РУССКИЙ ЛЕКСИКОГРАФИЧЕСКИЙ ЛАНДШАФТ: ИСТОРИЯ О 12 СЛОВАРЯХ

Киселёв Ю. А. (ykiselev.loky@gmail.com)[1, 2]
Крижановский А. А. (andrew.krizhanovsky@gmail.com)[3]
Браславский П. И. (pbras@yandex.ru)[1, 4]
Меньшиков И. Л. (unkmas@gmail.com)[1]
Мухин М. Ю. (mfly@sky.ru)[1]
Крижановская Н. Б. (nataly@krc.karelia.ru)[3]

[1]Уральский федеральный университет, Екатеринбург, Россия
[2]Яндекс, Екатеринбург, Россия
[3]ИПМИ КарНЦ РАН, Петрозаводск, Россия
[4]Kontur Labs, Екатеринбург, Россия

Ключевые слова: лексический ресурс, словарь, тезаурус, ворднет, русский язык

RUSSIAN LEXICOGRAPHIC LANDSCAPE: A TALE OF 12 DICTIONARIES

Yuri Kiselev (ykiselev.loky@gmail.com)[1, 2]
Andrew Krizhanovsky (andrew.krizhanovsky@gmail.com)[3]
Pavel Braslavski (pbras@yandex.ru)[1, 4]
Ilya Menshikov (unkmas@gmail.com)[1]
Mikhail Mukhin (mfly@sky.ru)[1]
Nataly Krizhanovskaya (nataly@krc.karelia.ru)[3]

[1]Ural Federal University, Ekaterinburg, Russia
[2]Yandex, Ekaterinburg, Russia
[3]Institute of Applied Mathematics Research, Karelian Research Center of RAS, Petrozavodsk, Russia
[4]Kontur Labs, Ekaterinburg, Russia

The paper reports on quantitative analysis of 12 Russian dictionaries at three levels: 1) headwords: the size and overlap of word lists, coverage of large corpora, and presence of neologisms; 2) synonyms: overlap of synsets in different dictionaries; 3) definitions: distribution of definition lengths and numbers of senses, as well as textual similarity of same-headword definitions in different dictionaries. The total amount of data in the study is 805,900 dictionary entries, 892,900 definitions, and 84,500 synsets. The study reveals multiple connections and mutual influences between dictionaries, uncovers differences in modern electronic vs. traditional printed resources, as well as suggests directions for development of new and improvement of existing lexical semantic resources.

Key words: lexical resource, dictionary, thesaurus, wordnet, Russian language

1. Introduction

The problem of analysis and comparison of existing lexical resources for Russian has arisen within the Yet Another RussNet (YARN) project[1]. YARN aims at creating an open thesaurus for Russian using crowdsourcing while maximizing the use of existing lexical-semantic resources (LSRs) [3]. From a linguistics point of view, YARN has rather traditional structure introduced in Princeton WordNet (PWN) [11] and adopted by its numerous successors and variants. YARN consists of synsets—groups of near-synonyms corresponding to a concept; synsets are linked to each other, primarily via hierarchical hyponymic/hypernymic relationships. The project is ongoing and expected to cover Russian nouns, verbs, and adjectives. The main difference from the previous projects is that it is based on crowdsourcing. We hope that crowdsourcing approach will make it possible to create a resource of satisfactory quality and size in foreseeable future and with limited financial resources. Our optimism is based both on international practice and recent examples of successful Russian NLP projects driven by volunteers.

The input information (synonymy and hierarchical relationships) to be validated by the "crowd" is a result of automatic processing of corpus and dictionary data. A brief description of the data sources and online tool that are used in the project at the moment can be found in [4].

The goal of this study is to create an inventory of available LSRs for Russian, to figure out how they relate to each other, what "gaps" in the description of Russian lexis exist and how data at hand can be incorporated into YARN. A big advantage to the study is that a large number of initially printed dictionaries are available today in machine-readable form[2]. As far as we know, no large-scale quantitative comparison of the body of Russian dictionaries has been conducted yet. We hope that our findings will be useful not only within YARN project, but also of interest for a wide lexicographic community as well.

For the study, we employed electronic versions of six printed explanatory dictionaries and three dictionaries of synonyms, online Russian Wiktionary, as well as electronic thesauri RuThes and Russian WordNet. The total amount of data in the study is 805,900 dictionary entries; 892,900 definitions, and 84,500 synsets. Despite the impressive amount of data used in the study, it still remains incomplete: not all Russian dictionaries that we would like to include in the study are available in machine-readable format, and we were not ready to conduct the whole routine of scanning, recognition, and post-processing. Moreover, available resources vary significantly in quality—both because of the structure and print layout of dictionary entries and the quality of recognition and subsequent processing (for example, we could not perform definitions analysis in one of the sources since it was impossible to parse it correctly).

We investigated the dictionary data at three levels: 1) headwords: size and overlap of headword lists, coverage of large corpora, and presence of neologisms; 2) synonyms: we attempted to align the meanings of synsets in different sources and analyze their intersections; 3) definitions: distribution of definition lengths and number of senses, as well as textual similarity of same-headword definitions in different dictionaries.

---

[1] http://russianword.net

[2] http://nlpub.ru/Ресурсы

2. Related work

In our study, we compare headword lists from different dictionaries, corpora coverage by respective word lists, make an attempt to directly compare synonym data contained in different dictionaries, as well as analyze various properties of definitions and their inter-dictionary similarity. First studies on automated analysis of dictionary data in machine-readable format can be dated back to 1980s. For example, an early paper [22] studied word frequency and length distributions of definitions in an English dictionary, distributions of semantic and part-of-speech marks, as well as coverage of definitions by the top-frequency words. Michiels and Yoshida [25, 31] proposed methods for identification of hierarchical relations between word senses based on dictionary data. Automatic thesaurus construction using existing dictionaries became widespread when open collaborative projects, primarily Wikipedia and Wiktionary, matured and accumulated sufficient data volumes. The latest example of a large multilingual thesaurus based on open data is BabelNet [27]. The current Babelnet version claims to comprise more than 40 mln glosses in 271 languages that form more than 13 mln synsets (http://babelnet.org/stats).

