Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2016»

Москва, 1–4 июня 2016

# АВТОМАТИЧЕСКАЯ МОРФОРАЗМЕТКА КОРПУСОВ РУССКОЯЗЫЧНЫХ СОЦИАЛЬНЫХ МЕДИА: ОБУЧЕНИЕ И ОЦЕНКА КАЧЕСТВА

**Селегей Д.** (danila-slg@yandex.ru)[1],
**Шаврина Т.** (rybolos@gmail.com)[1],
**Селегей В.** (Vladimir_S@abbyy.com)[2,3],
**Шаров С.** (s.sharoff@leeds.ac.uk)[4]

[1]Московский Государственный Университет, Россия
[2]Российский государственный гуманитарный университет, Россия
[3]ABBYY, Россия
[4]Университет Лидса, Великобритания

В статье описывается новый комплекс базовых средств для морфоразметки текстов русскоязычных social media, разработанный в рамках проекта создания мегакорпуса русскоязычного интернета ГИКРЯ. Этот комплекс включает новый tagset, полученный некоторым расширением и адаптаций тагсета, предложенного в Шаров et al., и тестового корпуса (золотого стандарта) современных social media с такой разметкой объемом около 2 млн токенов.

Тагсет, созданный с учетом самых популярных текущих тагсетов РЯ (MULTEXT-East, NLC, разметкой mystem) и сохраняющий совместимость с этими форматами, будет использован для морфоразметки корпуса ГИКРЯ.

Особенностью подхода к применению нового стандарта является полностью автоматический способ разметки тестовых и обучающих корпусов: для этой цели мы использовали результаты синтаксического разбора текстов разных разделов social media с помощью парсера Compreno, используемого по лицензии Compreno Based Research (CBR), с последующим уточнением разметки за счет автоматических процедур коррекции систематических ошибок, возникающих при анализе текстов этого сегмента. Показано, что имеющиеся парсеры (в частности, tnt) показывают при обучении на таком корпусе лучшие результаты для этого сегмента, чем при обучении на других ранее имевшихся в наличии размеченных корпусах (прежде всего, т. н. «снятника» НКРЯ).

Мы полагаем, что одним из наиболее полезных результатов данной работы является появившаяся возможность объективного сравнения результатов работы POS-тегеров на сегменте social media с помощью нового тестового корпуса, который размещен на сайте ГИКРЯ.

**Ключевые слова:** автоматическая морфоразметка, морфологическая разметка, морфотэггеры для русского языка, язык социальных медиа, обучение морфопарсера

# AUTOMATIC MORPHOLOGICAL TAGGING OF RUSSIAN SOCIAL MEDIA CORPORA: TRAINING AND TESTING

**Selegey D.** (danila-slg@yandex.ru)[1],
**Shavrina T.** (rybolos@gmail.com)[1],
**Selegey V.** (Vladimir_S@abbyy.com)[2,3],
**Sharoff S.** (s.sharoff@leeds.ac.uk)[4]

[1]Moscow State University, Russia
[2]Russian State University for the Humanities, Russia
[3]ABBYY, Russia
[4]University of Leeds, UK

This paper presents a new set of basic tools for morphosyntactic tagging of Russian texts coming from social media. This has been developed within GICR, a project for creating a very large corpus of the Russian-speaking Internet.

This toolset includes a new tagset, obtained via extending and adaptating the tagset proposed by Sharoff et al. It has been tested on a gold standard test corpus of modern social media of about 2 million tokens. A particular feature of our approach is a fully automated process for development of training corpora. Instead of manual annotation we started with the output of the syntactic parser of Compreno. This annotation has been subsequently improved by automatic correction of systematic errors detected through processing of texts from social media. In this paper we show that existing tagging tools (in particular, tnt) produce consistently better results if they are trained with our corpus rather with other available corpora, in particular, those using the disambiguated portion of the Russian National Corpus.

The resulting test corpus is available in open access.

**Keywords:** automatic morphotagging, morphological tagging, morphotagging for Russian, language of social media

## 1. Введение

Автоматическая морфоразметка корпусов стала сегодня важной и популярной задачей. Основная причина — желание использовать для лингвистических исследований большие размеченные интернет-корпуса актуального русского языка. Очевидно, что корпуса объемом в десятки миллиардов словоупотреблений практически непригодны для использования в отсутствие разметки, а об их ручной разметке не приходится и говорить. Имеется, однако, несколько серьезных проблем, которые не позволяют пока получить объективную оценку качества авторазметки и ее пригодности для лингвистики:

