КОРПУС С АВТОМАТИЧЕСКИ СНЯТОЙ МОРФОЛОГИЧЕСКОЙ НЕОДНОЗНАЧНОСТЬЮ: К МЕТОДИКЕ ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЙ

Шаров С. А. (s.sharoff@leeds.ac.uk)[1, 5],
Беликов В. И. (vibelikov@gmail.com)[1],
Копылов Н. Ю. (Nikolay_Ko@abbyy.com)[1, 2],
Сорокин А. А. (alexey.sorokin@list.ru)[1, 3, 4],
Шаврина Т. О. (rybolos@gmail.com)[3]

[1]РГГУ, Москва, Россия; [2]ABBYY, Москва, Россия;
[3]МГУ им. М. В. Ломоносова, Москва, Россия; [4]МФТИ, Москва, Россия; [5]University of Leeds, Великобритания

В данной статье поднимается вопрос о степени доверия к корпусам с автоматически снятой омонимией, а также метриках оценки качества автоматической морфологической разметки и интерпретации полученных результатов. Мы исследуем данный вопрос на примере метрик качества морфологической разметки и автоматического снятия омонимии, применяемых в Генеральном интернет-корпусе русского языка (ГИКРЯ). При создании модели использована система морфологических меток MULTEXT-east и свободно распространяемые анализаторы TNT-Russian и mystem.

Ключевые слова: морфологическая разметка, снятие омонимии, автоматически снятая омонимия, корпусная лингвистика

CORPUS WITH AUTOMATICALLY RESOLVED MORPHOLOGICAL AMBIGUITY: TO THE METHODOLOGY OF LINGUISTIC RESEARCH

Sharoff S. A. (s.sharoff@leeds.ac.uk)[1, 5],
Belikov V. I. (vibelikov@gmail.com)[1],
Kopylov N. Yu. (Nikolay_Ko@abbyy.com)[1, 2],
Sorokin A. A. (alexey.sorokin@list.ru)[1, 3, 4],
Shavrina T. O. (rybolos@gmail.com)[3]

[1]RSUH, Moscow, Russia; [2]ABBYY, Moscow, Russia;
[3]Lomonosov Moscow State University, Moscow, Russia;
[4]MIPT, Moscow, Russia; [5]University of Leeds, UK

The present article addresses the question of the credibility of morphological annotation for the corpora with automatically resolved morphological ambiguity with respect to various metrics. We study this problem for the metrics of annotation quality and ambiguity resolution used in General Internet Corpora of Russian Language (GICR). The system of morphological annotation under consideration includes MULTEXT-east morphological tags and the freeware parsers TNT-Russian and Yandex's mystem.

Keywords: morphological tagging, morphological disambiguation, automatic morphological disambiguation, corpus linguistics

1. Введение

С развитием компьютерной лингвистики растут как объемы корпусов, так и сложность анализа полученных на них результатов. В эру интернета как корпуса мы сталкиваемся с миллиардными корпусами [1], [2], и, следовательно, на них становится возможным применение лишь автоматической разметки. Можно ли ей доверять? С одной стороны, показатели качества TNT-Russian [3] на корпусе SynTagRus достигли точности (accuracy) 97,09 % [3], что сопоставимо с точностью в 96 %, полученной при ручной редактуре автоматической морфологической разметки в ХАНКО [4]. Как отмечается в [3],[13], модели автоматической разметки, основанные на правилах, обладают важным свойством устойчивости к шуму, так как достаточно успешно отделяют более вероятное от менее вероятного. С другой стороны, ошибки, встречаемые на корпусах с автоматической и ручной разметкой, носят системный и несистемный характер соответственно, поэтому стоит относиться к результатам, полученным на первых, с осторожностью, определяемой доверительным интервалом.

При разметке Генерального интернет-корпуса русского языка мы уделяли большое внимание исходному и итоговому качеству морфологической разметки, чтобы пользователь всегда мог рассчитать вероятность успешного получения требуемого результата. С целью добиться более высокой точности мы придерживаемся подхода, утверждающего, что на корпусе объемом в несколько миллиардов слов, омонимия, должна быть скорее снята автоматически, чем оставлены все возможные вариант. В противном случае исследователь рискует получить крайне «грязную» выдачу, непригодную для дальнейшего использования. Работа с неснятой омонимией в таких корпусах остается возможной на уровне словоформ.

В данном исследовании мы хотим обратить внимание на остро стоящую проблему корпусной лингвистики: все большее количество корпусов переходит на технологии автоматической разметки и автоматически снятой омонимии, однако информация о качестве полученной разметки зачастую остается недоступной для конечного пользователя. Отсутствие необходимой информации такого рода формирует впечатление о корпусах с автоматически снятой омонимией как о неком «черном ящике», на котором проблематично получать и интерпретировать результаты исследований.

Работая с большими корпусами с автоматическим снятием омонимии, лингвист должен предусмотреть следующие вопросы:

1) Является ли сегментный состав корпуса подходящим для задач исследования? Является ли выборка репрезентативной?

Как отмечается в [5], внимании следует также уделить следующим вопросам:

2) распределению результатов по числу вхождений, документов и авторов;
3) анализу временной динамики;
4) анализу распределения результатов по типам источников (параметрам метатекстовой разметки);
5) наличию в корпусе дубликатов или иных систематических факторов, искажающих значения счётчиков.

