A quantitative analysis of the English lexicon in Wiktionaries and WordNet

Andrew Krizhanovsky
Institution of the Russian Academy of Sciences St.Petersburg Institute for Informatics and Automation RAS, Russia

ABSTRACT

A quantitative analysis of the English lexicon was done in the paper. The three electronic dictionaries are under examination: the English Wiktionary, WordNet, and the Russian Wiktionary. It was calculated the quantity of English words and meanings (senses) in these dictionaries. 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: quantitative linguistics, lexicology, WordNet, Wiktionary, English language

INTRODUCTION

The richness of the language is hidden in the lexicon, in multiple meanings and shades of meanings, which are constantly changing over time in a subtle manner. It is one of the reasons of the existence of a kind of dictionaries named *thesaurus*, word has Latin roots signifying a "*treasure, hoard*". At a time when new big electronic dictionaries (containing tens and hundreds of thousands of entries) appeared, the real possibility to estimate these *treasures* numerically is brought into existence. The goal of this work will be to estimate numerically some properties of the dictionaries, to find out some language regularities and to compare dictionaries themselves.

An analysis and comparison of lexical resources will provide (1) an indication of which kind of resource is more suitable for dictionary users and software developers; (2) an indication of gaps which can be presented in the source material and the dictionary itself. This information should help to authors to improve their dictionaries.

All investigations will be performed on the basis of three electronic dictionaries: the English Wiktionary, WordNet, and the Russian Wiktionary. WordNet is a dictionary and a thesaurus for the English language in a machine-readable form. It is based on psycholinguistic theories to define word meaning. The WordNet data was used to solve many linguistics problems, e.g., word sense disambiguation (Montoyo, Palomar, & Rigau, 2001; Resnik & Yarowsky, 2000; Yarowsky, 1995), text coherence analysis (Harabagiu & Moldovan, 1995; Teich & Fankhauser, 2004), knowledge bases construction.

The Wiktionary is a multilingual and multifunctional dictionary and thesaurus. The Wiktionary contains not only word's definitions, semantically related words (synonyms, hypernyms, etc.),

translations, but also the pronunciations (phonetic transcriptions, audio files), hyphenations, etymologies, quotations, parallel texts (quotations with translations), and figures (which illustrate meaning of the words).

Wiktionary is popular since it is freely available and contains a huge database of words with translations to many languages. The salient properties of the Wiktionary are the multilinguality, the size, and the speed of evolution. It is difficult to compare dictionaries with the Wiktionary, since data quickly become outdated. E.g., the PanDictionary was compared with the Wiktionary data obtained in the year 2008, when it had 403 413 translations (Mausam, et. al., 2010). Two years later, in 2010, the English Wiktionary contained twice as many translations (964 019)[1]. So, the Wiktionary is permanently growing in number of entries and in the scope of languages. Now the English Wiktionary contains entries in about 800 different languages. There is an interesting paper by (Meyer & Gurevych, 2012) who investigated three Wiktionaries: English, German and Russian. The Wiktionary data are used:

- In *machine translation* between Dutch and Afrikaans (Otte & Tyers, 2011);
- In the *text parsing* system NULEX, where some Wiktionary data (verb tense) were integrated with WordNet and VerbNet (McFate & Forbus, 2011);
- In a *speech recognition and speech synthesis* as a basis for the rapid pronunciation dictionary creation (Qingyue He, 2009);
- In *ontology matching* (Lin & Krizhanovsky, 2011).

The paper has the following structure. In section 2, the quantity of English words and meanings, and the distribution of words for each part of speech are estimated. The question of "the ratio of polysemous and monosemous words" and the average polysemy across the three dictionaries is calculated in section 3. Section 4 presents the distribution of words by number of meanings.

2. EXPERIMENTS: PARTS OF SPEECH

There are two topics we will discuss in this section: (1) the quantity of English words and meanings and (2) the distribution of words for each part of speech. The following dictionaries are under consideration:

1. The English Wiktionary, the edition as of October 8, 2011
2. The Russian Wiktionary, the edition as of May 21, 2011
3. WordNet 3.0 (denoted as WN), the statistics data are taken from the WordNet project site.[2]

In multilingual dictionaries (the English Wiktionary and the Russian Wiktionary) only English entries were taken into account in this paper.

