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

Препринты ИПМ • Препринт № 4 за 2015 г.

Клышинский Э.С., Логачева В.К., Мансурова О.Ю., Максимов В.Ю., Карпик О.В., Зияздинов И.Б., Макеенко П.А.

Исследование неоднозначности употребления слов в европейских языках

Рекомендуемая форма библиографической ссылки: Исследование неоднозначности употребления слов в европейских языках / Э.С.Клышинский [и др.] // Препринты ИПМ им. М.В.Келдыша. 2015. № 4. 31 с. URL: http://library.keldysh.ru/preprint.asp?id=2015-4

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.К.Логачева, О.Ю.Мансурова,
В.Ю.Максимов, О.В.Карпик,
И.Б.Зиязтдинов, П.А.Макеенко

Исследование неоднозначности употребления слов в европейских языках

Москва — 2015

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

В работе рассматривается морфологическая и синтаксическая неоднозначности употребления слов в различных европейских языках. Для этого введены типы неоднозначности, разделенные в зависимости от того, по каким параметрам может быть неоднозначно данное слово. Также в работе приведен анализ неоднозначностей, возникающих при синтаксическом анализе текста. Для такого анализа используется информация о том, где может располагаться зависимое слово относительно главного в заданной паре с фиксированными частями речи: слева, справа или в любой из позиций. Анализ показал, что распределение слов по типам неоднозначности обладает уникальной формой для каждого из языков. Полученная информация позволяет перейти к анализу проблем языковой сложности.

Ключевые слова: морфологическая неоднозначность, синтаксическая неоднозначность, языковая сложность.

Klyshinsky E.S., Logacheva V.K., Mansurova O.Yu., Maximov V.Yu., Karpik O.V., Ziyaztdinov I.B., Makeyenko P.A.

Analysis of Words' Ambiguity in European Languages

In this paper, we investigated some properties of morphological and syntactical ambiguity of using of natural language words in several European languages. We introduced a set of ambiguity classes differentiated by predefined features resulting to lexical ambiguity. The syntactical ambiguity was investigated as well. In order to provide such analysis, we examined pairs of words with a given parts of speech. We examined the position of the governed word in the sentence: it precedes the heading word, it follows the heading word, or the governed word can be found in any position. We found out that resulting statistical distribution is unique for every language. It faces us the problem of language complexity, that should be investigated in a new project.

The research is partially supported by RFH grant №12-04-00060-a.

Key words: words co-occurrence, verbal government model, text style detection

Работа выполнена при поддержке РГНФ, грант № 12-04-00060-а.

1. Введение

На данный момент наиболее многообещающими методами обработки текстов на естественном языке являются методы статистического анализа данных. Это связано с тем, что в свободном доступе появляется всё больше как неразмеченных, так и размеченных, текстовых корпусов. В результате становится возможным не только описывать некоторые явления языка, но и давать им количественную оценку, что очень важно при разработке инструментов для автоматической обработки текстов. Практически все уровни анализа языка содержат как некоторое количество правил или регулярных структур, так и огромное количество частных случаев и исключений из этих правил. Обычно при создании компьютерных языковых моделей исключения требуют ручного анализа большого объема данных, в связи с чем их встраивание в системы анализа текстов занимает значительное время. С другой стороны, статистический анализ помогает определить частоты встречаемости подобных исключений и, как следствие, помочь принять решение о том, необходимо ли включить это новое правило в общую систему, или его можно опустить без значительной потери точности. Кроме того, обнаружение большого количества исключений может помочь обнаружить новую систему в их поведении и, как следствие, новые правила.

При разработке приложений, включающих в себя элементы обработки текстов на естественном языке, часто ставится вопрос о недостаточности имеющихся статистических данных для данного языка, например, статистики распределения частот употребления частей речи в текстах, количестве неоднозначно анализируемых слов, наиболее частых вариантов омонимии. На практике большая часть этой информации может быть легко извлечена из корпусов при условии создания соответствующих инструментов и методов. Однако насколько нам известно, до сих пор не было создано всеобъемлющего обзора, описывающего подобную информацию для большого числа языков. Чаще всего статистическая информация описывается небольшими фрагментами в различных статьях как незначимая часть предварительных исследований. Проекты, наподобие World Atlas of Language Structures Online (WALS)[1], собирают информацию для большого числа различных языков. При этом несмотря на полноту и большой объем собранной информации, содержащиеся в них данные являются скорее качественными, чем количественными. Подобная ситуация является скорее политикой разработчиков, чем упущением. Несмотря на это, мы считаем, что будучи собранной в одном месте, статистическая информация о некоторых явлениях в языке должна помочь при разработке практических приложений.

Среди прочего, подобное собрание статистической информации по ряду языков может быть полезно, например, при сравнении качества работы систем

[1] http://wals.info

автоматической обработки текстов, разработанных для разных языков. На данный момент результаты, получаемые для разных языков, довольно сложно сравнивать. С одной стороны, исследователи осознают наличие различий в фундаментальных основах языков. С другой стороны, для языков, входящих в одну группу по определенному признаку, результаты скорее всего будут также сходны. Как следствие, определить, что является причиной сравнительно низких результатов, полученных для нового языка, довольно сложно. Это может быть и недостаток обучающих данных, недостатки использованного метода или особенности самого исследуемого языка, накладывающие ограничения на возможный получаемый результат. Мы считаем, что статистическая информация о неоднозначности, наблюдаемой в текстах на естественном языке, может служить ключом к решению указанной проблемы.

Еще одним вопросом в автоматической обработке текстов, связанным с темой данного исследования, является перенос методов и техник на новые языки. Несмотря на то, что многие из методов декларируются как языконезависимые, они редко тестируются больше, чем на нескольких языках, а анализ применимости данных методов к другим языкам с учетом их особенностей не проводится. Чаще всего подобные ограничения выявляются для методов и систем, изначально ориентированных на применение для английского языка, но примененяемых к языкам с высокой степенью флексии, например, к русскому языку. Так, например, Протопопова и Бочаров в работе [Protopopova, 2013] описывают применение метода снятия омонимии Брилля (см. [Brill, 1995]) к русскому языку. Хотя сам по себе метод заявлен как применимый к любому языку, на практике авторам пришлось изменять набор грамматических параметров (добавив в него такие параметры, как род, число, падеж и др.) и учитывать богатую флексию русского языка, так как правила английского языка в чистом виде выдавали относительно низкий процент точности. В работе [Sharoff, 2011] показана аналогичная ситуация. Система морфологического анализа TnT, разработанная Брантсом (см. [Brants, 2000]), начинает выдавать приемлемые результаты только после расширения списка параметров. Анализ статистической информации о лексике нового языка может понять априори, необходимо ли будет вносить в метод или набор параметров те или иные изменения, чтобы он начал успешно работать. Он же должен помочь нам определить достаточный объем обучающей выборки, который необходим для различных языков, чтобы достигнуть приемлемый уровень качества.

Некоторые приложения в области автоматической обработки текстов могут требовать статистической информации для каждого из уровней обработки текста: от распределения фонем для общих характеристик текста (стиль, предметная область, время написания и проч.). Свое исследование мы решили начать с рассмотрения статистики в двух областях, морфологический и синтаксический анализ, поскольку эти два этапа используются наиболее часто. Задачей данной работы было оценить различия статистики распределения слов текстов по их грамматическим характеристикам (части речи, лемме и

грамматическим параметрам) и синтаксическим связям между ними. Здесь мы рассматривали грамматически неоднозначные слова, то есть слова, морфологический анализ которых возвращает более одного варианта для леммы, части речи или набора грамматических параметров. Подобное явление присуще всем языкам, но причины его возникновения различны. Так, например, в английском языке, известном своей полисемичностью, основную проблему составляет определение части речи слов, для которого необходимы особые инструменты. При этом в языках с развитой флексией неоднозначность по части речи гораздо ниже (общепризнанный факт, который до сих пор не был проверен количественным образом), а основной проблемой является определение корректной формы слова, употребленной в тексте. Эта разница заставляет исследователей использовать отличающиеся методы снятия омонимии в зависимости от рассматриваемого языка.

С точки зрения синтаксиса анализировалось взаимное расположение главного и зависимого слов. Обычно считается, что английский язык обладает строгим расположением слов в тексте, тогда как русский язык позволяет обращаться со структурой предложения весьма свободно, используя для ее восстановления информацию о согласовании и управлении (то есть опять высокую флективность языка). Однако данные положения также до сих представляют общепринятое ощущение, не подтвержденное цифрами.

В данной работе мы постарались заменить качественные оценки различий в степени неоднозначности текстов разных языков на количественные. Мы рассмотрели количественные различия слов, омонимичных по лемме, части речи и грамматическим параметрам. Мы также оценили соотношение синтаксических связей между главным и зависимым словом, в которых главное слово может находиться после зависимого, либо предшествовать ему. Подобная характеристика также отражает «предпочтения» разных языков к образованию тех или иных связей, их регулярности и количеству правил, необходимых для синтаксического анализа.

Эксперименты по оценке грамматической неоднозначности были проведены для английского, французского, испанского, итальянского, немецкого, русского и польского языков. Результаты для синтаксического анализа получены для немецкого, шведского, финского, итальянского и русского языков. Выбор языков был обоснован степенью доступности корпусов и словарей.

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

В большей части работ, связанных с разрешением грамматической неоднозначности, приводится информация лишь для одного языка. В общем случае, темой таких исследований является частное исследование, например, распределение падежей существительных в текстах некоторого жанра (см. [Копотев, 2008; Lyashevskaya, 2013]), разработка частотного словаря для фиксированного языка (см. [Bolshakov, 2002]) и др. Несмотря на приведенные численные данные, сравнение типов омонимии и подбор параметров для такого сравнения проводилось скорее с использованием эмпирических методов (и отражало интуицию исследователя). Статистика позволяет провести численную оценку влияния омонимии на другие параметры работы системы. Так, например, в работе [Fabricz, 1986] показано влияние частеречной омонимии на производительность систем машинного перевода, а в [Krovetz, 1997] показано, что разрешение частеречной омонимии повышает качество извлечения фактов.

