

Оценка точности работы методики автоматизированной идентификации лексем с применением виртуальных текстов

08, август 2013

DOI: 10.7463/0813.0583389

Тихомирова Е. А., Павлов Ю. Н., Каспрук В. В.

УДК 004.052

Россия, МГТУ им. Н.Э. Баумана

elizarti@mail.ru

pavlov@bmstu.ru

kasper.sochi@mail.ru

Введение

При проведении автороведческой экспертизы с целью идентификации авторских текстов с помощью частотного анализа требуется достаточно точное знание частот появления тех или иных лексем в текстах автора. Под лексемой понимается слово, рассматриваемое как единица словарного состава языка в совокупности всех его конкретных грамматических форм, а также всех возможных значений [1]. Термин «лексема» в этом смысле употребляется большинством отечественных авторов. Как показывают публикации последнего времени [1-4], проблема идентификации лексем остаётся актуальной.

Определение этих частот является достаточно сложной задачей по причине наличия у каждой лексемы множества словоформ, составляющих так называемую парадигму. С этой точки зрения лексема представляет собой результат абстракции отождествления словоформ, реально встречающихся в речи. Например, у лексемы «он» имеются словоформы «него», «нему», «его», «нем», «им» и другие. Выявить автоматизированным методом тот факт, что эти столь непохожие по написанию словоформы принадлежат одной лексеме, весьма затруднительно.

В предыдущей статье [5] автора предложена комплексная методика автоматизированной идентификации лексемы по различным словоформам, в которой были приведены результаты предварительной оценки точности работы комплексной методики на примере

короткого текста с общим объёмом около 350 слов. Под точностью работы методики, понимается процент правильно идентифицированных лексем. Для увеличения достоверности оценки точности работы комплексной методики необходимо иметь тексты, содержащие гораздо большее количество слов. Однако при этом встает вопрос о контрольном счете, который получить вручную практически невозможно. В этой связи в данной статье предлагается метод создания так называемых «виртуальных текстов». Суть этого метода заключается в следующем: каждая лексема русского языка имеет некоторое количество словоформ, сведения о которых содержатся в морфологических словарях. Сформировав с помощью словарей текст, содержащий определенное количество лексем и соответствующих им словоформ, можно получить для анализа тексты с достаточно большим количеством слов и имеющие при этом известные количественные характеристики, подлежащие определению с помощью комплексной методики.

Оценка точности работы предлагаемой методики [5] сводится к сравнительному анализу данных, содержащихся в виртуальном тексте, и данных, полученных с помощью автоматизированной процедуры идентификации. Поскольку в русском языке существуют различия в словообразовании различных частей речи, анализ ошибок необходимо проводить, составляя виртуальные тексты, однородные по частям речи, постепенно увеличивая сложность текста, соединяя последние, тем самым получая дифференциальную оценку точности работы комплексной методики идентификации.

Отрывок виртуального текста, составленный из имен существительных, выглядит следующим образом: «...представителем, представители, представитель, представителя, представителям, представителями, участие, участием, участии, участию, участия, картин, картина, картине, картиной, картину...» и так далее.

Комплексная методика содержит в своём составе два последовательно работающих блока:

- блок идентификации лексем по словоформам на основе словарной морфологии;
- блок идентификации лексем по словоформам аналитическим методом.

С целью локализации источника возникновения ошибок идентификации лексем был проведен анализ точности работы ее отдельных блоков.

Блок идентификации лексем по словоформам на основе словарной морфологии использует для своей работы словари русского языка Зализняка А.А. [6] и Про-Линг [7].

Для оценки точности работы блока были сформированы виртуальные тексты, составленные из имен существительных, глаголов и имен прилагательных, поскольку

данные части речи являются наиболее употребляемыми, как в русской разговорной речи, так и в художественных произведениях, и имеют наибольшее количество словоформ.

Тестовый виртуальный текст, содержал 109 лексем имён существительных с 509 словоформами, принадлежащими этим лексемам.

В таблице 1 приведены результаты работы блока автоматизированной идентификации лексем по словоформам на основе словарной морфологии.

