УДК 004.912

А.В. СМИРНОВ, В.М. КРУГЛОВ, А.А. КРИЖАНОВСКИЙ,
Н.Б. ЛУГОВАЯ, А.А. КАРПОВ, И.С. КИПЯТКОВА

КОЛИЧЕСТВЕННЫЙ АНАЛИЗ ЛЕКСИКИ РУССКОГО WORDNET И ВИКИСЛОВАРЕЙ*

Смирнов А.В., Круглов В.М., Крижановский А.А., Луговая Н.Б., Карпов А.А., Кипяткова И.С. **Количественный анализ лексики русского WordNet и викисловарей.**
Аннотация. В работе выполнен количественный анализ лексики русского языка по данным тезауруса Русский WordNet и двух электронных словарей (Русский Викисловарь и Английский Викисловарь). Сравнивается объём словарей и распределение слов русского языка по частям речи. Приводится соотношение многозначных слов и слов с одним значением, а также распределение русских слов по числу значений. Анализ распределения числа значений выявил проблему Английского Викисловаря – отсутствие или недостаточная проработка многозначных русских слов с числом значений больше четырёх (по сравнению со словами Русского Викисловаря). Эксперименты показывают, что лингвистические ресурсы, созданные энтузиастами, демонстрируют те же закономерности, что и традиционные словари.
Ключевые слова: вычислительная лингвистика, лексикография, лексический анализ, русский язык.

Smirnov A.V., Kruglov V. M., Krizhanovsky A.A., Lugovaya N.B., Karpov A.A., Kipyatkova I.S.
A quantitative analysis of the lexicon in Russian WordNet and Wiktionaries.
Abstract. A quantitative analysis of the Russian lexicon was performed in the paper. The thesaurus Russian WordNet and two electronic dictionaries are under examination: the Russian Wiktionary and the English Wiktionary. It was calculated the quantity of Russian words and meanings (senses). The distribution of words for each part of speech, the quantity of monosemous and polysemous words and the distribution of words by number of meanings were calculated and compared across these dictionaries. The analysis of the distribution of words by number of meanings revealed a problem that too few or no polysemy Russian words with number of meanings more than 4 are presented in the English Wiktionary (in comparison with the Russian Wiktionary). The analysis shows that the average polysemy, the number and the distribution of word senses follow similar patterns in both expert and collaborative resources with relatively minor differences.
Keywords: computational linguistics, lexicography, lexical analysis, Russian language.

1. Введение. Одноязычный толковый словарь – стремящееся к полноте комплексное лингвистическое описание словарного состава –

* Данный текст является препринтом статьи: Смирнов А.В., Круглов В.М., Крижановский А.А., Луговая Н.Б., Карпов А.А., Кипяткова И.С. Количественный анализ лексики русского WordNet и викисловарей // Труды СПИИРАН. 2012. Вып. 23. С. 231–253. http://www.mathnet.ru/php/archive.phtml?wshow=paper&jrnid=trspy&paperid=544&option_lang=rus

можно рассматривать как материал для изучения картины мира народа, говорящего на данном языке. В связи с анализом внутренних закономерностей построения лексической системы новые возможности может предложить машиночитаемый словарь, имеющий приложение в виде корпуса текстов, снабженный поисковой системой и являющийся, таким образом, хорошей базой для построения лингвистических моделей, проверки языковых закономерностей методами корпусной и квантитативной лингвистики.

В работе на основе двух больших онлайн-словарей, преобразованных в машиночитаемую форму (Русский Викисловарь и Английский Викисловарь), выполнена численная оценка количественных параметров лексики, проведен сопоставительный анализ названных словарей.

Викисловарь – это свободно пополняемый многофункциональный многоязычный словарь и тезаурус. В викисловаре[†] содержатся толкования и переводы слов, описание их фонетических и морфологических свойств, указания на системные связи слов внутри словарного состава[‡], на правила произношения и разбиения слов на слоги (см. транскрипцию и аудиофайлы). Кроме того, здесь представлены этимологические справки, цитаты из литературных произведений, иллюстрирующие употребление слов, и даже видео и фотографии, иллюстрирующие значения слов в прямом смысле.

Машиночитаемый тезаурус, построенный на основе данных викисловаря, занимает промежуточное положение между информационно-поисковыми тезаурусами и тезаурусами типа WordNet, описанными в работе Лукашевича [10], поскольку в викисловарях указаны парадигматическое отношение (например, синонимия) между значением слова и словарной статьёй, а не между значениями или между словами. Эта особенность викисловарей обсуждалась в предыдущей работе [6].

Популярности викисловарей способствует то обстоятельство, что они находятся в открытом доступе и содержат огромную базу данных

[†] Здесь и далее название конкретного проекта (Английский Викисловарь, Русский Викисловарь) пишется с заглавной буквы, название вообще словарей данного типа, т.е. викисловарей, пишется с маленькой буквы.

[‡] В среде редакторов Русского Викисловаря для обозначения отношений синонимии, антонимии, гиперонимии, холонимии и др. используют термин «Семантические отношения» (в Английском Викисловаре – “Semantic relations”)

с толкованиями слов, тезаурусом, с переводами на многие языки и другой лексикографической информацией. Наиболее привлекательными чертами словаря являются его многоязычность, огромный объём данных и высокая скорость пополнения.

Затруднительно сравнивать другие словари с Викисловарём, поскольку любые сравнения быстро устаревают. Например, в работе [11] был произведен сопоставительный анализ словаря PanDictionary с данными Викисловаря за 2008 год, когда в последнем было всего 403 413 переводов с английского на языки мира. Двумя годами позднее, в 2010, Английский Викисловарь содержал уже в два раза больше переводов (964 019)[§]. При этом Викисловарь растёт не только по числу переводов, но и по числу охватываемых языков. На конец 2011 года в Английском Викисловаре представлены переводы с английского на 274 языка, в нём содержатся словарные статьи о словах из, примерно, 800 языков.

Больша́я статья немецких учёных [15] посвящена сравнению трёх Викисловарей: английского, немецкого и русского. Здесь оценивается полнота словарей, отмечается неожиданно большой[**] объём неологизмов в словаре, количество словарных статей по частям речи, степень пересечения словарей, т.е. наличия общих слов с другими словарями. Сравниваются между собой словари по тому, какие значения слов и какое количество значений в них представлено, оценивается количество слов с одним и несколькими значениями.

В данной работе анализируется ещё один словарь-тезаурус. Это Русский WordNet. Словарь является закрытой разработкой (т.е. недоступен онлайн) и статистические данные по словарю, используемые для сравнения словарей в данной работе, были получены благодаря одной зарубежной публикации [15].

Очертим круг задач, связанных с автоматической обработкой текста, где используются данные Викисловаря:

- автоматизированный перевод:
    1. *машинный перевод* между нидерландским и бурским языками [16];

[§] См. http://en.wiktionary.org/wiki/User:AKA_MBG/Statistics:Translations

[**] Это неожиданное утверждение планируется проверить в следующей нашей работе, посвящённой автоматическому извлечению контекстных помет, т.е. условных сокращений (например, «устар.», «неол.»).