Конечная цель работы с любым корпусом — получение доверительного результата, с однозначным пониманием того, насколько этот результат можно масштабировать на весь язык в его текущем состоянии. Для лучшего понимания методики получения таких результатов мы рекомендуем пользователям ознакомиться с результатами статистической оценки точности, полноты и некоторых других метрик качества для автоматической разметки ГИКРЯ, приводимыми в данной работе.

2. Первичная оценка качества

Изначально модуль морфологической обработки в ГИКРЯ строился следующим образом: после токенизации слова поступали на вход программе TNT-Russian [3], использующей систему морфологических меток MULTEXT-east [12], в результате словоформе сопоставлялась лемма на основе словаря или правил, а также набор грамматических меток на основе n-граммной модели, обученной на подкорпусе НКРЯ со снятой омонимией.

Из текстов корпуса была составлена контрольная выборка в 7662 токена, сбалансированная относительно состава однородных сегментов в ГИКРЯ: в нее вошли тексты из новостных ресурсов, социальных сетей, блогов, Журнального Зала. В [5] подробно обсуждается, почему мы считаем такие сегменты дифференциальными в сравнении с прочими крупными корпусами, а значит, пригодными для масштабирования результата исследования в пределах некоторых дифференциальных признаков языка, например, жанровых.

При оценке мы использовали такие метрики как точность и полнота, так как они наиболее наглядно показывают вероятность успешного поиска при обращении к той или иной части речи.

При этом полнота словаря составила 95,6 %.

Как можно заметить, наибольшие проблемы затрагивают точность определения наречия и числительного, а также полнота определения междометия. Такие категории как «остальное» и «аббревиатура» трудно назвать проблемными, так как они не были задействованы при обучении, а значит, наличествовали только гипотетически.

Таблица 1. Первичное качество автоматической частеречной разметки

Часть речи	Точность	Полнота
1. Существительное	0,992	0,987
2. Глагол	0,989	0,991
3. Прилагательное	0,950	0,998
4. Местоимение	0,997	0,997
5. Наречие	0,808	0,947
6. Предлог	1,000	1,000
7. Союз	0,979	0,996
8. Числительное	0,815	0,963
9. Частица	0,996	0,959
10. Междометие	1,000	0,714
11. Остальное	0,000	0,000
12. Аббревиатура	0,000	0,000
Микроусреднение:	**0,794**	**0,796**
Макроусреднение:	**0,976**	**0,987**
Микроусреднение без 11–12:	**0,953**	**0,955**

3. Установленные проблемы

Помимо вышеупомянутых проблем точности и полноты отдельных частей речи, первоначальная оценка показала следующие системные проблемы корпуса:

1. Неполнота словаря — около 4,4% всех встречаемых слов получали леммы не на основании словаря, а с использованием модуля, прогнозирующего лемму по суффиксу.
2. Неправильные приоритеты при морфологической разметке — словарная лемма не имела большего веса в сравнении с несловарной, в результате на выходе n-граммной модели зачастую выдавались несловарные формы, в то время как верные словарные леммы в разметку не попадали.
3. Большой процент смешения омонимичных форм — неправильное автоматическое снятие омонимии создавало достаточно большую путаницу при анализе выдачи корпуса. Так, если при разведении омонимичных падежных форм существительного ошибочные вхождения занимают до 10% всей выдачи, то при поиске субстантивов доля неправильных примеров доходила до 38%. Более всего эта проблема коснулась таких форм, как: а) именительный и винительный падеж у неодушевленных существительных б) родительный и винительный падеж у одушевленных существительных в) краткие прилагательные, наречия и предикативы г) причастия и полные прилагательные д) субстантивы и прилагательные.

По данным направлениям (1–3) был проведен тщательный анализ и перестроение существующего модуля, не затрагивающие переобучение программ, но стремящиеся обеспечить значимый рост качества.

4. Методика работы

4.1. Неполнота словаря

Проблемы больших словарей подробно рассмотрены в исследованиях [7], [8], и словарь TNT-Russian, используемый в ГИКРЯ[1], не стал положительным исключением: несмотря на объеме более 7 млн вхождений, на сегментах социальных медиа он по-прежнему оказывался неполным, заставляя программу угадывать лемму для уже вошедших в обиход слов.

Автособираемые словари отличаются от собранных вручную следующим важным свойством: они хорошо отражают вариативность языка, но могут недостаточно хорошо покрывать наиболее частотную часть словаря. Поэтому было решено пополнить словарь ГИКРЯ за счет какого-либо точного словаря. Выбор пал на словарь свободно распространяемого морфологического анализатора mystem [6].

Все несловарные слова из корпуса были пропущены через mystem, и его вывод был с некоторыми потерями перекодирован в форму MULTEXT-east. Потери касались таких категорий, как разряды предлогов, частиц, союзов и междометий, которых нет в mystem. Были проигнорированы случаи с метками «часть композита», «аббревиатура», «топоним» и т.д. за их отсутствием в нашей кодировке. Таким образом было переведено 97% полученных вхождений.

Возникшие трудности:

1) Нам пришлось пожертвовать некоторым количеством граммем из кодировки MULTEXT-east (разряды предлогов, союзов, частиц, междометий).
2) 6 тысяч словарных вхождений пришлось исправить регулярными выражениями и вручную из-за замены е→ё, происходящей в mystem.