В столбце «Лексемы» представлены лексемы, словоформы которых формируют виртуальный текст, созданный из имен существительных. В столбце «Заданное количество» - заданное в виртуальном тексте количество словоформ данной лексемы. В столбце «Полученное количество» - количество словоформ данной лексемы, полученное в результате работы блока автоматизированной идентификации лексем на основе словарной морфологии [5].

Таблица 1.

Результаты автоматизированной идентификации лексем имен существительных на основе словарной морфологии.

Лексема	Заданное количество	Полученное количество	Лексема	Заданное количество	Полученное количество
Круг	9	9	Старик	5	5
Направление	8	8	Экономика	5	5
Регион	7	1	Литература	5	5
Изменение	7	7	Успех	5	5
Банк	7	7	Удар	5	5
Врач	7	7	Доллар	4	4
Договор	7	7	Рисунок	4	4
Объем	7	7	Течение	4	4
Специалист	7	7	Сцена	4	4
Требование	7	7	Население	4	4
Бумага	7	7	Памятник	4	4
Граница	7	7	Факт	4	4
Предмет	7	7	Хозяин	4	4
Магазин	7	7	Писатель	4	4
Текст	7	7	Род	4	4
Представитель	6	6	Солнце	4	4
Девочка	6	6	Вера	4	4
Команда	6	6	Процент	4	4
Самолет	6	6	Операция	4	4

Берег	6	6	Повод	4	4
Спектакль	6	6	Ребята	4	4
Фирма	6	6	Кабинет	4	4
Завод	6	6	Стихи	4	4
Журнал	6	6	Счастье	4	4
Руководитель	6	6	Сознание	4	4
Основание	6	6	База	4	4
Поэт	6	6	Название	4	4
Автомобиль	6	6	Выход	4	4
Надежда	6	6	Модель	4	4
Вариант	6	6	Миллион	4	4
Министр	6	6	Ум	4	4
Дух	6	6	Защита	4	4
Участие	5	5	Знание	4	4
Политика	5	7	Герой	3	3
Картина	5	5	Картинка	3	3
Спина	5	5	Церковь	3	3
Территория	5	5	Музыкант	3	3
Большинство	5	5	Море	3	3
Музыка	5	5	Пара	3	3
Правда	5	5	Пространство	3	3
Свобода	5	5	Степень	3	3
Союз	5	5	Поле	3	2
Дерево	5	5	Героиня	2	2
Природа	5	5	Представительство	2	2
Телефон	5	5	Политик	2	0
Позиция	5	5	Пол	2	3
Двор	5	5	Память	2	2
Способ	5	5	Цвет	2	2
Оценка	5	5	Господин	2	2
Песня	5	5	Площадь	2	2
Родитель	5	5	Политбюро	1	1
Половина	5	5	Спинка	1	1
Роман	5	5	Угол	1	1
Руководство	5	5	Сон	1	1
Анализ	5	5	Общее количество словоформ	509	

По результатам работы блока автоматизированной идентификации лексем, основанной на словарной морфологии, выявлены характерные ошибки идентификации (таблица 2).

Таблица 2.

Ошибки автоматизированной идентификации лексем при анализе имен существительных на основе словарной морфологии.

Лексема	Заданное количество	Полученное количество	Лексема	Заданное количество	Полученное количество
Регион	7	1	Политик	2	0
Политика	5	7	Пол	2	3
Поле	3	2			

Ошибки идентификации лексем, приведенные в таблице 2, возникают по двум причинам:

- 1) Морфологическая омонимия (разные по значению, но одинаковые по написанию единицы языка). В данном случае это относится к лексеме «политика», «политик», «пол» и «поле», поскольку в их парадигмах встречаются омонимы.
- 2) Отсутствие словоформы или лексемы в словарях, на основе которых происходит идентификация. В данном случае это относится к лексеме «регион».

Количество безошибочной идентификации при определении лексем блоком на основе словарной морфологии, подсчитанное по соотношению

$$\Delta = \left(1 - \frac{\sum_{i=1}^m \delta_i}{N}\right) \cdot 100\%, \quad (1)$$

составляет девяносто восемь процентов. Здесь Δ – процент безошибочной идентификации лексем, N – количество словоформ в виртуальном тексте, m – количество лексем в виртуальном тексте, δ_i – количество ошибок при идентификации i -ой лексемы:

$$\delta_i = |Q_{\text{полученное}} - Q_{\text{заданное}}|, \quad (2)$$

где $Q_{\text{полученное}}$ – количество словоформ, отнесенное блоком к i -ой лексеме, $Q_{\text{заданное}}$ – количество словоформ, относящееся к i -ой лексеме в виртуальном тексте.