The work by Meyer and Gurevych [24] is probably closest to ours. The main objective of the study was to compare collaboratively constructed language resources with traditional expert-built resources. The authors juxtaposed three different language editions of Wiktionary (English, German, and Russian) and corresponding thesauri—PWN, GermaNet, and Russian Wordnet. The paper presented basic statistics of resources—the total number of headwords, parts-of-speech and senses distributions, coverage of core vocabulary and neologisms in respective languages, overlap of headword lists, as well as presence of domain and register marks. The study did not analyze definitions and synonymy information presented in both kinds of resources.

A problem closely related to our research is sense alignment, i.e. matching of identical or similar senses in different LSRs. For example, an early work [20] compared PWN and printed dictionaries based on manual coding of meanings of 18 English verbs. Current approaches use fully automated methods: for example, Matuschek and Gurevych [23] combined graph-based distances between senses with textual similarity of definitions for aligning senses between Wiktionary and Wikipedia in English and German (the study also contains a nice overview of sense alignment methods and approaches). The paper [16] describes a task-oriented comparison (such as word and sentence relatedness problems) of synonymy information presented in PWN and different editions of Roget's thesaurus.

Large corpora are widely used for building modern dictionaries, in particular—to compile and update glossaries, extract collocations, and provide word usage examples. For example, Geyken and Lemnitzer [13] used Google Books Ngram Corpus to compile a wordlist for a new dictionary of German. A survey of corpus tools for lexicography can be found in [17]. In our study we handle an inverse problem: we investigate how existing dictionaries cover corpora, as well as how neologisms extracted from temporarily labeled subcorpora are presented in lexicographic resources.

Based on the literature review, we can conclude that our study is unprecedented in number of resources involved, volumes of data processed and aspects of dictionary data analyzed. Due to large volumes and wide diversity of data we employ mainly shallow processing techniques in our study.

3. Data

The resources in the study and their quantitative characteristics with brief descriptions are shown in Tables 1a and 1b (the editions of the printed dictionaries corresponding to the analyzed electronic version are specified).

Table 1a. Summary of lexical resources in the study: descriptions of dictionaries

Resource	Title [reference], year of the first edition	Editor(s)	Brief description and individual features
Explanatory dictionaries of classical type			
USH	Explanatory Dictionary of the Russian Language [9], 1935	D. N. Ushakov	influence of the Soviet ideology on definitions and examples; detailed system of style labels; obsolescence of the whole dictionary
OZH	Explanatory Dictionary of the Russian Language [10], 1949 (1992)	S. I. Ozhegov, N. Yu. Shvedova	popular normative dictionary; core vocabulary of the Russian literary language; brief examples
MAS	Small Academy Dictionary (Dictionary of the Russian language) [8], 1957	A. P. Evgenyeva	scientific approach, definitions with high accuracy; specific presentation of shades of meaning (à reduced number of isolated meanings); large number of usage examples
BTS	Big Dictionary of the Russian Language [14], 1998	S. A. Kuznetsov	MAS successor with a significantly extended word list; concise layout due to space limitations (one volume)
EFR	New dictionary of Russian [28], 2000	T. F. Efremova	large word list; extended number of meanings; systematic representation of regular polysemy; a large number of morphemes and MWEs; tendency to scientific definitions; no usage examples
ZLZ	Russian Grammar Dictionary [32], 1977	A. Zaliznyak	grammar dictionary (no definitions); one of the largest wordlists in the Russian lexicography by the time of first edition; the basis of almost all Russian lemmatizers
Synonym dictionaries			
ABR	Russian dictionary of synonyms and semantically similar expressions [1], 1900	N. Abramov	the oldest resource in the study, often used for Russian NLP
EVG	Dictionary of synonyms [6], 1970	A. Evgenyeva	large word list; significant number of usage examples, relies on the same initial data as BTS and MAS
BAB	Dictionary of synonyms of the Russian Language [7], 2011	L. Babenko	modern ideographic thesaurus

Resource	Title [reference], year of the first edition	Editor(s)	Brief description and individual features
Electronic lexical resources			
RWN	Russian Wordnet (http://wordnet.ru) [12], 2003		automatic translation of approx. 45% of PWN synsets based on parallel corpus, bilingual dictionaries and dictionaries of synonyms
WIKT	Machine-readable Wiktionary (http://ru.wiktionary.org) based on data from Russian Wiktionary [18], 2004		free multilingual online dictionary and thesaurus that can be collaboratively edited by users
RUT	Thesaurus RuThes-lite (http://www.labinform.ru/pub/ruthes) [21], 2014		linguistic ontology consisting of concepts and their relationships; same-root words (different POS) can belong to the same concept; concepts provided with definitions from WIKT

Table 1b. Summary of lexical resources in the study: quantitative characteristics (the values in parentheses in columns 3 and 4 correspond to synsets)[3]

Resource	# of entries, *10^3	# of unique lexical units, *10^3	# of MWE, *10^3	# of defs, *10^3
Explanatory dictionaries				
USH	88.8	87.1	0.0	130.5
OZH	41.2	40.3	0.0	n/a
MAS	83.5	81.6	0.0	135.8
BTS	76.3	103.2	0.0	111.8
EFR	135.2	123.7	2.3	219.0
ZLZ	93.4	93.4	0	0
Synonym dictionaries				
ABR	5.4	5.4 (16.0)	0.0 (2.1)	0
EVG	5.5	4.6 (16.4)	0.0 (0.3)	n/a
BAB	5.0	5.1 (19.6)	0.0 (1.2)	5.0
Electronic lexical resources				
RWN	51.7	30.8	9.3	74.6[3]
WIKT	193.5	192.0	5.8	161.2
RUT	26.4	96.7	46.6	54.9