1. Разметка корпусов Social Media делается сейчас парсерами, обученными на размеченном подкорпусе с существенно иной жанровой структурой — т. н. «снятнике» НКРЯ [Plungian, 2005] — сравнительно небольшом корпусе с ручной разметкой некоторой смеси текстов НКРЯ не вполне ясной жанрово-тематической принадлежности. Лексические и грамматические особенности этого подкорпуса определили также состав морфословарей и стандарт разметки MSD. Прямой перенос языковой модели на сегменты социальных медиа не позволяет получить морфоразметку нужного качества (Шаров, Беликов et al, 2015).
2. Отсутствует эталонная разметка (золотые стандарты) для оценки качества морфоанализа social media. Первое и последнее тестирование систем морфоразметки проводилось на Диалоге в 2010 году [Liashevskaia O., Astaf'eva I. et al 2010]. Были показаны неплохие результаты, но получены они были на «хороших» текстах. Попытка потестировать «грязные» не показалась тогда особо важной, соответствующая дорожка не собрала участников.
3. Имеются серьезные отличия в подходах к морфоразметке для нужд лингвистических корпусных исследований и для задач автоматического обучения морфопарсеров (POS-тегеров). Эти отличия, однако, явно не сформулированы и никак не учитываются, когда размеченные подкорпуса используются для обучения.

Всё в целом приводит к выводу, что уровень проработки этой проблемы не вполне соответствует ее важности для корпусной лингвистики. Актуальными оказываются две связанные задачи:

1. Создание и обоснование нового тагсета и синхронизированного по грамматической системе морфословаря.
2. Создание достаточно большого тестового корпуса, размеченного в соответствии с этим тагсетом, и выбор технологии его дальнейшего ведения и модификации в условиях большой динамики языка social media.

Необходима также оценка применимости для разметки динамичного и компактного тестового корпуса медленно, но достаточно надежно работающих парсеров «старших» категорий, умеющих снимать грамматическую омонимию на основании полных или локальных синтактико-семантических разборов.

## 2. Big Social Data и Big Social Corpora

Социальные медиа — такие как блоги (Живой Журнал, LiveInternet, Blogs Mail.ru, etc.), микроблоги (Twitter), социальные сети (VKontakte, Одноклассники, FaceBook, etc.), всевозможные форумы — на сегодняшний момент являются основным по объему данных[1] источником материала для корпусов актуального русского языка. Язык социальных сетей обладает отличиями в языке и типографике, осложняющими автоматическую разметку текста.

Авторы представляют проект Генерального интернет-корпуса русского языка (ГИКРЯ), сегмент social media которого включает сейчас около 20 млрд словоупотреблений. Результаты первой разметки этого сегмента с помощью стандартно обученных методов оказалась неоднозначными: несмотря на высокую точность определения частей речи и лемматизации (лучшие в сравнении с, например RuTenTen — см. приложение 1), качество снятия омонимии по отдельным категориям оказалось существенно ниже желаемого [Sharoff et al., 2015].

Мысль, что высокие морфо проценты скрывают серьезные проблемы, уже высказывалась в работе [Manning, 2011]. Так, при выборе метрики, оценивающей точность разбора на целых предложениях, у хороших морфопарсеров (например, Standford Part-of-Speech Tagger [Toutanova et al., 2003]) получилась точность в 55–57 %.

Использовавшиеся методы оценки (и в особености — точность приписывания полного морфо тега) оказались слабо интерпретируемы с точки зрения определения надежности результатов лингвистических исследований. Необходима дифференциальная оценка с учетом значимости отдельных категорий.

Но прежде чем оценивать точность приписывания морфотегов необходимо разобраться с теми грамматическими категориями, которые они реализуют.

## 3. Требования к грамматической системе корпусной морфоразметки

При разметке корпусов сегодня используются различные морфопарсеры для русского языка. Наиболее известные — mystem [Segalovich, 2003], Tnt-Russian [Sharoff, Nivre, 2011], Tree-Tagger [Schmid, 1994], Abbyy Compreno [Anisimovich et al 2011], морфопарсер системы ЭТАП.

Все вышеприведенные программы используют разные словари и наборы грамматических категорий: у mystem это тагсет системы ЭТАП [Apresian et al., 2003] и словарь Зализняка [Zalizniak, 1977], у TnT-Russian обычно тагсет системы Multext-east for Russian [Erjavec, 2010] и словарь, полученный

[1] Рунет в 2014 г. занимал как минимум 155 экзабайт, или 2,4% данных человечества (http://www.rg.ru/2013/05/14/infa-site.html). К 2020 году прогнозируется рост до 980 экзабайт (2,2 % мировых данных). В 2015 году, согласно данным internetworldstats, Россия занимает 6-ое место в мире по количеству интернет-пользователей — 103 147 691 при 70%-ной вовлеченности населения — http://www.cnews.ru/news/top/rossiya_sozdaet_24_mirovogo_obema_dannyh

автоматически на корпусе со снятой омонимией НКРЯ [Plungian, 2005], у системы Abbyy Compreno — словарь и тагсет собственной разработки, используемые в ряде проектов (Lingvo, FineReader, Compreno). Все эти системы развивались параллельно и результаты их работы серьезно не сравнивались до соревнования на Диалоге в 2010 (в котором наилучшие результаты показала система Compreno). По итогам соревнования был сделан небольшой тестовый корпус на объединенной разметке, но анализ особенностей разметок каждой отдельной системы не проводился.