При проведении расчетов по соотношению (2) словоформы повторно не учитывались. Таким образом, блок идентификации лексем на основе словарной морфологии при анализе виртуального текста, состоящего из имен существительных, допускает не более двух процентов ошибок.

При анализе виртуальных текстов, состоящих из глаголов и имен прилагательных, блоком идентификации на основе словарной морфологии были выявлены ошибки, представленные в таблицах 3, 4, соответственно.

Таблица 3.

Ошибки автоматизированного анализа лексем глаголов блоком идентификации на основе словарной морфологии.

Лексема	Заданное количество	Полученное количество	Лексема	Заданное количество	Полученное количество
Подать	16	15	Включать	5	1
Болезнь	13	12	Собирать	5	1
Выступить	12	12	Предполагать	5	1
Делаться	13	1	Осуществлять	5	1
Допустить	13	12	Производить	5	1
Менять	13	25	Учитывать	5	1
Меняться	13	1	Выбирать	5	1
Послать	13	12	Забывать	5	1
Замечать	5	1	Приносить	5	1
Встречать	5	1	Полагать	5	1
Готовить	5	1	Отдавать	4	1
Испытывать	5	1	Организовать	4	1
Поддерживать	5	1	Делать	0	12

Таблица 4.

Ошибки автоматизированного анализа лексем имен прилагательных блоком идентификации на основе словарной морфологии.

Лексема	Заданное количество	Полученное количество	Лексема	Заданное количество	Полученное количество
Большой	6	4	Любить	0	6
Будущий	6	0	Быть	0	6
Любимый	6	0	Большой	0	2

Ошибки, представленные в таблицах 3 и 4, вызваны теми же причинами, что и ошибки при идентификации лексем имен существительных (омонимия или отсутствие лексемы в

словаре). Количество ошибок при идентификации лексем глаголов составило 8%, а имен прилагательных сохранилось в пределах двух процентов.

Оценка эффективности работы блока идентификации лексемы по словоформам аналитическим способом проводилась на тех же виртуальных текстах, что и блока идентификации на основе словарной морфологии.

В таблицах 5-7 представлены ошибки автоматизированной идентификации лексем имен существительных, глаголов и имен прилагательных соответственно. Анализ результатов идентификации лексем имен существительных показал, что ошибки составляют шесть процентов, лексем глаголов – тринадцать процентов и лексем имен прилагательных – два процента.

Таблица 5.

Ошибки автоматизированного анализа лексем имен существительных блоком идентификации на основе аналитического метода.

Лексема	Заданное количество	Полученное количество	Лексема	Заданное количество	Полученное количество
Договор	7	0	Пол	2	5
Политика	5	7	Цвет	2	0
Способ	5	0	Договориться	0	7
Вера	4	0	Способный	0	5
Выход	4	0	Верить	0	4
Ум	4	0	Выходить	0	4
Поле	3	0	Уметь	0	4
Политик	2	0	Цветок	0	2

Таблица 6.

Ошибки автоматизированной идентификации лексем при анализе глаголов на основе
аналитического метода.

Лексема	Заданное количество	Полученное количество	Лексема	Заданное количество	Полученное количество
Подать	16	15	Выступить	12	0
Болезнь	13	9	Готовить	5	0
Вернуть	13	0	Вернуться	0	13
Взяться	13	0	Взять	0	13
Говориться	13	12	Учиться	0	13
Гореть	13	5	Выступать	0	12
Готовиться	13	12	Говорить	0	1
Делаться	13	1	Делать	0	12
Изменить	13	7	Следовать	0	12
Лечь	13	2	Лететь	0	11
Менять	13	25	Садить	0	2
Меняться	13	1	Боль	0	3
Отправить	13	0	Более	0	1
Отправиться	13	26	Присутствие	0	4
Повернуться	13	12	Пойти	0	4
Послать	13	9	Гора	0	8
Присутствовать	13	9	Измениться	0	6
Садиться	13	11	Готовый	0	6
Следить	13	1	Верный	0	1
Учить	13	0	Под	0	1

Таблица 7.