The experiments were conducted using the developed Wiktionary parser (*wikt_parser*), which is one of several tools that parse the Wiktionary data. Other tools include Zawilinski parser (Polish words in the English Wiktionary) (Kurmas, 2010), JWKTL (the English and the German versions of Wiktionary)[3]. Our parser *wikt_parser* transforms the Wiktionary database into the

[1] See http://en.wiktionary.org/wiki/User:AKA_MBG/Statistics:Translations (the tab "History")

[2] See http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html

[3] See http://www.ukp.tu-darmstadt.de/software/jwktl/

machine-readable dictionary and saves it as a smaller database (MySQL or SQLite) for later use (Krizhanovsky, 2010). So, all statistical data in this paper (related to Wiktionary) were calculated using two machine-readable databases based on the English Wiktionary and the Russian Wiktionary.
Table 1 contains the number of English words and the number of meanings for these words in the dictionaries. The same information in Fig. 1 clearly shows that the most number of English words (for every part of speech) and meanings is contained in the English Wiktionary. The number of *unique strings* (i.e. words, entries) in the English Wiktionary is larger by 1.78 times than in WordNet, and the number of *meanings* is larger by 1.79 times than in WordNet.

Table 1 Number of English words and senses

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**

The asterisk in the header of the row “Totals*” in Table 1 (and the field “Others” in Fig. 2) indicates that besides the parts of speech presented in WordNet (noun, verb, adjective, adverb), Wiktionaries also contain conjunctions, interjections, prepositions. Also “Others” contains a number of other lexical units presented in the Wiktionary but which are not (strictly speaking) parts of speech, e.g., prefixes, suffixes, idioms, acronym, abbreviation, etc.[4]

The number of words in the English Wiktionary (presented in Table 1) is smaller than in Table 1 in (Meyer & Gurevych, 2010), since an inflected word form is not considered as a full-fledged entry. The developed parser (Krizhanovsky, 2010) skips scanty Wiktionary entries, which contains a soft redirect to the canonical form of an inflected word (lemma).

[4] See http://en.wiktionary.org/wiki/Wiktionary:POS

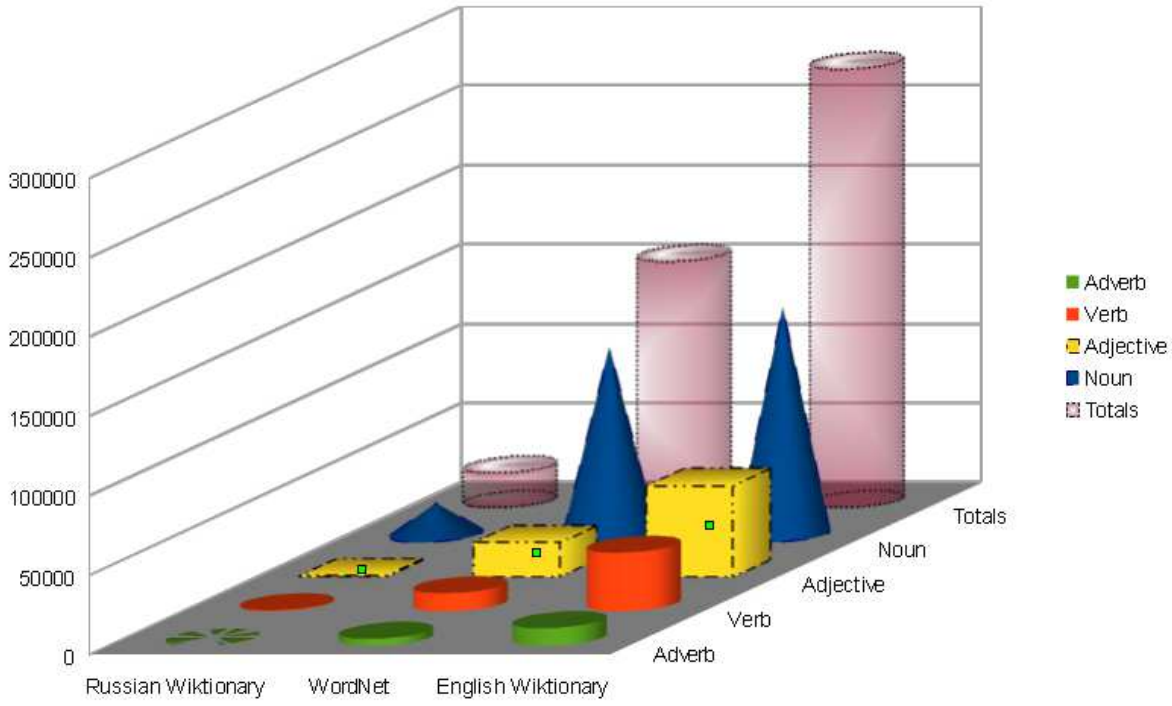

Fig. 1. Number of English words in different parts of speech in the English Wiktionary, WordNet and the Russian Wiktionary.

