УДК 004.912

А.А. КРИЖАНОВСКИЙ

КОЛИЧЕСТВЕННЫЙ АНАЛИЗ ЛЕКСИКИ АНГЛИЙСКОГО ЯЗЫКА В ВИКИСЛОВАРЯХ И WORDNET

Крижановский А.А. **Количественный анализ лексики английского языка в викисловарях и Wordnet.**
Аннотация. В работе выполнен количественный анализ лексики английского языка по данным трёх электронных словарей: Английского Викисловаря, WordNet и Русского Викисловаря. Сравнивается объём словарей и распределение слов английского языка по частям речи. Приводится соотношение многозначных слов и слов с одним значением, а также распределение английских слов по числу значений. Эксперименты показывают, что лингвистические ресурсы, созданные как экспертами, так и энтузиастами, подчиняются общим законам.
Ключевые слова: вычислительная лингвистика, лексикография, лексический анализ, английский язык.

Krizhanovsky A.A. **A quantitative analysis of the English lexicon in Wiktionaries and WordNet.**
Abstract. A quantitative analysis of the English lexicon was performed in the paper. The three electronic dictionaries are under examination: the English Wiktionary, WordNet, and the Russian Wiktionary. The quantity of English words and their meanings (senses) are calculated. 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 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, English language.

1. Введение. Богатство языка заключено в его лексиконе, в оттенках и значениях многозначных слов, меняющихся с ходом времени и зависящих от географического положения даже в пределах одной страны. Не случайно, один из видов словарей называют тезаурусом, ведь *thēsaurus* – это слово латинского происхождения, одним из значений которого является *сокровище*, *драгоценность*.

С появлением больших электронных словарей, содержащих десятки и сотни тысяч словарных единиц, появилась возможность численно оценить эти сокровища и вычислить количественные параметры лексики, выявить некоторые закономерности языка, что и является целью работы. Вычисления в работе производятся на основе данных трёх словарей, доступных в электронном виде: Английского Викисловаря, WordNet и Русского Викисловаря.

WordNet – это толковый словарь и тезаурус английского языка в машиночитаемой форме. В основе словаря лежат психолингвистические теории, с учётом которых были определены значения слов и связи между словами и значениями, а также связи между самими значениями [2].

Данные WordNet используются для решения многих задач, например, определения значения слова [10], [13], [15], вычисления логичности и связности предложений в тексте [3], [14], построения баз знаний и тезаурусов.

Викисловарь – это свободно пополняемый многофункциональный многоязычный словарь и тезаурус. В Викисловаре содержатся толкования и переводы слов, описание фонетических и морфологических свойств, семантические (парадигматические) отношения. Кроме того – произношение слов (транскрипция и аудио файлы), правила разбиения слов на слоги, ударения в словах, информация об этимологии слов. А также – цитаты из литературных произведения, иллюстрирующие употребление слов, и даже видео и фотографии, иллюстрирующие значения слов в прямом смысле.

Возможно, причинами популярности Викисловаря является то, что он находится в открытом доступе и содержит огромную базу данных слов с переводами на многие языки. Наиболее привлекательными чертами являются его многоязычность, огромный объём данных и высокая скорость развития.

Затруднительно сравнивать другие словари с Викисловарём, поскольку любые сравнения быстро устаревают. Например, в работе [7] сравнили словарь PanDictionary с данными Викисловаря за 2008 год, когда в нём было всего 403 413 переводов. 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 примерно 800 языков. Данные Викисловаря используются:

- в машинном переводе между нидерландским и бурским языками [11];

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

- в обработке текста парсером NULEX, где используется интеграция части данных Викисловаря (времена глаголов) с базой данных WordNet и VerbNet [8];
- в системе распознавания и синтеза речи, где Викисловарь – это основа для быстрого создания словаря произношений [12];
- при отображении онтологий [6].