Ошибки автоматизированного анализа лексем имен прилагательных блоком идентификации на основе аналитического метода.

Лексема	Заданное количество	Полученное количество	Лексема	Заданное количество	Полученное количество
Боевой	6	0	Боже	0	6
Божий	6	0	Бой	0	6
Будущий	6	0	Быстро	0	6
Быстрый	6	0	Быть	0	6
Верный	6	0	Вернуться	0	6
Дальнейший	6	0	Дальний	0	6
Долгий	6	0	Долго	0	6
Естественный	6	0	Естественно	0	6
Любимый	6	0	Любить	0	6
Открытый	6	0	Открыть	0	6
Постоянный	6	0	Постоянно	0	6
Правильный	6	0	Правильно	0	6
Прямой	6	0	Прямо	0	6
Пустой	6	0	Пусть	0	6
Тихий	6	0	Тихо	0	6
Точный	6	0	Точно	0	6
Частый	6	0	Часть	0	6
Ясный	6	0	Ясно	0	6

Природа ошибок, допущенных блоком идентификации аналитическим методом, объясняется наличием идентичности псевдооснов анализируемых лексем. Данный эффект объясняется тем, что часть речи при анализе не определяется, поэтому с помощью аналитического алгоритма последовательно происходит идентификация в словоформе морфем различных частей речи с последующим их отсечением. Таким образом, словоформы имени существительного «выход» и глагола «выходить» обладают одинаковыми псевдоосновами. При возникновении подобной ситуации блок ориентирован на выбор в пользу чаще встречающейся в русском языке лексемы. В данном примере блок отдал предпочтение лексеме «выходить» на основе частот употребления данных лексем в русском языке согласно частотному словарю русского языка

Шарова С.А.: «выход» - 131,48 ipm (ipm - instances per million words – вхождений на миллион слов), «выходить» - 275,27 ipm.[8]

Сравнительный анализ точности работы отдельных блоков комплексной методики по отдельным виртуальным текстам для имен существительных, глаголов и имен прилагательных представлен на рис. 1.