Fig. 2 shows the relative distribution of English words with respect to the part of speech. It contains the same data as in Table 1, but in percentage terms.

If we suppose that the largest size dictionary is the most elaborated one, then it may be supposed that the most elaborated is the English Wiktionary, WordNet is in the middle, and the Russian Wiktionary is at the beginning of the development (though English entries compose only a small fraction (8.7%) of all entries in the Russian Wiktionary). The analysis of Fig. 2 allows two conclusions to be drawn.

1. The largest part in all the dictionaries belongs to nouns (52-83%), then adjectives (6-20%), verbs (8-15%) and adverbs (1-4%).
2. The more complex and detailed dictionary is, the less proportion of nouns is presented, and other parts of speech become to occupy the more proportion in the dictionary. Fig. 2 shows that in the first place volunteers fill in nouns in Wiktionaries, and the possible reasons of that are (1) nouns are more in demand; (2) it is more simple to formulate definitions for nouns than for other parts of speech.

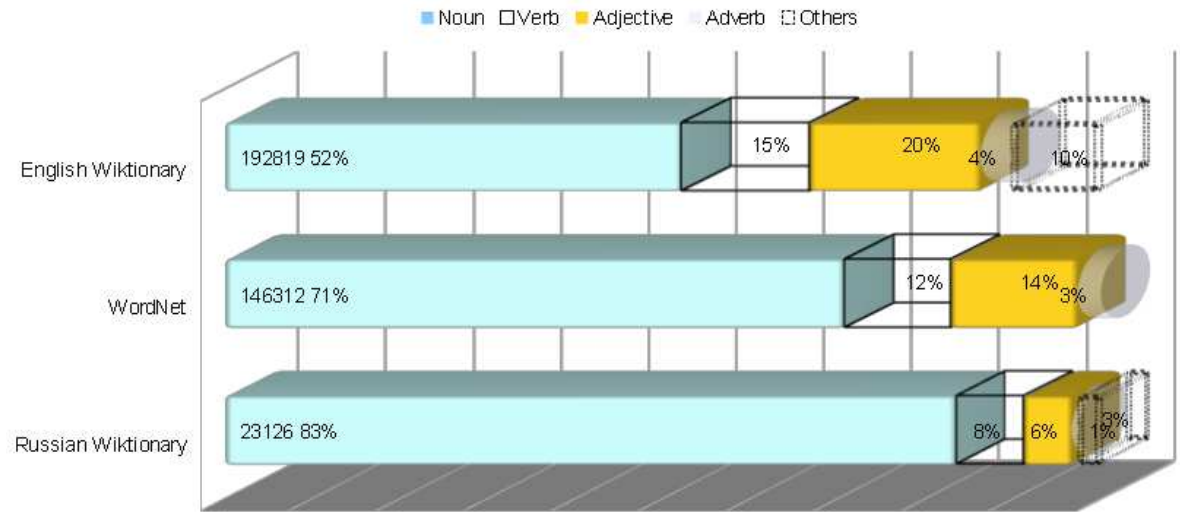

Fig. 2. The relative distribution of English words with respect to the part of speech in the English Wiktionary, WordNet and the Russian Wiktionary

3. EXPERIMENTS: POLYSEMY

An important characteristic of the dictionary is a proportion of polysemous to monosemous words and an average polysemy.

Table 2 contains the number of senses and the number of monosemous and polysemous words in total and for each part of speech. The same dictionaries are under consideration in this section: the Russian Wiktionary, WordNet and the English Wiktionary.

Table 2. Polysemy of English words in the Russian Wiktionary (Ru), WordNet (WN) and the English Wiktionary (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	**145 630**

Fig. 3 (built on the basis of the same data as Table 2) shows that both dictionaries (WordNet and the English Wiktionary) contain more monosemous words than polysemous, there are 81% of monosemous words in the English Wiktionary and 88% in WordNet. WordNet contains a relative large number of polysemous verbs – 46% (or 5252 words) in a comparison with 22% in the English Wiktionary (but 8070 words).