Итог: полнота словаря достигла 97–98,5% в зависимости от сегмента — 98,5% на сегментах с языком, наиболее близким к стандартному русскому (Новости), до 97% на сегментах с наибольшей вариативностью (блоги, микроблоги). Состав и статистику по сегментам ГИКРЯ см. в Приложении (таблицы 15–17).

4.2. Перестроение модуля морфологической разметки

Порядок присвоение леммы был изменен: если ранее лемма определялась уже после того, как n-граммная модель определяла наиболее вероятную грамматическую метку, то теперь, сначала производится порождение всех возможных лемм для слова, и лишь затем из них выбирается наиболее подходящая на основе n-граммной модели и вхождения в словарь. Таким образом удалось исключить частотные ошибки, при которых несловарная форма побеждала словарную.

---

[1] Словарь был автоматически собран на подкорпусе НКРЯ со снятой омонимией при разработке программы tnt-russian.

4.3. Улучшение качества автоматического снятия омонимии

Если исключить переобучение модели, у лингвиста остается не так много методов воздействия на качество снятия омонимии. В нашем исследовании самыми «шумными» категориями оказались прилагательные и причастия, а также краткие прилагательные и наречия, которые к тому же включали в себя предикативы, для которых отдельной категории выделено в исходной кодировке не было.

В результате были реализованы следующие решения:

А) все причастия были переведены в прилагательные. Такое радикальное объединение двух плохо разводимых категорий позволило обеспечить новообразованной категории качественную выдачу (полнота 0,997; точность 0,953).

Возникшие трудности:

1) всякое причастие имеет угаданную на основе правил лемму, так как предусмотреть все возможные причастия в языке нельзя;
2) причастия больше не имеют грамматической метки вида, времени, залога.

Б) Была введена категория предикативов, что позволило, отделив предикативы, точнее развести прилагательные и наречия. В результате точность определения наречий выросла с 80 % до 91 %.

4.4. Прочие работы

Как было показано в п.3, остались еще некоторые проблемы качества определения конкретных частей речи, которые были решены при помощи ручной обработки словарей:

1) были упразднены категории «аббревиатура» и «остальное», поскольку ни один токен в корпусе не получил подобной метки;
2) мы не были удовлетворены качеством определения междометий, поэтому внесли их в список неделимых вхождений токенизатора, что позволило безошибочно определять их по словарю: полнота определения междометий выросла с 71 % до 90 %;
3) качество определения числительных страдало из-за их смешения с наречиями типа «много», «мало» на уровне словаря. Такие вхождения были вычищены вручную, что позволило точности увеличиться с 82 % до 98 %.

5. Краткие итоги

В результате проведенной работы мы имеем несколько модифицированный набор кодировок MULTEXT-east и пополненный словарь, характеризующиеся лучшим покрытием современного русского языка, чем исходно применимая система TNT-Russian .

Итоги сравнения качества частеречной разметки ГИКРЯ представлены в таблице 2:

Таблица 2. Сравнение качества автоматической морфологической разметки ГИКРЯ до и после преобразований

Часть речи	**Точность**	**Полнота**	**Часть речи**	**Точность**	**Полнота**
1. Существительное	0,992	0,987	1. Существительное	0,997	0,990
2. Глагол	0,989	0,991	2. Глагол	0,998	0,998
3. Прилагательное	0,950	0,943	3. Прилагательное	0,953	0,997
4. Местоимение	0,997	0,997	4. Местоимение	1,000	1,000
5. Наречие	0,808	0,947	5. Наречие	0,974	0,913
6. Предлог	1,000	1,000	6. Предлог	1,000	0,998
7. Союз	0,979	0,996	7. Союз	0,993	0,993
8. Числительное	0,815	0,963	8. Числительное	0,980	1,000
9. Частица	0,996	0,959	9. Частица	0,996	0,996
10. Междометие	1,000	0,714	10. Междометие	1,000	0,900
11. Остальное	0,000	0,000	11. Предикатив	1,000	0,810
12. Аббревиатура	0,000	0,000			
Микроусреднение:	**0,794**	**0,796**	**Микроусреднение:**	**0,990**	**0,963**
Макроусреднение:	**0,976**	**0,987**	**Макроусреднение:**	**0,991**	**0,990**
Макроусреднение без 11–12:	**0,953**	**0,950**	**Словарь:**		**0,970**
Словарь:		**0,956**	**Словарь без учета причастий:**		**0,989**

Таблица 2 иллюстрирует примечательное улучшение качества частеречной разметки прежде проблемных категорий: 1) за счет перемещения причастий из категории глагола в категорию прилагательных уровень ошибок для обеих категорий заметно снизился (у глаголов точность возросла с 98% до 99%, у прилагательных полнота выросла с 94% до 99% 2) введение категории предикативов уменьшило уровень ошибок у наречий и прилагательных.

Важным примечанием будет то, что ГИКРЯ на сегодняшний момент является не единственным корпусом, где полностью автоматическая разметка получена при помощи использования TNT-Russian. Интернет-корпус русского языка университета Лидс использует этот морфологический анализатор в его исходном виде [3], поэтому качество разметки не отличается от исходной в ГИКРЯ; зато корпус ruTenTen проекта SketchEngine, собравший около 15 млрд слов в том числе из рунета, использует для своей разметки

результат работы двух конкурирующих анализаторов: Tree-Tagger[2] RF-Tagger[3] [2].