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

2. Эксперименты: части речи. Этот раздел отвечает на вопрос – в каком объёме представлены разные части речи в словарях. Рассматриваются три словаря:

- Английский Викисловарь (En), версия от 8 октября 2011 г.;
- Русский Викисловарь (Ru), версия от 21 мая 2011 г.;
- WordNet 3.0 (WN), статистические данные взяты со страницы WordNet проекта[†].

В многоязычных словарях (Английский Викисловарь и Русский Викисловарь) в этой работе учитываются только словарные статьи, описывающие английские слова.

Разрабатываемый 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) [5], систему JWKTL (работает с английской и немецкой версиями Викисловаря) [‡]. Наш парсер *wikt_parser* преобразует базу данных Викисловаря в машиночитаемый словарь и сохраняет результат в базу уже меньшего размера в формате MySQL или SQLite для последующего использования [12]. 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 приводится число английских слов и значений в словарях. Та же информация на рис. 1 наглядно показывает, что больше все-

[†] См. http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html

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

го английских слов и значений содержится в Английском Викисловаре. В нём словарных статей больше чем в WordNet в 1.78 раза, а значений – больше в 1.79 раза.

Табл. 1. **Число английских слов и значений**

POS	Unique Strings			Total Word-Sense Pairs		
	Ru	WN	En	Ru	WN	En
Noun	19 639	117 798	143 062	23 126	146 312	192 819
Verb	809	11 529	37 002	2 138	25 047	53 777
Adjective	831	21 479	57 525	1 530	30 002	72 320
Adverb	122	4 481	11 259	212	5 580	13 055
Totals*	21 946	155 287	**276 470**	27 719	206 941	**369 778**

Звёздочка в заголовке строки “Totals*” в табл. 1 (и значение ”Others” на рис. 2) указывают на то, что кроме частей речи, представленных в WordNet (существительное, прилагательное, глагол, наречие), в Викисловарях также представлены союзы, междометия, предлоги и т.д. Кроме того, в ”Others” попадает ряд других словарных единиц, представленных в Викисловаре, но, строго говоря, не являющихся частями речи, например: имя собственное, приставка, суффикс, фразеологизмы и т.д.[§]

Число слов для Английского Викисловаря в табл. 1 несколько меньше, чем аналогичные данные в табл. 1 в работе [9], поскольку было решено не рассматривать словоформы (inflected word forms) как полноценные словарные статьи, отдельные от основной статьи. Т.е. разработанный парсер [12] пропускает куцые словарные статьи, которые содержат только отсылку к основной словоформе.

Рис. 2 показывает пропорцию распределения английских слов по разным частям речи. Рисунок отображает данные той же табл. 1, но уже в процентном соотношении.

[§] См. http://en.wiktionary.org/wiki/Wiktionary:POS

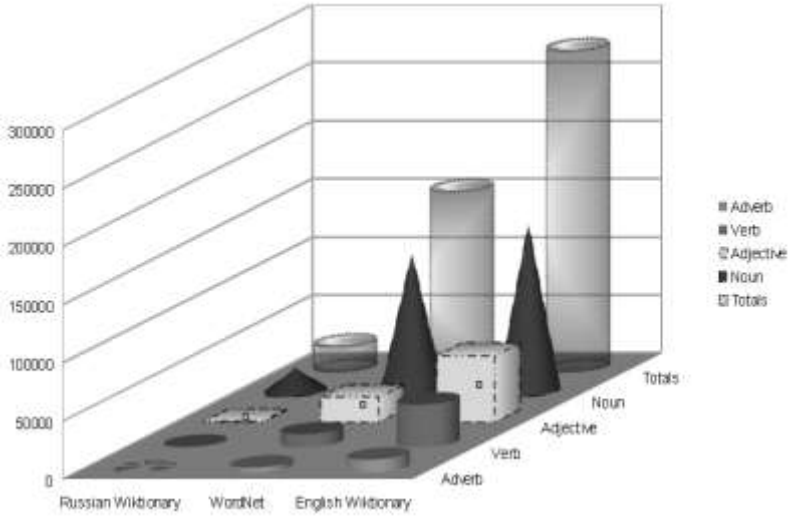


Рис. 1. Число английских слов по частям речи в Английском Викисловаре, WordNet и Русском Викисловаре.

