7935-2020-18-1-5-21

Введение

За последние десятилетия в квантитативной лингвистике произошла смена парадигмы исследований. Ранее для оценки языковых явлений чаще использовались эмпирические исследования, проводящие анализ отдельных примеров одного явления в большом количестве языков. С развитием параллельных корпусов с единой системой разметки всё чаще разрабатываются новые методы, количественно оценивающие подобные явления в разных языках (см., например, [Gibson et al., 2012]). Как следствие, появились методы, позволяющие оценить количественные параметры в морфологии [Hajič, Vidová-Hladká, 1998], синтаксисе [Köhler, 2012], сравнительном языкознании [Hawkins, 1983].

Статистическая информация приобретает большое значение при автоматической обработке текстов (АОТ). Так, этап разрешения грамматической неоднозначности (в области информационных технологий традиционно называющийся снятием омонимии) основывается на использовании статистики встречаемости групп слов в эталонном размеченном вручную корпусе. Результаты работы модуля будут зависеть от собранной статистики, но частота встречаемости комбинаций слов будут отличаться в зависимости от стиля и жанра текстов, привычек и школы авторов и редакторов и других факторов. Помимо этого, выясняется, что методы разрешения грамматической неоднозначности, изначально разрабатывавшиеся как языконезависимые, на практике нуждаются в дополнительной настройке на конкретный язык. Более того, прямое сравнение разных модулей снятия неоднозначности для разных языков становится невозможным, так как не совсем понятно, с чем связана разница в результатах: с языковыми особенностями, особенностями текста, полнотой набора текстов или различиями в программной реализации. Например, в работе [Protopopova, Bocharov, 2013] описывается применение метода снятия омонимии Брилля (см. [Brill, 1995]) к русскому языку. Хотя сам по себе метод заявлен как применимый к любому языку, авторам пришлось изменить набор грамматических параметров, так как правила английского языка давали низкую точность. В [Sharoff, Nivre, 2011] показано, что система TnT [Brants, 2000] также выдает приемлемые результаты лишь после расширения списка параметров.

В нашей работе мы исследуем различия в статистике распределения неоднозначных слов текстов для некоторых (в основном европейских) языков. Нашей задачей было исследовать возможность совпадения грамматических категорий у отдельных словоупотреблений, не вникая в семантику слова. Таким образом, мы говорим скорее об анализе омографов, чем омонимов, не подразумевая при этом анализ полноты совпадения парадигм двух слов. Вообще, вместо полного разнообразия лингвистических теорий в области омонимии, изложенных В. В. Виноградовым, Д. Э. Розенталем, М. И. Фоминой и другими авторами, в данной работе мы рассматриваем единственное явление – совпадение написания разных словоформ, обладающих различными значениями грамматических категорий, вне зависимости от их семантики или фонетики. Так как данная область лингвистики до сих пор является дискуссионной, а терминологический аппарат всё еще может трактоваться разными авторами по-разному, в подразделе 1.2 будет дано формальное (математическое) определение грамматической неоднозначности, являющееся основой для предложенного метода анализа.

Подобная постановка задачи важна для области АОТ, где снятие грамматической неоднозначности является одним из первых этапов автоматического анализа текста, точность работы которого во многом определяет точность результатов. Аналогичные исследования можно провести на материале текстов в устной фиксации, так как в речи неоднозначность проявляет себя иначе. Однако для получения сопоставимых результатов потребуются специальные размеченные корпусы длительностью в несколько тысяч часов звучащей речи. В целом такие эксперименты являются темой для самостоятельного исследования.

Результаты нашей работы позволят лучше предсказывать изменение точности работы различных методов АОТ (таких, как снятие грамматической неоднозначности, выделение терминов с использованием синтаксических шаблонов, синтаксической сегментации и других задач, так или иначе зависящих от точности определения значений грамматических категорий).

Эксперименты по количественной оценке грамматической неоднозначности проводились на материале английского, французского, испанского, каталанского, португальского, итальянского, немецкого, шведского, русского, польского, словенского и турецкого языков. Выбор языков был обусловлен доступностью и полнотой электронных грамматических словарей. Данная работа является продолжением работы [Клышинский и др., 2013]. По сравнению с предыдущим вариантом, мы вдвое увеличили количество рассматриваемых языков и провели несколько новых серий экспериментов, показывающих особенности языков и корректность наших выводов.

1. Метод и материалы для исследования грамматической неоднозначности

1.1. Существующие решения в области анализа грамматической неоднозначности в различных языках

Обычно в лингвистике исследуются вопросы, связанные с особенностями употребления грамматики языка, например, распределение падежей существительных в текстах некоторого жанра (см. [Копотев, 2008; Lyashevskaya, 2013]), разработка частотного словаря для фиксированного языка (см. [Bolshakov et al., 2002]) и др. Статистическая информация включается в качестве обоснования для разработки инструментов, ориентированных на определенные языки. Так, в работе [Hajič, Vidová-Hladká, 1998] приведена информация о частеречной неоднозначности в чешском языке. Авторы вводят понятие «класс неоднозначности», которое используется для представления множества словоформ одного слова с неоднозначной частью речи. Например, слову «process», которое в английском языке может являться как существительным, так и глаголом или прилагательным, присваивается класс POSNVA. Понятие класса неоднозначности использовалось в дальнейшем в некоторых работах при создании систем морфологического анализа с разрешением неоднозначности. В подобных исследованиях статистические данные использовались в качестве отправной точки. Некоторые статистические данные для венгерского и английского языков (количество неоднозначных токенов, среднее число словоформ на словоупотребление) приведены в работе [Oravecz, Dienes, 2002], а работа [Tufiş, 2000] содержит данные для румынского языка.

Также традиционно уделяется большое внимание таксономии явлений, обоснованию причин их появления, их сходству и отличиям. В работах В. В. Виноградова, Д. Э. Розенталя, М. И. Фомина и др. дается полное описание данного феномена. Однако развитие средств АОТ требует введения формального определения как минимум в связи с глобальностью подхода перечисленных авторов (они рассматривают совпадение слов на разных уровнях: фонетики, графики, грамматики, лексики, семантики и т. д.). Ниже мы описываем лишь один из частных случаев – грамматическую неоднозначность слов текста, не вникая в причины ее появления и не отвлекаясь на другие сходные явления. Здесь нас интересует тот факт, что при анализе письменного текста программа не может однозначно проанализировать его, так как в ее грамматическом словаре изначально хранится несколько вариантов анализа слова. Нашей основной целью было определение для разных языков доли текста, неоднозначного с точки зрения программы. Аналогичные исследования могут быть проведены, как сказано выше, и на материале устных текстов, что требует, однако, другой разметки корпусов, перестроения системы классов неоднозначности и изменения методики расчетов.