All dictionary data were converted to a uniform machine-readable representation. For each entry we kept headword (with variations), definitions, and synonyms. Headwords and synonyms were lowercased; diacritics removed. In rare cases it produced duplicate records that were then removed, e.g. (OZH):

[3] Translated synsets are provided with glosses from original PWN synsets

Ex. 1. «Заброни́ровать»—см. брони́ровать. (Zabronírovat'—sm. bronírovat').
Reserve, book.
Ex. 2. «Забронирова́ть»—см. бронирова́ть. (Zabronirovát'—sm. bronirovát').
Armor, armour.

Additionally, two corpora were used in the study: Russian National Corpus (RNC, http://www.ruscorpora.ru) and Google Books Ngram Corpus (GBN, https://books.google.com/ngrams). RNC [29], first published in 2004, contains nowadays more than 192 mln tokens. In our study, we employed pre-processed RNC frequency lists (http://ruscorpora.ru/corpora-freq.html). GBN [19] contains year-by-year n-gram frequencies (up to 5-grams) from about 6% of all ever-published books in different languages. The Russian subcorpus of GBN contains about 103 billion tokens according to our calculations, which is much more than indicated by the authors—about 67 billion tokens. It could be explained by differences in token counting. Only unigrams that contain letters (and possibly hyphens) were taken into account in our work. It should be noted that there are words written in Latin alphabet in the Russian subcorpus. Both corpora word lists were lemmatized with *mystem* (https://tech.yandex.ru/mystem).

4. Analysis of lexical resources

4.1. Word lists analysis

Word lists of resources under consideration cover different parts of the Russian lexicon. The size of the word list itself is not sufficient to draw any conclusions. For example, WIKT contains about 35,000 proper nouns (about 18% of the whole volume). Moreover, authors of lexicographic resources treat derivative words, including gender-specific variants, in different ways. E.g. MAS contains separate entries for *«второклассник» (vtoroklassnik, «second-grade school boy»)* and *«второклассница» (vtoroklassnitsa, «second-grade school girl»)*, whereas BTS contains *«второклассник» ("second-grade school boy»)* as headword, and *«второклассница» ("second-grade school girl»)* as a variant.

Dictionaries' overlap seems to be a more suitable measure. Table 2 shows pairwise overlaps (in thousands) above the main diagonal and share of the overlap in the whole dictionary for the smaller resource in the pair below the main diagonal.

It was anticipated that there is a high degree of overlapping (80–90% in average) between the classical explanatory dictionaries (Table 2). The most intersecting dictionaries are MAS and BTS that share the same initial data sources [14]: 90.5% of MAS word list was included in BTS. Also note that WIKT word list includes many words from the classical explanatory dictionaries. This finding could also be explained by the large number of WIKT entries. RuThes-lite (RUT) and Russian Wordnet (RWN) contain a lot of multiword expressions (about 50% and 30% of the word list, respectively, see Table 1), it leads to smaller overlaps with other explanatory dictionaries.

Table 2. Overlaps between dictionary word lists

	BTS	EFR	MAS	OZH	RUT	RWN	USH	WIKT	ZLZ
BTS		85.8	73.8	38.3	39.2	18.8	63.3	80.1	72.5
EFR	0.831		74.2	38.1	38.5	19.4	70.0	89.3	80.5
MAS	0.905	0.909		36.3	36.0	17.6	61.2	66.6	66.8
OZH	0.951	0.945	0.901		22.8	13.2	35.0	36.8	37.3
RUT	0.406	0.398	0.441	0.567		14.2	31.9	41.6	36.7
RWN	0.611	0.628	0.571	0.428	0.461		17.4	20.1	18.9
USH	0.727	0.803	0.750	0.868	0.366	0.564		62.1	68.6
WIKT	0.776	0.722	0.817	0.912	0.430	0.653	0.713		79.4
ZLZ	0.776	0.862	0.819	0.926	0.393	0.612	0.787	0.850	

While Table 2 quantifies the overlap between dictionaries, Fig. 1 depicts the number of unique words in the resources (words that are presented only in one dictionary). In order to make this comparison more fair for traditional dictionaries we filtered out proper names from WIKT and multiword expressions from all resources.

The analysis with and without proper names and MWEs resulted in several findings:

1) Proper names and MWEs constitute one third of all lexical units.
2) In RWN and RUT there are 9 and 50 thousand of unique headwords respectively, but there are only 1 and 6 thousand after the removal of MWEs (Fig. 1).
3) The filtering of proper names and MWEs in WIKT reduced to half the number of unique headwords—from 87 to 47 thousand (Fig. 1).
4) 5 traditional dictionaries contain 731 MWEs out of 60,000 lexical units.

Both charts illustrate succession in creation of Russian explanatory dictionaries (BTS, MAS, USH, OZH, and EFR): there are 107,800 words occurring in at least two of these dictionaries. In this regard, Russian Wiktionary (WIKT), Russian WordNet (RWN), and RuThes-lite (RUT) contain almost twice as less crossings (i.e. words that are represented in at least two dictionaries out of three): 46,200[4] words. 59,500 words (see Fig. 1) correspond to a union of words from traditional dictionaries not included in any other Russian dictionary. These data can be useful for creating new explanatory dictionaries and for the further developing of WIKT and RUT.

[4] This number is not presented on Fig. 1.