Если считать самый большой словарь так же и наиболее развитым, то выстраивается такая последовательность: самый проработанный – Английский Викисловарь, менее развитый – WordNet и ещё только начинающий своё развитие (в отношении слов английского языка) – Русский Викисловарь. Тогда рис. 2 позволяет сделать следующие выводы:

1. Наибольшую долю во всех словарях занимают существительные (52-83%), затем идут прилагательные (6-20%), глаголы (8-15%), наречия (1-4%).
2. Чем более полон словарь, тем меньшую долю в нём занимают собственно существительные и большую долю начинают занимать другие части речи. Рис. 2 показывает, что в первую очередь энтузиастами в викисловарях заполняются существительные, возможно, потому что существительные более востребованы и для них легче сформулировать толкования, чем для слов других частей речи.

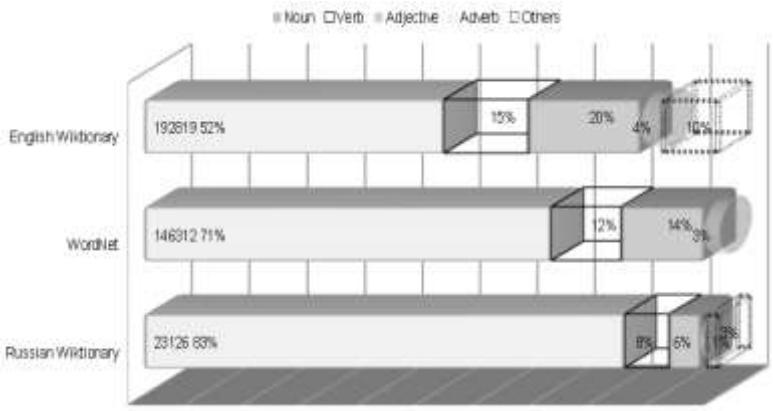


Рис. 2. Относительное распределение английских слов по частям речи в трёх словарях: Английский Викисловарь, WordNet, Русский Викисловарь.

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

В табл. 2 приведены данные о количестве слов с одним значением и с несколькими значениями. Также приведено суммарное число значений по всем многозначным словам. Данные приведены для тех же словарей, что и в предыдущем разделе, а именно: Русский Викисловарь, WordNet и Английский Викисловарь.

Табл. 2. **Многозначность английских слов в Русском Викисловаре (Ru), WordNet (WN) и в Английском Викисловаре (En)**

POS	Monosemous Words and Senses			Polysemous Words			Polysemous Senses		
	Ru	WN	En	Ru	WN	En	Ru	WN	En
Noun	18 036	101 863	115 772	1 603	15 935	27 290	5 090	44 449	77 047
Verb	264	6 277	28 932	545	5 252	8 070	1 874	18 770	24 845
Adjective	497	16 503	47 907	334	4 976	9 618	1 033	14 399	24 413
Adverb	74	3 748	9 931	48	733	1 328	138	1 832	3 124
Totals*	19 314	128 391	**224 148**	2 632	26 896	52 322	8 405	79 450	**145630**

Рис. 3, построенный на основе данных табл. 2, показывает, что оба словаря (WordNet и Английский Викисловарь) содержат больше слов с одним значением, чем с несколькими, 81% слов с одним значением в Английском Викисловаре и 88% в WordNet. В системе WordNet выделяется относительно большое число многозначных глаголов –

46% (5.2 тыс слов) по сравнению со всего 22% в Английском Викисловаре (8 тыс слов).

Рис. 3 показывает, что в Английском Викисловаре для существительных, глаголов и прилагательных многозначными являются примерно каждое пятое слово (17%-22%). Немного выделяются только наречия, многозначных слов всего 12%. Для WordNet тоже наблюдается некоторая стабильность, хотя и с немного большим разбросом, долю многозначных слов составляют от 14% до 23% слов (кроме глаголов).