1.2. Метод анализа неоднозначных словоупотреблений

Представим текст T как последовательность токенов (словоупотреблений), принадлежащих словарю V: $T = <w_1, w_2, \ldots, w_n>$, где $w_i \in V$ – это слово, находящееся в i-й позиции текста. Заметим, что словарь V содержит только слова, т. е. мы не рассматриваем знаки пунктуации, числа и прочие составные части текста.

Определим словоформу v как кортеж $v = <l, \pi, \mu>$, где l – это лемма данной словоформы, π – ее часть речи, μ – ее множество грамматических параметров. Результатами морфологического анализа токена w будет словоупотребление (множество словоформ) $\varphi(w) = \{v_1, v_2, \ldots, v_k\}$, где v_i – это один из вариантов разбора (словоформа). В соответствии с данным определением задачей морфологического анализа токена w является определение возможных вариантов и комбинаций для его леммы (начальной формы), части речи и набора (англ.: tag) грамматических параметров (категорий). Список приписываемых грамматических параметров зависит от языка, которому принадлежит данное слово, и части речи, полученной в результате анализа.

Назовем слово w несловарным, если оно не представлено в словаре, т. е. $\varphi(w) = \varnothing$ (или $k = 0$), в противном случае $k > 0$ и слово будет являться словарным. Если $\varphi(w)$ содержит более одной словоформы ($k > 1$), слово w будет называться неоднозначным. Мы выделили шесть типов грамматической неоднозначности в зависимости от отличий между частями кортежей. Описание указанных типов приведено в табл. 1.

Таблица 1

Классы неоднозначности

Table 1

Classes of Ambiguity

№ п/п	Часть речи	Лемма	Грамматические параметры	Описание
1	0	0	0	Однозначное (единственный вариант анализа токена)
2	0	0	1	**Неоднозначное по параметрам**. Все словоформы результата имеют одну и ту же часть речи и лемму, но их множества грамматических параметров различаются. *Пример*: немецкий глагол ‘wohnen’ (‘проживать’) имеет формы в инфинитиве, 1 л. мн. ч. наст. вр., 3 л. мн. ч. наст. вр. и вежливой формы 2 л. наст. вр.
3	1	0	–	**Неоднозначное по части речи**. Словоформы результата совпадают по лемме, но отличаются по части речи. Грамматические параметры сравниваться не могут. *Пример*: английское ‘close’ может быть существительным, глаголом и прилагательным. Русское ‘больной’ может быть существительным или прилагательным
4	0	1	0/1	**Неоднозначное по лемме**. Словоформы имеют одну часть речи, но различаются леммами. Совпадение множеств параметров неважно. *Пример с совпадающими параметрами*: русское ‘смели’ может быть формой 3 л. мн. ч. прош. вр. глаголов ‘сметь’ и ‘смести’. *Пример с различающимися параметрами*: русское ‘вина’ означает существительное ‘вина’ в им. п. ед. ч., или ‘вино’ в род. п. ед. ч. или им. п. мн. ч.
5	1	1	–	**Неоднозначное по части речи и лемме**. Словоформы отличаются как по части речи, так и по лемме, параметры не сравниваются. *Пример*: французское ‘est’ может быть существительным ‘est’ (‘восток’) или глаголом ‘être’ (‘быть’) 3 л. ед. ч. наст. вр.
6	–	–	–	**Несловарное**. Токен отсутствует в словаре. Данный класс позволяет оценить полноту используемого словаря

При автоматическом анализе текстов слова, принадлежащие разным классам неоднозначности, дают разное количество информации. Так, слово, неоднозначное по параметрам, обладает единственной частью речи и леммой. В этом случае у нас нет всего контекста, определяющего роль слова в тексте, но эта роль ограничена известными частью речи и леммой. Некоторые задачи АОТ, решаемые при помощи модели «мешка слов» или синтаксических шаблонов, не используют грамматические параметры. Слова, неоднозначные по параметрам, в таких задачах могут рассматриваться как однозначные.

Слова, неоднозначные по части речи, хуже определяют синтаксическую структуру предложения. Но если в текстах встречается мало слов с неоднозначной частью речи, некоторые задачи АОТ могут решаться без снятия неоднозначности. Помимо этого, неоднозначные по части речи слова являются полисемичными. Например, «больной», «красный», «зеленый» могут быть прилагательным или существительным и являются полными омонимами. В противоположность им, слова, попадающие к класс неоднозначных по части речи и лемме, являются омографами.

Класс несловарных слов нужен, чтобы, с одной стороны, оценить полноту выбранного морфологического словаря, а с другой стороны, чтобы каждое проанализированное слово могло быть отнесено к тому или иному классу неоднозначности.

Разделив подобным образом все неоднозначные слова на классы, мы можем перейти к расчету частот встречаемости этих классов в текстах на разных языках. Так как результаты будут зависеть от многих параметров (язык, тематика, стиль или жанр текста, его авторы и редакторы, выбранный морфологический словарь и др.), рассмотрим более подробно использованные в нашей работе словари и коллекции.

1.3. Используемые данные и инструменты

Описание использованных системы морфологического анализа и коллекций текстов дано в табл. 2. Для унификации набора грамматических параметров мы нормализовали системы разметки, убрав также синтаксическую информацию. Для испанского, итальянского и каталанского языков набор параметров не менялся.

Для того чтобы устранить влияние зависимости от стиля коллекции, мы проводили основные эксперименты на новостных текстах. Чтобы устранить влияние тематики текстов, проверка также осуществлялась на параллельных новостных корпусах News Commentary для английского, французского, немецкого и испанского языков. Дополнительная серия экспериментов проводилась на текстах разных жанров: новости, научная публицистика и техническая периодика, беллетристика. Незнакомые слова во всех экспериментах не предсказывались, а помещались в класс несловарных.