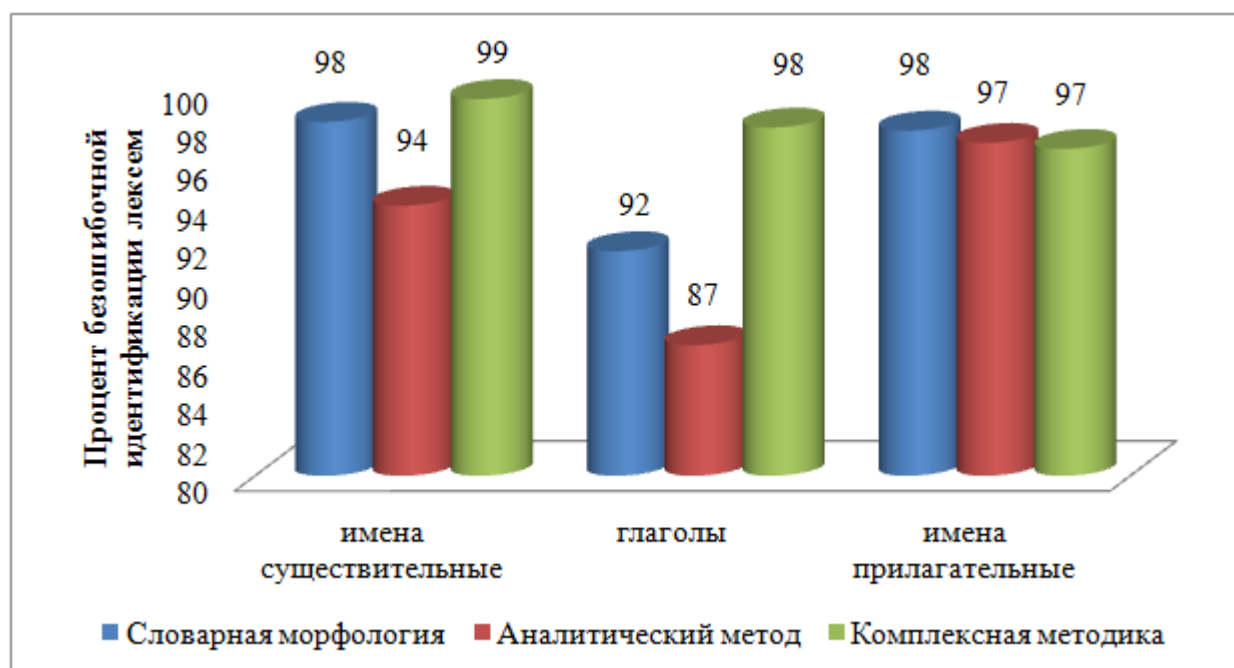


Рис. 1. Барограмма точности работы алгоритмов автоматизированной идентификации лексем на отдельных виртуальных текстах

Для оценки качества совместной работы блоков идентификации лексем по их словоформам, то есть комплексной методики в целом, был сформирован виртуальный текст, содержащий 2256 словоформ имен существительных, глаголов и имен прилагательных.. Полученные результаты представлены на рис. 2.



Рис. 2. Барограмма точности работы алгоритмов автоматизированной идентификации лексем

Таким образом, комплексная методика автоматизированной идентификации лексем, предложенная автором [5] при проведении тестового эксперимента совершила два процента ошибок двух типов. Первый тип ошибок (морфологическая омонимия) характерен и для человека, если он имеет дело с отдельными словами, вырванными из текста (например, отдельно взятое слово «замок»). При возникновении подобной ситуации, комплексная методика решает эту задачу, обратившись к частотному словарю русского языка, взяв в качестве лексемы наиболее часто употребляемое слово.

Второй тип ошибок совершается в случае отсутствия лексемы в используемых для анализа словарях.

Выводы

1) Блок идентификации лексем на основе морфологического анализа обеспечивает уровень безошибочной работы в 98 процентов на текстах, содержащих только имена существительные и прилагательные. На текстах, содержащих глаголы, количество ошибок возрастает, и процент безошибочной работы снижается до 92 процентов.

2) Блок идентификации лексем аналитическим методом обеспечивает уровень безошибочной работы на текстах, содержащих только имена существительные 94 процента, на текстах, содержащих только глаголы – 87 процентов и на текстах, содержащих только имена прилагательные – 97 процентов.

3) Комплексная методика идентификации лексем по их словоформам в тестовых текстах, состоящих из имен существительных, глаголов и имен прилагательных обеспечивает уровень безошибочной работы равный 98 процентам.

На основе проведенного анализа можно сформировать следующие рекомендации:

1) Для уменьшения процента ошибок в случае морфологической омонимии и совпадения псевдооснов следует учитывать предпочтения автора текста, то есть выбор лексемы должен быть основан не только на частоте употребления лексем в русском языке, но и частоте употребления в анализируемом тексте. Таким образом, так же учитывается и тематика текста.

2) В случае отсутствия лексемы в используемых для анализа словарях необходимо добавить анализ найденных словоформ аналитическим методом независимо от словарей.

Список литературы

1. Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие. М.: МИЭМ, 2011. 272 с.
2. Вороной С.М. Егошина А.А. Обобщенный алгоритм работы со словарем при морфологическом анализе // Искусственный интеллект. 2007. № 3. С. 303-308.
3. Большакова Е.И., Васильева Н.Э., Морозов С.С. Лексико-синтаксические шаблоны для автоматического анализа научно-технических текстов // Десятая Национальная конференция по искусственному интеллекту с международным участием КИИ-2006 : труды. В 3 т. М.: Физматлит, 2006. Т. 2. С. 506-524.
4. Крылов С.А. Некоторые уточнения к определениям понятий словоформы и лексемы // Семиотика и информатика : сб. науч. ст. М.: ВИНТИ, 1982. вып. 19. С. 118-136.
5. Тихомирова Е.А. Минимизация ошибок идентификации лексем в текстах, написанных на естественном языке// Вестник МГТУ им. Н.Э. Баумана. Сер. Приборостроение. 2012. Спец. вып. Информатика и системы управления. С. 131-139.
6. Архив форума «Говорим по-русски». Парадигма // <http://speakrus.ru/> . Режим доступа: <http://speakrus.ru/dict/#paradigma> (дата обращения 13.04.2012).
7. Архив форума «Говорим по-русски». Про-Линг // <http://speakrus.ru/> . Режим доступа: <http://speakrus.ru/dict/#proling> (дата обращения 20.04.2012).
8. Шаров С.А. Частотный словарь русского языка. Режим доступа: <http://www.artint.ru/projects/frqlist.asp> (дата обращения 20.04.2012).