Некоторая статистическая информация по отдельным языкам включалась в статьи в качестве обоснования для разработки инструментов, ориентированных на определенные языки. Так, например, в работе [Hajič, 1998] приведена информация о частеречной омонимии в чешском языке. Авторы вводят понятие «класс неоднозначности», которое используется для представления множества словоформ одного слова с неоднозначной частью речи. Так, например, слову «process», которое в английском языке может являться как существительным, так и глаголом или прилагательным, присваивается класс POS_{NVA}. Понятие класса неоднозначности использовалось в дальнейшем в некоторых работах при создании систем морфологического анализа с разрешением неоднозначностей. Подобные работы использовали статистические данные в качестве отправной точки своих исследований. Некоторые статистические данные для венгерского и английского языков (количество неоднозначных токенов, среднее число мловоформ на словоупотребление) приведены в работе [Oracecz, 2002], а работа [Tufiş, 2000] содержит данные для румынского языка. Работа [Pinnis, 2011] также использует классы неоднозначности, применяемые для исследования размеров и свойств этих классов в эстонском, литовском и латышском языках.

Ряд работ по дискурсивному анализу предоставляет иную статистическую информацию, например, в [Li, 2012] приведено распределение частей речи в тексте. Поскольку эта информация весьма важна при определении стиля текста или его авторства, подобные данные используются как основа для дальнейшего исследования. Диссертация на соискание степени PhD, написанная Перцовой [Pertsova, 2007], уже содержит в себе основу для исследования шаблонов, выявляющихся при изучении различных языков. К сожалению, данное

исследование также скорее является качественным, чем количественным, так как опирается на уже упоминавшийся проект WALS. В целом, имеющиеся на данный момент исследования не выработали какой-либо межъязыковой шкалы для выделения типов неоднозначностей и их количественного сравнения.

Краткое сравнение грамматической неоднозначности слов для русских и английских текстов можно найти в [Клышинский, 2013], однако двух языков очевидно недостаточно для полноценного сравнения. Чтобы исправить текущую ситуацию, в данной работе мы включили в рассмотрение еще несколько языков из разных языковых групп, а также дали собственное формальное определение классам неоднозначности.

3. Метод анализа омонимичной лексики

Представим текст T как последовательность токенов (слов), 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, то есть мы не рассматриваем знаки пунктуации, числа и прочие составные части текста.

Задачей морфологического анализа токена w является определение его леммы (начальной формы), его части речи и набора грамматических параметров (англ.: tag). Список приписываемых грамматических параметров будет зависеть от языка, которому принадлежит данное слово, и части речи, полученной в результате анализа. Одному и тому же слову может быть приписано несколько таких наборов. В связи с этим определим словоформу v как кортеж $v=<l, \pi, \mu>$, где l – это лемма данной словоформы, π – ее часть речи и 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). Назовем слово w несловарным, если оно не представлено в словаре, т.е. $\varphi(w) = \varnothing$ (или $k = 0$), в противном случае слово будет являться словарным и $k > 0$. Если $\varphi(w)$ содержит более одной словоформы ($k > 1$), слово w будет называться неоднозначным. Подобные слова представляют основную проблему в ходе морфологического анализа и являются предметом этапа снятия омонимии: каждое слово должно быть соотнесено с единственной словоформой, т.е. с единственной леммой, частью речи и набором параметров. Подобный подход принят при решении большого количества практических задач. Заметим, что разные виды неоднозначности по-разному влияют на результаты решения практических задач. Кроме того, различные виды неоднозначности требуют применения различных (более простых или более сложных) средств. В ряде задач омонимия может вовсе не сниматься.

В данной работе мы выделили шесть типов омонимии в зависимости от того, какие части кортежей отличаются между собой. Описание указанных типов приведено в табл. 1.

Таблица 1. Классы омонимии (ч.р. = часть речи, пар. = грамматические параметры).

	Ч.р.	**Лемма**	**Пар.**	**Описание**
1	0	0	0	**Однозначное** (слово имеет один вариант анализа)
2	0	0	1	**Неоднозначное по параметрам**: В анализе присутствуют словоформы с различными множествами грамматических параметров. **Пример:** Немецкий глагол ‘wohnen’ (‘проживать’) имеет совпадающие формы ‘wohnen’ в инфинитиве, настоящем времени первого лица множественного числа, третьего лица множественного числа и вежливой формы второго лица.
3	1	0	–	**Неоднозначное по части речи**: У слова есть словоформы, совпадающие по лемме, но отличающиеся по части речи. Грамматические параметры в данном случае сравниваться не могут. **Пример:** В английском языке слово ‘close’ может быть существительным, глаголом и прилагательным.
4	0	1	0/1	**Неоднозначное по лемме:** Словоформы обладают одной и той же частью речи, но различаются леммами. Параметры при этом могут как совпадать, так и нет. **Пример с совпадающими параметрами:** Русское слово ‘смели’ может быть формой третьего лица множественного числа прошедшего времени как глагола ‘сметь’, так и глагола ‘смести’. **Пример с различающимися параметрами:** В русском языке слово ‘вина’ может означать существительное ‘вина’, находящееся в именительном падеже единственного числа, или слово ‘вино’ в родительном падеже единственного числа или именительный падеж множественного числа.
5	1	1	–	**Омонимичное по части речи и лемме**: Данный класс включает слова, в которых словоформы отличаются как по части речи, так и по лемме. Параметры в данном случае несравнимы. **Пример:** Французское ‘est’ может быть существительным ‘est’ (‘восток’) или глаголом ‘être’ (‘быть’) в настоящем времени третьего лица единственного числа.
6	–	–	–	**Несловарное**: Слово не содержится в словаре.

Рассмотрим приведенные в табл. 1 типы омонимии более подробно. В некотором смысле, если слово принадлежит любому из указанных типов, это означает, что мы не можем быть до конца уверенными в лемме, части речи или параметрах данного слова. Однако мы можем предположить, что одни виды омонимии обрабатываются проще чем другие.

Например, слово, **омонимичное по параметрам**, обладает однозначно определяемой частью речи и леммой. В этом случае мы не обладаем всем контекстом, определяющим роль слова в предложении, но эта роль ограничена некоторым набором, присущим данной части речи или лемме. На практике многие задачи в области автоматической обработки текстов, опирающиеся,

например, на модель «мешка слов» (такие как извлечение фактов, анализ тональности текста и проч.), обычно не используют морфологическую информацию. Для них более важна информация о части речи, так как именно часть речи чаще всего влияет на изменение смысла слова, тогда как падеж существительного или время у глагола влияют на смысл не так сильно. В связи с этим слова, которым приписан данный вид омонимии, могут рассматриваться в некоторых задачах как однозначные.

Слова, **омонимичные по части речи**, не дают нам подобной подсказки, помогающей определить синтаксическую или семантическую структуру предложения. Для того чтобы извлечь подобную информацию, необходимо применять методы разрешения омонимии в форме правил или статистической информации. Как первые, так и вторая обычно извлекаются из размеченных корпусов. Однако подобные корпуса обычно содержат ошибки, влияющие на качество итогового снятия омонимии. При этом здесь наблюдается некоторое противоречие. Если размер обучающей выборки слишком мал, велика вероятность, что некоторые языковые явления не будут в ней представлены (подобная ситуация, к сожалению, характерна для многих современных языков). Большая же выборка скорее всего будет содержать больше ошибок.

Для того чтобы продолжить изложение, введем некоторые понятия, касающиеся систем морфологической разметки текстов. Будем считать, что система морфологической разметки текстов содержит в себе два компонента: *словарь*, который однозначно определяет все возможные варианты анализа для каждого поданного на вход слова, и *модуль снятия неоднозначности*, который выбирает один из вариантов анализа слова, используя при этом его контекст. Некоторые системы (например, Tree-Tagger) не строятся по этой двучастной структуре. Несмотря на это, они требуют для обучения размеченного корпуса, то есть обучаются и формируют из него свой словарь. Таким образом, для системы морфологического анализа с множеством наборов грамматических параметров $\boldsymbol{T}$ (англ.: tagset), условная вероятность $P(t \mid w)$ для слова w определяется не для каждого набора параметров из $\boldsymbol{T}$, а для некоторого подмножества наборов $\boldsymbol{T'}$, встретившихся в обучающей выборке: $\boldsymbol{T'} \subseteq \boldsymbol{T}$, $\forall\ t' \in \boldsymbol{T'}$ $P(w, t') > 0$. Подобные нюансы становятся более очевидными, если мы рассмотрим определение условной вероятности:

$$P(A \mid B) = P(A, B)/P(B).$$

Если A и B никогда не встречаются вместе, то $P(A, B) = 0$, и, как следствие, $P(A \mid B) = 0$. В нашем случае, если в обучающей выборке не встретилось слово w с набором параметров t, то $P(w, t) = 0$ и $P(t \mid w) = 0$. Как следствие, словарь вероятностной системы морфологической разметки содержит в себе только те наборы параметров, которые встретились в обучающей выборке. Наборы параметров, составленные вручную, чаще всего будут более полны, так как скорее отражают регулярности языка, чем частоты встречаемости наборов параметров в фиксированном тексте.

В соответствии с нашими рассуждениями об информации, получаемой из неоднозначных слов, можно сделать следующий вывод. Если в текстах на данном языке встречается мало слов с омонимией по части речи, некоторые задачи по обработке текстов могут быть решены без использования системы морфологической разметки. В таких задачах точность разметки превалирует над ее полнотой, то есть, например, в ходе анализа не будут рассматриваться грамматические параметры (и в этом смысле система может допускать ошибки), но точная атрибуция по части речи очень важна для принятия решения.