Таблица 2

Характеристики словарей и корпусов

Table 2

Taggers and Corpora

Язык	Система разметки	Размер словаря (в лексемах)	Коллекция	Размер корпуса (в миллионах словоупотреблений)
Английский	Расширенная АОТ.ru [1]	105 000	Reuters	46,9
Французский	Morphalu [2]	68 000	Le Parisien	43,1
Шведский	Saldo Morphology [3]	118 000	Dagens Nyheter	24,4

Окончание табл. 2

Язык	Система разметки	Размер словаря (в лексемах)	Коллекция	Размер корпуса (в миллионах словоупотреблений)
Испанский	FreeLing [4]	76 000	Abc.es	15,2
Каталанский	FreeLing	76 000	El Periódico de Catalunya	39,3
Португальский	FreeLing	110 000	Expresso	41,6
Итальянский	FreeLing	40 000	Corriere dela Serra	7,9
Немецкий	FreeLing	155 000	Die Zeit	7,1
Русский	Расширенная AOT.ru	167 000	Lenta.ru	32,4
Польский	Morfologik [5]	> 400 000	Different sources	21,2
Словенский	FreeLing	95 000	Dnevnik.si	57,9
Турецкий	Расширенная AOT.ru	64 000	Haberler	28,6
Параллельный корпус			News Commentary Corpus	~ 2

[1] http://aot.ru [Sokirko, Toldova, 2004].
[2] http://www.cnrtl.fr/lexiques/morphalou/.
[3] https://spraakbanken.gu.se/eng/resource/saldom [Pilán, 2015].
[4] http://devel.cpl.upc.edu/freeling/downloads?order=time&desc=1 [Padró, Stanilovsky, 2012].
[5] http://morfologik.blogspot.ru/2013/02/morfologik-20-rc2.html/.

2. Результаты экспериментов по определению грамматической неоднозначности

2.1. Распределение словоупотреблений по типам грамматической неоднозначности

Результаты экспериментов показаны в табл. 3 и на рис. 1. Процент однозначных словоупотреблений колеблется между 30 и 55 % за исключением двух выбросов для польского языка с 19 % и немецкого языка с 14 %. Языки, в которых развита флексия (русский, польский, словенский, немецкий и турецкий), показали высокий уровень неоднозначности по грамматическим параметрам (25–40 % против 0–5 % для других языков). В английском языке ожидаемо половина слов неоднозначна по части речи.

Если рассматривать ряды в табл. 3 как числовые последовательности, то можно обнаружить высокую степень корреляции внутри языковых семей. Немецкий, итальянский и испанский языки показывают корреляция от 0,83 до 0,91. При этом французский сходен с итальянским (корреляция 0,93) и меньше похож на немецкий и испанский (0,79 и 0,7). Романские, английский и русский языки обладают самым большим процентом однозначных слов (35–55 %). Самым отличающимся от всех языков является польский: средняя корреляция 0,21 с выбросом для русского языка 0,5. Следом идут английский (средняя корреляция 0,45) и русский (средняя корреляция 0,5).

Таблица 3

Распределение словоформ по классам неоднозначности, %

Table 3

Distribution of words by classes of ambiguity, %

Язык	Одно-значные	Неоднозначные				Несловар-ные
		по части речи	по лемме и пара-метрам	по лемме и части речи	по пара-метрам	
Русский	49,25	7,41	2,60	8,11	29,98	2,64
Польский	18,76	13,67	4,39	20,98	38,51	3,68
Словенский	23,51	8,44	2,25	24,35	38,62	2,82
Английский	36,29	43,89	0,32	7,42	2,89	9,20
Английский NC	39,82	46,78	0,30	6,88	3,17	3,05
Немецкий	13,39	21,66	1,53	24,89	25,54	12,99
Немецкий NC	14,24	23,55	1,08	29,69	25,71	5,73
Шведский	29,56	12,83	3,33	26,70	17,19	10,40
Испанский	42,01	25,79	0,46	20,29	4,21	7,24
Испанский NC	42,44	26,44	0,33	21,39	4,07	5,34
Каталанский	28,51	25,79	0,62	22,88	3,17	19,04
Португальский	39,14	16,78	0,95	25,16	5,80	12,17
Итальянский	50,11	13,75	1,07	16,85	5,95	12,26
Французский	54,48	7,07	5,70	18,95	3,78	10,01
Французский NC	54,91	7,75	7,50	20,15	4,37	5,31
Турецкий	28,26	7,19	2,03	16,45	27,10	18,98

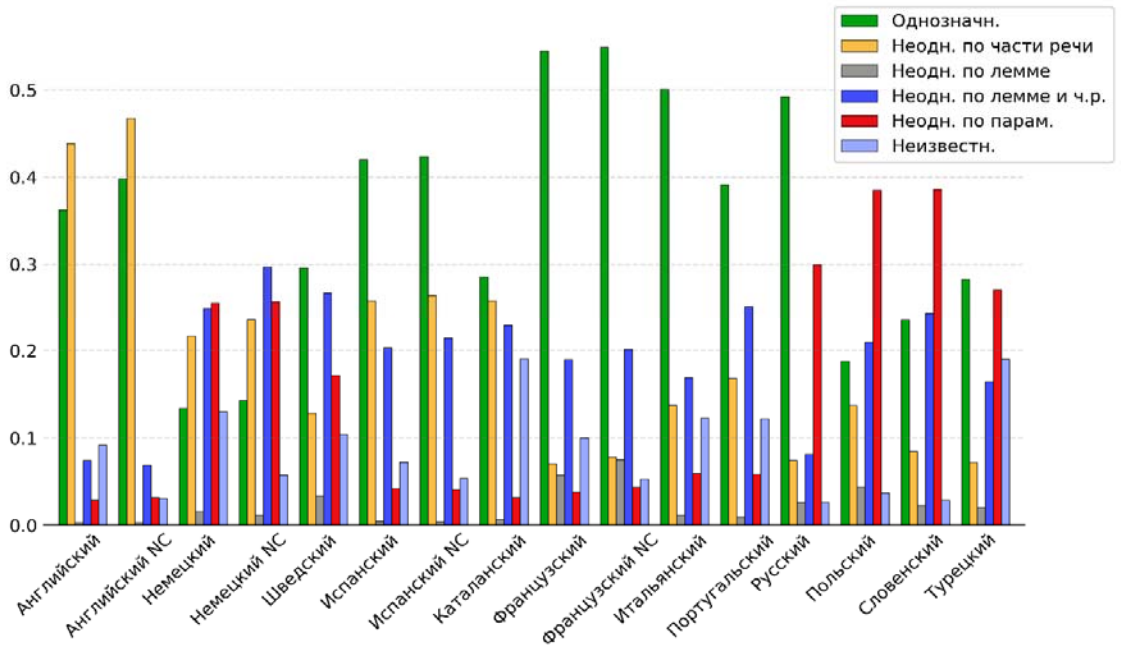

Рис. 1. Распределение словоупотреблений по типам неоднозначности

Fig. 1. Distribution of words by classes of ambiguity

Для параллельных и непараллельных корпусов распределения получились сходными. Но если корреляция для двух корпусов французского языка составила 0,999, а для испанского – 0,998, то для английского и немецкого языков снизилось количество несловарных слов и слов, неоднозначных по лемме и части речи, а число однозначных слов и слов, неоднозначных только по части речи, снизилось. Корреляция при этом остается выше 0,9.