Чтобы сравнение разных разметок стало возможным, организаторы приняли в 2010 г. несколько упрощающих соглашений, в частности сократили систему частеречных признаков до 6 граммем: существительные, прилагательные, глаголы (в том числе причастия и деепричастия), предлоги, и союзы. Все прочие неизменяемые слова, наречия, вводные слова, частицы, попали в одну категорию. Местоимения и числительные, а также ряд других объектов вовсе не размечались. Соответствующим образом выглядел и получившийся золотой стандарт.

Для корпусной разметки такое решение, объединяющее в одну кучу всё, что Пушкин относил к той самой столь важной для языка «мелкой сволочи»[2], не годится ни с точки зрения обучения парсеров, ни для корпусного исследования.

В современных корпусах наблюдается сейчас некоторое единство в сфере стандартов разметки. Такие проекты, как НКРЯ, RuTenTen, корпус университета Лидс и Araneum Russicum используют стандарт Multext-East for Russian и морфопарсер TnT-Russian. В проекте ГИКРЯ также использовался данный парсер, однако, с модификациями, дающими некоторый выигрыш в качестве на сегменте social media [Sharoff et al., 2015] — см. приложения 1 и 2.

Но проблема «мелкой сволочи» в этой разметке решается непоследовательно, что в значительной степени отражает непоследовательность в разметке самого снятника, в частности при анализе вводных конструкций типа «по большому счету» и т. п.

При формулировании требований к грамматической системе корпусной морфоразметки приходится считаться как с нуждами обучения, так и с потребностями исследователя языка:

1) В интересах обучения парсера грамматическая система должна обеспечивать явное маркирование любых частотных дистрибуционных различий. Только в этом случае возможно эффективное снятие грамматической омонимии.
2) Грамматическая система должна соответствовать запросам пользователей корпуса (ее можно изучать на основании статистики запросов к ГИКРЯ и НКРЯ).

---

[2] *…А подбирать союзы да наречья;*
*Из мелкой сволочи вербую рать.*
*Мне рифмы нужны; все готов сберечь я,*
*Хоть весь словарь; что слог, то и солдат —*
*Все годны в строй: у нас ведь не парад.*
(А. С. Пушкин Домик в Коломне)

3) Граммемы этих категорий должны быть потенциально «надежно определяемыми» парсерами, иначе их выставление будет лишь сбивать с толку исследователя (такие граммемы могут все же быть полезными, но только при явном понимании пользователем степени их надежности. Это, например, относится к грамматической интерпретации глагольных форм на *-ся*).

Таким образом, граммемы неоднородны с точки зрения их использования:

- часть граммем не нужна для обучения, но требуется исследователю («информационные» категории), например граммемы одушевленности;
- часть граммем может вводиться в систему исключительно для нужд учета специфики дистрибуции высокочастотной лексики замкнутых классов; например, такие псевдограммемы могут быть приписаны формам «нет» в глагольно-предикативном использовании.

На практике приходится также считаться с особенностями парсеров, которые могут быть чувствительны к росту индекса, увеличению числа низкочастотных комбинаций граммем и проч. Желательно, однако, чтобы «архитектурные» ограничения парсера не влияли на разметку золотого стандарта — это вопрос его адаптации к конкретному алгоритму снятия омонимии.

## 4. Особенности использования грамматической системы синт. парсеров

Кажется разумным использовать разметку парсера, с помощью которого анализировался текстовый корпус. Имеется, однако, проблема, связанная с тем, что парсер Compreno, например, для которого важна полнота морфо- и синтаксических описаний, использует большое число конвенций, являющихся по сути артефактами конкретной модели описания. Кроме того, необходимость описания синтаксических явлений порождает большой класс синтаксических граммем, которые «отменяют» морфологические, возникают проформы разных типов и т. д.

Это делает мэппинг грамматической системы Compreno и любой другой системы полного анализа на новый стандарт морфоразметки достаточно сложным делом, но зато позволяет в дальнейшем использовать эту «тяжелую» разметку для новых тестовых подкорпусов.