Разница в подходах хорошо чувствуется в языках с малым количеством языковых ресурсов. Разработка словаря системы морфологического анализа требует значительных усилий. Ручная разметка новых ресурсов позволяет использовать эти ресурсы для решения новых задач, увеличивая, например, точность результатов для существующих систем машинного обучения или предоставляя новые данные для их тестирования. В этом смысле ручная разметка и сбор статистики более выгодны. Но с другой стороны, морфологический словарь, составленный вручную, будет обладать более высокими полнотой и точностью, чем словарь, составленный автоматически.

Зачастую создание множества наборов грамматических параметров (tagset) вручную для языков с малым количеством языковых ресурсов может дать больший эффект. Правила образования слов в словаре, составленном вручную, являются детерминистическими и скорее всего не войдут в противоречие, как это часто случается при автоматическом сборе вероятностных правил. Само заполнение морфологического словаря автоматизируется уже после того, как в него вводятся основные шаблоны словоизменения (лингвисту достаточно собрать слова в группы с одинаковыми шаблонами и выделить этот шаблон). Словарь, составленный вручную, обладает значительно большим покрытием, так как содержит в себе «запасные» формы слов, не встреченные в сравнительно небольшом обучающем корпусе. Особенную актуальность этот вопрос получает в языках с развитой флексией, когда вероятность найти на маленьком корпусе (или даже корпусе среднего размера) все формы заданного слова стремится к нулю.

Итак, разработка системы морфологической разметки, содержащей в себе только словарь и не содержащей модуля снятия омонимии, является более простой, но всё еще полезной задачей. Если задача предполагает анализ информации только о части речи слова, для некоторых языков она может быть решена с более низким покрытием, но высокой точностью, без снятия омонимии. Точность в этом случае будет близка к 100% (с учетом возможных ошибок в словаре или отсутствия омонимов), а процент покрытия определяется числом однозначных по части речи слов (однозначных, неоднозначных по параметрам и неоднозначных по лемме).

4. Используемые данные и инструменты

Для получения результатов мы использовали информацию по следующим языкам: английский, французский, испанский, итальянский, немецкий, русский и польский. Мы использовали существующие системы морфологического анализа, описание которых приведено в табл. 2. Предсказание незнакомых слов было отключено, не найденные системой слова помещались в класс несловарных.

Наши предыдущие эксперименты показали, что результаты анализа зависят от стиля корпуса [Клышинский и др., 2013]. Для того чтобы устранить влияние этого фактора, мы проводили эксперименты только на новостных корпусах. Характеристики корпусов также приведены в табл. 2.

Таблица 2. Характеристики систем разметки и корпусов

Язык	**Система разметки**	**Размер словаря в лексемах**	**Корпус**	**Размер корпуса в словоупотреблениях**
Английский	Расширенная АОТ.ru[2]	105 000	Reuters	300 млн
Французский	Morphalu[3]	68 000	Le Parisien	43.1 млн
Испанский	FreeLing[4]	76 000	Abc.es	15.2 млн
Итальянский	FreeLing	40 000	Corriere dela Serra	7.9 млн
Немецкий	TreeTagger[5]	n/a	Die Zeit	7.1 млн
Русский	Расширенная АОТ.ru	167 000	Lenta.ru	32.4 млн
Польский	Morfologik[6]	> 400 000	Different sources	21.2 млн

Для того чтобы провести межъязыковой анализ более точно и устранить разницу между текстами, мы также провели анализ на параллельных корпусах. Мы использовали французский, немецкий и испанский корпуса News Commentary, созданный для тестов в ходе семинара по статистическому машинному переводу.[7]

Каждая система анализа имела свой собственный набор параметров, изучение которых показало на существенные различия. Например, в немецком словаре содержатся тэги для одной и той же части речи, обозначающие различные роли слов в предложении (ADJA для атрибутивных прилагательных и ADJD для предикативных). В русском языке наблюдалась противоположная ситуация: часть речи для прилагательного была одна, его свойства отражались самостоятельными параметрами, но при этом не включались в тэг. Для унификации мы нормализовали все тэги, убрав всю синтаксическую информацию или извлекая грамматическую информацию из тэга и перемещая

[2] http://aot.ru (Sokirko and Toldova, 2004)
[3] http://www.cnrtl.fr/lexiques/morphalou/
[4] http://devel.cpl.upc.edu/freeling/downloads?order=time&desc=1 (Padró and Stanilovsky, 2012)
[5] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (Schmid, 1995)
[6] http://morfologik.blogspot.ru/2013/02/morfologik-20-rc2.html
[7] http://statmt.org/wmt14/news-commentary-v9-by-document.tgz

ее в грамматические параметры. Для таких языков, как испанский и итальянский, нормализованный набор параметров не отличался от исходного. Полученный список частей речи и число различных грамматических параметров для каждого языка приведены в табл. 3. Число параметров приведено во второй строке. Для более подробного анализа рекомендуется ознакомиться с документацией на программные модули и описанием их словарей.

Таблица 3. Список частей речи для исследуемых языков

Рус.	Англ.	Исп.	Итал.	Нем.	Франц.	Польск.	Часть речи
12	9	19	10	4	5	12	
adj	adj	ADJ	ADJ	Adjektive	adjective	adjective	Прилагательное
adv	adv	ADV	ADV	Adverbien	adverb	adverb	Наречие
conj	conj	CONJ	CONJ	Konjunktionen		conjuction	Соединительное слово
	article	DET	DET	Artikel	functionWord		Определитель
interj	int	INTERJ	INTERJ	Interjektionen	interjection	interjection	Междометие
noun	noun	NOUN	NOUN	Nomina	commonNoun	noun	Существительное
prep	prep	PREP	PREP	Adpositionen	functionWord	preposition	Предлог
pers_pron	pronoun	PRON	PRON	Pronomina		pronoun	Местоимение
verb	verb	VERB	VERB	Verben	verb	verb	Глагол
particle	part			Partikeln		particle	Частица
card_num	card			Kardinalzahlen		numeral	Количественное числительное
deepr							Деепричастие
dem_pron	pn_dem						Указательное местоимение
ord_num	ord						Счетное числительное
participle						participle	Причастие
poss_pron	pn_poss						Притяжательное местоимение

5. Результаты экспериментов по определению морфологической неоднозначности

5.1. Распределение словоупотреблений по типам омонимии

Для выбранных языков мы провели серию экспериментов. Результаты экспериментов показаны в табл. 4. Те же данные в визуальной форме представлены на рис. 1. Процент словоформ с единственным вариантом анализа (неомонимичных) колеблется между 30% и 50% за исключением выброса для польского языка с 20%. Два языка с развитой флексией (русский и польский) показали высокий уровень омонимии по грамматическим параметрам (примерно 30-40% против 0-5% для других языков). Английский язык показывает очень высокий процент слов, омонимичных по части речи.

Таблица 4. Распределение словоформ по классам неоднозначности

	Однозначные	Неоднозн. по параметрам	Неоднозн. по части речи	Неоднозн. по лемме и параметрам	Неоднозн. по лемме и части речи	Несловарн.
Рус.	48.28%	27.68%	5.26%	4.67%	9.92%	4.38%
Польск.	18.94%	39.13%	13.49%	7.23%	18.44%	2.78%
Англ.	38.87%	2.79%	50.35%	0.32%	0.69%	7.65%
Итал.	41.66%	0.14%	13.36%	0.98%	23.10%	20.77%
Франц.	51.21%	3.69%	7.40%	7.96%	19.59%	10.15%
Франц. NC	54.54%	2.80%	8.21%	9.54%	19,83%	5.08%
Исп.	33.00%	0.09%	21.48%	0.47%	28.97%	15.98%
Исп. NC	35.53%	0.05%	21.51%	0.33%	29.65%	12.93%
Нем.	33.27%	4.51%	22.40%	1.36%	13.07%	25.39%
Нем. NC	44.53%	9.35%	23.91%	0.91%	5.26%	16.03%

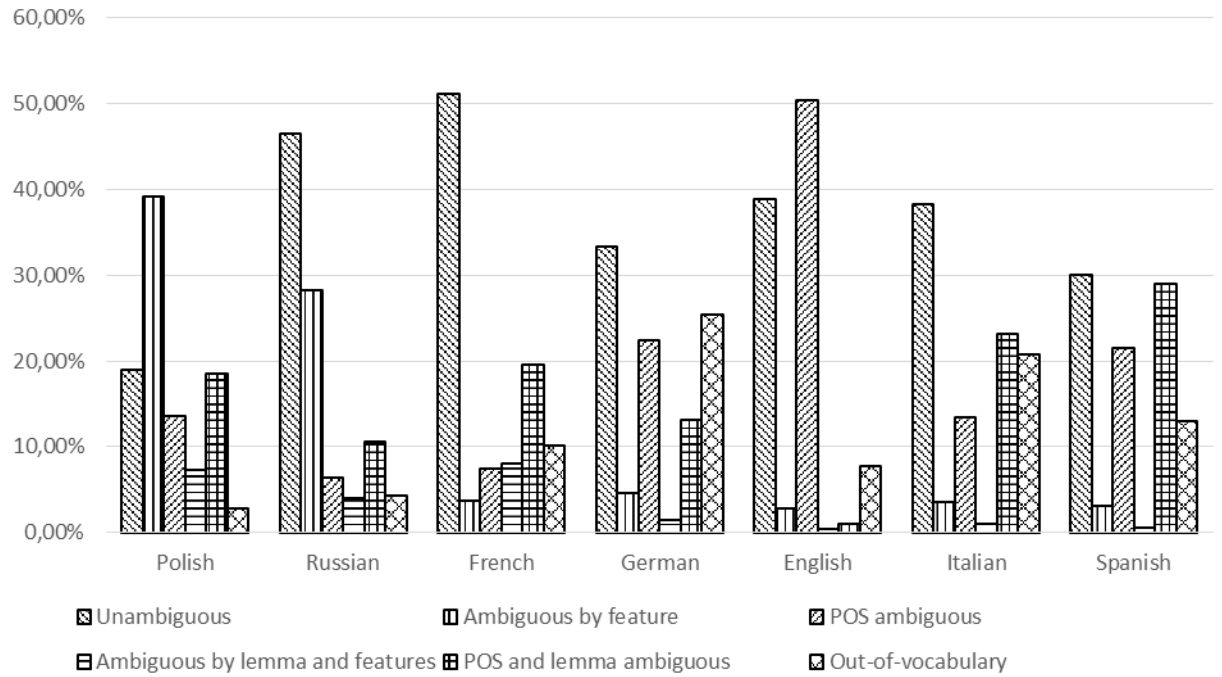

Рис. 1. Распределение словоупотреблений по типам неоднозначности для различных языков

Для анализа сходства распределений мы использовали коэффициент корреляции. Распределения для немецкого, итальянского и испанского языков имеют сходную форму (корреляция от 0,83 до 0,91) и показывают высокий процент однозначных слов. При этом французский сходен с итальянским (корреляция 0,93), и меньше похож на остальные два языка (0,79 и 0,7 для немецкого и испанского).