2. добавления переводов к единицам хранения (организованным в таксономии) в электронном каталоге на английском языке (на примере данных Английского, Немецкого и Венгерского викисловарей) [2], используются семантические категории викисловарей;
3. пословный перевод (lemmatic translation) между таким языками, как: английский, японский, русский и словенский; был использован словарь PanDictionary, включающий викисловари вместе с более чем 600 машиночитаемыми двуязычными словарями [20];

- обработка текста парсером NULEX, где используется интеграция части данных Викисловаря (времена глаголов) с базой данных WordNet и VerbNet [12];
- распознавание и синтез речи, где Викисловарь выступает в роли источника данных для автоматического построения словаря произношений [18], [19];
- отображение онтологий [9];
- в качестве источника данных при извлечении семантических отношений (из толкований слов в Викисловаре) [17];
- в качестве источника данных для построения баз знаний, например Concept Net[††];
- для оценки адаптированных (упрощённых) текстов. В работе [13] по тексту словарной статьи из Английского Викисловаря подсчитывается количество омонимов, число переводов, длина толкования для данного слова. Эти числа позволяют построить для текста трёхмерный вектор и различить тексты стандартной Английской Википедии и Википедии на упрощённом английском языке (Simple English Wikipedia).

Статья имеет следующую структуру. Во второй главе оценивается объём рассматриваемых словарей и покрытие слов в списках Сводеша для русского языка и списка Штейнфельдт. В третьей главе обсуждается распределение слов русского языка по частям речи. В четвёртой главе затрагивается вопрос многозначности, приводится соотношение многозначных слов и слов с одним значением – как в целом, так и для каждой части речи отдельно. В пятой главе представлено распределение русских слов по числу значений.

[††] См. http://conceptnet5.media.mit.edu

2. Эксперименты: размер словаря, покрытие словарей. В этом разделе сравниваются размеры словарей, оценивается степень покрытия словарями списков слов Сводеша и Штейнфельдт.

Здесь и далее рассматриваются три словаря:

- Английский Викисловарь (En), версия от 8 октября 2011 г.;
- Русский Викисловарь (Ru), версия от 21 мая 2011 г.;
- Русский WordNet, разработанный компанией «Новософт» [4], данные по данному тезаурусу представлены в работе [15].

В многоязычных словарях (Русский Викисловарь и Английский Викисловарь) в этой статье будут учитываться только словарные статьи, описывающие русские слова. Анализ слов английского языка в этих словарях представлен в предыдущей работе [5].

Разрабатываемая система автоматического анализа, разбора и преобразования текстов Викисловаря в машиночитаемый вид (далее кратко – парсер Викисловаря, *wikt_parser*) – это один из нескольких инструментов, предназначенных для обработки данных Викисловаря. Среди других программ можно отметить парсер Zawilinski (обрабатывает польские слова в Английском Викисловаре) [8], систему JWKTL (работает с английской и немецкой версиями Викисловаря) [‡‡]. Разрабатываемый парсер *wikt_parser* преобразует корпус текстов Викисловаря в машиночитаемый словарь и сохраняет результат в базу уже меньшего размера в формате MySQL или SQLite [21]. Таким образом, все последующие расчёты в этой работе были выполнены на основе двух машиночитаемых словарей, построенных по данным Английского Викисловаря и Русского Викисловаря.

В табл. 1 представлены размеры исследуемых словарей и степень покрытия двух списков русских слов: Сводеша и Штейнфельдт. Список слов Сводеша [21] выражает «основной список значений», обязательно представленный в любом языке, например: «рука», «птица», «я». Список Сводеша для русского языка включает 207 слов. Список Штейнфельдт[§§] содержит 2500 наиболее употребительных слов русского литературного языка.

[‡‡] См. http://www.ukp.tu-darmstadt.de/software/jwktl/

[§§] http://ru.wiktionary.org/wiki/Приложение:Список Штейнфельдт

Табл. 1. **Число русских слов в словарях, покрытие списков слов**

Русские слова	Русский Викисловарь		Русский WordNet [15]		Английский Викисловарь	
		%		%		%
всего	135 396***	100	130 062	100	16 654†††	100
с пустыми значениями	81 761‡‡‡	60,4	0	0	928§§§	5,6
со значениями	53 635	39,6	130 062	100	15 726	94,4
Список Сводеша	100%****		84,4%		93,7%††††	
Штейнфельдт	100%‡‡‡‡		67,9%		–	

Заметим, что в строках «список Сводеша» и «список Штейнфельдт» для викисловарей оценивалось только наличие словарной статьи, в независимости от присутствия в статье толкования.

Эти данные позволили визуально представить относительное число русских слов в двух Викисловарях (рис. 1). Наглядно видно, что в Английском Викисловаре мало словарных статей о русских словах без толкований (5.57%) по сравнению с количеством статей-заготовок в Русском Викисловаре (60.39% статей без толкований). Тем не менее, в Русском Викисловаре представлено почти в 3.4 раза больше словарных статей с толкованиями для русских слова, чем в Английском Викисловаре: 53.6 тысячи против 15.7 тысяч.

*** См. http://ru.wiktionary.org/wiki/User:AKA_MBG/Статистика:Семантические_отношения

††† См. http://en.wiktionary.org/wiki/User:AKA_MBG/Statistics:Semantic_relations

‡‡‡ См. http://ru.wiktionary.org/wiki/User:AKA_MBG/Статистика:POS

§§§ См. http://en.wiktionary.org/wiki/User:AKA_MBG/Statistics:POS

**** См. http://ru.wiktionary.org/wiki/Приложение:Списки_Сводеша

†††† См. http://en.wiktionary.org/wiki/Appendix:Russian_Swadesh_list

‡‡‡‡ См. http://ru.wiktionary.org/wiki/Приложение:Список_Штейнфельдт

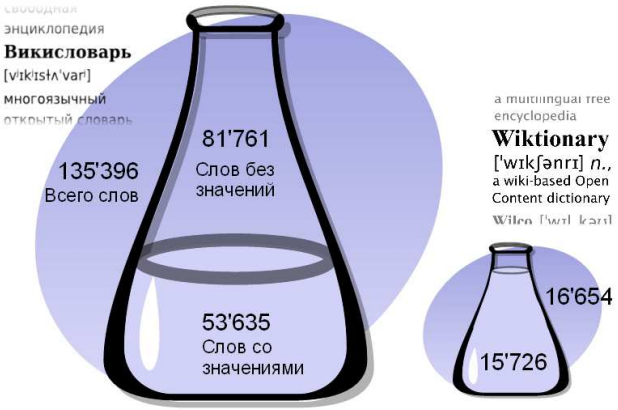


Рис. 1. Относительное число русских слов в Русском Викисловаре (слева) и в Английском Викисловаре (справа).

3. Эксперименты: части речи. Этот раздел отвечает на вопрос – в каком объёме представлены разные части речи в словарях.