Нельзя не коснуться вопроса о перспективах полной синтаксической разметки больших корпусов социальных медиа. Анализ разборов Compreno показывает что качество построения полных семантико-синтаксических структур на специфических текстах этого сегмента все же заметно хуже, чем результаты на текстах non-fiction, на которых, например, в условиях тестирования 2012 года были показаны очень высокие результаты [Toldova S., Sokolova E. et al. 2012]. Кроме того, ресурсы на такую разметку оказались бы слишком велики (не говоря уже о вопросах некоммерческого использования этих технологий).

При этом статистика запросов к ГИКРЯ показывает, что для лингвистических исследований на мегакорпусах, сильно смещенных в сторону лексических явлений, гораздо чаще не нужен полный синтаксический разбор, и при хорошем качестве снятия омонимии можно использовать локальные Sketch-грамматики.

Таким образом, в данной работе мы используем морфословарь и синтаксическую разметку системы Compreno (от которой берем по соглашению CBR только снятую грамматическую омонимию). Оправданность такого решения хорошо видна при сравнении качества снятия омонимии на одних и тех же текстах социальных медиа в разметке ГИКРЯ сейчас (TNT-парсер) и в разметке Compreno (после мэппинга, но до оптимизации). Данные см. в таблице 1.

**Таблица 1.** Точность автоматического снятия омонимии на некоторых граммемах имени существительного

| граммема | Точность при разметке TnT | Точность при разметке Compreno |
|---|---|---|
| Неодушевленный номинатив | 0,792 | 0,947 |
| Неодушевленный аккузатив | 0,858 | 0,884 |
| Одушевленный аккузатив | 0,661 | 0,980 |
| Одушевленный генитив | 0,890 | 0,890 |
| Субстантивы (на выборке из омонимичных вхождений) | 0,680 | 0,916 |
| Прилагательные (на выборке из омонимичных вхождений) | 0,900 | 0,918 |

## 5. Принципиальные проблемы автоматической морфоразметки social media

При автоматическом создании текстового корпуса приходится решать 2 проблемы:

1. «Мэппинг» — уже описанная выше задача мэппирования грамматической системы используемого «тяжелого» парсера на новый морфостандарт;
2. «Оптимизация» — исправление частотных систематических ошибок, допускаемых этим парсером.

Несмотря на высокую точность, заметно превосходящую результаты, получаемые обычными тегерами, в непосредственно полученном после мэппирования корпусе есть еще, с чем можно побороться. Речь идет о систематических ошибках, которые исправляются полностью автоматически в любом новом цикле «разбор парсером — мэппирование — финальное улучшение». Примерные результаты разбора после мэппирования (они незначительно зависят от разбираемого подкорпуса социальных медиа) в Таблице 2:

**Таблица 2**

| Формат | Среднее кол-во граммем на слово | Точность разметки |
|---|---|---|
| Abbyy Compreno | 7 | ABBYY 0,942 |
| MSD | 4 | TnT-Russian 0,887 |
| ЭТАП | 4 | Mystem 0,927 |

Статистика остаточных ошибок по типам в GICRMorpho и ГИКРЯ в таблице 3:

**Таблица 3**

| | |
|---|---|
| неправильная лемма | 38,18% |
| неправильная часть речи | 23,64% |
| падежная омонимия | 23,64% |
| неправильная собственность | 5,45% |
| неправильная транзитивность | 5,45% |
| неправильная одушевленность | 1,82% |
| омонимия формы | 1,82% |

Большая часть ошибок (ошибки на лемму и часть речи) вызваны опечатками, сленгом и именами собственными, неизвестными словарю (более подробно — см. Таблицу 4).

**Таблица 4**

| Причина ошибки | Штук | Доля |
|---|---|---|
| опечатка | 7 | 17,50% |
| сленг | 14 | 35% |
| неизменяемое слово | 1 | 2,50% |
| имя собственное | 3 | 7,50% |
| омонимия | 15 | 37,50% |

## 6. Данные о новом тестовом корпусе GICRMorth

Тестовый корпус Social Media представляет собой выборку текстов из Живого Журнала объемом в 2 млн слов. Данные тексты были очищены от html-разметки и затем направлены на анализ в систему ABBYY Compreno. Полученный материал был очищен от синтаксической разметки и смэппирован в новый тагсет MSD-GICR. Из обучающей выборки были извлечены триграммные частоты.

На данном этапе распространяемый по лицензии корпус следует рассматривать как тестовый: безусловно, по результатам тестирования можно ожидать как изменений в тагесете, так и в разметке и в самом подборе текстов.