Fig. 3 shows that there is a regularity for nouns, verbs and adjectives in the English Wiktionary, about every fifth word (17%-22%) is a polysemous. Adverbs are a little bit outside this regularity, there are only 12% of polysemous adverbs. There is some stability in this proportion

in WordNet too, though with more spread in the proportion of polysemous words in the range of 14 to 23% (except verbs).

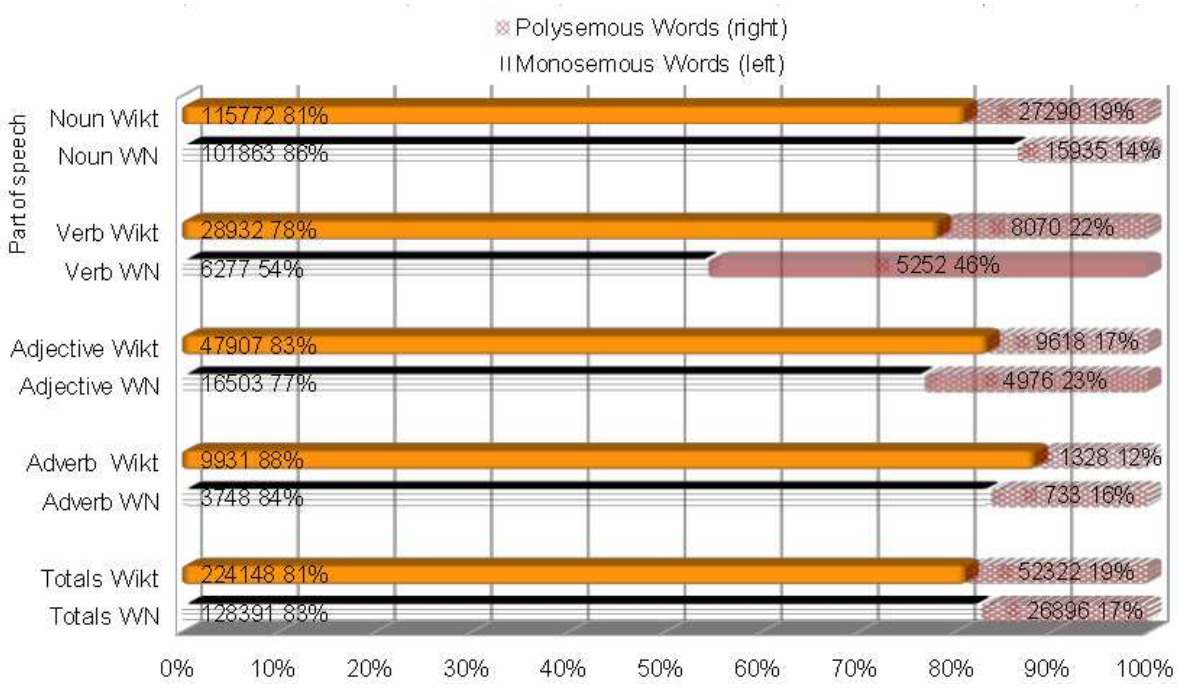

Fig. 3. The relative distribution of monosemous and polysemous English words (for each part of speech) in the English Wiktionary (Wikt) and WordNet (WN)

Table 3 contains values of average polysemy of English words:

- Including monosemous words (left part of Table 3 and Fig. 4a);
- Excluding monosemous words, i.e. only polysemous words (right part of Table 3 and Fig. 4b).

Table 3. Average polysemy of English words in the Russian Wiktionary (Ru) WordNet (WN) and the English Wiktionary (En)

	Average Polysemy Including Monosemous Words			Average Polysemy Excluding Monosemous Words		
POS	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