Средняя корреляция для языков колеблется в широких пределах. Так, наиболее «похожими» на все языки оказались немецкий (средняя корреляция 0,68), итальянский (корреляция 0,65) и французский (корреляция 0,61). Наиболее выделяющимся оказался польский язык (средняя корреляция 0,21 с

выбросом у сходства с русским языком – корреляция 0,5). Следом за ним идут английский (средняя корреляция 0,45) и русский (средняя корреляция 0,5).

Если рассмотреть отдельно немецкий, испанский, французский и итальянский языки, то их средняя корреляция между собой возрастает до значений между 0,7 для французского и 0,83 для немецкого языка. Польский и русский языки показывают корреляцию на уровне 0,5. Таким образом, мы можем сказать, что в распределении омонимии романские и немецкий языки ведут себя сходно, английский и русский отличаются от них, а польский не коррелирует ни с одним языком, кроме слабого сходства с русским).

Как это отмечалось выше, мы провели проверку с использованием параллельных корпусов текстов, переведенных с английского языка. Для французского и испанского корпусов распределения получились сходными с распределениями, полученными для корпусов, изначально написанных на этих языках. Корреляция в распределениях омонимов для двух корпусов французского языка составила 0,999, для испанского – 0,998. Распределения для немецких корпусов имеют заметные различия, их корреляция достигла лишь 0,94. Анализ этих различий представляет интерес для самостоятельной работы и не будет рассмотрен здесь подробно. Однако можно предположить, что одним из объяснений может быть особенность перевода, когда переводчик не пишет текст заново, а пытается подобрать более простые слова и конструкции, сходные с таковыми в языке оригинала.

5.2. Разметка текста без разрешения неоднозначности

В разделе 3 мы предположили, что в ряде задач может быть полезной обработка только тех слов, которые однозначны по части речи. Описанный метод анализа позволяет оценить процент слов текста, которые могут быть использованы при подобном анализе. Для этого необходимо взять все слова, для которых однозначно определена часть речи, то есть процент подобных слов определяется как сумма процентов однозначных слов, слов, неоднозначных по параметрам, и слов, неоднозначных по лемме.

На рис. 2 посредством сплошной линии показан процент слов, однозначных по части речи, для различных языков. Из графика видно, что языки с более развитой флексией обладают большим процентом слов, однозначных по части речи. Для русского, польского и французского языков более чем две трети слов могут быть использованы без разрешения неоднозначности по части речи. Из этого можно сделать вывод о большей эффективности применения подобных методов в разных языках. Более того, становится очевидным, что основные проблемы, возникающие при разрешении неоднозначности, в разных языках принципиально отличаются. Если при анализе английских и испанских текстов основную проблему составляют слова, омонимичные по части речи, то для русского и польского языка большую проблему составляет неоднозначность определения набора грамматических параметров (причем зачастую не всего набора, а лишь отдельных параметров).

Подобная разница иллюстрируется рисунком 3, на котором показан процент словоупотреблений, неоднозначных по грамматическим параметрам.

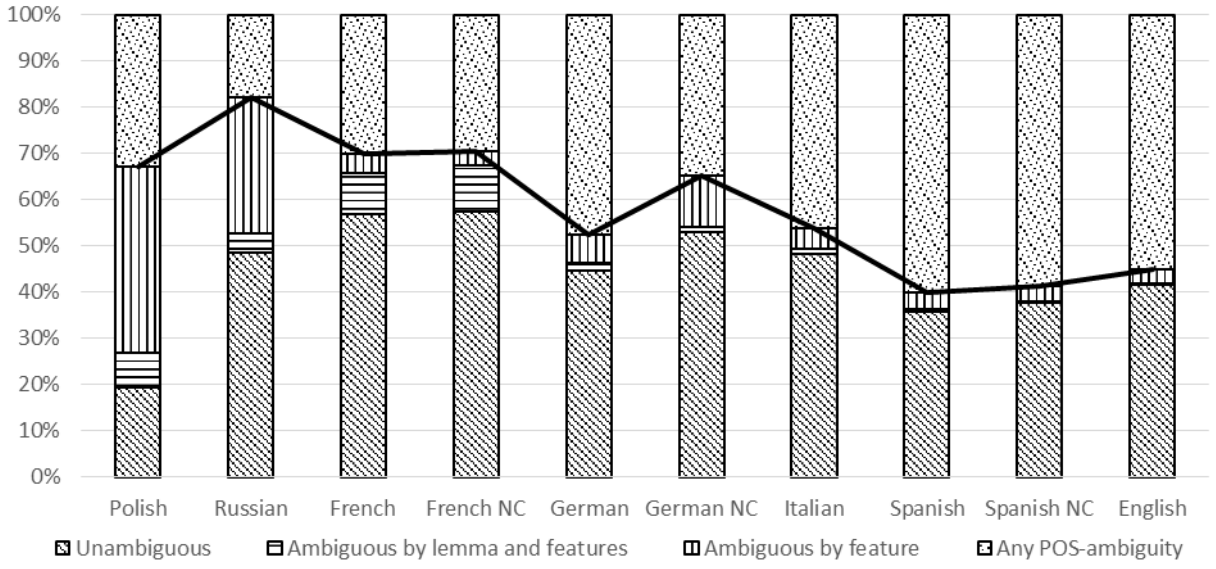

Рис. 2. Процент словоупотреблений, который может быть использован в различных языках без разрешения неоднозначности по части речи

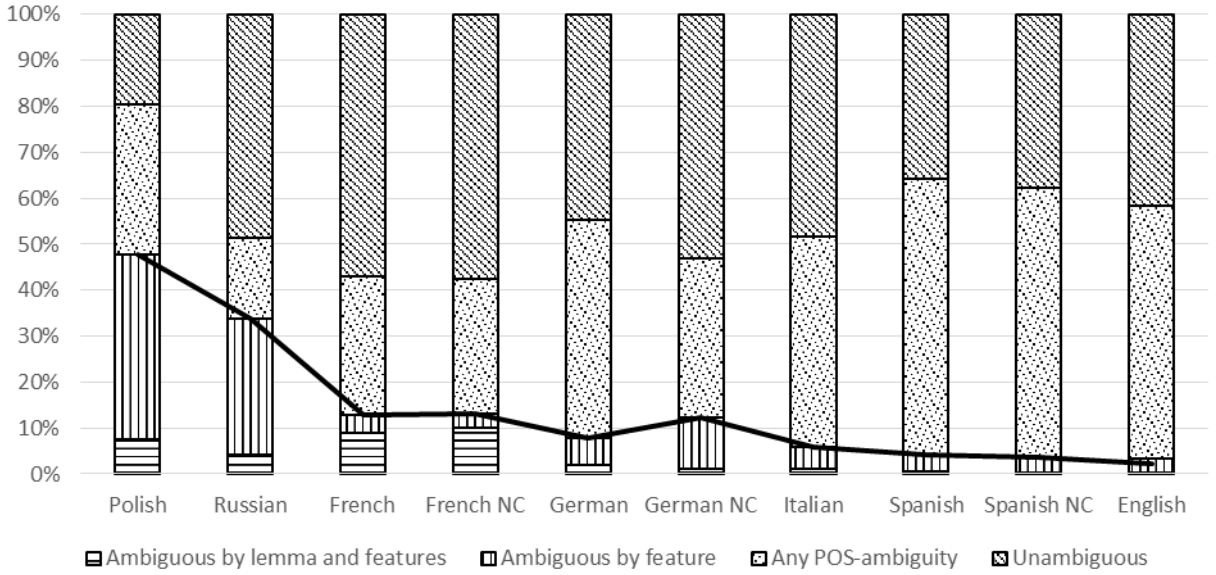

Рис. 3. Процент словоупотреблений, обладающих неоднозначностью по грамматическим параметрам

5.3. Дополнительные эксперименты

Следует заметить, что для своих экспериментов мы использовали весьма разнородные ресурсы и инструменты: корпуса и словари отличаются друг от друга по размеру, количество параметров не совпадает и т.д. В связи с этим необходимо оценить влияние этой неоднородности на распределение неоднозначных слов по типам, так как между ними может быть найдена какая-либо зависимость. Для того чтобы показать отсутствие подобной зависимости, мы провели дополнительную серию экспериментов.

Первая гипотеза, которую мы решили отвергнуть, состоит в том, что наблюдаемые различия между распределениями для различных языков могут быть объяснены отличиями в размерах морфологических словарей. Для этого мы решили проверить, как изменяется распределение в зависимости от размеров словаря. Эксперименты проводились для русского и английского языков. Мы отсортировали слова в словарях в соответствии с частотами их употребления в рассматриваемых корпусах. После этого распределение словоформ по типам было рассчитано лишь для фрагментов словарей, содержащих в себе наиболее встречающиеся 1000, 3000 и 5000 слов.