Важно, что методика получения корпуса является полностью автоматической и может быть повторена с другой выборкой данных (например, для другого сегмента Social Media).

## 7. Результаты, открытые вопросы

Основные результаты нашей работы сводятся к двум важным составляющим: первая — это первый золотой стандарт морфологической разметки в 2 миллиона словоформ, обладающий высоким качеством морфологической разметки и являющийся общедоступным источником, на котором другие исследователи смогут обучать свои морфологические и синтаксические парсеры, ориентированные на обработку текстов social media.

Пример разметки в новом тагсете приведен в таблице 5:

**Таблица 5**

| Vertical text | Lemma | MSD-GICR | |
|---|---|---|---|
| Если | [если] | C | #союз |
| хочешь | [хотеть] | V-ip2s-a-p-ym | #глагол, личная форма, наст.вр., 2-е лицо, ед. ч., активный залог, несов., перех. |
| тусить | [тусить] | V-n----a-p-nm | #глагол, инфин., активный залог, несов., неперех. |
| — | | | |
| туси | [тусить] | V-m-2s-a-p-nm | #глагол, пов.накл., 2-е лицо, ед. ч., активный залог, несов., неперех. |
| . | | | |
| Если | [если] | C | |
| хочешь | [хотеть] | V-ip2s-a-p-ym | |
| бухнуть | [бухнуть] | V-n----a-e-ym | #глагол, пов.накл., 2-е лицо, ед. ч., активный залог, соверш., перех. |
| — | | | |
| бухни | [бухнуть] | V-m-2s-a-e-ym | #глагол, пов.накл., 2-е лицо, ед. ч., активный залог, соверш., перех. |
| . | | | |

Вторая составляющая — это n-граммная статистика для POS-тегера, которая будет применена к соответствующим сегментам ГИКРЯ — социальным сетям и блогам.

Качество разметки нового парсера в сравнении с качеством стандарта — в таблице 6:

**Таблица 6.** Итоговое качество частеречной разметки

| Часть речи | Точность | Полнота | F-мера |
|---|---|---|---|
| 1. Существительное | 0,989 | 0,960 | 0,974 |
| 2. Глагол | 0,995 | 0,979 | 0,987 |
| 3. Прилагательное | 0,978 | 0,978 | 0,978 |
| 4. Местоимение | 1,000 | 0,896 | 0,945 |
| 5. Наречие | 0,940 | 0,935 | 0,937 |
| 6. Предлог | 1,000 | 0,972 | 0,986 |
| 7. Союз | 0,927 | 0,962 | 0,944 |
| 8. Числительное | 0,957 | 0,978 | 0,967 |
| 9. Частица | 0,964 | 0,891 | 0,926 |
| 10. Междометие | 1,000 | 0,585 | 0,738 |
| 11. Предикатив | 1,000 | 0,900 | 0,947 |
| 12. Вводное слово | 0,954 | 0,807 | 0,874 |
| **Макроусреднение** | **0,975** | **0,904** | **0,934** |

Анализ ошибок результирующей разметки показал, что основная масса ошибок приходится на:

- Омонимию разрядов местоимений («*его*»-«*его*»)
- Омонимию субстантивов и причастий («*данные*»)
- Омонимию форм у неизменяемых существительных («бананы, груши, *манго*»)
- Омонимию форм частиц и союзов («*однако*»)
- Опечатки
- Неправильную лемматизацию несловарных слов

Открытыми остаются вопросы о включении в цепочку морфологической обработки блока псевдолемматизации и исправления орфографии. В дальнейшей работе над морфологической разметкой мы планируем включить наработки из работы [Sorokin, Khomchenkova, 2016], которые позволят улучшить качество обработки несловарных слов, а также результаты работы [Sorokin, Shavrina, 2016], позволяющие стабилизировать качество разметки независимо от грамотности авторов текстов и тем самым избежать сдвигов статистики употреблений отдельных слов.

В процессе разработки мы создали правила меппинга, позволяющие перекодировать существующие форматы MSD, MSD-GICR [Sharoff et al., 2015], mystem и ABBYY Compreno в новый формат разметки, доступный исследователям.

## 8. Заключение

В данной статье мы представляем результаты создания нового обучающего корпуса и нового стандарта морфоразметки social media для русского языка.. Модель обучения на триграммах, представленная в [Sharoff, Nivre, 2011] была

воспроизведена на основе новой обучающей выборки. В рамках этой работы результаты автоматической морфологической разметки ABBYY Compreno были перенесены на новый тагсет, который представляет из себя дальнейшее развитие Multext. Полученный морфотаггер показывает возможность успешного снятия морфологической омонимии и лемматизации на материале социальных сетей.