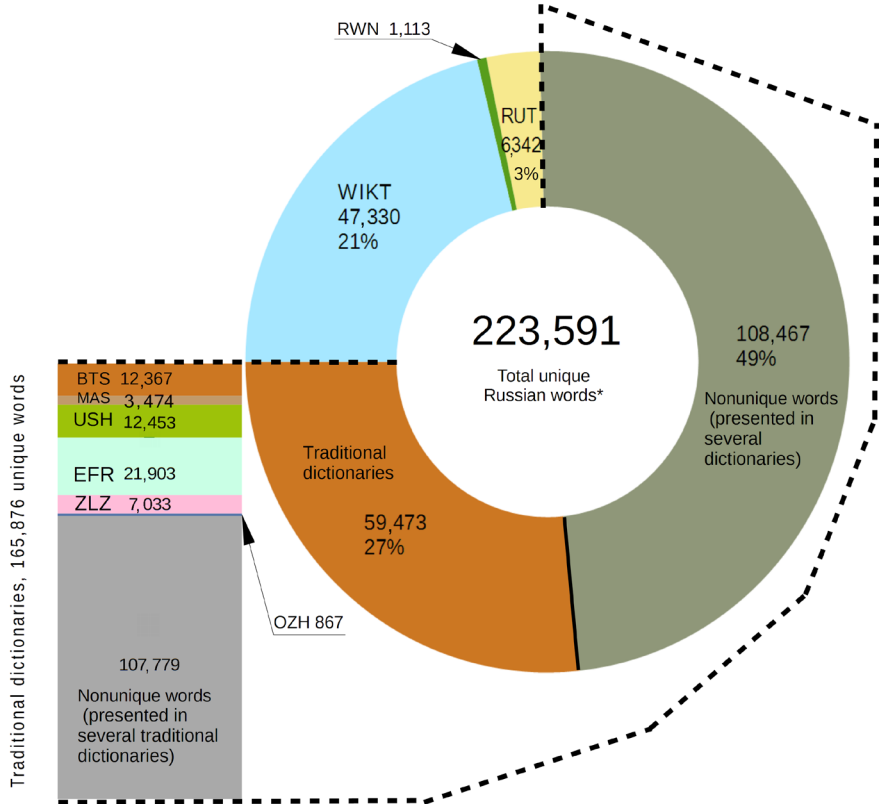

Fig. 1. Number of unique words in traditional dictionaries (vertical stripe) and in all dictionaries (pie chart)

*) Three smallest values—ABR (450 unique words), BAB (290) and EVG (130)—are not depicted in the pie chart, but accounted for in the total.

4.2. Corpora coverage

On the next stage of our study we quantified coverage of RNC and GBN with respective dictionary headwords. We employed two approaches to measure corpora coverage: 1) overlap of dictionary word lists and top-frequency lists extracted from corpora and 2) direct coverage of corpora (excluding stopwords).

The first approach simply measures intersection of word lists, the second one takes word frequencies in corpora into account. After lemmatization of the RNC frequency list around 263,000 unique terms remain; the number of unique lemmas in GBN corpus is more than 1.7 mln. The latter large number is partially due to a high level of misprints and systematic OCR errors [19]. Both corpora contain also a significant portion of proper names. We considered only top-100k most frequent lemmas for each corpus (the overlap of these two lists is 68,000 terms). Fig. 2 shows the presence

of the most frequent words from each of the corpora in dictionaries. For example, from the 1,000 most frequent RNC words 95.3% of them are presented in WIKT (i.e. 47 words are absent).

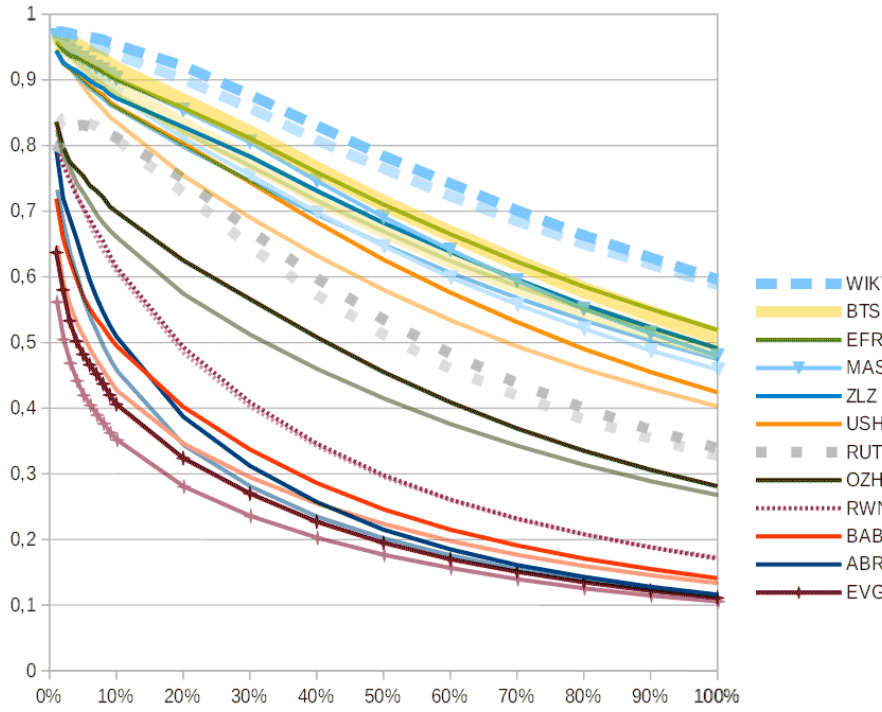

Fig. 2. Coverage of top-100k terms from RNC (solid lines) and GBN (softer lines)

On the right-side pane in Fig. 2 the dictionaries are listed in descending order of coverage for both corpora wordlists. The Figure clearly shows three groups of resources: 1) modern large dictionaries with good coverage (WIKT, BTS, EFR, and MAS); 2) borderline dictionaries (USH, RUT, and OZH); 3) synonym dictionaries with a lower coverage (RWN, BAB, ABR, and EVG). It is important to note that the dictionaries' ranks are the same for both corpora. This allows us to be more confident in generalizing the conclusions obtained from the data of either of two corpora.