Результат, представленный на рис. 4, показывает, что изменение размеров словаря не изменяет формы распределения, то есть для русского языка сохраняется относительно большее количество слов с неоднозначностью по грамматическим параметрам, тогда как для английского языка на первом месте остаются слова с неоднозначностью по части речи. Корреляция между распределениями для русского языка не опускалась ниже 0,993, для английского – 0,999, тогда как между собой они коррелируют не более, чем со значением 0,42.

Заметим, что корреляция между распределениями для различных языков и русским с усеченным словарем увеличивалась по мере уменьшения словаря, тогда как для английского языка – колебалась с минимумом в районе размера словаря в 3000 слов.

В связи с этим мы можем утверждать, что вид распределения словоупотреблений по типам омонимии не изменяется в зависимости от размеров словаря. Более того, можно сформулировать утверждение, что форма распределения для языка является свойством некоторого базового ядра, содержащего в себе наиболее частотные слова. Однако проверка данного утверждения на большем числе языков является задачей для отдельного исследования.

Мы также проверили наличие зависимости от количества используемых частей речи. Для того чтобы отвергнуть гипотезу о подобной зависимости, мы постарались устранить разницу в наборах параметров для русского и английского языков. Так, для русского словаря мы заменили личные местоимения существительными, указательные местоимения на прилагательные, а деепричастия и причастия на глаголы. Изменения в

полученном в результате процентном соотношении составило около 0,1%, так как большинство местоимений в русском языке имеют уникальную лемму, а все причастия и деепричастия обладают характерной парадигмой изменения.

Наконец, мы также проверили наличие зависимости от количества используемых параметров, то есть флективности языка. Для этого мы последовательно удалили из рассмотрения такие параметры, как одушевленность, падеж, лицо и число, в словах из словаря русского языка. Удаление одушевленности изменило значения на 0,005%, удаление лица не изменило результирующие числа вообще. Устранение остальных параметров ожидаемо привело к уменьшению процента слов, неоднозначных по грамматическим параметрам и увеличило процент однозначных слов.

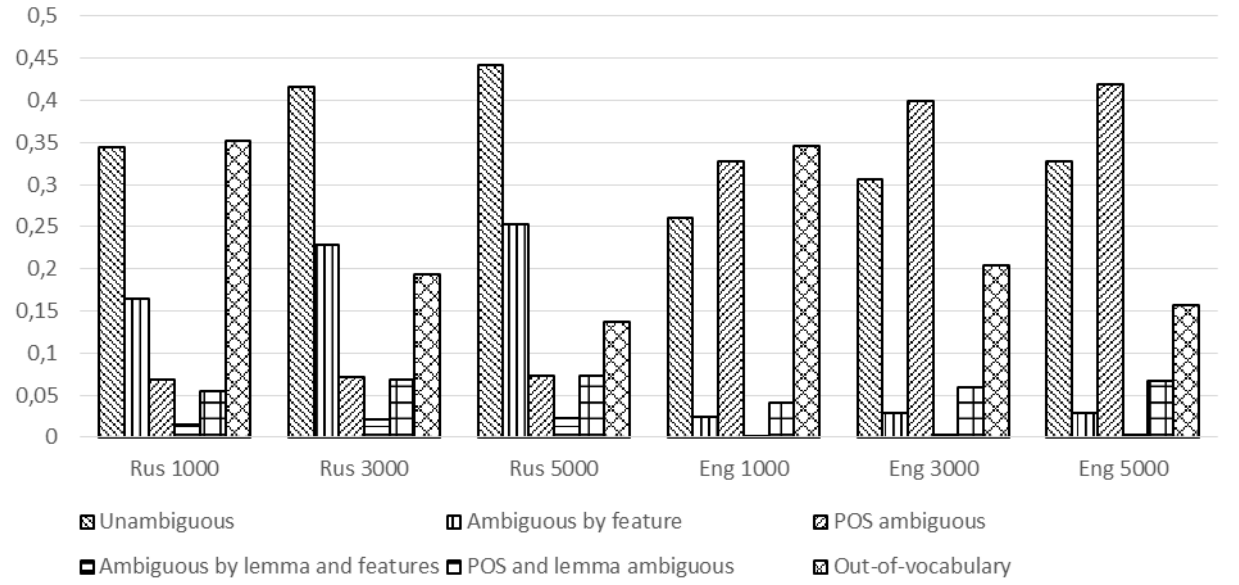

Рис. 4. Влияние размера словаря на форму распределения

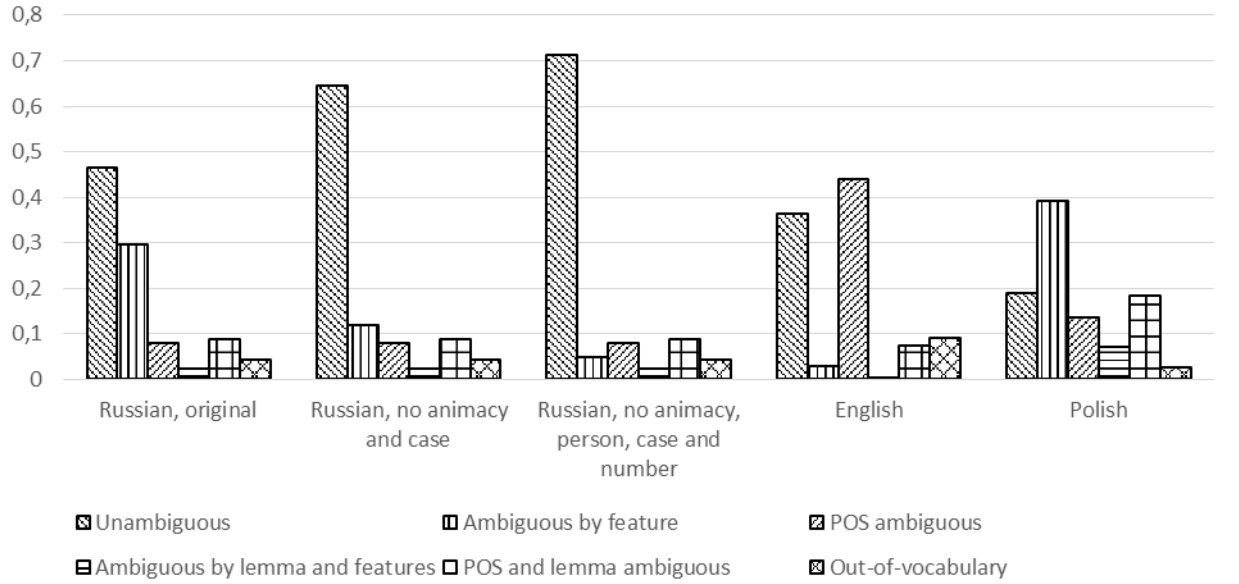

Рис. 5. Влияние набора параметров на форму распределения

Результаты для этих двух серий экспериментов приведены на рис. 5. Распределения для английского и польского языков даны для сравнения результатов. Из них хорошо видно, что несмотря на устранение частей речи и параметров из словаря русского языка распределение не стало ближе к другим языкам.

5.4. Некоторые выводы из экспериментальных данных

Заметим, что все представленные параметры влияют на форму распределения словоформ по типам неоднозначности. Тем не менее опробованные комбинации параметров словарей не приводят к значительным изменениям в форме распределений, ни одно из распределений не стало больше походить на другое. В любом случае, в русском языке сохранялось большее число однозначных словоупотреблений, чем в английском, при том, что английский язык сохранял больший процент словоупотреблений с неоднозначной частью речи. Заметим, что размеры словарей для русского и английского языков сравнимы. Сокращение числа грамматических параметров сокращает число слов, относящихся к омонимичным по грамматическим параметрам в русском языке с его богатой флексией. Однако разбор причин, приводящих к снятию такого рода омонимии, здесь рассматриваться не будет.

Представленный анализ не является исчерпывающим. Так, например, для анализа использовалась лишь одна система морфологического анализа для каждого языка, тогда как алгоритм работы и, что самое главное, размер и строение словаря очевидно будут влиять на результаты работы (хотя и не так заметно, как это следует из проведенных экспериментов). Эксперименты проводились только на новостных текстах, тогда как полное исследование следовало бы провести еще и, как минимум, на беллетристике и текстах делового стиля. Наши предыдущие исследования показали, что лексика и синтаксические конструкции, используемые в таких текстах, существенно отличаются. С другой стороны, наши эксперименты показали, что размер словаря не оказывает значительного влияния на форму распределения. Более того, мы можем утверждать, что 5000 наиболее частотных слов определяют форму распределения, причем менее частотные слова не изменяют ее существенно.

6. Обсуждение результатов анализа распределения словоформ по типам омонимии

Результаты наших экспериментов показали значительные отличия в рассматриваемых языках. Нахождение причин подобных отличий является темой для отдельного исследования. Однако полученные значения позволяют сделать некоторые выводы. Нами не было обнаружено корреляции между распределением слов по типам омонимии и количеством частей речи и грамматических параметров, за исключением того факта, что при уменьшении числа параметров снижается процент словоформ, неоднозначных по грамматическим параметрам, и увеличивается процент однозначных слов. Также отсутствует корреляция между размером словаря и формой распределения. Из этого мы можем сделать вывод, что форма распределения слов по типам омонимии зависит от других свойств языка.