2.2. Разметка текста без разрешения неоднозначности

Обладая статистической информацией о распределении слов по классам неоднозначности, мы можем оценить, насколько применимы к разным языкам методы, основанные на анализе слов, однозначных по части речи. На рис. 2 красной линией показан процент слов, однозначных по части речи (сумма процентов слов, являющихся однозначными, неоднозначными по параметрам и неоднозначными по лемме). Как видно из рисунка, для русского, польского, словенского и французского языков более чем две трети слов могут быть использованы без разрешения неоднозначности по части речи.

Рисунок также демонстрирует известное положение о том, что в разных языках существуют разные проблемы с разрешением неоднозначности: если в романских и германских языках основную проблему составляют слова с неоднозначностью части речи, то в славянских языках основной проблемой является определение корректного набора грамматических параметров (чаще не набора целиком, а отдельных его параметров). Подобная разница иллюстрируется разницей между красной и желтой линией на рис. 2 (последняя показывает долю слов, неоднозначных по параметрам).

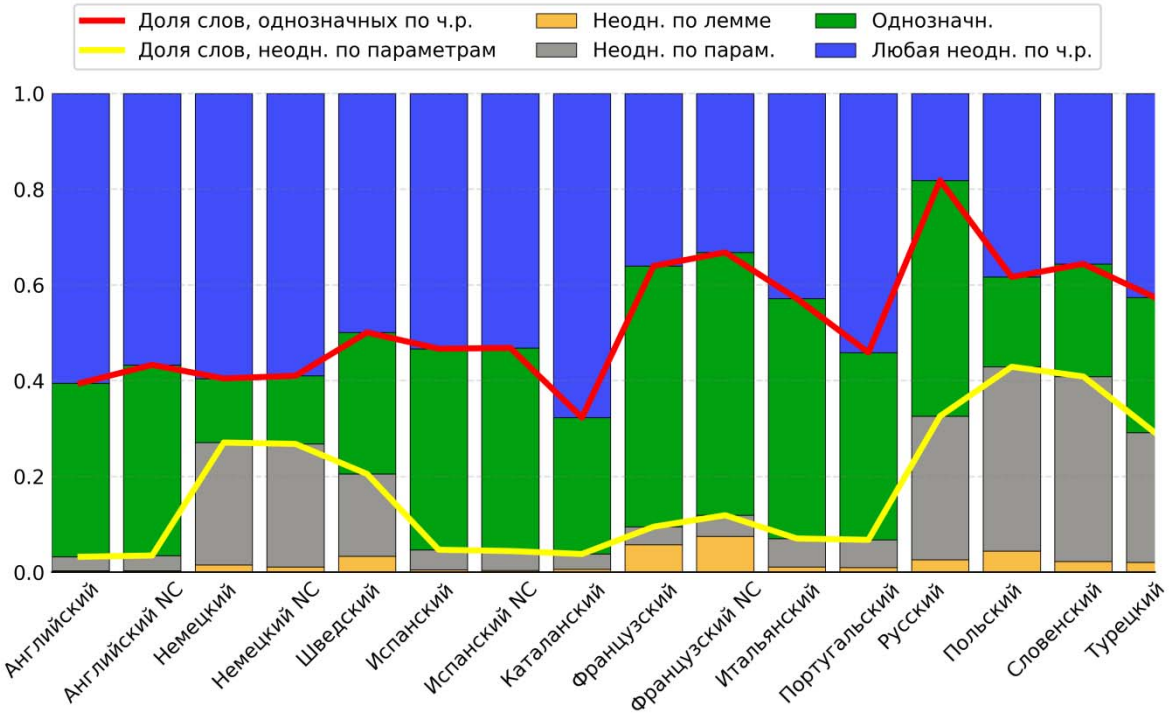

Рис. 2. Процент словоупотреблений, которые могут быть использованы без разрешения неоднозначности по части речи, и однозначных по части речи

Fig. 2. Percentage of part-of-speech unambiguous words and words which are ambiguous by features

2.3. Влияние набора грамматических параметров и словаря на результаты экспериментов

В ходе экспериментов использовались разнородные ресурсы и инструменты: корпуса и словари различаются по размеру, набор грамматических параметров в разных языках не совпадает и т. д. Мы провели серию дополнительных экспериментов, чтобы исследовать, как зависит результат от подобных скрытых параметров модели.

В первом эксперименте мы исследовали зависимость распределения от размеров морфологического словаря выбранного языка. Для этого мы отсортировали словарные леммы по частоте употребления в выбранных корпусах (в случае неоднозначности по лемме увеличивалось значение частоты для всех вариантов леммы). Далее распределение словоупотреблений по классам неоднозначности рассчитывалось для первой 1 000, 3 000 и 5 000 самых частотных лемм русского, словенского, английского, немецкого, французского и испанского языков.

Как видно из рис. 3, изменение размеров словаря не меняет форму распределения кардинальным образом. Например, в русском языке всегда превалируют слова с неоднозначностью по грамматическим параметрам, а в английском языке – неоднозначные по части речи. Корреляция между распределениями для русского языка не опускалась ниже 0,993, для английского – 0,999, тогда как между собой они коррелируют со значением не более чем 0,42.