В табл. 3 приведено среднее число значений:

- с учётом слов с одним значением (левая часть табл. 3, рис. 4a);
- без учёта слов с одним значением, т.е. только для многозначных слов (правая часть табл. 3, рис. 4b).

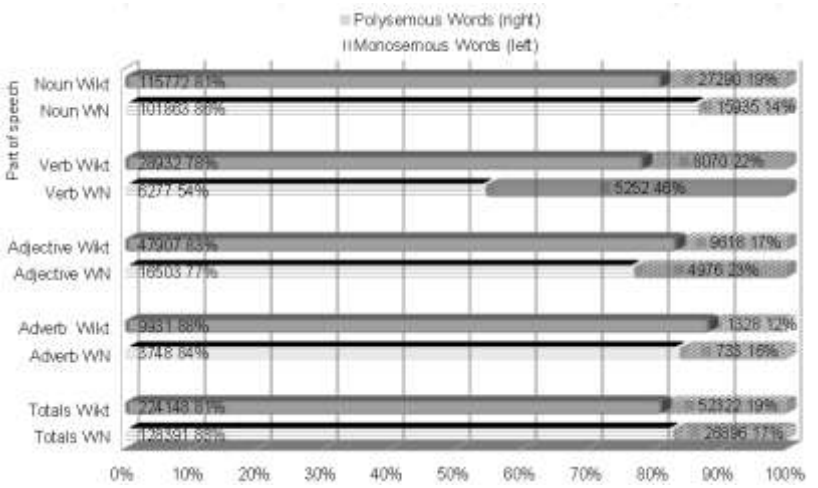


Рис. 3. Относительное число однозначных и многозначных слов (по частям речи) в Английском Викисловаре (Wikt) и WordNet (WN).

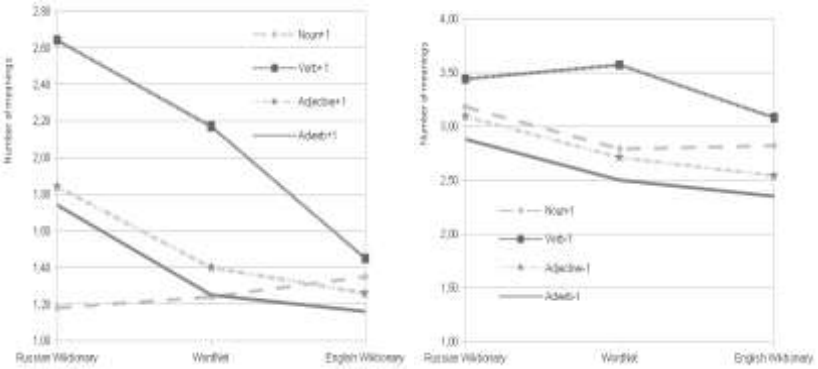
Рис. 4 позволяет сделать некоторые выводы:

- Наиболее многозначными являются глаголы (верхняя кривая на обоих рисунках), без учёта слов с одним значением среднее число значений у глаголов больше трёх, а именно: 3.08-3.57 (рис. 4b).

- Есть наглядное совпадение формы кривых у прилагательных и наречий (т.е. по всем трём словарям), причём у прилагательных больше значений, чем у наречий.
- Меньше всего значений (без учёта слов с одним значением) у наречий, диапазон 2.35-2.88 (нижняя кривая, рис. 4b), за исключением Русского Викисловаря, где меньше всего значений у наречий и существительных (рис. 4a).

Табл. 3. **Среднее число значений у многозначных английских слов в Русском Викисловаре (Ru), WordNet (WN) и в Английском Викисловаре (En)**

POS	Average Polysemy Including Monosemous Words			Average Polysemy Excluding Monosemous Words		
	Ru	WN	En	Ru	WN	En
Noun	**1.18**	1.24	1.35	3.18	2.79	2.82
Verb	2.64	2.17	1.45	3.44	**3.57**	3.08
Adjective	1.84	1.40	1.26	3.09	2.71	2.54
Adverb	1.74	1.25	**1.16**	2.88	2.5	2.35



a) b)

Рис. 4. Среднее число значений для английских слов с одним и более значениями (a), среднее число значений только для многозначных слов (b).