Рассчитанные значения показывают, что применимость методов анализа текстов, не опирающихся на снятие неоднозначности, будет колебаться от языка к языку. Если часть речи определяется корректно и без снятия неоднозначности, как это можно наблюдать в языках с развитой флексией, применение подобных методов позволяет существенно повысить точность получаемых результатов, пожертвовав при этом их полнотой. При этом, с одной стороны, можно надеяться, что увеличение размеров корпуса позволит компенсировать снижение полноты. С другой стороны, слова, омонимичные во всех своих формах, останутся таковыми вне зависимости от размеров корпуса. Таким образом, мы можем безвозвратно потерять определенный процент лексики (для работы с которой можно использовать снятие омонимии). Следовательно, устранение модуля снятия омонимии из состава системы морфологической разметки текстов может оказаться приемлемым для языков с развитой флексией. С другой стороны, подобное преобразование применимо лишь для ограниченного числа задач, тогда как для остальных снятие омонимии чаще всего необходимо.

Задача морфологической разметки текстов со снятием неоднозначности зачастую ставится как задача последовательной разметки, в которой разметка текущего словоупотребления производится в соответствии с метками, присвоенными предыдущим. Так, например, скрытая модель Маркова максимизирует следующую вероятность:

$$P(t_k) = P(t_1)\prod_i P(t_i|t_{i-1})P(w_i|t_i).$$

Для ее оценки нам необходимо определить значения в множестве вероятностей перехода из одного состояния в другое $P(t_i|t_{i-1})$, которое содержит k*k элементов, где k – количество грамматических помет (содержащих как часть речи, так и уникальный набор параметров), и множество вероятностей обнаружить слово $P(w_i|t_i)$, чей размер равен количеству грамматических помет, умноженному на количество слов, имеющих данный набор помет.

Заметим, что чем больше в языке грамматических параметров и частей речи, тем больше будет набор помет для обозначения различных форм слов. Так, если система TreeTagger для английского языка содержит всего две пометы для существительного (NN для единстенного числа и NNS для множественного), то для флективных языков с двумя значениями числа и развитой системой, содержащей *c* падежей, нам потребуется как минимум 2**c* помет (не говоря о других параметрах). Таким образом, количество значений вероятности, которые нам необходимо получить из обучающего корпуса, может вырасти экспоненциально при переносе алгоритма снятия неоднозначности с одного языка в другой. То есть система снятия морфологической неоднозначности, разрабатываемая для языка с развитой флексией, требует обучения на гораздо большем корпусе, чем в случае языка со слабой флексией.

С другой стороны, в предыдущих работах, посвященных иностранным языкам (см. [Hajič, 1998; Pinnis, 2011]), или в нашей работе, посвященной русскому языку [Клышинский и др. 2013], указывается, что некоторые сочетания частей речи и параметров чаще образуют омонимичные пары. Так, например, существительные в русском языке могут быть омонимичны прилагательным, тогда как причастиям или предлогам они омонимичны довольно редко. Очевидно, что подобные сочетания зависят от языка. В любом случае, подобная статистическая информация позволяет надеяться, что разница в размерах обучающих корпусов для флективных и нефлективных языков будет не настолько чувствительна.

На практике необходимая статистика о парах омонимии может быть получена при помощи уже имеющихся корпусов или программных средств для каждого из рассмотренных языков. Данная информация должна помочь улучшить качество морфологической разметки лишь за счет изменения формы хранимых статистических данных. В ряде случаев возможен переход к правилам, показывающим, как снимается омонимия в конкретных случаях при фиксированном сочетании набора грамматических параметров.

Закончив рассмотрение явления омонимии, перейдем теперь к рассмотрению особенностей синтаксического анализа.

7. Синтаксическая неоднозначность

В ходе синтаксического анализа часто проявляется явление, называемое синтаксической неоднозначностью, когда для одного и того же предложения может быть построено несколько вариантов разбора в виде деревьев зависимостей или составляющих. Подобное явление связано с тем, что слова могут связываться при помощи различных типов зависимостей или составлять связи с другими словами. Подобное смещение связей может быть объяснено, например, наличием лексической омонимии, когда разные связи образуются с разными словоформами. С другой стороны, слова обладают набором валентностей, то есть связей, которые могут или должны быть заполнены другими словами предложения, или требовать при присоединении управления

(согласования лексических параметров соединяемых слов). Валентность предъявляет требования к присоединяемому слову. Так, например, глагол требует наличия определенного предлога, а присоединяемое существительное должно находиться в заданном падеже. В связи с этим неоднозначность может появляться в связи с перемещением зависимых слов к другим главным словам для заполнения свободных валентностей. Еще одним вариантом появления неоднозначности может являться возможность правого и левого ветвления для определенных связей. Под левым ветвлением будем понимать ситуацию, когда зависимое слово находится слева от главного, под правым – когда зависимое слово находится слева. (Подробнее см. [Тестелец, 2002]).

В данной работе была поставлена задача исследовать статистические характеристики подобного ветвления для различных языков. Так, например, принято считать, что английский язык обладает строгими правилами относительно правого или левого расположения зависимых слов, тогда как в русском языке зависимые слова более свободны в смысле выбора своего расположения. Численный анализ данного явления должен помочь разобраться в данном вопросе.

Для анализа синтаксической регулярности языка удобнее использовать синтаксически размеченный корпус текстов, так как автоматический синтаксический анализ представляет собой лишь некоторую модель, отражающую представление ее разработчиков о языке. В связи с этим использование той или иной системы синтаксического анализа позволит лишь оценить внесенные в нее правила, а не закономерности языка. Кроме того, на данный момент на реальных текстах системы синтаксического анализа показывают качество, лишь несколько превышающее 90%. Столь значительный процент ошибок может серьезно повлиять на качество собираемой статистической информации. Для данного исследования были выбраны синтаксически размеченные корпуса русского (СинТагРус [Apresjan et al., 2006; Богуславский и др., 2002]), финского (Turku Dependency Treebank [Haverinen et al., 1998]), шведского (TalBanken [Nivre et al., 2006]), немецкого (TiGer [Brants et al., 2004]) и итальянского (Turin University Treebank [Bosco et al., 2008]) языков. Нами были выбраны корпуса, размеченные с использованием деревьев зависимостей, так как задачей являлся анализ связей между отдельными словами. Деревья, полученные с использованием грамматик составляющих, не способны предоставить подобную информацию в явном виде.

В данном исследовании нашей задачей являлось проверить для различных видов синтаксических связей соотношение числа правого и левого расположения подчиненных слов. Так, например, в русском языке прилагательное может располагаться как до, так и после управляющего им существительного. В первом случае будем говорить о левом ветвлении, во втором случае – о правом.

Будем также говорить, что данная связь обладает симметрией, если для нее наблюдается как левое, так и правое ветвление, причем количество для обоих

случаев сопоставимо. Будем говорить, что данная связь обладает асимметрией, если в корпусе присутствует либо правое, либо левое ветвление, либо присутствуют оба вида ветвления, но количество примеров для одного в 20 и более раз превышает количество примеров для другого.

Метод анализа заключается в следующем. На вход метода поступает множество синтаксически размеченных предложений заданного языка, представленных в виде деревьев зависимостей. Для всех зависимых слов в корпусе составляются тройки вида <POS_l, *dir*, POS_r>, где POS_l – часть речи слова, расположенного в предложении слева, POS_r – часть речи слова, расположенного в предложении справа, и *dir* – обозначение расположения главного и зависимого слова (будем использовать знаки ← и → для обозначения левого и правого ветвления соответственно). В этом случае конструкция *красивая девушка* образует тройку < adj, ←, noun >, а конструкция *девушка красивая* – тройку < noun, →, adj >.

Для всех конструкций собирается статистика их встречаемости во всём корпусе. Далее проводится сравнение частот встречаемости для троек вида < POS_1, →, POS_2 > и < POS_2, ←, POS_1 >. По соотношению частот встречаемости и количестве симметричных и несимметричных конструкций может быть сделан вывод о строгости конструкций языка.

8. Результаты экспериментов по расчету синтаксической неоднозначности

Количество различных троек, выделенных для выбранных корпусов, показано в табл. 5 в графе «Количество сочетаний». Заметим, что количество сочетаний во многом определяется числом частей речи, используемых при разметке корпуса. Так, например, в корпусе СинТагРус личные местоимения размечены как существительные, а притяжательные – как прилагательные.

В остальных графах показан процент связей в тексте, для которых отношение разницы между встречаемостью симметричных связей к максимуму встречаемости находится в заданных пределах. То есть, если считать, что f_1 – это количество связей вида < POS_1, →, POS_2 >, а f_2 – количество связей вида < POS_2, ←, POS_1 >, то значение симметричности вычисляется как $abs(f_1-f_2)/max(f_1, f_2)$.

Графа «Единственный вид связи» табл. 5 показывает процент связей, не имеющих симметричной пары (симметрия равна 1), связи являются полностью асимметричными. Графа «Преобладание асимметрии» показывает процент связей, в которых имеются пары, однако число встречаемости одной связи в 20 и более раз преобладает над другой (симметрия от 0,95 до 1). Графа «Слабое преобладание асимметрии» показывает процент симметричных связей, в которых преобладание составляет 10-20 раз (симметрия от 0,9 до 0,95). Наконец, графа «Симметричные связи» показывает процент связей с

преобладанием одной над другой менее 10 раз (симметрия менее 0,9). Те же данные, но с несколько большей детализацией, представлены на рис. 2.

Также для оценки можно использовать диаграммы другого вида. Пусть по оси абсцисс будет отложена встречаемость троек, а по оси ординат – мера их симметрии. В этом случае можно визуально оценить, какие конструкции встречаются чаще всего – симметричные или несимметричные. Подобные диаграммы для различных языков приведены на рис. 7-11. Из графиков видно, что в русском языке существует четыре вида связи, на которые приходится значительная часть связей в предложениях, причем связь глагол-существительное симметрична, тогда как связь существительное-прилагательное, существительное-существительное и (в особенности) существительное-предлог весьма асимметричны. На диаграммах по оси абсцисс приведена сумма встречаемости для сравниваемых троек, а не их максимум.