The second approach accounts for all words presented in the corpora dataset (see Sec. 2) along with their frequencies except for stopwords that account for 34.5% and 28.8% tokens in RNC and GBN, respectively. To make the comparison fair for wordnets that typically do not contain functional words, we excluded stopwords from calculation. Corpora coverage with the dictionary words lists is shown in Fig. 3.

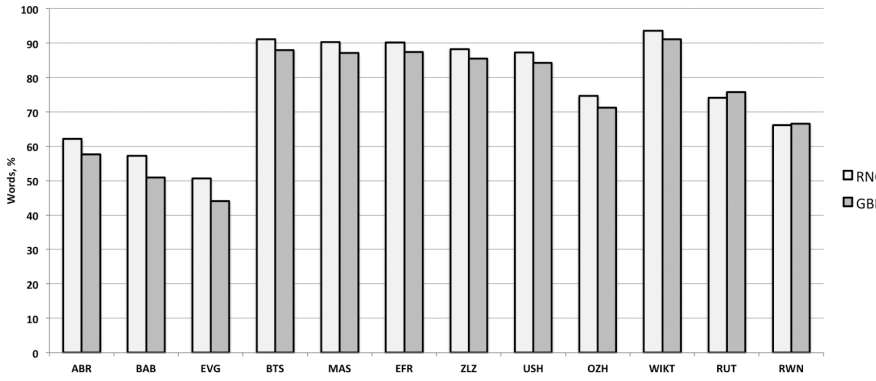

Fig. 3. RNC and GBN coverage

4.3. Analysis of modern lexicon coverage

As the results in the previous section show, all lexical resources cover core lexis quite well. We used GBN to evaluate how dictionaries under consideration reflect neologisms. To this end, we chose two 20-year intervals: 1970–1989 and 1990–2009, and selected lemmas that appeared at least in one thousand books in each time period. Then we ordered them by descending ratios of frequencies in the newer / older subcorpora and took top-2k words. The list contains many proper names, OCR errors, spelling variants and results of incorrect lemmatization. However, according to our manual evaluation, about a half of the list can be regarded as good 'headword candidates'. Fig. 4 shows how many neologisms from the 2,000 are presented in the dictionaries (Fig. 4 shows only dictionaries covering at least 50 lemmas). It is interesting to note that the attempt to create a list of obsolete words in the same simple way (by ordering the list by ascending frequency ratios) did not succeed: all top words were OCR errors or typos.

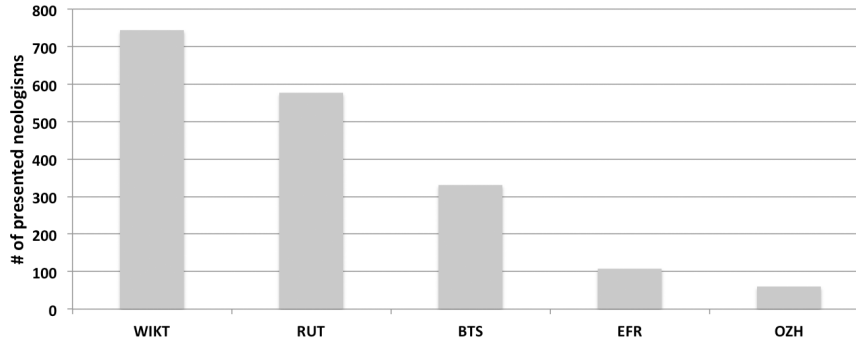

Fig. 4. Neologisms presented in dictionaries (from the list of 2,000 words)

4.4. Synonymy analysis

Synonymic resources are presented by three printed dictionaries of synonyms (ABR, EVG, and BAB), two thesauri (RWN and RUT), and WIKT. The latter combines properties of explanatory dictionary and dictionary of synonyms. All these dictionaries form their synsets/concepts uniquely, except WIKT, whose synsets are attached to respective headwords and are not necessarily coordinated, cf.:

Ex. 1. ***«Собака» (sobaka, «dog»)**, «пёс» (pyos, «dog»), «псина» (psina, «dog»), «друг человека» (drug cheloveka, «friend of a man»), «четвероногий друг» (chetveronogiy drug, «four-legged friend»).*
Ex. 2. ***«Пёс» (pyos, «dog»)**, «собака» (sobaka, «dog»), «кобель» (kobel', «male dog»).*

These two synsets, created from two different entries (for headwords «*Собака*» and «*Пёс*» respectively), will be treated as reflecting different meanings. However, synsets represented by the same set of words are considered as equal. Note that we did not consider one-word synsets[5] as well. For every synset we counted all pairs formed from its words (i.e., a synset consisting of 4 words forms 4 * 3/2 = 6 synonym pairs). Table 3 summarizes data for six dictionaries[6].

Table 3. Quantitative characteristics of synsets from different dictionaries

	ABR	BAB	EVG	RUT	RWN	WIKT
Total # of synsets, thousands	7.5	4.9	5.4	22.7	11.0	33.0
Average synset size, words	6.8	6.0	3.9	4.9	2.2	2.9
Total # of synonym pairs, thousands	125.9	107.5	45.7	378.1	15.0	121.1

For synsets comparison we made an assumption that any pair of synset's terms defines roughly the meaning of the synset. This is quite a strong assumption that is often violated. Two following examples (from BAB) show that even when one synset is a subset of another synset they still may have different meanings:

Ex. 1. *«Начинающий» (Nachinayushchiy, «beginner»), «дебютант» (debyutant, «debutant»), «новенький» (novenʹkiy, «newcomer»), «новичок» (novichok, «novice»)—тот, кто впервые выступает на сцене, участвует в соревнованиях; делает первые шаги на каком-либо публичном поприще (a person who performs on stage for the first time, participates in competition; makes his/her first steps in any public arena).*
Ex. 2. *«Начинающий» (Nachinayushchiy, "beginner"), «дебютант» (debyutant, "debutant")—недавно приступивший к какому-либо роду деятельности (о человеке, группе лиц и т.п.) (a person who recently begun any kind of activity (about a person, group of individuals, etc.)).*

---

[5] E.g. RWN contains 14,000 one-word synsets.