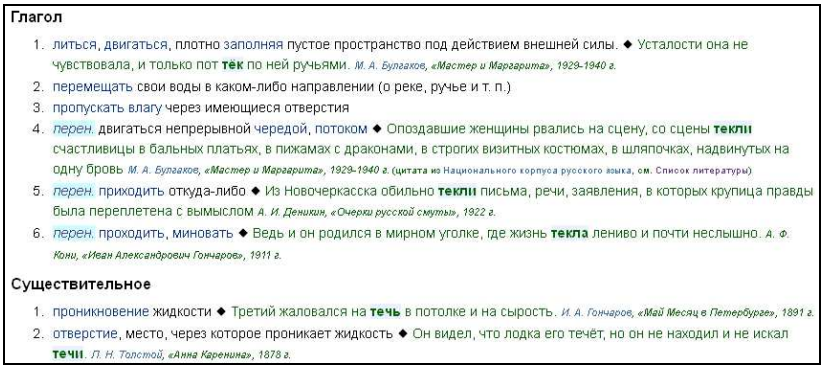
В табл. 2 приводится число русских слов и значений в словарях. Больше всего лексем в тезаурусе Русский WordNet, но нет информации о числе пар слово-значение в отличие от классического принстонов-ского тезауруса WordNet.

Табл. 2. **Число русских слов и значений по частям речи**[§§§§]

Часть речи	Лексема (уникальная строка)			Общее число пар (слово-значение)	
	Русский Викисл.	Русский WordNet	Англ. Викисл.	Русский Викисл.	Англ. Викисл.
Существ.	22 281	**97 257**	9 232	**35 320**	13 197
Глагол	6 654	8 995	1 508	16 637	2 996
Прилагат.	5 288	16 087	1 895	9 069	2 768
Наречие	1 215	3 885	636	1 805	894
Фраза	2409	0	117	2 768	122
Аббревиатура	151	0	0	258	0
Местоимение	109	0	146	180	206
Междометие	93	0	167	138	219
Числительное	65	0	72	96	79
Предлог	43	0	78	114	141
Союз	37	0	51	70	67
Частица	36	0	31	71	44
Имя собств.	0	0	1366	0	1514
Другие	0	0	**311**	0	528
Всего*	**38 381**	**126 224**	**15 610**	**66 526**	**22 775**

Поясним табл. 2. Одна словарная статья в Викисловаре (рис. 2) может включать несколько омонимов, например, частеречные омонимы: «течь» (протекать) и «течь» (протекание). В этом случае вклад словарной статьи «течь» в табл. 2 будет равен одному глаголу и одному существительному (*лексемы* – левая часть табл. 2), шести значениям для глаголов и двум – для существительных (*значения* – правая часть табл. 2).

[§§§§] В Викисловаре к так называемому «заголовку третьего уровня» (он же тип “*Part of speech (POS)*”) относятся не только настоящие части речи (сущ., гл., …), но и такие понятия, как: приставка, суффикс, фразеологизмы и т.д. См. http://en.wiktionary.org/wiki/Wiktionary:POS



Глагол

1. литься, двигаться, плотно заполняя пустое пространство под действием внешней силы. ◆ Усталости она не чувствовала, и только пот **тёк** по ней ручьями. *М. А. Булгаков, «Мастер и Маргарита», 1929-1940 г.*
2. перемещать свои воды в каком-либо направлении (о реке, ручье и т. п.)
3. пропускать влагу через имеющиеся отверстия
4. *перен.* двигаться непрерывной чередой, потоком ◆ Опоздавшие женщины рвались на сцену, со сцены **текли** счастливицы в бальных платьях, в пижамах с драконами, в строгих визитных костюмах, в шляпочках, надвинутых на одну бровь *М. А. Булгаков, «Мастер и Маргарита», 1929-1940 г.* (цитата из Национального корпуса русского языка, см. Список литературы)
5. *перен.* приходить откуда-либо ◆ Из Новочеркасска обильно **текли** письма, речи, заявления, в которых крупица правды была переплетена с вымыслом *А. И. Деникин, «Очерки русской смуты», 1922 г.*
6. *перен.* проходить, миновать ◆ Ведь и он родился в мирном уголке, где жизнь **текла** лениво и почти неслышно. *А. Ф. Кони, «Иван Александрович Гончаров», 1911 г.*

Существительное

1. проникновение жидкости ◆ Третий жаловался на **течь** в потолке и на сырость. *И. А. Гончаров, «Май Месяц в Петербурге», 1891 г.*
2. отверстие, место, через которое проникает жидкость ◆ Он видел, что лодка его течёт, но он не находил и не искал **течи**. *Л. Н. Толстой, «Анна Каренина», 1878 г.*

Рис. 2. Фрагмент словарной статьи для слова «течь» из Русского Викисловаря

Табл. 2 наглядно показывает, что в Викисловарях представлены все части речи, в отличие от Русского WordNet, в том числе: числительное, местоимение, предлог, союз, частица и междометие. В строке «Другие» табл. 2 в Английском Викисловаре указано 311 статей, которые соответствуют таким отдельно выделяемым классам словарных статей, как: акроним, приставка, аффикс, суффикс, детерминатив, буква, слог Пиньинь, символ, причастие, предикатив, идиома, пословица или поговорка.

Для строки «Имя собственное» (Proper noun) в табл. 2 для Русского Викисловаря указан ноль, для Английского Викисловаря – 1366 словарных статей о русских именах собственных. Из этого отнюдь не следует, что в Русском Викисловаре нет имён собственных. Например, в обоих словарях есть словарные статьи «Москва» (Рис. 3). Техническая разница только в том, что

- в статье Английского Викисловаря указан заголовок третьего уровня “Proper Noun”,
- в статье Русского Викисловаря информация о том, что это имя собственное, спрятана в шаблон *{{топоним|}}*.

Парсер Русского Викисловаря ориентируется на фрагмент шаблона «*{{сущ*» и, следовательно, распознаёт слово как существительное (табл. 3).

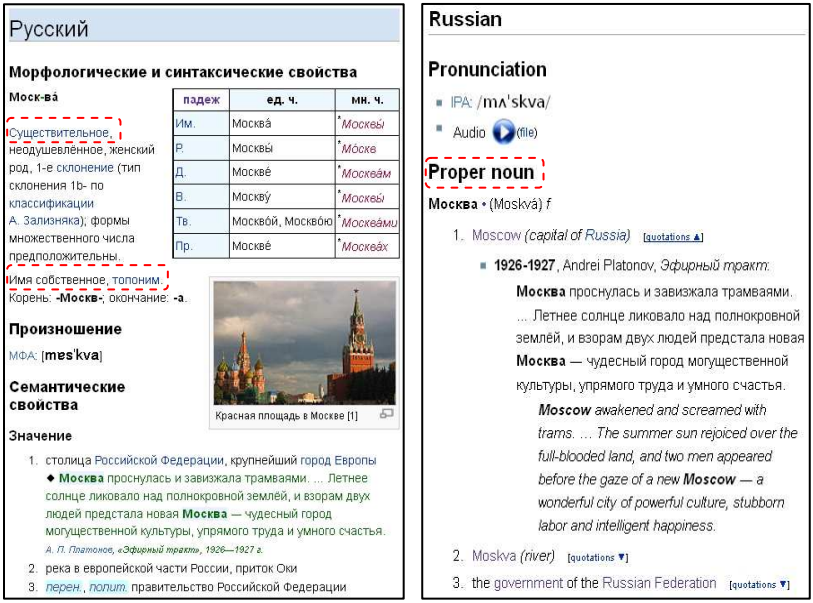
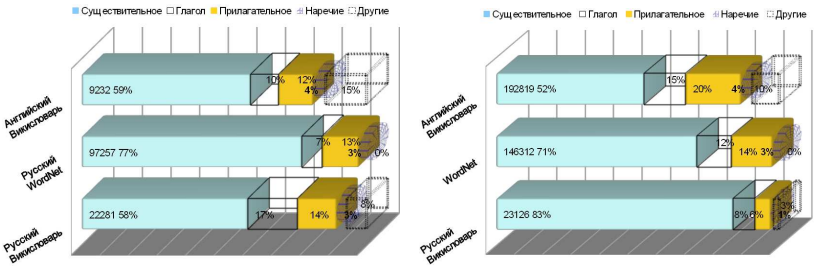