Этот морфотаггер будет использован для разметки и развития в рамках проекта «Генеральный интернет-корпус русского языка». Подкорпус из 2 млн словоупотреблений social media, размеченный в новом наборе категорий с автоматически снятой омонимией, доступен для свободного использования [http://www.webcorpora.ru/news/282]. Авторы надеются, что их разработка поможет NLP-community в развитии и обучении их морфологических и синтаксических парсеров на базе нового стандарта разметки.

## Благодарности

Авторы выражают благодарность компании ABBYY за возможность использования парсера Compreno в рамках академической лицензии CBR. Мы благодарим Анастасию Сиротину, чьи комментарии были крайне полезны при анализе морфоразметки. Выражаем особую благодарность Валерию Новицкому и Андрею Андрианову из ABBYY, чья помощь и консультации сделали возможными морфологические эксперименты, описанные в статье.

## Литература

1. *Anisimovich K. V., Druzhkin K. Ju., Minlos F. R., Petrova M. A., Selegey V. P. Zuev K. A.* Syntactic and semantic parser based on ABBYY Compreno linguistic technologies. In: Computational linguistics and intellectual technologies. 2012, Vol. 11. http://www.dialog-21.ru/digests/dialog2012/materials/pdf/Anisimovich.pdf
2. *Apresian J., Boguslavskii I., Iomdin L., Lazurskii A., Sannikov V., Sizov V., Tsinman L.* (2003) ETAP-3 Linguistic Processor: a Full-fledged NLP Implementation of the MTT. First International Conference on Meaning-Text Theory: 279–288.
3. *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).
4. *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.
5. *Dimitrova, L., T. Erjavec, N. Ide, H.-J. Kaalep, V. Petkevic, and D. Tufis* (1998), MULTEXT-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In COLING-ACL '98. Montreal, Quebec, Canada.
6. *Erjavec T.* (2010) Multext-east Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10).

7. *Gareyshina A., Ionov M., Lyashevskaya O., Privoznov D., Sokolova E., Toldova S.* (2012) RU-EVAL-2012: Evaluating Dependency Parsers for Russian. Proceedings of COLING 2012: Posters. P. 349–360. URL: http://www.aclweb.org/anthology/C12–2035.
8. *Jongejan B. and Dalianis H.* (2009) Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Suntec, Singapore: Association for Computational Linguistics, 2009. s. 145–153
9. *Kilgarriff A., Baisa V., Busta J., Jakubicek M., Kovar V., Michelfeit J., Rychly P., Suchomel V.* (2014). "The Sketch Engine: ten years on". Lexicography (Springer Berlin Heidelberg) 1 (1): 7–36.
10. *Kilgarriff A., Rychly P., Smrz P., Tugwell D.* (2004) The Sketch Engine/ Proc. Euralex. Lorient, France
11. *Liashevskaia O., Astaf'eva I., Bonch-Osmolovskaia A., Gareishina A., Iu., G., D'iachkov V., Ionov M., Koroleva A., Kudrinski M., Litiagina A., Luchina E., Sidorova E., Toldova S., Savchuk S., and Koval' S.* (2010) Evaluation of Automatic Text Parsing Methods: Morphological Parsers in Russian [Otsenka Metodov Avtomaticheskogo Analiza Teksta: Morfologicheskie Parsery Russkogo Iazyka]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2010" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2010"): 318–326.
12. *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.
13. *Plungian V. A.* (2005) What do We Need Russian National Corpus for? [Zachem Nuzhen Natsionalnii Korpus Russkogo Iazyka?] Natsionalnii Korpus Russkogo Iazyka: 6–20.
14. *Schafer 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)
15. *Schmid H.* (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.
16. *Segalovich I.* (2003) A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine.
17. *Sharoff S. A., Belikov V. I., Kopylov N. Y., Sorokin A. A., Shavrina T. O.* (2015) Corpus with Automatically Resolved Morphological Ambiguity: to the Methodology of Linguistic Research. In Dialogue, Russian International Conference on Computational Linguistics, Moscow.
18. *Sharoff S., Nivre J.* (2011) The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2011" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2011"], Bekasovo, pp. 591–605.