[6] Dictionaries of synonyms cannot include synsets consisting of just one word by design, yet thesauri (RWN, RUT) can. So we considered only 2+ word synsets.

So we calculated the number of synset pairs between dictionaries, that Jaccard similarity coefficient is no less that 0.5 (Table 4). We analyzed synsets consisting of two or more words; so all "overlapped" pairs have two or more words in common. Note that this method takes both within-dictionary (main diagonal) and intra-dictionary overlaps into account.

Table. 4. Synset overlapping

	ABR	BAB	EVG	RUT	RWN	WIKT
ABR	20,370	400	590	90	410	1,290
BAB		880	2,100	410	840	2,680
EVG			830	440	1,210	4,080
RUT				1,380	350	1,290
RWN					1,810	4,390
WIKT						12,620

As we can see from Table 4, EVG has greatly influenced the later Russian dictionaries of synonyms.

4.5. Quantitative analysis of definitions

It is natural to expect that comprehensive dictionaries differ not only by the size of their word lists, but also by a number of definitions in them. In order to compare the resources in this regard, we analyzed seven dictionaries out of 12 discussed in the paper. We treated shades of meaning (usually separated by a double vertical line) as separate definitions.

However because of ambiguous formatting of electronic versions of explanatory dictionaries we had, sometimes it was impossible to get all definitions for an entry. This is particularly true for «noticeable shift in meaning» [8], labeled by a single vertical line. So it could lead to slight inaccuracies in measurements, caused by detecting not all meanings for headwords. Nevertheless we suppose that it did not significantly affect the result of our experiments.

Table 5 shows the quantitative characteristics of definitions from dictionaries under consideration.

Table 5. Quantitative characteristics of definitions in dictionaries

	USH	MAS	BTS	EFR	WIKT	BAB	RUT
Unique definitions, thousands	110.4	121.5	97.3	155.6	81.0	4.1	10.4
Avg. definition length, words	5.43	5.37	5.10	5.51	6.87	10.19	7.21
# of words, thousands	72.2	77.2	67.6	94.7	45.4	4.1	28.4
Avg. # of definitions per entry	1.56 (1.47)	1.64 (1.63)	1.50 (1.47)	1.72 (1.62)	1.89 (0.83)	1.00 (1.00)	1.25 (0.57)

When calculating characteristics in Table 5 we considered only headwords presented in at least two resources and having at least one definition. Values corresponding to the whole set of entries (i.e. without filtration) are presented in parentheses (see the last row in Table 5).

One can see from Table 5 that the average definition length and the average number of definitions are similar for all traditional dictionaries, yet the same characteristics of electronic resources differ significantly. E.g., WIKT and RUT contain many entries without definitions at all, which obviously lowers the average number of definitions per entry. WIKT word list includes proper names that usually have only one meaning and do not occur in other dictionaries. The average number of word meanings (definitions) for an entry in WIKT is 1.89 essentially depends on part-of-speech. By 2011, the average number of definitions for verbs was 2.5, whereas for nouns, adjectives and adverbs the value laid in range 1.5–1.7 [30].

The distribution of entries by the number of definitions is shown in Fig. 5.

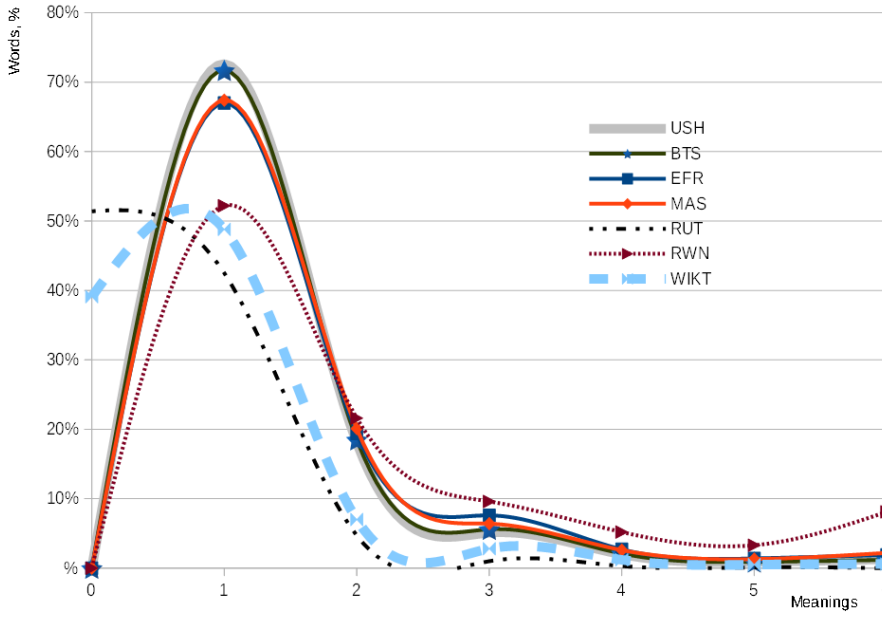

Fig. 5. Distribution of entries by the number of meanings (definitions)

Fig. 5 allows for interesting observations. Firstly, the distributions for USH and BTS almost coincide (both dictionaries have the largest share of monosemantic headwords). Secondly, the number of single-meaning words divides resources into two classes: academic (BTS, MAS, EFR and USH—a share of unambiguous words is about 70%) and other dictionaries (RUT, RWN and WIKT, drawn by dashed line, where unambiguous words comprise less than a half.