Была проведена сравнительная оценка качества разметки в корпусах ГИКРЯ и ruTenTen: использовавшаяся ранее для проверки ГИКРЯ выборка была загружена и обработана в SketchEngine; полученные метрики приведены в таблице 3:

Таблица 3. Сравнительная оценка качества автоматической частеречной разметки ruTenTen и ГИКРЯ

RuTenTen			**ГИКРЯ**		
Часть речи	**Точность**	**Полнота**	**Часть речи**	**Точность**	**Полнота**
1. Существительное	0,948	0,987	1. Существительное	0,997	0,990
2. Глагол	0,966	0,976	2. Глагол	0,998	0,998
3. Прилагательное	0,942	0,969	3. Прилагательное	0,953	0,997
4. Местоимение	0,988	0,975	4. Местоимение	1,000	1,000
5. Наречие	0,927	0,914	5. Наречие	0,974	0,913
6. Предлог	1,000	0,997	6. Предлог	1,000	0,998
7. Союз	0,993	0,991	7. Союз	0,993	0,993
8. Числительное	0,797	0,911	8. Числительное	0,980	1,000
9. Частица	0,986	0,983	9. Частица	0,996	0,996
10. Междометие	1,000	0,551	10. Междометие	1,000	0,900
11. Остальное	0	0	11. Предикатив	1,000	0,810
12. Аббревиатура	0	0			
Микроусреднение:	**0,979**	**0,975**	**Микроусреднение:**	**0,990**	**0,963**
Макроусреднение:	**0,796**	**0,771**			
Макроусреднение без 11–12:	**0,955**	**0,925**	**Макроусреднение:**	**0,991**	**0,990**

Как видно из таблицы, среднее качество автоматической разметки в ГИКРЯ выше и по полноте, и по точности, что можно объяснить как разницей алгоритмов, так и качеством словарей: в ruTenTen для морфологической разметки с автоматически снятой омонимией используется комбинация двух n-граммных анализаторов: RF-Tagger и Tree-Tagger — для каждого слова выбирается наиболее вероятный вариант из предоставленных данными программами; такой подход, безусловно, позволяет снизить эффект статистических ошибок каждого из анализаторов, однако качество обеих программ находится примерно на одном уровне несколько ниже качества TnT-Russian.

---

[2] Tree-Tagger является по сравнению с TnT-Russian, «голой» n-граммной моделью, которая потом была обучена на корпусе со снятой омонимией НКРЯ, чтобы получить сам TnT [10].

[3] RF-Tagger основан на аналогичной статистической модели, более подробно см. [11]

5.1. Другие оценки качества

5.1.1. Качество снятия омонимии по другим граммемам кодировки

Помимо более общих оценок по частям речи, были проведены более подробные исследования по оценке качества граммем более низкого порядка. Приведем таблицу для граммем некоторых существительного:

Таблица 4. Оценка качества автоматической морфологической разметки существительного

	точность	**полнота**	**f-мера**
нарицательные	0,9891	0,9946	0,9919
собственные	0,9033	0,8219	0,8607
Муж.р	0,9902	0,9922	0,9912
Ж.р.	0,9967	0,9931	0,9949
с.р.	0,9891	0,9979	0,9935
о.р.	0,2941	0,1923	0,2326
ед.ч	0,9978	0,9987	0,9982
мн.ч	0,9962	0,9939	0,9951

Примечание: по таблице 4 заметно плохое качество определения существительных общего рода. Однако стоит отметить, что частота их встречаемости в тексте (в среднем по корпусу 0,0006) гарантирует незначительность вклада общего рода в процент ошибок после макроусреднения.

Качество таких граммем, как род, число, падеж, залог и т. д. может быть ниже заявленного качества определения самой части речи, и пользователю рекомендуется осознавать, что с увеличением количества граммем в запросе вероятность получить идеальную выдачу будет снижаться.

5.1.2. Оценка качества автоматического снятия омонимии

Автоматическое снятие омонимии остается одной из «ахиллесовых пят» автоматической разметки: качество разведения отдельных категорий, таких как отмеченные в п. 3.3, находится в среднем на уровне 70–80%.

Для оценки снятия омонимии использовалась выборка из 30 тыс. слов, в результате которой были получены следующие данные: см. таблицу 5.

Таблица 5. Точность автоматического снятия омонимии на некоторых граммемах имени существительного

Граммема:	**Точность**
Неодушевленный номинатив	0,792
Неодушевленный аккузатив	0,858
Одушевленный аккузатив	0,661
Одушевленный генитив	0,890
Субстантивы (на выборке из омонимичных вхождений)	0,680
Прилагательные (на выборке из омонимичных вхождений)	0,900

Точность по отдельно взятым падежам так же оставляет желать лучшего: частности, где аккузатива у неодушевленных существительных она составляет приблизительно 86%, а 14% приходится на ошибочно определенный номинатив. При этом тот же аккузатив, но у одушевленных, определяется правильно в 66% случаем в выдаче, а остальную долю занимают генитив (26% выдачи), номинатив (5%), локатив (2%) и прочие случаи (1%). Графики с процентным соотношением смешений по падежам см. в «Приложении» (таблицы 6–9).