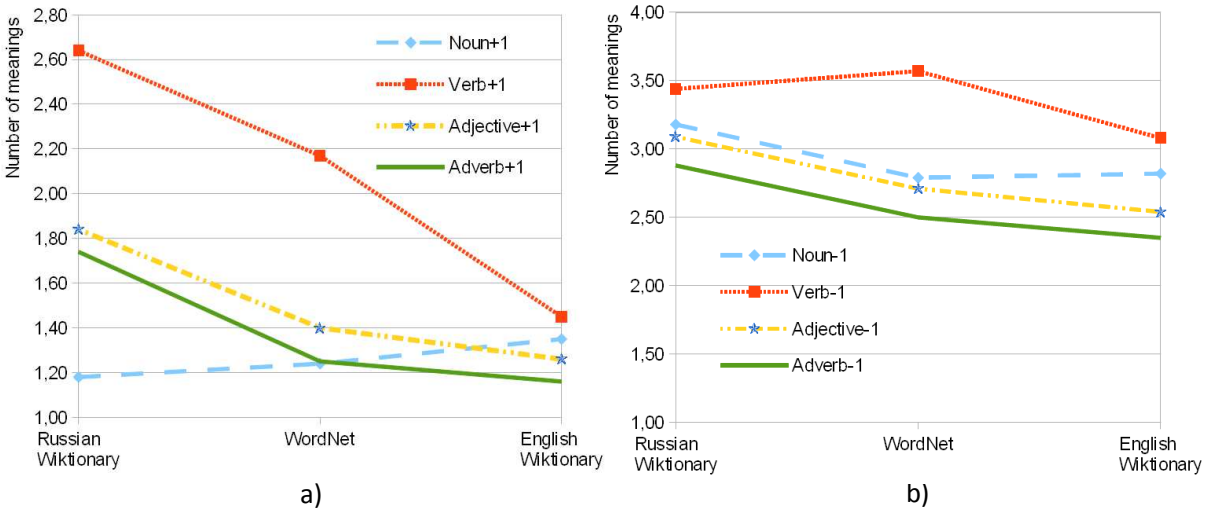

Fig. 4. Average polysemy of English words (a – including monosemous words, b – excluding monosemous words): in the Russian Wiktionary, WordNet and the English Wiktionary

The analysis of Fig. 4 allows some conclusions to be drawn:

- The most polysemous words are verbs (the upper curve in both figures). Excluding monosemous words (Fig. 4b) the average polysemy of verbs is more than three in the dictionaries (the range 3.08-3.57).
- There is the visual correspondence of curves of adjectives and adverbs (across all the three dictionaries), and adjectives have more meanings than adverbs.
- Adverbs (excluding monosemous words) have the minimum number of meanings, the range 2.35-2.88 (the lower curve in Fig. 4b), except the Russian Wiktionary, where adverbs and nouns have the minimum value (Fig. 4a).

4. EXPERIMENTS: DISTRIBUTION OF MEANINGS

The distribution of words with respect to the number of meanings was constructed for two wiktionaries (Russian and English). That is, it was counted the number of words without definitions (i.e., with 0 meanings), the number of words with 1 meaning, with 2 meanings, etc. Fragments of two tables with distribution of meanings are available online for the English Wiktionary[5] and for the Russian Wiktionary[6].

The distribution of English words is presented in Fig. 5. The maximum number of meanings in the figure was constrained by 22 for the English Wiktionary and 12 meanings for the Russian Wiktionary, because:

[5] See http://en.wiktionary.org/wiki/User:AKA_MBG/Statistics:POS

[6] See part of speech statistics for the Russian Wiktionary http://bit.ly/IxF7f5

1) There are no words with some greater number of meanings (e.g., with 23 and 13 meanings in the English and Russian Wiktionaries, respectively). The approximation is better with these constraints.
2) The developed parser does not always correctly count the number of meanings for some very long articles that encompass many meanings. A reason for this is that in these articles editors deviate from the Wiktionary strict formatting rules (which are followed by our parser), e.g., in order to present the article in a more useable form. E.g., the meanings of the English preposition *of* [7] are grouped into more common meanings, but this approach is not reflected in the Wiktionary formatting rules[8].