Note that we did not remove entries with zero definitions. Presence of some word means that a dictionary reflects it, but the quality of this reflection may vary and depend among other on definitions (some quantitative characteristics of definitions were discussed earlier).

4.6. Analysis of textual similarity of definitions

On the next stage of our study we compared textual similarity of same-word definitions in different dictionaries (note that we did not try to align the meanings of definitions). To this end we employed Monge-Elkan string similarity measure that combines word- and character-level similarity, demonstrates high performance and good balance between precision and recall [15, 5, 26].

Monge-Elkan similarity is not symmetrical, so we used the year of the first dictionary edition for selection of the direction of comparison (see Table 1). This direction reflects how definitions in newer resources resemble their predecessors' ones. In our study we used *DKPro Similarity* implementation of Monge-Elkan method [2].

Dictionaries contain a large number of 2–3 word definitions, which can skew similarity measurements, since such definitions are rather "standard" and occur in many dictionaries. So we filtered out such definitions, which resulted in exclusion of 176,000 lexical units. Typical examples of excluded definitions are:

- A widely used synonym;
- Ex. *«Помешкаться»—задержаться. (Pomeshkatsya—zaderzhatsya). Delay, linger.*
- lists of synonyms;
- Ex. *«Утопист»—мечтатель, фантазер. (Utopist—mechtatel›, fantazer). Dreamer, visionary.*
- A gloss without examples.
- Ex. *«Манка»—манная крупа. (Manka—mannaya krupa). Semolina.*

As similar we considered definitions with similarity value above 0.9. Fig. 6 depicts textual similarity of definitions in different resources as a graph: vertices are dictionaries; edge thickness is proportional to the number of similar definitions in a pair of resources; borrowings from an older to a newer dictionary are displayed clockwise; numbers reflect the percent of borrowings in recipient dictionary.

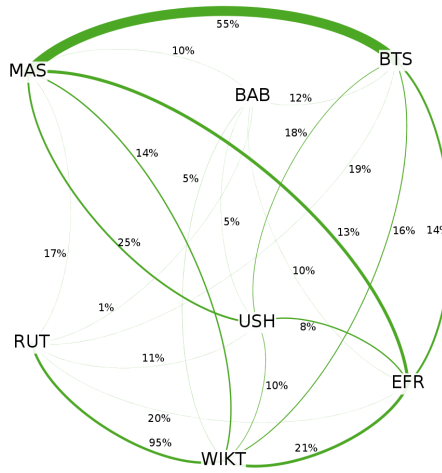

Fig. 6. Graph of textual similarity of definitions

5. Conclusion

Our results let us make the following conclusions.

1. Overlaps between dictionary word lists. A developed tradition and succession of different dictionary creation projects explain significant overlaps between word lists of traditional dictionaries. USH dictionary stands out in this regard, which could be explained by the fact that USH project is has not been developing anymore. A low overlap between RWN and other dictionaries, on the contrary, indirectly confirms the idea that a straightforward translation of a thesaurus into another language significantly reduces lexicon.

2. Number of unique words in dictionaries. As we found out, there are relatively few unique words and phrases (i.e. contained in only one dictionary). This fact is partly due to the choice of representation of derivatives—as a separate headword or inside an entry. At the same time in these dictionaries (even in EFR) there is a significant lack of multiword expressions, which are presented in electronic resources much better. In addition, a large number of unique terms in WIKT can be explained by the fact that its word list includes proper names (35,000 words out of 193,500).

3. Corpora coverage. Judging by the share of unique words, one would assume that the traditional dictionaries do not have good corpora coverage. However that is not true—especially with respect to BTS, EFR, and MAS. A noticeable "deficiency" of dictionaries of synonyms is quite clear and expected. The obtained results can give a raw estimate of Russian lemmas that are involved in synonymy relationships—about 60–70%.

4. Quantitative analysis of definitions. The share of monosemantic words, contained in traditional dictionaries, was significantly higher than in the electronic resources. This fact indicates the orientation of the latter towards actual word usage and a tendency to represent specific meanings.

5. Analysis of modern lexicon coverage. Finally, a comparison of dictionaries by presence of neologisms shows a great potential of modern electronic resources that can be dynamically modified. It does not mean that traditional dictionaries are obsolete. The lag from changes in a language gives an opportunity to reflect in the dictionary not just random, but established language phenomena: words, meanings, variations, etc.

The current situation in modern Russian lexicography reflects the transition period from traditional printed editions to large-scale projects based on large corpora and crowdsourcing. Traditional dictionaries based on manual sampling and data processing are regarded as high-quality sources, yet they are clearly behind the resources like WIKT, considering their volume and coverage of modern lexicon. At the same time, the specifics of electronic projects are often criticized for their quality.

We expect that our findings will be helpful for lexicographic practice—no matter what form will be chosen by dictionary authors.

Acknowledgment

Pavel Braslavski's and Mikhail Mukhin's contribution to the study was supported through grant #13-04-12020 "New Open Electronic Thesaurus for Russian" from the Russian Foundation for the Humanities. Mikhail Mukhin's work is also supported through Ural Federal University Competitiveness Enhancement Program # 02.A03.21.0006. Some parts of the research of Andrew Krizhanovsky are carried out in the project supported by grant # 15-04-12006 from the Russian Foundation for the Humanities, and the project "Veps corpus: computer morphological base development" of the basic research program of the Literature and language section of Department of history and philology RAS "Language and information technology" 2015–2017. We also thank Natalia Loukachevich for granting us access to RuThes-lite data.