Может ли такой уровень качества быть достаточным для конкретных исследовательских задач или нет, он должен быть эксплицитно прописан разработчиками, чтобы не создавать путаницы между качеством автоматической частеречной разметки (которое обычно достаточно высоко и колеблется в пределах 97–99%), и всей автоматической морфологической разметки (разброс качества гораздо больше — от 60 до 99% в среднем).

5.1.3. Оценка качества определения согласованных групп с помощью mystem

Нами была проведена также автоматическая оценка качества определения согласованных именных групп (существительное + прилагательное), предложных групп (предлог + существительное, предлог + прилагательное) по биграммам и триграммам (предлог + прилагательное + существительное). Для каждой биграммы вида предлог + прилагательное, предлог + существительное и прилагательное + существительное (с учётом порядковых числительных и местоименных существительных) сравнивались падежи у членов биграммы. Для триграмм сравнивалось совпадение троек падежа, числа и рода. В случае несовпадения какой-либо граммемы у пары или тройки проводилась проверка по mystem.

Замер по биграммам и триграммам из 30-тысячной выборки показал, что доля совпадения падежей у групп «прилагательное — существительное» = 48,7%, а доля совпадения падежа у групп «предлог-существительное», «предлог — прилагательное — существительное» = 99,6%.

Более подробную выкладку см. в «Приложении» (таблицы 13–14).

6. Открытые вопросы

Одной из возможностей для решения проблем с выбором падежа существительного или разведения прилагательных и субстантивов (см. ошибки в Таблице 5) является использование полноценного синтаксического анализатора. Традиционная модель анализа подразумевает последовательность, начинающуюся с лучшего анализа нижнего уровня, например, выбора части речи, и применение к его результату анализа более высокого уровня, например, синтаксического анализатора. Подобно тому, как был объединен процесс частеречного анализа и лемматизации, мы планируем объединить процесс частеречного анализа и синтаксического разбора. Есть несколько моделей для такого объединения, например, описанные в [13].

Нерешенной так же остается проблема улучшения качества некоторых граммем. Нередко возникает ситуация, приводящая к распространенному заблуждению, когда метрики качества частеречной разметки понимаются как усредненные метрики качества всех граммем по частям речи. Упомянутые в [14] 97% качества нередко принимаются пользователем безо всякой проверки, а между тем, 97% качества определения, например, категории существительных, вовсе не противоречит качеству определения субстантивов мужского рода единственного числа в 60%. Скрупулезное и однозначное прописывание всех метрик качества — необходимый аспект работы с современными корпусами, который способен обнажить как систематические недостатки автоматической морфологической разметки, так и необходимость переработки методологии исследования.

Благодарность

Исследовательский коллектив выражает свою благодарность за важную методологическую помощь в подготовке материала Владимиру Павловичу Селегею.

Литература

1. *Schäfer R.* (2015) FYI: COW: Free, Large Web Corpora in European Languages. LINGUIST List 26.2114 (web resource: http://linguistlist.org/issues/26/26-2114.html)
2. *Kilgarriff A., Baisa V., Bušta J., Jakubíček M., Kovář V., Michelfeit J., Rychlý P., Suchomel V.* (2014). “The Sketch Engine: ten years on”. Lexicography (Springer Berlin Heidelberg) 1 (1): 7–36.
3. *Sharoff S. and Nivre J.* (2011), The proper place of men and machines in language technology: Processing {Russian} without any linguistic knowledge. In Proc. Dialogue, Russian International Conference on Computational Linguistics, Bekasovo.
4. *Копотев М. В.* (2008), К построению частотной грамматики русского языка: падежная система по корпусным данным. Инструментарий русистики: корпусные подходы, Хельсинки
5. *Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S.* (2013), Corpus as language: from scalability to register variation. In Proc. Dialogue, Russian International Conference on Computational Linguistics, Bekasovo.
6. *Segalovich I.* (2003), A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine
7. *Peterson, J. L.* (1986), A note on undetected typing errors. Commun. ACM 29, 7 (July), 633–637.
8. *Kukich K.* (1992), Techniques for Automatically Correcting Words in Text. ACM Computing Surveys, Vol. 24, No. 4.

9. *Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S.* (2013), Big and diverse is beautiful: A large corpus of Russian to study linguistic variation. In Proc. Web as Corpus Workshop (WAC-8).
10. *Schmid H.* (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.
11. *Schmid H., Laws F.* (2008), Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging, COLING 2008, Manchester, Great Britain.
12. *Dimitrova, L., T. Erjavec, N. Ide, H.-J. Kaalep, V. Petkevič, and D. Tufiş* (1998), MULTEXT-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In COLING-ACL '98. Montreal, Quebec, Canada.
13. *Finkel J. R., Manning C. D., Ng A. Y.* (2006), Solving the Problem of Cascading Errors: Approximate Bayesian Inference for Linguistic Annotation Pipelines. Computer Science Department Stanford University Stanford, CA 94305
14. *Manning C. D.* (2011), Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? In CICLing Conference on Intelligent Text Processing and Computational Linguistics