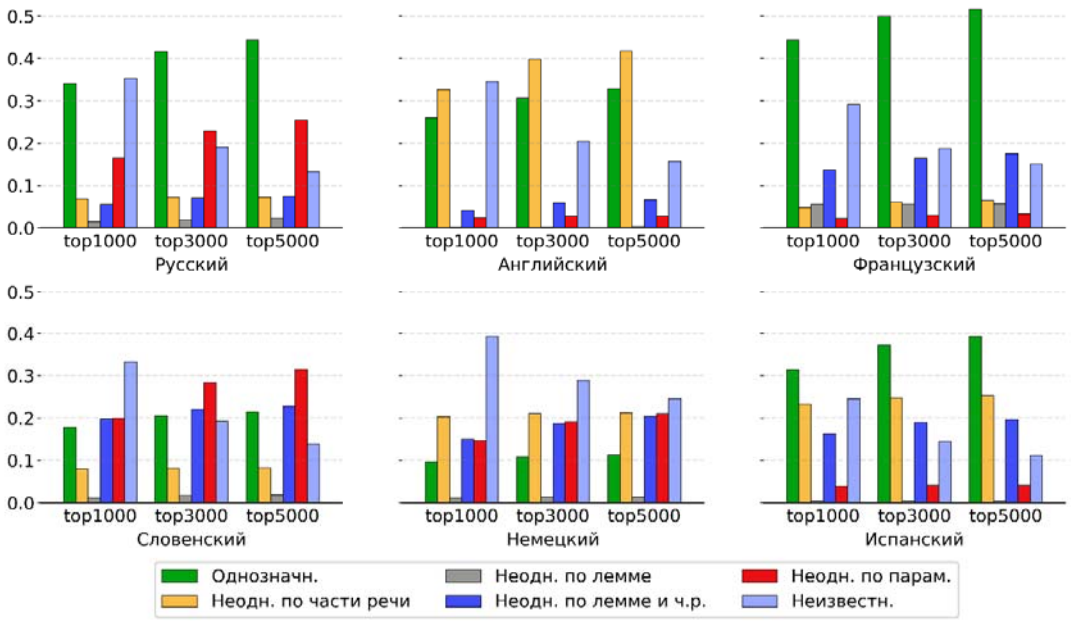

Рис. 3. Влияние размера словаря на форму распределения

Fig. 3. Dependency of a distribution on the number of most frequent words

В связи с этим мы можем утверждать, что вид распределения словоупотреблений по типам неоднозначности существенно не зависит от размеров словаря. Более того, можно сформулировать утверждение, что форма распределения для языка является свойством некоторого базового ядра лексики, содержащего наиболее частотные слова.

С другой стороны, степень корреляции изменяется с ростом словаря. Например, для русского языка изменения между 1 000 и 3 000 наиболее частотных слов больше, чем между 3 000 и 5 000. В связи с этим мы решили исследовать соотношение классов неоднозначности в зависимости от их частоты в словаре. Для этого, как и в предыдущем случае, мы отсорти-

ровали леммы по частоте встречаемости и выделили наиболее частотные 10 000 начальных форм. Они были разделены на десять равных частотных групп: с первой по тысячную позицию, с тысячной по двухтысячную и т. д. После этого распределение словоупотреблений по классам неоднозначности было рассчитано отдельно по частотным группам. Итоговые распределения для всех анализируемых языков показаны на рис. 4 (неизвестные слова отбрасывались, в связи с чем значения частот несколько сдвинуты).

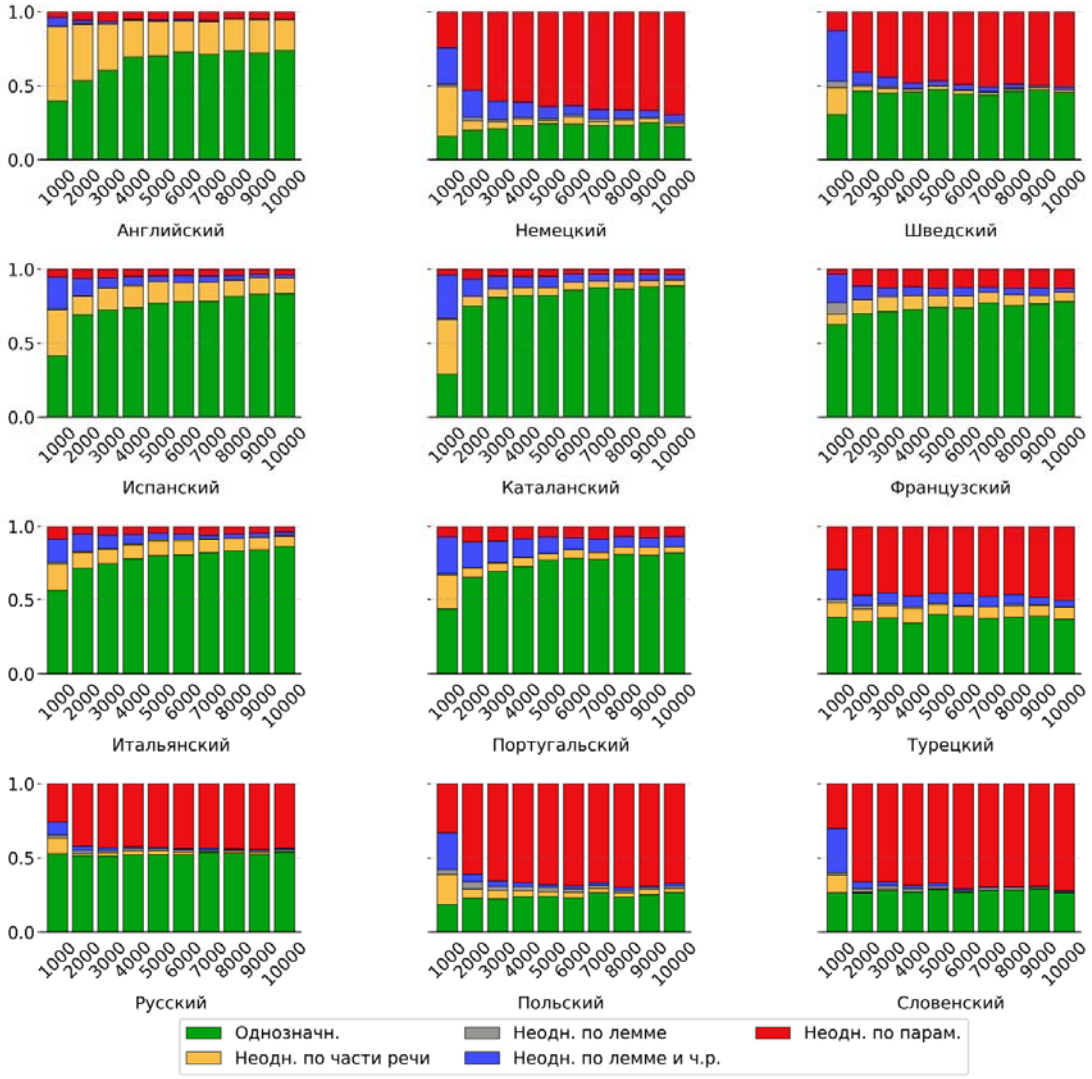

Рис. 4. Влияние частоты токена на форму распределения

Fig. 4. Deprendency of a distribution on the frequency of a word

Как видно из рис. 4, распределение слов по классам неоднозначности для первой тысячи наиболее частотных лемм отличается от прочих слов. Например, в ней есть неоднозначность по части речи или части речи и лемме (тогда как для славянских языков процент таких слов падает, особенно резко – начиная со второй тысячи). Для русского, польского, немецкого и турецкого языков количество однозначных слов примерно сохраняется, тогда как языки романской группы показывают существенное сокращение числа однозначных слов у наибо-

лее частотных слов. Основное количество слов, неоднозначных по лемме и параметрам, также находится в первой тысяче.