Рис. 3. Фрагмент словарной статьи «Москва» в Русском Викисловаре (слева) и в Английском Викисловаре (справа), пунктиром выделено указание части речи и имени собственного

Табл. 3. **Различие в представлении лексикографической информации в двух викисловарях**

	Русский Викисловарь	Английский Викисловарь
Словарная статья	Москва	Москва
Фрагмент текста статьи, содержащего информацию о части речи (в режиме редактирования статьи)	= {{-ru-}} = ===Морфологические и синтаксические свойства === {{**сущ** ru f ina 1b- |основа=Москв |основа1=Мо́скв |слоги={{по-слогам|Моск|ва́}} }} {{топоним}}	==Russian== ===**Proper noun**=== {{ru-proper noun| tr=Moskvá|g=f}}
Парсер извлёк “Part of speech (POS)”	pos = сущ	pos = proper noun
Результат в машиночитаемом Викисловаре для слова Москва	таблица: part_of_speech id = 38 name = noun	таблица: part_of_speech id = 59 name = proper noun

Относительное распределение русских и английских слов по разным частям речи представлено на рис. 4. Левая часть рис. 4 для слов русского языка отображает данные той же табл. 2, но уже в процентном соотношении. Правая часть рис. 4 – распределение английских слов по частям речи – получена в предыдущей работе [5].



а) русский язык б) английский язык

Рис. 4. Относительное распределение а) *русских слов* по частям речи в Английском Викисловаре, Русском WordNet'е и Русском Викисловаре, б) распределение *английских слов* в Английском Викисловаре, WordNet'е и Русском Викисловаре по данным за 2011 г. [5]

Проанализируем данные рис. 4. Если считать наиболее представительным из этих словарей для английского языка Английский Викисловарь (по объёму), а для русского языка - Русский Викисловарь (по числу частей речи), тогда обнаруживается следующая закономерность:

> *В большем по объёму и более проработанном словаре — деление слов по частям речи является более равномерным. Чем более полон словарь, тем меньшую долю в нём занимают собственно существительные и большую долю начинают занимать другие части речи.*

Рассмотрим долю существительных в словаре для разных языков. Английские существительные в Английском Викисловаре составляют 52% против 71% в WordNet. Русские существительные в Русском Викисловаре составляют 58% против 77% в Русском WordNet. Чем меньшую долю занимают существительные, тем в большей пропорции представлены другие части речи.

Отметим закономерный факт сходства пропорции разных частей речи в WordNet и Русском WordNet на рис. 4, что можно объяснить построением Русского WordNet на основе принстоновского. Сравнение

же викисловарей (друг с другом и с WordNet) значительно сложнее, так как:

1. существует непропорциональное представление языка в том викисловаре, для которого данный язык является главным относительно представления этого же языка в любом другом из викисловарей. Например, английских слов в Английском Викисловаре больше, чем английских слов в Русском Викисловаре в 12.6 раза, а значений в 13.3 раза (на 2011 г., см. табл. 1 в [5]). Русских слов больше в Русском Викисловаре, чем в Английском Викисловаре в 2.5 раза, значений больше в 2.9 раза (табл. 2). Т.е. язык, являющийся родным (главным) для викисловаря, описан в нём полнее, в данном случае – русский язык является «главным» для Русского Викисловаря и представлен большим числом лексем и значений, чем в Английском Викисловаре.
2. крайне мало́ пересечение множества лексем двух словарей, например только 11% лексем Английского Викисловаря (данные за 2011 год) было найдено среди лексем WordNet [15].

4. Эксперименты: многозначность. Важной характеристикой словаря являются: доля многозначных слов по отношению к количеству однозначных слов, среднее число значений у многозначных слов.

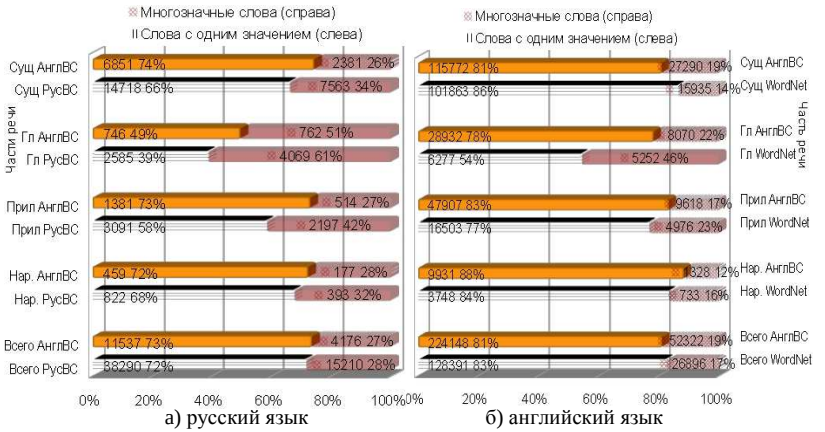
В табл. 4 приведены данные о количестве русских слов в викисловарях с одним значением и с несколькими значениями (информация по Русскому WordNet недоступна). Данные в таблице приведены для (1) каждой части речи (о «заголовках третьего уровня» в викисловарях, например «фразы», см. выше) и (2) суммарно для всех слов.

Интересно отметить, что русских многозначных глаголов в полтора раза больше, чем с одним значением в Русском Викисловаре (4069 и 2585 слов соответственно). Для остальных частей речи преобладают слова с одним значением.

Табл. 4. **Многозначность русских слов в Русском Викисловаре и в Английском Викисловаре**

Часть речи и другие заголовки «третьего уровня» викисловарей	Число слов с одним значением		Число многозначных слов	
	Русский Викисловарь	Английский Викисловарь	Русский Викисловарь	Английский Викисловарь
Существ.	14 718	6 851	7 563	2 381
Глагол	**2 585**	746	**4 069**	762
Прилагат.	3 091	1 381	2 197	514
Наречие	822	459	393	177
Междометие	59	129	34	38
Местоимение	79	99	30	47
Числительное	47	70	18	2
Фраза	2 139	112	270	5
Другие	14 750	1 690	622	250
Всего	**38 290**	11 537	**15 210**	4 176

Левая часть рис. 5 (русские слова) построена на основе данных табл. 4, правая часть (английские слова) – на основе предыдущей работы [5]. Рис. 5 показывает, что оба языка в трёх словарях (Русский Викисловарь, Английский Викисловарь и WordNet) содержат больше слов с одним значением, чем с несколькими, всего 72-73% русских слов с одним значением, 81-83% английских слов с одним значением. Этот рисунок позволяет обнаружить интересную закономерность, а именно: для каждой части речи доля многозначных слов больше для русского языка, чем для английского в этих словарях.



а) русский язык б) английский язык

Рис. 5. Относительное число однозначных и многозначных русских слов (слева) и английских слов (справа) *по частям речи* в Русском Викисловаре (РусВС), Английском Викисловаре (АнглВС) и WordNet, данные за 2011 год.