Приложение

Таблица 6. Смешение номинатива у неодушевленных

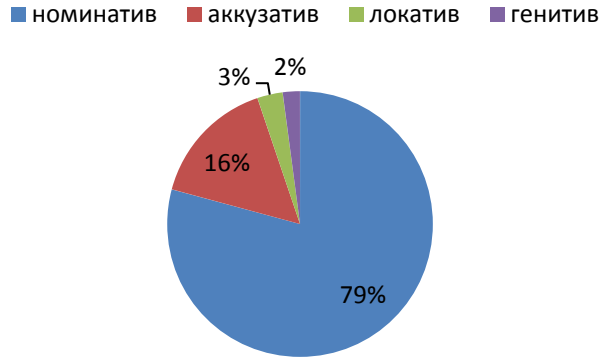

Таблица 7. Смешение аккузатива с другими падежами у неодушевленных

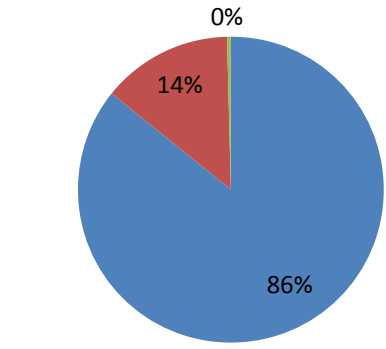

Таблица 8. Смешение аккузатива с другими падежами (у одушевленных)

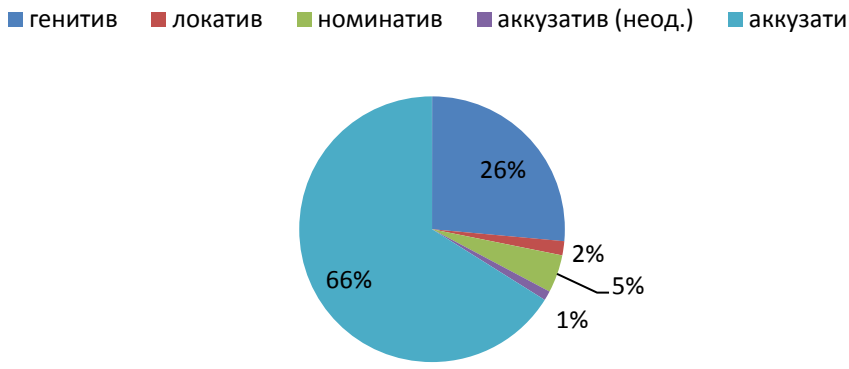

Таблица 9. Смешение генитива с другими падежами (у одушевленных)

Таблица 10. Смешение генитива с другими падежами (у одушевленных)

Таблица 11. Смешение прилагательных с омонимичными субстантивами

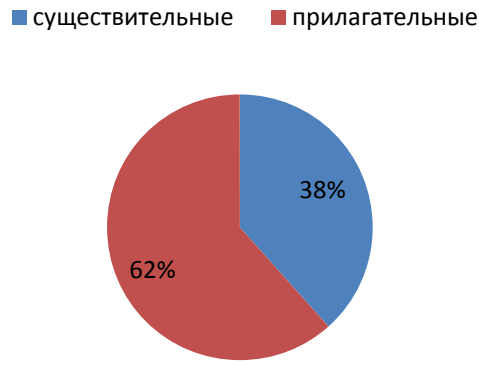

Таблица 12. Смешение субстантивов с омонимичными прилагательными

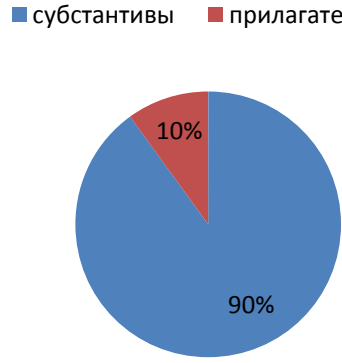

Таблица 13. Совпадение граммем в парах прилагательное-существительное

				прилагательное						существительное					
	количество	совпадение	процент	падеж	число		род			падеж	число		род		
ap-	2,2	0	0	0,2	11,1	2,2	100	0	0	0,2	11,1	2	88,9	0,3	14,8
apf	63,4	13	20,5	20,5	32,3	63,4	100	13	20,5	27,2	42,8	48,2	76	63,4	100
apm	45	31	68,89	31,5	70	45	100	45	100	43	95,6	45	100	43,5	96,7
apn	21,8	4	18,39	5,2	24,1	21,8	100	4	18,4	6,2	28,7	17,8	81,6	21,8	100
asf	123	119	96,75	123	100	123	100	123	100	119	96,8	123	100	123	100
asm	153,5	104	67,75	139,5	90,9	154	100	148	96,1	114	74,3	154	100	154	100
asn	74,5	53	71,14	61	81,9	74,5	100	74,5	100	66,5	89,3	74,5	100	74,5	100
dp-	0,7	0	0	0,3	50	0	0	0	0	0,7	100	0,7	100	0,7	100
dpf	12,8	2	15,58	11,3	88,3	7,3	57,14	2	15,6	12,8	100	12,8	100	12,3	96,1
dpm	22,8	13	56,93	15,3	67,2	13,3	58,39	22,8	100	22,8	100	22,8	100	22,5	98,5
dpn	7,3	2	27,27	6,3	86,4	4,3	59,09	3	40,9	7,3	100	7,3	100	7	95,5
dsf	58,8	32	54,39	36,3	61,8	58,8	100	57,5	97,7	41,8	71,1	58,2	98,9	58,8	100
dsm	34	34	100	34	100	34	100	34	100	34	100	34	100	34	100
dsn	18	8	44,44	18	100	18	100	8	44,4	18	100	18	100	18	100
gp-	3,1	0	0	3,1	100	3,1	100	0	0	2,8	91,9	3,1	100	0,3	10,8
gpf	141,5	33	23,32	131,5	92,9	142	100	33	23,3	140,5	99,3	142	100	142	100
gpm	129,8	111	85,49	119,8	92,3	130	100	129	99,2	122,3	94,2	130	100	129	99
gpn	61,8	6	9,7	58,8	95,2	61,8	100	6	9,7	61,8	100	61,8	100	60,8	98,4
gsf	231,3	199	86,02	211,3	91,4	231	100	231	99,9	213,3	92,2	224	96,7	227	98,3
gsm	170,5	108	63,34	162,5	95,3	171	100	119	69,5	168	98,5	171	100	170	99,4
gsn	94	71	75,53	92	97,9	94	100	74	78,7	89	94,7	90	95,7	92	97,9