4. Эксперименты: распределение слов по числу значений. В Для обоих викисловарей (кроме WordNet) получилось построить распределение слов по числу значений, т.е. подсчитать число слов без толкований (т.е. с нулём значений), число слов с одним значением, число слов с двумя значениями и т.д. Фрагменты двух таблиц распре-

деления значений доступны в интернете и для Английского Викисловаря[**], и для Русского Викисловаря[††].

На рис. 5 представлено распределение слов по числу значений. На этом рисунке представлены данные по Английскому Викисловарю только до слов с 22 значениями, по Русскому Викисловарю – до слов с 12 значениями, т.к. (1) при больших числах начинают встречаться нулевые значения (т.е. таких слов нет в словаре), непригодные для аппроксимации, (2) для больших словарных статей, описывающих много значений, не всегда точно получается подсчитать число значений, вероятно, из-за того, что авторы таких статей вследствие разных причин (например, удобство подачи материала) отклоняются от жёсткого формата Викисловаря, на который настроен наш парсер. Например, значения в статье[‡‡] об английском предлоге *of* разбиты на подзначения, что не отражено в правилах оформления статей[§§].

Степенные функции хорошо аппроксимируют распределения значений английских слов как в Английском Викисловаре, так и в Русском Викисловаре с высоким коэффициентом детерминации 0.99. Рис. 5 хорошо показывает, что развитие викисловарей идёт достаточно равномерно. И распределение значений английских слов в Русском Викисловаре, основанном в 2004 г., подчиняется близкому по показателю степенному закону, что и в огромном Английском Викисловаре, начатом в 2002 г.

5. Заключение. Для численного анализа лексики английского языка были построены машиночитаемые версии Английского Викисловаря и Русского Викисловаря [12]. В многоязычных викисловарях учитывались только словарные статьи, описывающие английские слова. Третьим использованным словарём был WordNet. Были проведены эксперименты, в ходе которых сравнили:

- число английских слов и значений в словарях. Больше всего английских слов (276470) и значений (369778) содержится в Английском Викисловаре. В нём словарных статей больше чем в WordNet в 1.78 раза, а значений – больше в 1.79 раза.

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

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

[‡‡] См. http://en.wiktionary.org/wiki/of#Preposition

[§§] См. http://en.wiktionary.org/wiki/Wiktionary:ELE

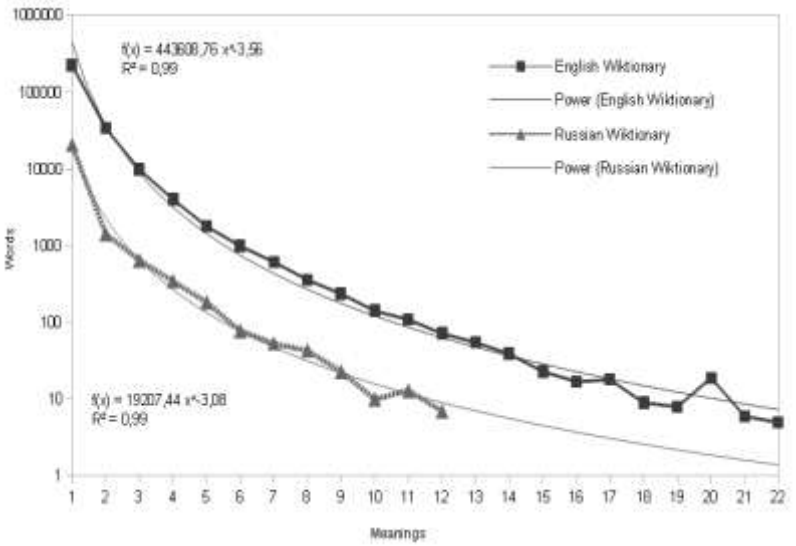


Рис. 5. Распределение английских слов по числу значений в Английском Викисловаре (верхняя кривая) и в Русском Викисловаре (нижняя кривая), а также их аппроксимации степенными функциями.