В табл. 5 приведено среднее число значений:

- с учётом слов с одним значением (левая часть табл. 5, рис. 6а);
- без учёта слов с одним значением, т.е. только для многозначных слов (правая часть табл. 5, рис. 6б).

Рис. 6 позволяет сделать некоторые выводы:

- Наиболее многозначными являются русские глаголы (верхняя кривая на обоих рисунках), без учёта слов с одним значением среднее число значений у глаголов равно трём и более, а именно: 2.95 и 3.45 в Английском Викисловаре и Русском Викисловаре соответственно (рис. 6б).
- Меньше всего значений (нижняя кривая на обоих рисунках) у русских наречий, причём средние значения в обоих викисловарях оказались достаточно близкими: 1.41 и 1.49 (рис. 6а), 2.46 и 2.5 (рис. 6б).

Табл. 5. **Среднее число значений у многозначных русских слов в Русском Викисловаре и в Английском Викисловаре**

Среднее значение полисемии (учитывая слова с одним знач.)			Среднее значение полисемии (без учёта слов с одним знач.)		
Часть речи	Русский Викисл.	Английский Викисл.	Часть речи	Русский Викисл.	Английский Викисл.
Сущ + 1	1,59	1,43	Сущ - 1	2,72	2,67
Гл + 1	2,50	1,99	Гл - 1	3,45	2,95
Прил + 1	1,72	1,46	Прил - 1	2,72	2,70
Нареч+1	1,49	1,41	Нареч -1	2,50	2,46

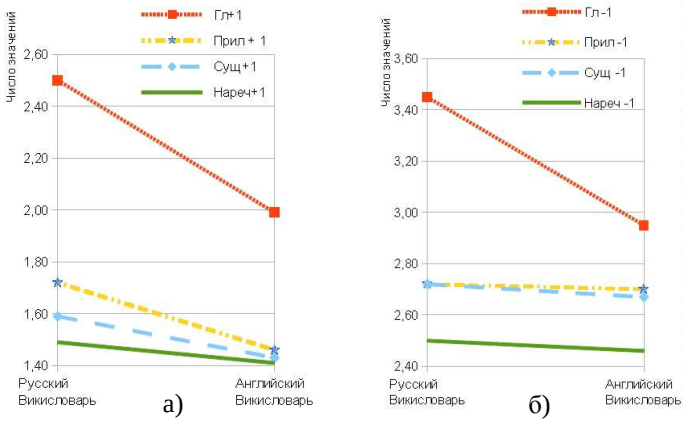


Рис. 6. Среднее число значений для русских слов
(а) с одним и более значениями, (б) с двумя и более значениями.

Сравнение с аналогичными графиками для слов английского языка в работе [5] указывает на то, что в викисловарях и WordNet больше всего значений (в среднем, для обоих языков) представлено у глаголов, меньше всего – у наречий.

5. Эксперименты: распределение слов по числу значений. В Для обоих викисловарей (кроме Русского WordNet) было построено распределение русских слов по числу значений, т.е. подсчитано число слов без толкований (с нулём значений), число слов с одним значением, число слов с двумя значениями и т.д. Фрагменты двух таблиц распределения значений доступны в Интернете и для Русского Викисловаря[*****], и для Английского Викисловаря[†††††].

На рис. 7 представлено распределение слов по числу значений. На этом рисунке представлены данные по Русскому Викисловарю только до слов с 17 значениями (например, глаголы: «*поставить*» – 16 значений, «*развернуть*» – 17), по Английскому Викисловарю – до слов с 13 значениями (например, существительные: «*ход*» – 13 значений, «*рожок*» – 11), т.к. (1) при бóльших числах начинают встречаться нулевые значения (т.е. таких слов нет в словаре), непригодные для аппроксимации, (2) для больши́х словарных статей, описывающих много значений, не всегда точно получается подсчитать число значений, вероятно, из-за того, что авторы таких статей в виду разных причин (например, удобство подачи материала) отклоняются от жёсткого формата Викисловаря, на который настроен наш парсер. Например, в статье Английского Викисловаря о русском предлоге «*по*» – значения разбиты на подзначения, что не отражено в правилах оформления статей Викисловаря.

Распределение значений русских слов в Русском Викисловаре хорошо аппроксимируется *степенной* функцией, квадрат коэффициента корреляции Пирсона (R^2) равен 0.96 (рис. 7). Однако распределение значений русских слов в Английском Викисловаре лучше аппроксимируется *экспоненциальной* функцией, R^2 равен 0.91. Распределение значений английских слов в обоих викисловарях отлично аппроксимируется *степенными* функциями (R^2 равно 0.99 [5]). Таким образом, из четырёх вариантов (два языка в двух словарях) выделяется один – распределение значений русских слов в Английском Викисловаре. В этом

[*****] См. http://ru.wiktionary.org/wiki/Участник:AKA_MBG/Статистика:POS

[†††††] См. http://en.wiktionary.org/wiki/User:AKA_MBG/Statistics:POS

варианте можно отметить (нижняя кривая на рис. 7) более стремительное падение числа слов с большим числом значений (проблемный диапазон: 5-9 значений), хотя и с небольшим всплеском числа слов с 10-11 значениями. Можно обобщить эту, очевидно, временную проблему Английского Викисловаря так – отсутствие или недостаточная разработанность многозначных русских слов с числом значений больше четырёх по сравнению со словами Русского Викисловаря.

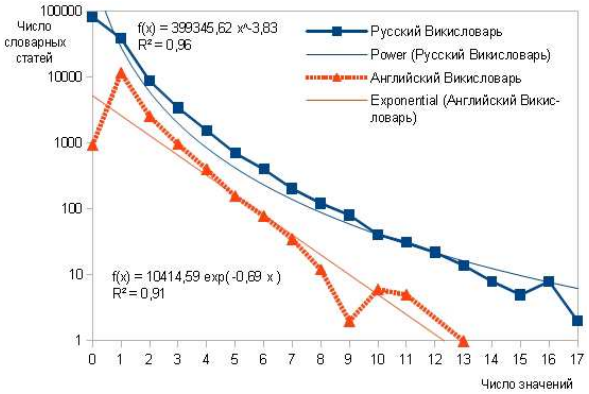


Рис. 7. Распределение русских слов по числу значений в Русском Викисловаре (верхняя кривая, аппроксимация степенной функцией) и в Английском Викисловаре (нижняя кривая, аппроксимация экспоненциальной функцией).

6. Заключение. Для численного анализа лексики русского языка были построены машиночитаемые версии Английского Викисловаря и Русского Викисловаря [21]. В многоязычных викисловарях учитывались только словарные статьи, описывающие русские слова. Третьим использованным словарём был Русский WordNet. Было выполнено:

- *сравнение числа русских слов и значений* в словарях. Больше всего русских слов на 2011 год (т.е. самый большой словник) содержится в Русском Викисловаре: 135 396 слов (табл. 1). Для сравнения – «Словарь современного русского литературного языка» (Большой академический словарь) насчитывает около 120 тыс. слов. В Русском Викисловаре словарных статей о русских словах больше, чем в Русском WordNet в 1.04 раза (табл. 1), однако значений 66 526 – меньше в 2.7 раза, чем Русском WordNet, где их всего 182 448 [15].