19. *Shavrina T., Sorokin A.* (2015) Modeling Advanced Lemmatization for Russian Language Using TnT-Russian Morphological Parser. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2015”
20. *Sokirko A.* (2004) Morphological Modules on the web-site www.aot. ru [Morphologicheskie Moduli na saite www.aot.ru]. Komp’iuternaia Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii “Dialog 2004” (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference “Dialog 2004”).
21. *Sokirko A., Toldova S.* (2005) Sravnenie Effektivnosti Dvukh Metodik Sniatiia Lexicheskoi i Morfologicheskoi Neodno znachnosti dlia Russkogo Iazyka. Internet-matematika.
22. *Sorokin A. A., Baitin A. V., Galinskaya I. E., Shavrina T. O.* (2016) SpellRuEval: The First Competition On Automatic Spelling Correction For Russian. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2016”
23. *Sorokin A. A., Shavrina T. O.* (2016) Automatic spelling correction for Russian social media texts. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2016”
24. *Sorokin A. A., Khomchenkova I. A.* (2016) Automatic detection of morphological paradigms using corpora information.
25. *Toldova S., Sokolova E. et al.* NLP evaluation 2011–2012: Russian syntactic parsers. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2012”
26. *Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer.* (2003) Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252–259.
27. *Zalizniak A.* (1977) Russian Grammar Dictionary [Grammaticheskii Slovar’ Russkogo Iazyka. Russki Iazyk].

## Приложения

**Приложение 1** — Table 3 from [Sharoff et al., 2015, 8]

| **RuTenTen** | | | **GICR** | | |
|---|---|---|---|---|---|
| **Part of Speech** | **Precision** | **Recall** | **Part of Speech** | **Precision** | **Recall** |
| 1. Noun | 0,948 | 0,987 | 1. Noun | 0,997 | 0,99 |
| 2. Verb | 0,966 | 0,976 | 2. Verb | 0,998 | 0,998 |
| 3. Adjective | 0,942 | 0,969 | 3. Adjective | 0,953 | 0,997 |
| 4. Pronoun | 0,988 | 0,975 | 4. Pronoun | 1,000 | 1,000 |
| 5. Adverb | 0,927 | 0,914 | 5. Adverb | 0,974 | 0,913 |
| 6. Preposition | 1,000 | 0,997 | 6. Preposition | 1,000 | 0,998 |
| 7. Conjunction | 0,993 | 0,991 | 7. Conjunction | 0,993 | 0,993 |
| 8. Numeral | 0,797 | 0,911 | 8. Numeral | 0,98 | 1,000 |

| RuTenTen | | | GICR | | |
|---|---|---|---|---|---|
| **Part of Speech** | **Precision** | **Recall** | **Part of Speech** | **Precision** | **Recall** |
| 9. Particle | 0,986 | 0,983 | 9. Particle | 0,996 | 0,996 |
| 10. Inetrjection | 1,000 | 0,551 | 10. Inetrjection | 1,000 | 0,9 |
| 11. Other | 0 | 0 | 11. Predicative | 1,000 | 0,81 |
| 12. Abbreviation | 0 | 0 | | | |
| Microaverage: | 0,979 | 0,975 | Microaverage: | 0,99 | 0,963 |
| Macroaverage: | 0,7956 | 0,7712 | | | |
| Macroaverage without 11–12: | 0,955 | 0,925 | Macroaverage: | 0,991 | 0,99 |

**Приложение 2** — Table 2 from [Sharoff et al., 2015, 7]

| Old morphology of GICR | | | New morphology of GICR | | |
|---|---|---|---|---|---|
| **Part os speech** | **Precision** | **Recall** | **Part os speech** | **Precision** | **Recall** |
| 1. Noun | 0,992 | 0,987 | 1. Noun | 0,997 | 0,99 |
| 2. Verb | 0,989 | 0,991 | 2. Verb | 0,998 | 0,998 |
| 3. Adjective | 0,95 | 0,943 | 3. Adjective | 0,953 | 0,997 |
| 4. Pronoun | 0,997 | 0,997 | 4. Pronoun | 1 | 1 |
| 5. Adverb | 0,808 | 0,947 | 5. Adverb | 0,974 | 0,913 |
| 6. Preposition | 1 | 1 | 6. Preposition | 1 | 0,998 |
| 7. Conjunction | 0,979 | 0,996 | 7. Conjunction | 0,993 | 0,993 |
| 8. Numeral | 0,815 | 0,963 | 8. Numeral | 0,98 | 1 |
| 9. Particle | 0,996 | 0,959 | 9. Particle | 0,996 | 0,996 |
| 10. Inetrjection | 1 | 0,714 | 10. Inetrjection | 1 | 0,9 |
| 11. Other | 0 | 0 | 11. Predicative | 1 | 0,81 |
| 12. Abbreviation | 0 | 0 | | | |
| **Microaverage:** | **0,794** | **0,796** | **Microaverage:** | **0,99** | **0,963** |
| **Macroaverage:** | **0,976** | **0,987** | **Macroaverage:** | **0,991** | **0,99** |
| **Macroaverage without 11–12:** | **0,953** | **0,95** | | | |