- распределение слов английского языка по частям речи в Английском Викисловаре, Русском Викисловаре и WordNet. Найдено, что наибольшую долю во всех словарях занимают существительные (52-83%), затем идут прилагательные (6-20%), глаголы (8-15%), наречия (1-4%).
- число слов с одним значением (81% в Английском Викисловаре и 88% в WordNet) и многозначных слов. Оба словаря (WordNet и Английский Викисловарь) содержат больше слов с одним значением, чем с несколькими. В Английском Викисловаре для существительных, глаголов и прилагательных многозначными являются примерно каждое пятое слово (17%-22%), только для наречий многозначных слов всего 12%.
- среднее число значений для слов, принадлежащих разным частям речи. Во всех трёх словарях наиболее многозначными оказались глаголы, среднее число значений у глаголов (без учёта слов с одним значением) больше трёх (диапазон 3.08-

3.57). Меньше всего значений (без учёта слов с одним значением) у наречий (диапазон 2.35-2.88).

Также для Английского Викисловаря и для Русского Викисловаря были вычислены распределения английских слов по числу значений. Для распределений были построены аппроксимирующие кривые, соответствующие экспериментальным данным по экспоненциальному закону с высоким коэффициентом детерминации 0.99.

В работе [9] оценивается распределение слов по числу значений именно на основе WordNet и Английского Викисловаря. Разница в том, что там исследуются не все словарные статьи (т.е. 276 тысяч в Английском Викисловаре и 155 тысяч в WordNet), а только пересечение Викисловаря и WordNet, равное 76 тысячам слов.

Полученные результаты (количество английских слов по частям речи – рис. 1, распределение значений английских слов – рис. 5) наглядно показывают последовательность и закономерность в развитии викисловарей от только ещё начинающего своё развитие (в отношении слов английского языка) – Русского Викисловаря до наиболее проработанного и большого – Английского Викисловаря.

Вслед за авторами статьи [9], а также на основе анализа рис. 1 (число английских слов по частям речи), рис. 2 (относительное распределение английских слов по частям речи) и рис. 4 (среднее число значений) можно утверждать, что лингвистические ресурсы, созданные как экспертами, так и энтузиастами, подчиняются общим законам.

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, а именно – Английский Викисловарь, содержит 61 тыс. словарных статей с пустыми, т.е. пока что незаполненными толкованиями, что составляет 5% от числа всех статей. При этом быстрый рост числа статей в Викисловаре позволяет предположить, что ещё не скоро наступит насыщение, а значит, словари ещё далеко не полны.

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.

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. *Ferrer-i-Cancho R.*. The structure of syntactic dependency networks: insights from recent advances in network theory // In: V. Levickij and G. Altmann (Eds.), Problems of quantitative linguistics, 2005. P.60-75.
3. *Harabagiu S., Moldovan D.* A marker-propagation algorithm for text coherence. // In Working Notes of the Workshop on Parallel Processing at the 14th International Joint Conference on Artificial Intelligence. Montreal. 1995. P.76-86.
4. *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 17.11.2011)
5. *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.
6. *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.
7. *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.
8. *McFate C., Forbus K.* NULEX: An Open-License Broad Coverage Lexicon. (accepted). In: The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA. June, 2011.
9. *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 17.11.2011)
10. *Montoyo A., Palomar M., Rigau G.* Method for WordNet enrichment using WSD // In Proceedings of 4th International Conference on Text Speech and Dialogue TSD'2001. Selezna Ruda - Spieak, Czech Republic. Published in Lecture Notes in Artificial Intelligence 2166, Springer-Verlag. 2001
11. *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.
12. *Qingyue He*. Automatic Pronunciation Dictionary Generation from Wiktionary and Wikipedia. // Thesis. Karlsruhe Institute of Technology. 2009.
13. *Resnik P., Yarowsky D.* Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation // Natural Language Engineering. Vol. 5, No. 2, 2000. P.113-133.
14. *Teich E., Fankhauser P.* WordNet for lexical cohesion analysis // In Proceedings of the Second Global WordNet Conference. Brno, Czech Republic. January, 2004. P.326-331.
15. *Yarowsky D.* Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics. Cambridge, MA. 1995. P.189-196.