- *оценка распределения слов русского языка по частям речи* в Русском Викисловаре, Русском WordNet и Английском Викисловаре. Найдено, что наибольшую долю во всех словарях занимают существительные (58-77%), на втором месте глаголы (7-17%) и прилагательные (12-14%), затем наречия (3-4%) и другие части речи.
- *оценка числа многозначных русских слов и слов с одним значением*. 72% русских слов (или 38 тыс. слов) из всех слов в Русском Викисловаре с непустыми толкованиями и 73% (11,5 тыс) слов в Английском Викисловаре имеют ровно одно значение (рис. 5а). Оказалось, что русских многозначных глаголов в полтора раза больше, чем с одним значением в Русском Викисловаре (4069 и 2585 слов соответственно). Для остальных частей речи преобладают слова с одним значением. Анализ Русского Викисловаря, Английского Викисловаря и WordNet показал, что для каждой части речи доля многозначных слов больше для русского языка, чем для английского.
- *оценка среднего числа значений* для слов, принадлежащих разным частям речи. В викисловарях наиболее многозначными оказались глаголы, среднее число значений у глаголов (без учёта слов с одним значением) равно трём и более, а именно: 2.95-3.45 (рис. 6б). Меньше всего значений у русских наречий, причём средние значения в обоих викисловарях оказались достаточно близкими: 1.41 и 1.49 (рис. 6а), 2.46 и 2.5 (рис. 6б).

Сравнение со средним числом значений слов английского языка в работе [5] указывает на то, что в викисловарях и WordNet больше всего значений (в среднем, для русских и английских слов) представлено у глаголов, меньше всего – у наречий.

Также для Русского Викисловаря и для Английского Викисловаря были вычислены распределения русских слов по числу значений, которые хорошо аппроксимируется степенной и экспоненциальной функциями соответственно. Анализ распределений и аппроксимирующих кривых позволил выявить проблему Английского Викисловаря – отсутствие или недостаточная проработка многозначных русских слов с числом значений больше четырёх (по сравнению со словами Русского Викисловаря).

Анализ русского и английского языка в четырёх словарях (два викисловаря и два WordNet) позволил найти следующую закономерность. В большем по объёму и более проработанном словаре — деление слов по частям речи является более равномерным. Чем более по-

лон словарь, тем меньшую долю в нём занимают, в частности, существительные и большую долю начинают занимать другие части речи.

Полученные результаты для русских слов (количество слов по частям речи – рис. 4а, относительное число однозначных и многозначных слов – рис. 5а, среднее число значений – рис. 6, распределение числа значений русских слов – рис. 7) наглядно показывают последовательность и закономерность в развитии викисловарей от только ещё начинающего своё развитие (в отношении слов русского языка) – Английского Викисловаря, до наиболее проработанного и большого – Русского Викисловаря.

Вслед за авторами статьи [14], а также на основе анализа табл. 2 (число русских слов по частям речи) и рис. 4а (относительное распределение русских слов по частям речи) можно утверждать, что лингвистические ресурсы, созданные энтузиастами, демонстрируют те же закономерности, что и традиционные словари.

При этом необходимо отметить, что, независимо от результатов экспериментов, оценивается только часть слов данного языка, т.к. ни один из рассматриваемых словарей на данный момент не является сколько-нибудь полным. Даже самый большой из этих словарей по размеру словника, а именно – Русский Викисловарь, содержит 82 тыс. словарных статей с пустыми, т.е. пока что незаполненными толкованиями, что составляет 60% от числа всех статей. При этом быстрый рост числа статей в Викисловаре не говорит однозначно о том, что скоро наступит насыщение, поскольку вместе с увеличением числа «хороших», полноценных словарных статей может увеличиваться и число статей с незаполненными толкованиями. Необходимы дополнительные исследования для анализа динамики роста викисловарей.

Кроме того, интересным продолжением этой работы будет измерение семантического расстояния между разными языками [1]. Английский Викисловарь содержит 83 языка с числом словарных статей больше 1000, а значит, на его основе вполне можно рассчитать и построить карту этих языков.

Нельзя не упомянуть ещё одну важную задачу, связанную с будущим викисловарей. Востребованной задачей является объединение лексикографических данных Викисловарей. В каждом из 170 викисловарей указаны строгие правила оформления словарных статей, чётко прописана структура словарных статей. К сожалению, каждый викисловарь создаётся своим сообществом лексикографов-энтузиастов, что влечёт «разброд и шатание» в правилах и оформлении статей разных словарей, даже при наличии общей цели – «создания свободно попол-

няемого многофункционального многоязычного словаря и тезауруса». Поэтому крайне важной задачей является стандартизация и выработка единых правил для разных викисловарей. Некоторой подвижкой в этом направлении можно считать работу [2], в которой представлен инструментарий, позволяющий преобразовывать статьи английского, немецкого и венгерского языковых версий викисловаря в формат TEI (Text Encoding Initiative), широко используемый в гуманитарных исследованиях, связанных с цифровой обработкой данных.

Литература

1. *Cooper M*. Measuring the Semantic Distance between Languages from a Statistical Analysis of Bilingual Dictionaries // Journal of Quantitative Linguistics, 2008. т. 15. № 1. С. 1-33.
2. *Declerck T., Morth K., Lendvai P*. Accessing and standardizing Wiktionary lexical entries for the translation of labels in Cultural Heritage taxonomies // In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey. May 23-25, 2012. Pp. 2511-2514. <http: // www.lrec-conf.org/proceedings/lrec2012/pdf/820_Paper.pdf> (по состоянию на 23.09.2012)
3. *Esuli A., Sebastiani F*. SentiWordNet: A publicly available lexical resource for opinion mining // In: Proceedings of Language Resources and Evaluation (LREC), 2006.
4. *Гельфейнбейн И. Г., Гончарук А. В., Лехельт В. П., Липатов А. А., Шило В. В.* Автоматический перевод семантической сети WordNet на русский язык // Труды Международного семинара Диалог по компьютерной лингвистике и её приложениям, Протвино, Россия, 2003. <http: // www.dialog-21.ru/Archive/2003/Goncharuk.pdf> (по состоянию на 23.09.2012)
5. *Крижановский А. А.* Количественный анализ лексики английского языка в викисловарях и Wordnet // Труды СПИИРАН. 2011. Вып. 19. С. 87–101.
6. Крижановский А. А. Машинная обработка Русского Викисловаря. // Викиконференция 2009. 24–25 октября, Санкт-Петербург. <http: // ru.wikipedia.org/wiki/Википедия:Вики-конференция_2009/Программа/Доклады/Машинная обработка Русского Викисловаря > (по состоянию на 23.09.2012)
7. *Krizhanovsky A. A.* Transformation of Wiktionary entry structure into tables and relations in a relational database schema. 2010. <http: // arxiv.org/abs/1011.1368> (по состоянию на 23.09.2012)
8. *Kurmas Z.* Zawilinski: a library for studying grammar in Wiktionary. // In: Proceedings of the 6th International Symposium on Wikis and Open Collaboration, Gdansk, Poland, July 2010.
9. *Lin F., Krizhanovsky A.* Multilingual ontology matching based on Wiktionary data accessible via SPARQL endpoint // In: Proceedings of the 13th Russian Conference on Digital Libraries RCDL'2011. Voronezh, Russia. October, 2011. P.19-26.
10. *Лукашевич Н. В.* Тезаурусы в задачах информационного поиска. // М.: Издательство Московского университета, 2011. 512 с.
11. *Mausam, Soderland S., Etzioni O., Weld D. S., Reiter K., Skinner M., Sammer M., Bilmes J.* Panlingual Lexical Translation via Probabilistic Inference // Artificial Intelligence Journal (AIJ). Vol. 174, No. 9-10, 2010. P.619-637.
12. *McFate C., Forbus K.* NULEX: An Open-License Broad Coverage Lexicon. In: The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA. June, 2011. Vol. 2. Pp. 363-367.
13. Medero J. and Ostendorf M. "Analysis of vocabulary difficulty using wiktionary" // In: Proceedings SLaTE Workshop, 2009. <http: // www.eee.bham.ac.uk/SLaTE2009/papers%5CSLaTE2009-41-v2.pdf (по состоянию на 23.09.2012)>
14. *Meyer C. M., Gurevych I.* How Web Communities Analyze Human Language: Word Senses in Wiktionary // In: Proceedings of the WebSci10: Extending the Frontiers of Society On-Line, Raleigh, NC: US. April, 2010. <http: // journal.webscience.org/349/> (по состоянию на 23.09.2012)
15. *Meyer C. M., Gurevych I.* Wiktionary: a new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography // Electronic Lexicography. Oxford: Ox-