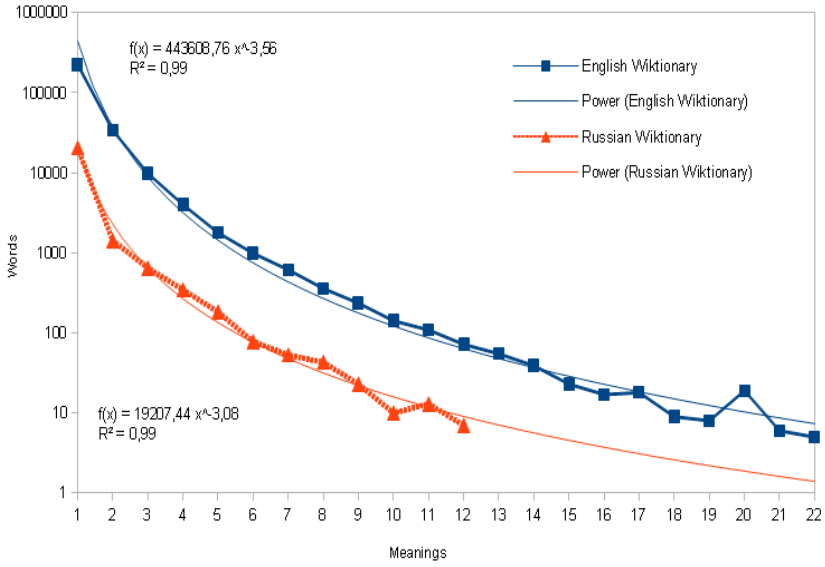

Fig. 5. The distribution of English words with respect to the number of meanings in the English Wiktionary (the upper curve), in the Russian Wiktionary (low curve), and approximations to the power functions

The distributions (for both wiktionaries) were approximated using power law functions where coefficient of determination is 0.99. Fig. 5 explicitly shows 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).

CONCLUSION

The machine-readable dictionaries on the basis of the English Wiktionary and the Russian Wiktionary were constructed (Krizhanovsky, 2010) in order to perform a quantitative analysis of the English lexicon. In multilingual wiktionaries only English entries were taken into account in

[7] See http://en.wiktionary.org/wiki/of#Preposition

[8] See http://en.wiktionary.org/wiki/Wiktionary:ELE

this analysis. The third dictionary examined in the experiment was WordNet. In the experiment it was calculated and compared:

- A quantity of English words and meanings (senses) in the dictionaries. The most number of English words (276 470) and meanings (369 778) is contained in the English Wiktionary. The number of *unique strings* (i.e. words, entries) in the English Wiktionary is larger by 1.78 times than in WordNet, and the number of *meanings* is larger by 1.79 times than in WordNet.
- A distribution of English words with respect to the part of speech in the English Wiktionary, the Russian Wiktionary and WordNet. The largest part in all dictionaries belongs to nouns (52-83%), then adjectives (6-20%), verbs (8-15%) and adverbs (1-4%).
- A quantity of monosemous and polysemous words. WordNet and the English Wiktionary dictionaries contain more monosemous words (81% in the English Wiktionary and 88% in WordNet) than polysemous. There is a regularity for nouns, verbs and adjectives in the English Wiktionary, about every fifth word (17%-22%) is a polysemous. Adverbs are a little bit outside this regularity, there are only 12% of polysemous adverbs.
- An average polysemy of English words belonging to different parts of speech. Across all three dictionaries the most polysemous words are verbs. The average polysemy of verbs (excluding monosemous words) is more than three in dictionaries (the range 3.08-3.57). Adverbs (excluding monosemous words) have the minimum number of meanings, the range 2.35-2.88, except the Russian Wiktionary, where adverbs and nouns have the minimum value.

Also the distributions of English words with respect to the number of meanings in the English Wiktionary and the Russian Wiktionary were calculated. These distributions were approximated using power law functions where coefficient of determination is 0.99.

In paper (Meyer & Gurevych, 2010) the word sense distribution is estimated in WordNet and the English Wiktionary. Their approach differs in that not all entries were examined (i.e. 276 thousands in the English Wiktionary and 155 thousands in WordNet in our research), but only the intersection of WordNet and the Wiktionary, which is equal to 76 000 words.

The obtained results (a quantity of English words in different parts of speech in Fig. 1, a distribution of meanings of English words in Fig. 5) clearly show consistency and regularity in the development of wiktionaries from the green Russian Wiktionary (it is concerned only with the English entries) to the most elaborated English Wiktionary.

The analysis of the Fig. 1, Fig. 2 and Fig. 4 supports the conclusion of Meyer & Gurevych (2010) 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.

There are many measures in the quantitative linguistics which could be used in order to compare lexical resources, e.g.: phoneme frequencies, quantitative syllable structure, the length-frequencies of words, the frequency of polysemy, the degree of popularity of slang words, etc. But there are some constraints. The project of data extraction from the Wiktionary is in its early stage (as the Wiktionary itself). Thus, now only the following data can be extracted from the English and Russian Wiktionaries: definitions, thesaurus and translations. When the parser will be extended (e.g. to extract transcriptions, hyphenations, context labels from the Wiktionary) then lexical resources could be analyzed more thoroughly.