Табл. 5. Процентное соотношение симметричных и несимметричных связей

Язык	**Кол-во сочетаний**	**Единственный вид связи, =1**	**Преобладание асимметрии [0,95;1)**	**Слабое преобладание асимметрии [0,9;0,95)**	**Симмет-ричные связи, >0,9**
русск.	191	13,09%	20,94%	12,57%	53,40%
финск.	2656	34,56%	2,79%	3,61%	59,04%
итал.	212	35,85%	23,58%	7,55%	33,02%
шведск.	1876	31,02%	14,50%	7,68%	46,80%
немецк.	494	19,84%	22,27%	6,48%	51,42%

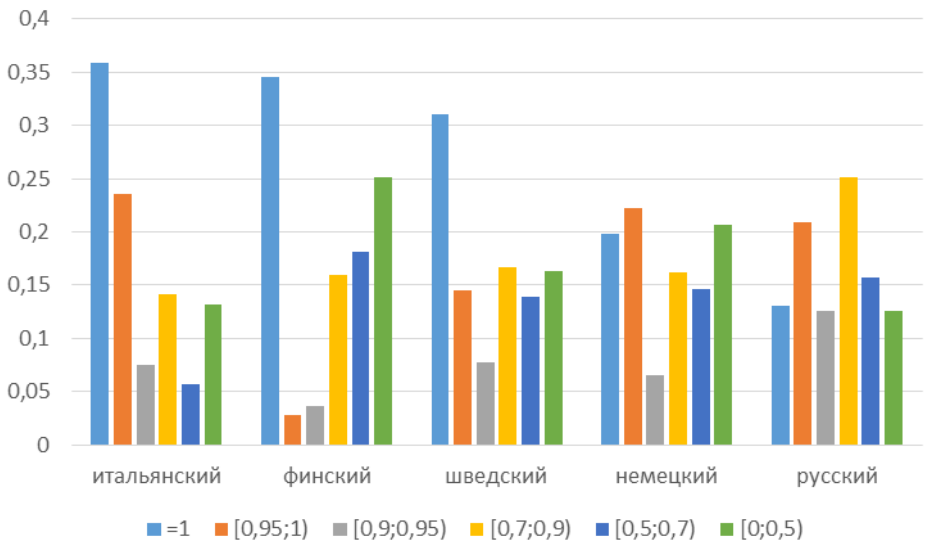

Рис. 6. Доля различных видов симметричности в корпусах

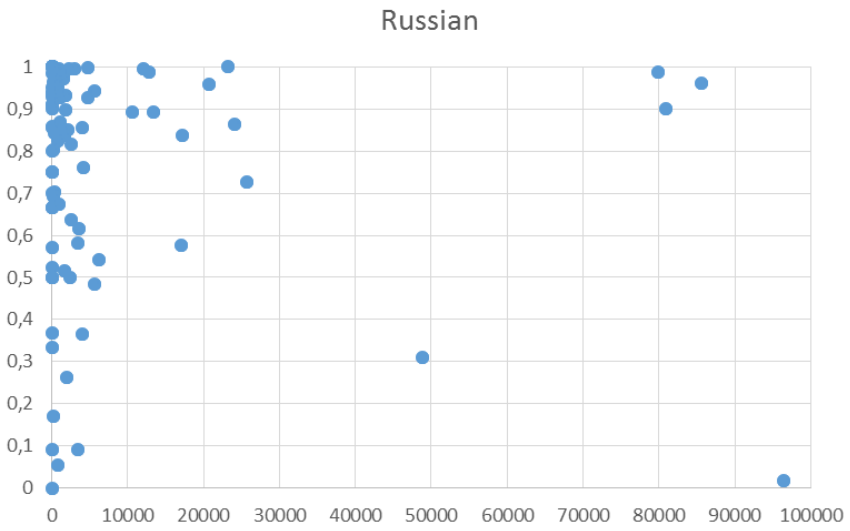

Рис. 7. Распределение троек связности для русского языка

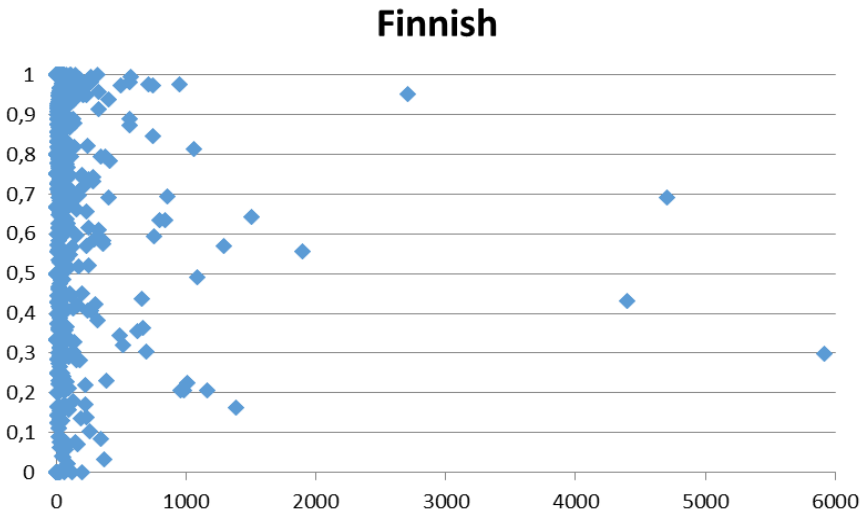

Рис. 8. Распределение троек связности для финского языка

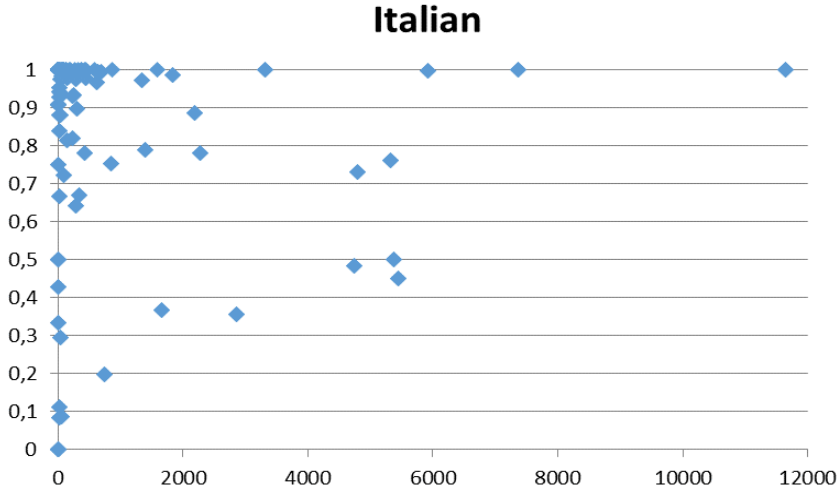

Рис. 9. Распределение троек связности для итальянского языка

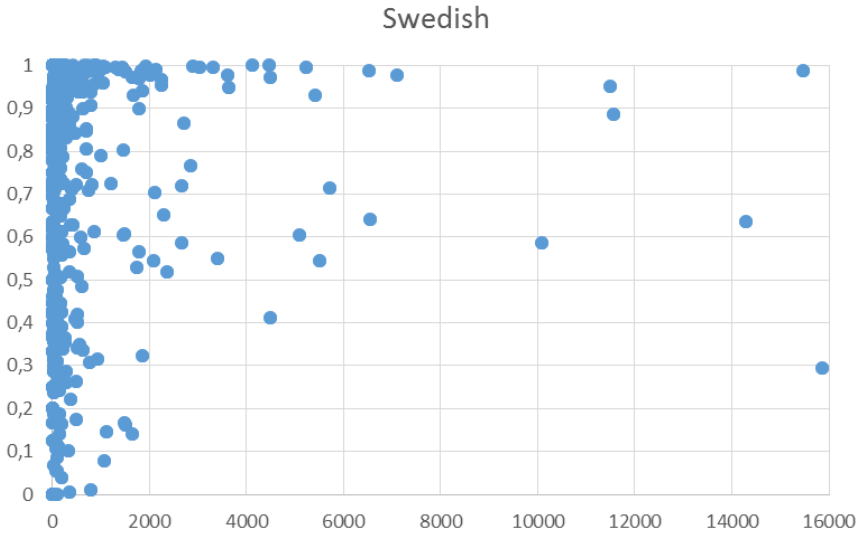

Рис. 10. Распределение троек связности для шведского языка

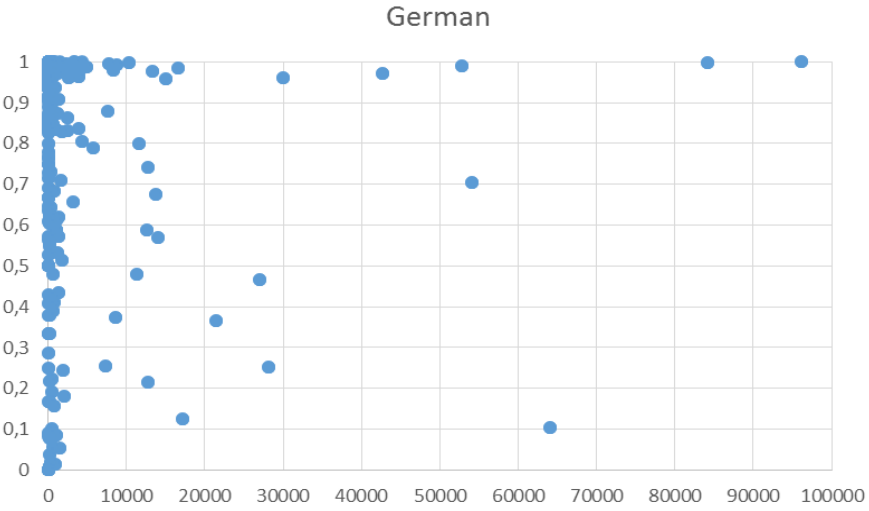

Рис. 11. Распределение троек связности для немецкого языка

В финском языке наиболее частые сочетания скорее симметричны, однако язык обладает большим количеством низкочастотных асимметричных связей. При этом наиболее частотной вновь является связь глагола и существительного, а на втором месте располагается связь глагола со знаком препинания.