Крижановский Андрей Анатольевич — к.т.н.; старший научный сотрудник лаборатории интегрированных систем автоматизации Учреждения Российской академии наук Санкт-Петербургского института информатики и автоматизации РАН (СПИИРАН).

Область научных интересов: автоматическая обработка текста, корпусная лингвистика. Число научных публикаций — 66. andrew.krizhanovsky@gmail.com, whinger.narod.ru; СПИИРАН, 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т. +7(812)328-8071, факс +7(812)328-0685.

Krizhanovsky Andrew Anatoliyevich — PhD; senior researcher, Computer-aided integrated systems laboratory of Institution of the Russian Academy of Sciences, St.Petersburg Institute for Informatics and Automation RAS (SPIIRAS). Research Interest: information retrieval, corpus linguistics. The number of scientific publications — 66. andrew.krizhanovsky@gmail.com, whinger.narod.ru; SPIIRAS, 39, 14-th Line V.O., St. Petersburg, 199178, Russia; phone +7(812)328-8071, fax +7(812)328-0685.

Поддержка исследований. Работа выполнена при финансовой поддержке РФФИ (проект № 11-01-00251, № 12-01-00481, № 12-01-00499), поддержке РГНФ (проект № 12-04-12062) и проекта Программы фундаментальных исследований Президиума РАН «Интеллектуальные информационные технологии, математическое моделирование, системный анализ и автоматизация»

Рекомендовано лабораторией интегрированных систем автоматизации, заведующий лабораторией Смирнов А.В., д. т. н., проф.
Статья поступила в редакцию 12.11.2011.

РЕФЕРАТ

Крижановский А.А. **Количественный анализ лексики английского языка в викисловарях и Wordnet.**

Разработана компьютерная система анализа лексикографических данных, позволившая выполнить количественный анализ лексики английского языка по данным трёх электронных словарей: Английского Викисловаря, WordNet и Русского Викисловаря. В многоязычных викисловарях (Английский Викисловарь, Русский Викисловарь) учитывались только словарные статьи, описывающие английские 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 английских слов по частям речи, распределение значений английских слов) наглядно показывают последовательность и закономерность в развитии викисловарей от только ещё начинающего своё развитие (в отношении слов английского языка) – Русского Викисловаря, до наиболее проработанного и большого – Английского Викисловаря.

С помощью степенных функций получена аппроксимация распределения значений английских слов в Английском Викисловаре и в Русском Викисловаре, при этом аппроксимации соответствует высокий коэффициент детерминации 0.99. Анализ показывает, что развитие викисловарей идёт достаточно равномерно. Таким образом, распределение значений английских слов в Русском Викисловаре, основанном в 2004 г., подчиняется близкому по показателю степенному закону, что и в огромном Английском Викисловаре, который начали разрабатывать в 2002 г.

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

SUMMARY

Krizhanovsky A.A.. **A quantitative analysis of the English lexicon in Wiktionaries and WordNet.**

A software system of analysis of lexicographic data was developed. It was used in order to perform a quantitative analysis of the English lexicon in the three electronic dictionaries: the English Wiktionary, WordNet, and the Russian Wiktionary. Only English 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 English 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 green Russian Wiktionary (only in relation of English entries) to the most thoroughly developed and huge the English Wiktionary.

The distributions (for both wiktionaries) were approximated using power law functions where coefficient of determination is 0.99. It could be concluded that wiktionaries are developed in a relatively uniform manner. And the distribution of English words in the Russian Wiktionary (launched in 2004), accomplishes to the same power law with similar exponent that in the huge English Wiktionary (launched in 2002).

Therefore, it was calculated the quantity of English 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 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.