Estimation of working precision of the method of automated identification of lexical items by using virtual texts

08, August 2013

DOI: 10.7463/0813.0583389

Tihomirova E.A., Pavlov Y.N., Kaspruk V.V.

Bauman Moscow State Technical University, 105005, Moscow, Russian Federation

elizarti@mail.ru

pavlov@bmstu.ru

kasper.sochi@mail.ru

This article describes a method for analyzing working precision of algorithms of automated identification of lexical items of natural language word forms on the basis of virtual texts. Estimation of working precision of this complex automated method was carried out with lexical items of Russian word forms. Expediency of applying this comprehensive solution to the problem was proved. The nature of identification errors was revealed, and on the basis of the results of the conducted analysis recommendations were given on modification of the proposed method in order to increase accuracy of automated identification of Russian lexical items.

Publications with keywords: [morphological analysis](#), [lexeme](#), [wordform](#), [the automated identification lexemes of wordforms](#)

Publications with words: [morphological analysis](#), [lexeme](#), [wordform](#), [the automated identification lexemes of wordforms](#)

References

1. Bol'shakova E.I., Klyshinskiy E.S., Lande D.V., Noskov A.A., Peskova O.V., Yagunova E.V. *Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i komp'yuternaya lingvistika* [Automatic processing of natural language text and computational linguistics]. Moscow, MIEM Publ., 2011. 272 p.
2. Voronoy S.M. Egoshina A.A. Obobshchennyy algoritm raboty so slovarem pri morfologicheskom analize [The generalized algorithm of work with dictionary at morphological analysis]. *Iskusstvennyy intellekt*, 2007, no. 3, pp. 303-308.
3. Bol'shakova E.I., Vasil'eva N.E., Morozov S.S. Leksiko-sintaksicheskie shablony dlya avtomaticheskogo analiza nauchno-tekhnicheskikh tekstov [Lexico-syntactic templates for

automatic analysis of scientific and technical texts]. *Desyataya Natsional'naya konferentsiya po iskusstvennomu intellektu s mezhdunarodnym uchastiem KII-2006 : trudy. V 3 t.* [Tenth National Conference on Artificial Intelligence with International Participation CAI-2006: proceedings. In 3 vols]. Moscow, Fizmatlit, 2006, vol. 2, pp. 506-524.

4. Krylov S.A. Nekotorye utochneniya k opredeleniyam ponyatiy slovoformy i leksemy [Some refinements to definition of notions of word-form and lexeme]. *Semiotika i informatika : sb. nauch. st.* [Semiotics and informatics : collected scientific articles]. Moscow, VINITI, 1982, iss. 19, pp. 118-136.

5. Tikhomirova E.A. Minimizatsiya oshibok identifikatsii leksem v tekstakh, napisannykh na estestvennom yazyke [Minimization of errors of identification of lexemes in texts written in natural language]. *Vestnik MGTU im. N.E. Baumana. Ser. Priborostroenie* [Herald of the Bauman MSTU. Ser. Instrument Engineering], 2012, Spets. vyp. *Informatika i sistemy upravleniya* [Spec. iss. Informatics and control systems], pp. 131-139.

6. Arkhiv foruma "Govorim po-russki". *Paradigma* [Archive forum "Speak in Russian". Paradigm]. Website <http://speakrus.ru> . Available at: <http://speakrus.ru/dict/#paradigma> , accessed 13.04.2012.

7. Arkhiv foruma "Govorim po-russki". *Pro-Ling* [Archive forum "Speak in Russian". Pro-Ling]. Website <http://speakrus.ru> . Available at: <http://speakrus.ru/dict/#proling> , accessed 20.04.2012.

8. Sharov S.A. *Chastotnyy slovar' russkogo yazyka* [The frequency dictionary for Russian]. Available at: <http://www.artint.ru/projects/frqlist.asp> , accessed 20.04.2012.