Заметим, что объем выборки для расчета разных частотных групп сильно отличается, так как частоты слов распределены по закону Ципфа. Если словоупотребления из первой тысячи самых частотных лемм составляют 60–70 % корпуса, то для пятой тысячи это лишь 2–3 %. Мы не можем здесь отрицать влияние логарифмической шкалы распределения, как следствие, данный вопрос нуждается в дальнейшем исследовании.

Мы проверили зависимость от списка используемых частей речи. Для этого одни части речи заменялись другими. Для русского языка личные местоимения заменялись существительными, указательные местоимения – прилагательными, а деепричастия и причастия – глаголами. Изменения в распределении составило около 0,1 %. Это связано с тем, что местоимения в русском языке чаще всего имеют уникальную лемму, а парадигма для причастий и деепричастий содержит характерные окончания.

В еще одном эксперименте проверялось влияние набора грамматических параметров, фактически зависимость распределения слов от флективности языка. Из набора грамматических категорий русского языка последовательно удалялись одушевленность, род, падеж, лицо и число. Удаление одушевленности изменило распределение на 0,005 %, удаления категории числа – на 0,04 %, удаление лица не изменило результаты. Удаление прочих параметров привело к снижению доли слов, неоднозначных по грамматическим параметрам, и увеличило процент однозначных слов (рис. 5).

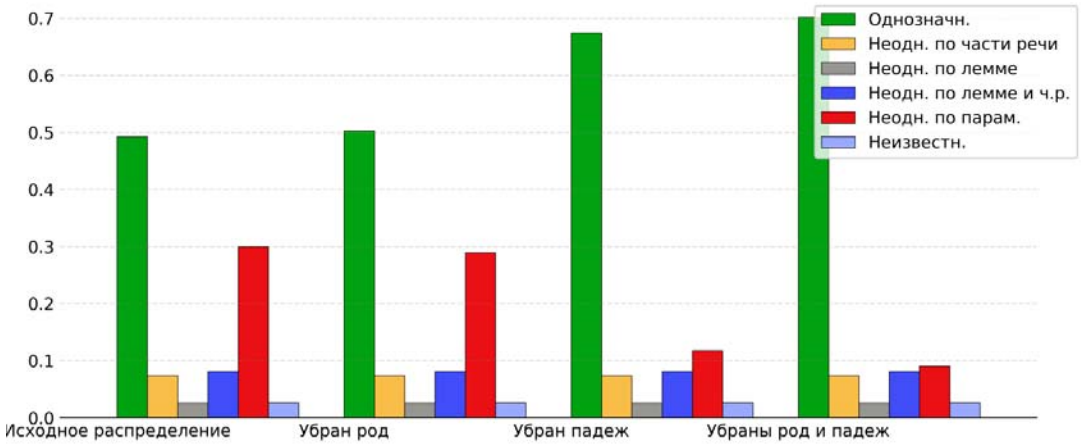

Рис. 5. Влияние набора параметров на форму распределения

Fig. 5. Dependency of a distribution on a reduced set of grammatical features

Для проверки поведения распределения слов по типам неоднозначности в зависимости от всего комплекса параметров мы использовали два независимых морфологических словаря немецкого языка. Один из них, FreeLing, является полноценным словарем, хранящим как основы слов, так и их флексии; второй, TreeTagger, хранит только словарь флексий и предсказывает набор грамматических параметров и начальную форму слова. Результаты показаны в табл. 4. В таблице также указана разница между показателями для словарей TreeTagger и FreeLing.

Как видно из таблицы, словарь TreeTagger обладает значительно меньшим словарем: он показывает примерно на 10 % больше несловарных слов и на 12–25 % меньше слов, неоднозначных по лемме. При этом за счет более бедной системы грамматических параметров ниже процент слов, неоднозначных по параметрам. Поведение слов во TreeTagger сопоставимо с поведением слов в первой тысяче самых частотных слов немецкого языка для новостей.

Таблица 4

Сравнение результатов для словарей FreeLing и TreeTagger, %

Table 4

Results for FreeLing and TreeTagger dictionaries, %

Язык	Одно-значные	Неоднозначные				Несловар-ные
		по пара-метрам	по части речи	по лемме и пара-метрам	по лемме и части речи	
Немецкий FreeLing	13,39	25,54	21,66	1,53	24,89	12,99
Немецкий NC FreeLing	14,24	25,71	23,55	1,08	29,69	5,73
Немецкий TreeTagger	33,27	4,51	22,40	1,36	13,07	25,39
Немецкий NC TreeTagger	44,53	9,35	23,91	0,91	5,26	16,03
Разница новости	19,88	–21,03	0,74	–0,17	–11,82	12,40
Разница NC	30,29	–16,36	0,36	–0,17	–24,43	10,30

В завершение приведем результаты для корпусов различной тематики, написанных на русском языке. Мы провели расчеты для новостей общей направленности, околокомпьютерных новостей, профильного журнала «САПР и графика», сборника фэнтези, сборника любовных историй. Как видно из результатов, показанных на рис. 6, лексика беллетристики отличается большим разнообразием (больше несловарных слов и слов, неоднозначных по лемме), но при этом более четко выражена роль слова (неоднозначность по грамматическим параметрам ниже на 10–15 %). Общая форма распределения сохраняется.