It should be noted that no one of these dictionaries is matured and completed. Even the biggest of these dictionaries – the English Wiktionary – contains 61 thousand entries with empty definitions (it is 5% of all entries), which are expected to be formulated one day. At the same, time the speed of the growing of the number of entries in the Wiktionary indicates that the goal is reachable, though there is a long way to go before making a dictionary which contains "all words in all languages".

An interesting continuation of these experiments may lead to a measure of the semantic distance between languages (Cooper, 2008). The English Wiktionary contains 83 languages, where there are more than 1000 entries. Thus, it is possible to construct the map of these languages using an algorithm.

ACKNOWLEDGEMENT

Some parts of the research were carried out under projects funded by grants # 11-01-00251, # 12-01-00481 and # 12-07-00070 of the Russian Foundation for Basic Research, grant # 12-04-12062 of the Russian Foundation for Humanities and project of the research program "Intelligent information technologies, mathematical modelling, system analysis and automation" of the Russian Academy of Sciences.

REFERENCES

Cooper M. (2008). Measuring the Semantic Distance between Languages from a Statistical Analysis of Bilingual Dictionaries. *Journal of Quantitative Linguistics*, 15 (1). 1-33. Retrieved from ftp://ftp.irit.fr/pub/IRIT/ADRIA/hs2.pdf

Harabagiu S. & Moldovan D. (1995). 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. 76-86. Retrieved from http://www.seas.smu.edu/~sanda/papers/parai.ps.gz

Krizhanovsky A. A. (2010). Transformation of Wiktionary entry structure into tables and relations in a relational database schema. Retrieved from http://arxiv.org/abs/1011.1368

Kurmas Z. (2010, July). Zawilinski: a library for studying grammar in Wiktionary. In: *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*, Gdansk, Poland. Retrieved from http://www.cis.gvsu.edu/~kurmasz/Software/#Zawilinski

Lin F. & Krizhanovsky A. (2011, October) 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. 19-26.

Mausam, Soderland S., Etzioni O., Weld D. S., Reiter K., Skinner M., Sammer M., & Bilmes J. (2010). Panlingual Lexical Translation via Probabilistic Inference. *Artificial Intelligence Journal (AIJ)*. 174 (9-10), 619-637. Retrieved from http://www.aaai.org/ocs/index.php/AAAI/AAAI10/paper/viewFile/1688/2281

McFate C., & Forbus K. (2011, June). 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. Retrieved from http://www.aclweb.org/anthology/P/P11/P11-2063.pdf

Meyer C. M. & Gurevych I. (2010, April). 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. Retrieved from http://journal.webscience.org/349/

Meyer C. M. & Gurevych I. (2012) Wiktionary: a new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography // Electronic Lexicography. Oxford: Oxford University Press. 2012. (to appear). Retrieved from http://www.informatik.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2011/oup-elex2012-meyer-wiktionary.pdf

Montoyo A., Palomar M., & Rigau G. (2001). 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. Retrieved from http://www.lsi.upc.es/~nlp/papers/2001/tsd01-mpr.ps.gz

Otte P. & Tyers F. M. (2011). Rapid rule-based machine translation between Dutch and Afrikaans. In: *16th Annual Conference of the European Association of Machine Translation*, EAMT11. Retrieved from http://xixona.dlsi.ua.es/~fran/publications/eamt2011a.pdf

Qingyue He. (2009). Automatic Pronunciation Dictionary Generation from Wiktionary and Wikipedia. *Thesis*. Karlsruhe Institute of Technology. Retrieved from http://csl.anthropomatik.kit.edu/index.php?id=25

Resnik P. & Yarowsky D. (2000). Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. *Natural Language Engineering*. 5(2), 113-133. Retrieved from http://www.cs.jhu.edu/~yarowsky/pubs.html

Teich E. & Fankhauser P. (2004, January). WordNet for lexical cohesion analysis. In *Proceedings of the Second Global WordNet Conference*. Brno, Czech Republic. 326-331. Retrieved from http://www.fi.muni.cz/gwc2004/proc/77.pdf

Yarowsky D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*. Cambridge, MA. 189-196. Retrieved from http://www.cs.jhu.edu/~yarowsky/pubs.html