ford University Press. 2012. (to appear). <http: // www.informatik.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2011/oup-elex2012-meyer-wiktionary.pdf> (по состоянию на 23.09.2012)
16. *Otte P., Tyers F. M.* Rapid rule-based machine translation between Dutch and Afrikaans // In: 16th Annual Conference of the European Association of Machine Translation, EAMT11. 2011.
17. *Panchenko A., Adeykin S., Romanov P., Romanov A.* Extraction of Semantic Relations between Concepts with KNN Algorithms on Wikipedia // In: Concept Discovery in Unstructured Data Workshop (CDUD) of International Conference On Formal Concept Analysis, Belgium, 2012. P.78-88.
18. *Qingyue He*. Automatic Pronunciation Dictionary Generation from Wiktionary and Wikipedia. // Thesis. Karlsruhe Institute of Technology. 2009.
19. *Schlippe T., Ochs S., Schultz T.* Wiktionary as a Source for Automatic Pronunciation Extraction // In: Proceedings of the 11th Annual Conference of Interspeech, Makuhari, Japan, 2010. P.2290–2293
20. *Soderland S., Lim C., Mausam, Bo Qin, Etzioni O., Pool J.* Lemmatic machine translation // In: Proceedings of Machine Translation Summit XII, Ottawa, Canada, 2009.
21. Старостин С. А. Сравнительно-историческое языкознание и лексикостатистика // Лингвистическая реконструкция и древнейшая история Востока. Материалы к дискуссиям на Международной конференции (Москва, 29 мая — 2 июня 1989 г.). Часть 1. М. <http: // altaica.ru/LIBRARY/glotto.pdf> (по состоянию на 23.09.2012)

Смирнов Александр Викторович — д-р техн. наук, проф.; заместитель директора по научной работе Федерального государственного бюджетного учреждения науки Санкт-Петербургского института информатики и автоматизации РАН (СПИИРАН), заведующий лабораторией интегрированных систем автоматизации. Область научных интересов: интеллектуальное управление конфигурациями виртуальных и сетевых организаций, логистика знаний, поддержка принятия решений. Число научных публикаций — 304. Адрес: smir@iias.spb.su; СПИИРАН, 14-я линия В. О., д. 39, Санкт-Петербург, 199178, РФ; р.т. +7(812)328-2073, факс +7(812)328-4450.

Smirnov Alexander Victorovich — D.Sc., Prof.; a Deputy Director for Research and a Head of Computer Aided Integrated Systems Laboratory at St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), a full professor of St.Petersburg State Electrical Engineering University. Research interests: intelligent configuration management of virtual and network organizations, knowledge logistics, decision support. The number of publications — 304. Address: smir@iias.spb.su; SPIIRAS, 39, 14th Line V. O., St. Petersburg, 199178, Russia; tel. +7(812)328-2073, fax: +7(812)328-4450.

Круглов Василий Михайлович — д-р фил. наук, проф.; ведущий научный сотрудник Федерального государственного бюджетного учреждения науки Института лингвистических исследований РАН (ИЛИ РАН), руководитель лаборатории информационных лингвистических технологий. Область научных интересов: русская лексикология и лексикография, компьютерная лингвистика, корпусная лингвистика, электронные картотеки, компьютерная лексикография. Число научных публикаций — 30. Адрес: vmkruglov@yandex.ru; ИЛИ РАН, Тучков переулок, д. 9, Санкт-Петербург, 199053, РФ; р.т. +7(812)328-1612, факс +7(812)328-4611.

Kruglov Vasil Mikhailovich — D.Sc., Prof.; leading senior researcher of Institute for Linguistic Studies of the Russian Academy of Sciences (ILI RAS), a head of Information Linguistics

Technologies laboratory. Research interests: Russian lexicology and lexicography, computational linguistics, corpus linguistics, electronic library catalogues, computational lexicology. The number of publications — 30. Address: vmkruglov@yandex.ru; ILI RAS, 39, Tuchkov pereülok 9, St. Petersburg, 199053, Russia; tel. +7(812)328-1612, fax: +7(812)328-4611.

Крижановский Андрей Анатольевич — к.т.н.; старший научный сотрудник лаборатории интегрированных систем автоматизации Федерального государственного бюджетного учреждения науки Санкт-Петербургского института информатики и автоматизации РАН (СПИИРАН), старший научный сотрудник лаборатории информационных компьютерных технологий Федерального государственного бюджетного учреждения науки Института прикладных математических исследований Карельского научного центра Российской академии наук (ИПМИ КарНЦ РАН). Область научных интересов: автоматическая обработка текста, корпусная лингвистика. Число научных публикаций — 67. andrew.krizhanovsky@gmail.com, code.google.com/p/wikokit; ИПМИ КарНЦ РАН, ул. Пушкинская, д. 11, г. Петрозаводск, 185910, РФ; р.т. +7(8142)76-63-13, факс +7(8142)76-63-13.

Krizhanovsky Andrew Anatoliyevich — PhD; senior researcher, Computer Aided Integrated Systems Laboratory at St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), senior researcher, Laboratory for Information Computer Technologies of Institute of Applied Mathematical Research of the Karelian Research Centre of the Russian Academy of Sciences (IAMR). Research Interest: information retrieval, corpus linguistics. The number of scientific publications — 67. andrew.krizhanovsky@gmail.com, code.google.com/p/wikokit; IAMR KRC RAS, 11, Pushkinskaya str., Petrozavodsk, Karelia, 185910, Russia; phone +7(8142)76-63-13, fax +7(8142)76-63-13.