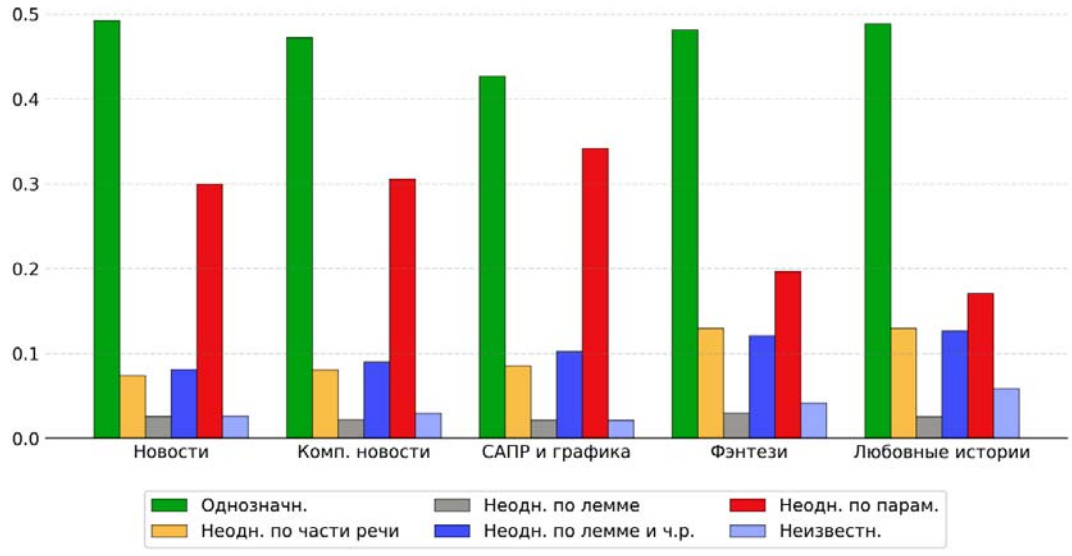

Рис. 6. Влияние тематики корпуса на форму распределения

Fig. 6. Dependency of a distribution on the text's genre

Выводы

В данной статье мы ввели шесть классов грамматической неоднозначности слов. Цель нашей работы состояла в том, чтобы дать количественную оценку степени грамматической неоднозначности текстов на разных языках и продемонстрировать наличие некоторых феноменов в таком распределении. Результаты экспериментов показали значительные отличия в распределении слов по классам неоднозначности в рассмотренных языках. Эксперименты показали, что форма распределения слов по типам неоднозначности определяется наиболее частотными словами языка, хотя само распределение существенно зависит от частотности слов, т. е. более частотные слова несколько иначе образуют классы неоднозначности, чем менее частотные. Данный вопрос должен быть исследован отдельно, однако уже сейчас можно предположить, что более частотные слова относятся к более старым слоям лексики, имеющим меньшую длину слова, и, как следствие, имеющих больше шансов совпасть с другими словами. В этом случае менее частотная лексика является более молодой, образовывавшейся с учетом возможности совпадения с более старыми словами. Однако без детального изучения лексики в зависимости от ее частоты предыдущие утверждения остаются лишь гипотезами.

Наши эксперименты показали, что простое изменение набора грамматических категорий не позволяет привести распределение по классам грамматической неоднозначности одного языка к такому же для другого. Судя по всему, различия между языками лежат глубже, в области статистики использования отдельных слов и синтаксиса. Самым важным параметром является размер словаря: до тех пор, пока процент несловарных слов не достигнет уровня 10–15 %, сложно говорить о свойствах языка. Даже при достижении такого уровня можно говорить скорее о свойствах коллекции, хотя сравнение нескольких коллекций, написанных в различных стилях, позволяет выйти на уровень свойств неоднозначности в языке.

Нами было показано, что стилистические особенности текста отражаются на его распределении неоднозначных слов. По всей видимости, разные стили и жанры накладывают определенные требования на понятность изложения. В научном стиле подобная понятность выражается логикой высказываний, а сама лексика должна быть однозначной. Беллетристика же предъявляет менее строгие требования к определению объекта, но роль, которую он играет, должна быть указана явно.

В данной работе нашел количественное подтверждение общеизвестный эмпирический факт, что разным языкам присуща различная грамматическая неоднозначность: почти любое слово английского языка может быть любой частью речи, а в русском языке основную проблему составляет неоднозначность наборов грамматических параметров. Количественная оценка позволила определить степень различий. Если в английском языке неоднозначно по части речи примерно каждое второе слово, то в русском языке таких слов лишь 15–25 %. Внутри языковых групп наблюдается определенное сходство в распределениях, но и различия также велики. Так, с точки зрения грамматической неоднозначности английский язык больше похож на романские языки, чем на германские.

Полученная информация должна помочь в улучшении точности работы систем морфологического анализа. Наши эксперименты (без применения нейронных сетей), не вошедшие в данную работу, показали, что учет частотности слова позволяет выиграть 0,5–1 % точности. Кроме того, мы показали, что для некоторых приложений в языках с богатой флексией будут работать методы, не предполагающие снятия неоднозначности, но опирающиеся на использование большого корпуса.

В ходе сравнения систем морфологического анализа обычно сравнивается точность их работы на всем тексте, при этом для разных языков получаются результаты, которые сложно сравнивать. Если проводить сравнение на неоднозначной части текстов, оценка работы таких систем будет более корректной, а результаты более сравнимыми. Так как между текстами разных стилей и жанров также наблюдаются определенные различия, сравнение только

по неоднозначной части текстов даже для одного языка сделает результаты также более сравнимыми.

Список литературы

Клышинский Э. С., Кочеткова Н. А., Мансурова О. Ю., Ягунова Е. В., Максимов В. Ю., Карпик О. В. Формирование модели сочетаемости слов русского языка и исследование ее свойств // Препринты ИПМ им. М. В. Келдыша. 2013. № 41. 23 с.

Копотев М. В. К построению частотной грамматики русского языка: падежная система по корпусным данным // Инструментарий русистики: корпусные подходы / Под ред. А. А. Мустайоки, М. В. Копотева, Л. А. Бирюлина, Е. Ю. Протасовой. Хельсинки, 2008. С. 136–150.

Сокирко А. В., Толдова С. Ю. Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка // Интернет-математика 2005. Автоматическая обработка веб-данных. М.: Яндекс, 2005. С. 80–94.

Bolshakov, I. A., Galicia-Haro, S. N., Gelbukh, A. Quantitative Comparison of Homonymy in Spanish EuroWordNet and Traditional Dictionaries. In: Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science, 2002, vol. 2276, p. 280–284.