## Приложение 3

Руководствуясь приведенными выше принципами, мы создали итоговый тагсет, основанный на позиционной системе MSD, но качественно дополненный граммемами Abbyy Compreno, помогающими снимать омонимию, и лишенный граммем, сложных в определении и отягчающих реализацию разметки:

| Position | Code | Meaning |
|---|---|---|
| 0 | N | Noun |
| 1 | p/c | Type: proper/common |

| Position | Code | Meaning |
|---|---|---|
| 2 | m/f/n/c/- | Gender: Masculine/Feminine/Neuter/Common/Undefined (for pluralia tantum) |
| 3 | s/p | Number: Singular/Plural |
| 4 | n/g/d/a/l/i/v | Case:Nominative/(Genitive\|Partitive)/Dative/Accusative/(Locative\|Prepositional)/Instrumental/Vocative |
| 5 | n/y | Animatedness: Inanimate/Animate |
| 6 | p/l/- | Case2: Partitive/Locative/(Nominative\|Genitive\|Dative\|Accusative\|Prepositional\|Instrumental\|Vocative) |
| 0 | V | Verb |
| 1 | - | |
| 2 | i/m/n/p/g/x | GrammaticalType: Indicative/Imperative/Infinitive/Participle/Adverb/WordNet |
| 3 | p/f/s/-/* | Tense: present/future/past/Undefined/Cannot be disambgued |
| 4 | 1/2/3/- | Person: First/Second/Third/Undefined |
| 5 | s/p/- | Number: Singular/Plural/Undefined |
| 6 | m/f/n/- | Gender: Masculine/Feminine/Neuter/Undefined |
| 7 | a/p/s | Voice: Active/Passive/VoiceSya |
| 8 | s/f/- | ParticipleShortness: ShortForm/FullForm/Undefined |
| 9 | p/e/-/* | Aspect: Imperfective/Perfective/Undefined/Cannot be disambgued |
| 9 | n/g/d/a/l/i/- | Case: Nominative/Genitive/Dative/Accusative/Prepositional/Instrumental/Undefined |
| 10 | n/y | Transitivity: Intransitive/Transitive |
| 11 | m/b | Pairness: (MonoAspectual\|Paired)/BiAspectual |
| 0 | A | Adjective |
| 1 | - | |
| 2 | p/c/s | DegreeOfComparison: Positive/Comparative/Superlative |
| 3 | m/f/n/- | Gender::Masculine/Feminine/Neuter/Undefined |
| 4 | s/p | Number: Singular/Plural/ |
| 5 | n/g/d/a/l/i | Case: Nominative/Genitive/Dative/Accusative/Prepositional/Instrumental |
| 6 | s/f | AdjectiveShortness: ShortForm/FullForm |
| 0 | P | Pronoun |
| 1 | p/d/i/s/q/x/z/n | ReferenceClass::RCPersonal/RCDemonstrative/RCIndefinite/RCPossessive/RCInterrogative/RCReflexive/RCNegative/RCAttributive |
| 2 | 1/2/3/- | Person: First/Second/Third/Undefined |
| 3 | m/f/n/-/* | Gender: Masculine/Feminine/Neuter/Undefined/»все» and «всё» cannot be disambgued |

| Position | Code | Meaning |
|---|---|---|
| 4 | s/p/-/* | Number: Singular/Plural/Undefined/»все» and «всё» cannot be disambgued |
| 5 | n/g/d/a/l/i | Case: Nominative/Genitive/Dative/Accusative/ Prepositional/Instrumental |
| 6 | n/a/r | Syntactic Type: Nominal/adjectival/adverbial |
| 7 | s/f/- | Shortnesss: ShortForm/FullForm/Undefined |
| 0 | R | Adverb |
| 1 | p/c/s | DegreeOfComparison: Positive/Comparative/ Superlative |
| 0 | W | Predicative |
| 0 | S | Preposition |
| 1 | p | Type: preposition |
| 2 | - | |
| 3 | g/d/a/l/i | Case: Genitive/Dative/Accusative/Prepositional/ Instrumental |
| 0 | C | Conjunction |
| 0 | M | Numeral |
| 1 | c/l/o | Type: cardinal/collect/ordinal |
| 2 | m/f/n/- | Gender: Masculine/Feminine/Neuter/Undefined |
| 3 | s/p/- | Number: Singular/Plural/Undefined |
| 4 | n/g/d/a/l/i | Case: Nominative/Genitive/Dative/Accusative/ Prepositional/Instrumental |
| 5 | l/d/r | Form: numeral/arabic digit/roman digit |
| 0 | Q | Particle |
| 0 | I | Interjection |
| 0 | H | Parenthetical phrase |
| 0 | X | Residual |