Луговая Наталья Борисовна — ведущий инженер-программист лаборатории информационных компьютерных технологий Федерального государственного бюджетного учреждения науки Института прикладных математических исследований Карельского научного центра Российской академии наук (ИПМИ КарНЦ РАН). Область научных интересов: разработка информационных систем для поддержки научных исследований и образования с использованием Интернет-технологий. Число научных публикаций — 29. nataly@krc.karelia.ru, http://nataly.krc.karelia.ru; ИПМИ КарНЦ РАН, ул. Пушкинская, д. 11, г. Петрозаводск, 185910, РФ; р.т. +7(8142)76-63-12, факс +7(8142)76-63-13.

Lugovaya Natalia Borisovna – leading programmer, Laboratory for Information Computer Technologies of Institute of Applied Mathematical Research of the Karelian Research Centre of the Russian Academy of Sciences (IAMR). Research Interest: developing of the information system for scientific research and education using internet-technologies. The number of scientific publications — 29. nataly@krc.karelia.ru, http://nataly.krc.karelia.ru; IAMR KRC RAS, 11, Pushkinskaya str., Petrozavodsk, Karelia, 185910, Russia; phone +7(8142)76-63-12, fax +7(8142)76-63-13.

Карпов Алексей Анатольевич — канд. техн. наук, старший научный сотрудник лаборатории речевых и многомодальных интерфейсов Федерального государственного бюджетного учреждения науки Санкт-Петербургского института информатики и автоматизации РАН (СПИИРАН). Область научных интересов: автоматическое распознавание речи, многомодальные интерфейсы, аудиовизуальное распознавание речи. Число научных публикаций — 150. karpov@iias.spb.su; СПИИРАН, 14-я линия В.О., д. 39, Санкт-Петербург, 199178, РФ; р.т. +7(812)328-7081, факс +7(812)328-7081.

Karpov Alexey Anatolyevich — PhD, senior researcher, Laboratory of Speech and Multimodal Interfaces St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS). Research interests: automatic speech recognition, multimodal interfaces, audio-visual speech recognition. The number of publications — 150. karpov@iias.spb.su; SPIIRAS, 39, 14th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-7081, fax +7(812)328-7081.

Кипяткова Ирина Сергеевна — канд. техн. наук, старший научный сотрудник лаборатории речевых и многомодальных интерфейсов Федерального государственного бюджетного учреждения науки Санкт-Петербургского института информатики и автоматизации РАН (СПИИРАН). Область научных интересов: автоматическое распознавание речи, статистические модели языка. Число научных публикаций — 35. kipyatkova@iias.spb.su; СПИИРАН, 14-я линия В.О., д. 39, Санкт-Петербург, 199178, РФ; р.т. +7(812)328-7081, факс +7(812)328-7081.

Kipyatkova Irina Sergeevna — PhD, senior researcher, Laboratory of Speech and Multimodal Interfaces St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS). Research interests: automatic speech recognition statistical language models. The number of publications — 35. kipyatkova@iias.spb.su; SPIIRAS, 39, 14th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-7081, fax +7(812)328-7081.

Поддержка исследований. Работа выполнена при финансовой поддержке РФФИ (проект № 11-01-00251, № 12-01-00481, № 12-07-00070, № 12-08-01265), РГНФ (проект № 12-04-12062), проекта № 213 Программы фундаментальных исследований Президиума РАН «Интеллектуальные информационные технологии, математическое моделирование, системный анализ и автоматизация», проекта № 2.2 Программы ОНИТ РАН «Интеллектуальные информационные технологии, системный анализ и автоматизация», Министерства образования и науки РФ (ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007-2013 годы», госконтракт No. 07.514.11.4139), Совета по грантам Президента РФ (проект МК-1880.2012.8).

Рекомендовано лабораторией интегрированных систем автоматизации, заведующий лабораторией Смирнов А.В., д. т. н., проф.
Статья поступила в редакцию 12.10.2012.

РЕФЕРАТ

Смирнов А.В., Круглов В.М., Крижановский А.А., Луговая Н.Б., Карпов А.А., Кипяткова И.С. **Количественный анализ лексики русского WordNet и викисловарей.**

Разработана компьютерная система анализа лексикографических данных, позволившая выполнить количественный анализ лексики русского языка по данным трёх электронных словарей: Русского Викисловаря, Русского WordNet и Английского Викисловаря. В многоязычных викисловарях (Русский Викисловарь и Английский Викисловарь) учитывались только словарные статьи, описывающие русские слова.

Разработанный программный комплекс, включающий базу данных машиночитаемого Викисловаря, имеет большую научную и практическую значимость: (1) в научных прототипах и приложениях (автоматизация построения онтологий и баз знаний, распознавание значения слова), (2) в офисных приложениях (проверка правописания, перевод), (3) в промышленных системах (выявление интересов пользователей, построение профиля клиента), (4) в системах мониторинга и кластеризация текстовых потоков, выявлении текстов заданной тематики.

Полученные результаты (количество русских слов по частям речи, распределение значений русских слов) наглядно показывают последовательность и закономерность в развитии викисловарей от только ещё начинающего своё развитие (в отношении слов русского языка) – Английского Викисловаря, до наиболее проработанного – Русского Викисловаря.

Построена аппроксимация распределения числа значений русских слов в Русском Викисловаре с помощью степенной функции и в Английском Викисловаре – с помощью экспоненциальной. Анализ распределений и аппроксимирующих кривых позволил выявить временную трудность Английского Викисловаря – отсутствие или недостаточную проработку многозначных русских слов с числом значений больше четырёх (по сравнению со словами Русского Викисловаря).

Анализ словарей показал, что лингвистические ресурсы, созданные как экспертами (WordNet), так и энтузиастами (Викисловарь), подчиняются общим закономерностям (по пропорции распределения слов русского языка по частям речи, по соотношению слов с одним значением к числу многозначных слов, по среднему числу значений, по распределению числа значений русских слов).

SUMMARY

Smirnov A.V., Kruglov V.M., Krizhanovsky A.A., Lugovaya N.B., Karpov A.A., Kipyatkova I.S. **A quantitative analysis of the lexicon in Russian WordNet and Wiktionaries.**

A software system of analysis of lexicographic data was developed. It was used in order to perform a quantitative analysis of the Russian lexicon in the three electronic dictionaries: the Russian Wiktionary, Russian WordNet, and the English Wiktionary. Only Russian entries were taken into account in multilingual dictionaries.

The developed software system has great scientific and practical value. It can be used (1) in scientific prototypes (an automatic building of ontologies and knowledge bases, word sense disambiguation), (2) in office applications (check spelling, translation), (3) in industry systems (user tacit preferences acquire, building of a user profile), (4) in systems of monitoring of texts flow, filtering and extracting of texts related to some topics.

The obtained results (the quantity of Russian words for each part of speech, the distribution of meanings of words) shows clearly that there is succession and regularity in the evolution of Wiktionaries from the English Wiktionary to the most thoroughly developed and huge the Russian Wiktionary (in relation of Russian entries).

The distribution of number of meanings of Russian words in the Russian Wiktionary was approximated using power law function, in the English Wiktionary – exponent law function. The analysis of the distribution of words by number of meanings revealed a problem that too few or no polysemy Russian words with number of meanings more than 4 are presented in the English Wiktionary (in comparison with the Russian Wiktionary).

It was calculated the quantity of Russian words and meanings (senses). The analysis shows that the distribution of words for each part of speech, the quantity of monosemous and polysemous words, the average polysemy, the distribution of word senses number follow similar patterns in both expert and collaborative resources with relatively minor differences.