Brants, T. TnT – a statistical part-of-speech tagger. In: Proc. of 6th Applied Natural Language Processing Conference, 2000, p. 224–231.

Brill, E. Unsupervised Learning of Disambiguation Rules for Part Of Speech Tagging. In: Proc. of the Third Workshop on Very Large Corpora, 1995, p. 1–13.

Gibson, E., Piantadosi, S. T., Fedorenko, E. Quantitative methods in syntax / semantics research: A response to Sprouse and Almeida. *Language and Cognitive Processes*, 2012, p. 229–240.

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, p. 483–490.

Hawkings, J. A. Word Orders Universalis. Academic Press, 1983.

Köhler, R. Quantitative syntax analysis. De Gruyter Mouton, 2012.

Lyashevskaya, O. Frequency Dictionary of Inflectional Paradigms: Core Russian Vocabulary. *Preprints of HSE. Series: Humanity*, WP BRP 35/HUM/2013, 2013.

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, p. 710–717.

Padró, L., Stanilovsky, E. FreeLing 3.0: Towards Wider Multilinguality. In: Proc. of the Language Resources and Evaluation Conference (LREC'12) ELRA, 2012, p. 2473–2479.

Pilán, I. Helping Swedish words come to their senses: word-sense disambiguation based on sense associations from the SALDO lexicon. In: Proc. of NODALIDA, 2015, p. 275–279.

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, p. 655–675.

Sharoff, S., Nivre, J. The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. In: Proc. of Computational Linguistics and Intellectual Technologies (Dialog-2011), 2011, p. 657–670.

Tufiş, D. Using a large set of EAGLES-compliant morpho-syntactic descriptors as a tagset for probabilistic tagging. In: Proc. of Second International Conference on Language Resources and Evaluation, 2000.

References

Bolshakov, I. A., Galicia-Haro, S. N., Gelbukh, A. Quantitative Comparison of Homonymy in Spanish EuroWordNet and Traditional Dictionaries. In: Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science, 2002, vol. 2276, p. 280–284.

Brants, T. TnT – a statistical part-of-speech tagger. In: Proc. of 6th Applied Natural Language Processing Conference, 2000, p. 224–231.

Brill, E. Unsupervised Learning of Disambiguation Rules for Part Of Speech Tagging. In: Proc. of the Third Workshop on Very Large Corpora, 1995, p. 1–13.

Gibson, E., Piantadosi, S. T., Fedorenko, E. Quantitative methods in syntax / semantics research: A response to Sprouse and Almeida. *Language and Cognitive Processes*, 2012, p. 229–240.

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, p. 483–490.

Hawkings, J. A. Word Orders Universalis. Academic Press, 1983.

Klyshinsky, E. S., Kochetkova, N. A., Mansurova, O. Yu., Iagounova, E. V., Maximov, V. Yu., Karpik, O. V. Development of Russian subcategorization frames and its properties investigation. *Keldysh IAM Preprints*, 2013, no. 41, 23 p. (in Russ.)

Köhler, R. Quantitative syntax analysis. De Gruyter Mouton, 2012.

Kopotev, M. Towards the frequency grammar of Russian: corpus evidence on the grammatical case system. In: Mustayoki A., Kopotev M. V., Biryulin L. A., Protasova E. Yu. (eds.). Instruments of Russian linguistics: corpus approaches. Helsinki, 2008, p. 136–150. (in Russ.)

Lyashevskaya, O. Frequency Dictionary of Inflectional Paradigms: Core Russian Vocabulary. *Preprints of HSE. Series: Humanity*, WP BRP 35/HUM/2013, 2013.

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, p. 710–717.

Padró, L., Stanilovsky, E. FreeLing 3.0: Towards Wider Multilinguality. In: Proc. of the Language Resources and Evaluation Conference (LREC'12) ELRA, 2012, p. 2473–2479.

Pilán, I. Helping Swedish words come to their senses: word-sense disambiguation based on sense associations from the SALDO lexicon. In: Proc. of NODALIDA, 2015, p. 275–279.

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, p. 655–675.

Sharoff, S., Nivre, J. The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. In: Proc. of Computational Linguistics and Intellectual Technologies (Dialog-2011), 2011, p. 657–670.

Sokirko, A., Toldova, S. Comparing a stochastic tagger based on Hidden Markov Model with a rule-based tagger for Russian. In: Internet-mathematics. Moscow, 2005, p. 80–94. (in Russ.)

Tufiş, D. Using a large set of EAGLES-compliant morpho-syntactic descriptors as a tagset for probabilistic tagging. In: Proc. of Second International Conference on Language Resources and Evaluation, 2000.

Материал поступил в редколлегию
Date of submission
25.10.2019

Сведения об авторах / Information about the Authors

Клышинский Эдуард Станиславович, кандидат технических наук, доцент школы лингвистики Национального исследовательского университета «Высшая школа экономики» (ул. Мясницкая, 20, Москва, 101000, Россия)

Eduard S. Klyshinsky, PhD in Tech. Sci., Associate Professor, National Research University Higher School of Economics (20 Myasnitskaya Str., Moscow, 101000, Russian Federation)

eklyshinsky@hse.ru
ORCID 0000-0002-4020-488X

Логачева Варвара Константиновна, кандидат физико-математических наук, научный сотрудник Сколковского института науки и технологий (Территория инновационного центра «Сколково», Большой бульвар, д. 30, стр. 1, Москва, 121205, Россия)

Varvara K. Logacheva, PhD in Appl. Math., Skolkovo Institute of Science and Technology (30 Bolshoy Blvrd., bld. 1, Moscow, 121205, Russian Federation)

v.logacheva@skoltech.ru

Карпик Олеся Владимировна, младший научный сотрудник Института прикладной математики им. М. В. Келдыша РАН (Миусская пл., 4, Москва, 125047, Россия)

Olesya V. Karpik, Junior Researcher, Keldysh Institute of Applied Mathematics RAS (4 Miusskaya Sq., Moscow, 125047, Russian Federation)

parlak@mail.ru
ORCID 0000-0002-0477-1502

Бондаренко Александр Викторович, доктор физико-математических наук, заместитель генерального директора Государственного научно-исследовательского института авиационных систем (ул. Викторенко, 7, Москва, 125319, Россия)

Alexander V. Bondarenko, Dr. Hab. in Appl. Math., Deputy Director, State Research Institute of Aviation Systems (7 Viktorenko Str., 125319, Moscow, Russian Federation)

bond@fgosniias.ru