				прилагательное						существительное					
	количество	совпадение	процент	падеж	число		род			падеж	число		род		
ip-	2	0	0	2	100	2	100	0	0	2	100	2	100	0	0
ipc	0,5	0	0	0,5	100	0,5	100	0	0	0,5	100	0,5	100	0,5	100
ipf	29,2	1	3,43	29,2	100	29,2	100	1	3,43	29,2	100	29,2	100	26,7	91,4
ipm	46,2	44	95,31	46,2	100	46,2	100	46,2	100	46,2	100	46,2	100	44	95,3
ipn	16,2	5	30,93	16,2	100	16,2	100	5	30,9	16,2	100	16,2	100	15,5	95,9
isf	89	68	76,4	68	76,4	89	100	88	98,9	89	100	89	100	89	100
ism	69,7	62	89	68,3	98,1	69,7	100	63,3	90,9	69	99	69	99	69	99
isn	33,7	15	44,55	30,3	90,1	31,7	94,06	15,3	45,5	33	98	33	98	33	98
lp-	0,6	0	0	0	0	0,6	100	0	0	0,2	42,9	0,6	100	0,3	57,1
lpf	18,5	5	27,03	15,5	83,8	18,5	100	5	27	18,5	100	18,5	100	18,5	100
lpm	26,8	26	97,2	26,5	99,1	26,8	100	26,8	100	26,8	100	26,8	100	26,2	98,1
lpn	10	4	40	8	80	10	100	4	40	10	100	10	100	10	100
lsf	77,8	51	65,52	56,3	72,4	77,8	100	76,5	98,3	60	77,1	77,2	99,1	77,8	100
lsm	60,5	48	79,34	60,5	100	60,5	100	54	89,3	54,5	90,1	60,5	100	58,5	96,7
lsn	25,5	16	62,75	25,5	100	25,5	100	16,5	64,7	25,5	100	25,5	100	25	98
np-	2,7	1	37,5	2,4	90,6	2,7	100	1	37,5	2,2	81,3	2,4	90,6	1	37,5
npf	59,9	16	26,7	54,4	90,8	59,9	100	16	26,7	35,5	59,3	47,7	79,6	59,9	100
npm	105,2	97	92,23	104,7	99,5	105	100	105	99,7	99	94,1	104	99,1	102	97
npn	20,8	4	19,28	19,5	94	20,8	100	4	19,3	15,5	74,7	16,8	80,7	20,8	100
nsf	146	136	93,15	146	100	146	100	146	100	136	93,2	146	100	146	100
nsm	128	68	53,12	75,5	59	128	100	123	96,1	119,5	93,4	128	100	128	100
nsn	97,5	75	76,92	89,5	91,8	97,5	100	97,5	100	82	84,1	97,5	100	97,5	100
		Среднее:	**48,7393**	**51,777**	**82,7**	**58,5**	**94,62**	**47**	**57,7**	**53,0581**	**87,6**	**57,8**	**97,3**	**58,2**	**90**

Таблица 14. Совпадение граммем в парах предлог-существительное

Предлог	Кол-во биграмм с ним	% совпадения падежа	доля ошибочных примеров по mystem	% ошибочных примеров по mystem	доля ошибочных примеров	% ошибочных примеров					
против	1	100	0	0	0	0	1	0	0	0	0	0
без	12	100	0	0	0	0	2	0	0	10	0	0
благодаря	2	100	0	0	0	0	0	0	0	2	0	0
в	768	99,1	6	0,8	1	0,13	240	1	0	528	5	1
ввиду	1	100	0	0	0	0	1	0	0	0	0	0
вдоль	1	100	0	0	0	0	0	0	0	1	0	0
включая	2	100	0	0	0	0	1	0	0	1	0	0
вместо	2	100	0	0	0	0	0	0	0	2	0	0
вне	2	100	0	0	0	0	0	0	0	2	0	0
внутри	9	100	0	0	0	0	1	0	0	8	0	0
во	32	100	0	0	0	0	3	0	0	29	0	0
возле	1	100	0	0	0	0	1	0	0	0	0	0
вокруг	3	100	0	0	0	0	0	0	0	3	0	0
вопреки	1	100	0	0	0	0	1	0	0	0	0	0
вроде	2	100	0	0	0	0	0	0	0	2	0	0
вследствие	1	100	0	0	0	0	1	0	0	0	0	0
для	80	98,8	1	1,2	0	0	16	0	0	64	1	0
до	44	100	0	0	0	0	16	0	0	28	0	0