Шведский и немецкий языки показывают сходное распределение при том, что немецкий можно назвать менее строгим (то есть более симметричным). В шведском языке чаще всего встречаются связи глагола с местоимением, предлога с существительным и глагола с причастием. В немецком – существительное с артиклем, существительное с предлогом и глагол с существительным.

Тройки для итальянского языка практически идеально лежат выше прямой, соединяющей точку (0,0) с точкой с максимальной встречаемостью. То есть чем чаще в итальянских текстах встречается какая-либо связь, тем больше шансов, что она будет асимметричной. К наиболее частым относятся связи артикля с

существительным (которое в рассмотренном корпусе занимает подчиненную позицию), предлога с артиклем и существительного с предлогом.

Заметим, что количество видов связей не влияет на распределение связей. Русский и итальянский языки обладают наименьшим количеством типов связей, но при этом находятся на двух различных полюсах по числу асимметричных связей.

9. Выводы

Как это видно из данных, представленных в табл. 5, русский и немецкий языки обладают большей степенью симметрии, то есть более свободным порядком слов. Наиболее асимметричным языком является итальянский, то есть он предъявляет более строгие требования к порядку следования слов. Эти данные вполне согласуются с имеющимися представлениями о синтаксической структуре языков.

С другой стороны, если сложить значения в колонках «Единственный вид связи» и «Преобладание асимметрии» (что даст нам представление о принятых по умолчанию значениях асимметрии), картина несколько изменится (см. табл. 6). Так, в финском языке мы наблюдаем либо практически полностью симметричные связи, либо связи, обладающие асимметрией. Число последних сопоставимо с аналогичным для русского языка. Отрыв же итальянского языка после сложения данных по асимметричности конструкций значительно увеличится. Аналогично увеличится и «строгость» немецкого языка.

Табл. 6. Соотношение симметричных и несимметричных связей

Язык	Асимметрия	Симметрия
русск.	34,03%	65,97%
финск.	37,35%	62,65%
итал.	59,43%	40,57%
шведск.	45,52%	54,48%
немецк.	42,11%	57,89%

Серьезным недостатком проведенного исследования является отсутствие глубокого лингвистического анализа выделенных конструкций. Помимо этого, проведенный анализ не учитывает целый ряд важных параметров. Для более строгого анализа русского языка необходимо выделение более развернутого списка частей речи. Также, например, в русском языке прилагательное в роли существительного, которому подчиняются другие существительные, всегда будет стоять в правой позиции, тогда как в большинстве остальных случаев оно будет находиться слева от главного слова: *речь лучшего в классе* vs *речь лучшего ученика в классе*. При сравнении связи глагола и существительного необходимо рассчитывать частоты встречаемости отдельно для именительного и косвенных падежей в связи с зависимостью роли существительного в предложении от его падежа. Наличие отрицательных частиц также сказывается на роли слов в предложении.

Таким образом, хотя исследование вполне корректно показало различия в исследуемых языках, в дальнейшем необходимо ввести большее число параметров при расчете статистики: лексические параметры слов, наличие зависимых слов у обоих участников связи, часть речи этих третьих слов и так далее. Это позволит лучше отличить, например, изменение стиля текста (за счет перенятой из древнегреческого языка традиции менять порядок слов при переходе на возвышенный стиль) от свободного порядка следования слов при использовании данной конструкции.

Несмотря на это, вслед за [Даль, 2009] можно предположить, что сложность синтаксиса языка определяется количеством синтаксических конструкций. Если одна и та же пара слов может быть расставлена различным образом, это означает, что либо язык располагает соответствующими правилами, либо существуют правила, оговаривающие контекст, в котором зависимое слово должно располагаться справа или слева. Таким образом, полученные значения могут использоваться как мера для оценки языковой сложности. Заметим при этом, что корреляции между распределением слов по типам морфологической и синтаксической неоднозначности нами обнаружено не было. С одной стороны, мы провели расчеты обоих видов неоднозначности всего лишь для трех языков. С другой стороны, немецкий и итальянский языки обладают примерно сходной степенью флексии, процентом слов, омонимичных по части речи, но различные распределения симметрии синтаксических связей. Мы думаем, что данный вопрос нуждается в дополнительных опытах и дополнительных исследованиях.

Список литературы

[Apresjan et al., 2006] Apresjan J., Boguslavsky I., Iomdin B., Iomdin L., Sannikov A., and Sizov V. A syntactically and semantically tagged corpus of russian: State of the art and prospects. // In Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006), May 24-26, 2006, Genoa, Italy.

[Bolshakov et al., 2002] 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, 2276, 280 – 284.

[Bosco et al., 2008] Bosco C., Mazzei A., Lombardo V., Attardi G., Corazza A., Lavelli A., Lesmo L., Satta G., Simi M. Comparing Italian parsers on a common treebank: the Evalita experience. // In Proceedings of LREC'08, pp. 2066-2073. Marrakesh, Morocco

[Brants, 2000] Brants T. TnT - a statistical part-of-speech tagger. // In Proc. of 6th Applied Natural Language Processing Conference, pp. 224 – 231. Seattle.

[Brants et al., 2004] Brants A., Dipper S., Eisenberg P., Hansen S., König E., Lezius W., Rohrer C., Smith G., and Uszkoreit H. TIGER: Linguistic Interpretation of a German Corpus. // Journal of Language and Computation, 2004 (2), pp. 597-620.

[Brill, 1995] Brill E. Unsupervised Learning of Disambiguation Rules for Part Of Speech Tagging // In Proc. of the Third Workshop on Very Large Corpora. Cambridge, Massachusetts, USA.

[Fábricz, 1986] Fábricz K. Particle Homonymy and Machine Translation. // In Proc. of International Conference on Computational Linguistics, pp. 59 – 61.

[Hajič, 1998] Hajič J., Vidová-Hladká B. Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. // In Proc. of the COLING-ACL Conference, pp. 483 – 490. Montreal, Canada.

[Haverinen et al., 2013] Haverinen K., Nyblom J., Viljanen T., Laippala V., Kohonen S., Missilä A., Ojala S., Salakoski T., Ginter F. Building the essential resources for Finnish: the Turku Dependency Treebank. // Language Resources and Evaluation. 2013.

[Krovetz, 1997] Krovetz R. Homonymy and Polysemy in Information Retrieval. // In Proc. of EACL '97 Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics archive, pp. 72 – 79.

[Li et al., 2012] Li Y., Yu Y., Fung P. A Mandarin-English Code-Switching Corpus. // In Proc. of Eighth International Conference on Language Resources and Evaluation (LREC-2012).

[Lyashevskaya, 2013] Lyashevskaya O. Frequency Dictionary of Inflectional

Paradigms: Core Russian Vocabulary. // Preprints of HSE, Series: Humanity, WP BRP 35/HUM/2013. Retrieved from http://www.hse.ru/data/2013/06/27/1285976210/35HUM2013.pdf

[Nivre et al., 2006] Nivre J., Nilsson J. and Hall J. Talbanken05: A Swedish Treebank with Phrase Structure and Dependency Annotation. // In Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006), May 24-26, 2006, Genoa, Italy.

[Oravecz et al., 2002] 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).

[Padró et al., 2012] Padró L., Stanilovsky E. FreeLing 3.0: Towards Wider Multilinguality. // In Proc. of the Language Resources and Evaluation Conference (LREC'12) ELRA. Istanbul, Turkey.

[Pertsova, 2007] Pertsova K. Learning Form-Meaning Mappings in Presence of Homonymy: a linguistically motivated model of learning inflection (PhD Thesis in Linguistics). Retrieved from http://linguistics.ucla.edu/people/grads/pertsova/pertsovaThesis.pdf

[Pinnis et al., 2011] 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). Zurich, Switzerland.

[Protopopova et al., 2013] Protopopova E.V., Bocharov V.V. Unsupervised learning of part-of-speech disambiguation rules // In Proc. of Computational Linguistics and Intellectual Technologies (Dialog 2013). Bekasovo, Russia.

[Sharoff et al., 2011] 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). Bekasovo, Russia.

[Schmid, 1995] Schmid H. Improvements in Part-of-Speech Tagging with an Application to German. // Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland.

[Tufiş, 2000] 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.

[Богуславский и др., 2002] Богуславский И.М., Иомдин Л.Л., и др. Разработка синтаксически размеченного корпуса русского языка. // Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных» // СПб, изд-во Санкт-Петербургского университета, 2002, сс. 40–50.

[Даль, 2009] Возникновение и сохранение языковой сложности / Эстен Даль; пер. с англ. Д. В. Сичинавы. - Москва : ЛКИ, 2009. - 558 с.

[Клышинский и др., 2013] 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. URL: http://library.keldysh.ru/preprint.asp?id=2013-41 .

[Копотев, 2008] Копотев М.В. К построению частотной грамматики русского языка: падежная система по корпусным данным // В Мустайоки А.А. Копотев М.В., Бирюлин Л.А., Протасова Е.Ю. (ред.) Инструментарий русистики: корпусные подходы с. 136 – 150). Хельсинки.

[Сокирко и др., 2004] Сокирко А.В., Толдова С.Ю. Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка // Сб. трудов конференции «Корпусная лингвистика-2004» – СпБ, 2004.

[Тестелец, 2002] Тестелец Я.Г. Введение в общий синтаксис / Я.Г. Тестелец. - М.: Российский государственный гуманитарный университет, 2001. - 798 с.

Оглавление