Предлог	Кол-во биграмм с ним	% совпадения падежа	доля ошибочных примеров по mystem	% ошибочных примеров по mystem	доля ошибочных примеров	% ошибочных примеров					
за	89	97,8	0	0	2	2,2	21	0	1	68	0	1
из	99	99	1	1	0	0	33	0	0	66	1	0
из-за	8	100	0	0	0	0	1	0	0	7	0	0
изо	1	100	0	0	0	0	1	0	0	0	0	0
к	152	100	0	0	0	0	43	0	0	109	0	0
ко	8	100	0	0	0	0	0	0	0	8	0	0
кроме	3	100	0	0	0	0	3	0	0	0	0	0
между	51	100	0	0	0	0	16	0	0	35	0	0
мимо	2	100	0	0	0	0	0	0	0	2	0	0
на	367	98,7	4	1,1	1	0,27	103	0	0	264	4	1
над	10	100	0	0	0	0	4	0	0	6	0	0
надо	1	100	0	0	0	0	0	0	0	1	0	0
о	72	98,6	1	1,4	0	0	15	0	0	57	1	0
об	11	100	0	0	0	0	1	0	0	10	0	0
обо	1	100	0	0	0	0	1	0	0	0	0	0
около	2	100	0	0	0	0	1	0	0	1	0	0
от	86	100	0	0	0	0	33	0	0	53	0	0
перед	15	100	0	0	0	0	3	0	0	12	0	0
передо	2	100	0	0	0	0	0	0	0	2	0	0
по	120	98,4	1	0,8	1	0,82	48	1	0	72	0	1

Предлог	Кол-во биграмм с ним	% совпадения падежа	доля ошибочных примеров по mystem	% ошибочных примеров по mystem	доля ошибочных примеров	% ошибочных примеров						
под	35	97,2	0	0	1	2,78	4	0	0	31	0	1
помимо	2	100	0	0	0	0	1	0	0	1	0	0
после	42	100	0	0	0	0	11	0	0	31	0	0
посреди	1	100	0	0	0	0	0	0	0	1	0	0
посредством	1	100	0	0	0	0	0	0	0	1	0	0
прежде	12	100	0	0	0	0	0	0	0	12	0	0
при	33	100	0	0	0	0	6	0	0	27	0	0
про	17	100	0	0	0	0	6	0	0	11	0	0
против	8	100	0	0	0	0	6	0	0	2	0	0
путём	2	100	0	0	0	0	0	0	0	2	0	0
ради	3	100	0	0	0	0	2	0	0	1	0	0
с	283	99	2	0,7	1	0,35	103	1	0	180	1	1
сквозь	2	100	0	0	0	0	0	0	0	2	0	0
со	35	100	0	0	0	0	17	0	0	18	0	0
спустя	5	83,3	0	0	1	16,7	3	0	0	2	0	1
среди	2	100	0	0	0	0	2	0	0	0	0	0
у	86	100	0	0	0	0	9	0	0	77	0	0
через	16	100	0	0	0	0	7	0	0	9	0	0
Итог:	**2649**	**99,4625**	**0,28571**	**0,125**	**0,14286**	**0,41518**	**14,0714**	**0,05357**	**0,01786**	**33,2321**	**0,23214**	**0,125**

Таблица 15. Распределение документов по сегментам ГИКРЯ

	2000	2001	2002	2003	2004	2005	2006	2007
Лента	16289	22219	22742	21885	24947	31939	37014	36160
Регнум	0	0	5588	17395	37383	55131	106416	73379
Риа	0	0	0	0	107470	86884	66683	77471
Росбалт	260	2707	3954	13146	13303	12247	24378	47704
БМР	0	0	0	0	0	3684	151396	368233
ЖЖ	1703	2313	47919	241014	774369	1861157	3431508	4265158
ВК старый	0	0	0	0	0	0	0	0
ЖЗ	1838	2190	2742	3267	3409	3917	4032	3845
	2008	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**
Лента	50308	50517	42809	42545	43986	22820	0	0
Регнум	115853	50413	63411	61949	69347	90365	0	0
Риа	99735	111586	122082	132389	124124	127228	0	0
Росбалт	77734	83597	98023	116939	130282	69034	0	0
БМР	428747	689814	843903	970300	1231772	1179789	0	
БМР НМ	412183	657109	792761	903198	1131433	1071341	0	0
ЖЖ	3951446	4893123	6626766	7171622	6429387	8155294	0	0
ВК старый	0	0	0	0	0	0	61295938	0
ЖЗ	4125	4105	3955	4146	4284	4072	825	0

Таблица 16. Объем сегментов ГИКРЯ, млн слов

Объем сегментов:	**млн слов:**
Вконтакте	10000
Блоги Мейлру	707
Живой Журнал	9556
Новости	626
Журнальный Зал	313

Таблица 17. Распределение по полу на сегментах ВКонтакте и Журнальный Зал

Вк	**всего док-тов:**	**%**
Женщины	30854835	50,34
Мужчины	18002868	29,37
NA	12438235	20,29
Всего:	61295938	100

ЖЗ	**всего док-тов:**	**%**
Женщины	13202	28,60
Мужчины	32964	71,40
NA	0	0
Всего:	46166	100