References

1. *Abramov N.* Russian dictionary of synonyms and similar expressions on sense, M.: Russian dictionaries, 1999.
2. *Bär D., Zesch T., Gurevych I.* DKPro Similarity: An Open Source Framework for Text Similarity //ACL (Conference System Demonstrations).—2013.—P. 121–126.
3. *Braslavski, P., Mukhin, M. Y., Lyashevskaya, O. N., Bonch-Osmolovskaya, A. A., Krizhanovsky, A. A., & Egorov, P.* (2013). Yarn Begins. Proceedings of Dialog-2013.
4. *Braslavski, P., Ustalov, D., & Mukhin, M.* (2014). A spinning wheel for YARN: user interface for a crowdsourced thesaurus. In Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden (pp. 101–104).
5. *Cohen W., Ravikumar P., Fienberg S.* A comparison of string metrics for matching names and records //KDD Workshop on Data Cleaning and Object Consolidation.—2003.—T. 3.—C. 73–78.
6. Dictionary of synonyms, ed. A. Evgenyeva, L.: Nauka, 1975.
7. Dictionary of synonyms of the Russian Language, ed. L. Babenko. M.: AST: Astrel, 2011.
8. Dictionary of the Russian Language (Malyy akademicheskiy slovar'). Four volumes. RAS, Institute for Linguistic Studies, ed. A. P. Evgenyeva, M: Russkiy yazyk; Poligrafresursy, 1999.
9. Dictionary of the Russian Language (Tolkovyy slovar' russkogo yazyka). Four volumes, ed. D. N. Ushakov, editor. 1935–1940, State Publishing House of Foreign and National Dictionaries.
10. Explanatory Dictionary of Russian Language: 80,000 words and set phrases. Eds. S. I. Ozhegov and N. Yu. Shvedova. Moscow: Azbukovnik, 1999.
11. Fellbaum, Christiane. WordNet. Blackwell Publishing Ltd, 1998.
12. *Gel'fejn'bejn I. G., Goncharuk A. V., Lehel't V. P., Lipatov A. A., Shilo V. V.* (2003), Automatic translation of Wordnet in russian [Avtomaticheskij perevod semanticheskoj seti Wordnet na russkij jazyk], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2003"

[Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2003"], Protvino.
13. *Geyken, Alexander, and Lothar Lemnitzer.* «Using Google books unigrams to improve the update of large monolingual reference dictionaries.» In Proceedings of the 15th EURALEX International Congress, pp. 362–366. 2012.
14. Great Dictionary of the Russian Language (Bol'shoy tolkovyy slovar'), ed. S. A. Kuznetsov, St. Petersburg, Norint (1998).
15. *Jimenez S.* et al. Generalized Mongue-Elkan method for approximate text string comparison //Computational Linguistics and Intelligent Text Processing.—Springer Berlin Heidelberg, 2009.—C. 559–570.
16. *Kennedy, Alistair, and Stan Szpakowicz.* «Evaluating Roget's Thesauri.» In ACL, pp. 416–424. 2008.
17. *Kilgarriff, Adam, and Iztok Kosem.* «Corpus tools for lexicographers.» Electronic lexicography (2012): 31–56.
18. *Krizhanovsky A., Smirnov A.* An approach to automated construction of a general-purpose lexical ontology based on Wiktionary // Journal of Computer and Systems Sciences International, 2013, Vol. 52, No. 2, pp. 215–225.
19. *Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman, Slav Petrov.* Syntactic annotations for the Google Books Ngram Corpus. Proc. of the ACL 2012 System Demonstrations, p. 169–174, July 10–10, 2012, Jeju Island, Korea
20. *Litkowski, Kenneth C.* Towards a meaning-full comparison of lexical resources // Association for Computational Linguistics SIGLEX Workshop. 1999.
21. *Loukachevitch N. V., Dobrov B. V., Chetviorkin I. I.* RuThes-lite, A publicly available version of thesaurus of russian language RuThes // Computational Linguistics and Intellectual Technologies: Conference "Dialogue", Issue 13 (20), pp. 340–349, Moscow, RGGU, 2014.
22. *Luk, Robert WP, and Venus MK Chan.* «A Quantitative Analysis of Word-Definition in a Machine-Readable Dictionary.» (1995).
23. *Michael Matuschek and Iryna Gurevych.* "High Performance Word Sense Alignment by Joint Modeling of Sense Distance and Gloss Similarity." In Proceedings of the the 25th International Conference on Computational Linguistics (COLING 2014), pp. 245–256.
24. *Meyer, Christian M., and Iryna Gurevych.* «Wiktionary: a new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography." Electronic Lexicography (2012): 259–291.
25. *Michiels, Archibald, and Jacques Noël,* Approaches to Thesaurus Production // In Proceedings of the 9th conference on Computational linguistics-Volume 1, pp. 227–232. Academia Praha, 1982.
26. *Monge A. E.* et al. The Field Matching Problem: Algorithms and Applications // KDD.—1996.—C. 267–270.
27. *Navigli, Roberto, and Simone Paolo Ponzetto.* "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network." Artificial Intelligence 193 (2012): 217–250.
28. New dictionary of Russian: the sensible and word-formation (Novyy tolkovo-slovoobrazovatel'nyy slovar' russkogo yazyka), ed. T. F. Efremova, 1996.

29. *Plungyan, V. A., T. I. Reznikova, and D. V. Sichinava*. «The National Corpus of the Russian Language: general characteristics.» Nauchno-tekhnicheskaia informatsiia, ser. 2 (2005): 913.
30. *Smirnov A. V., Kruglov V. M., Krizhanovskiy A. A., Lugovaya N. B., Karpov A. A., Kipyatkova I. S.* A quantitative analysis of the lexicon in Russian WordNet and Wiktionaries (In Russian) // In Trudy SPIIRAN (St. Petersburg, 2012), Issue 23, pp. 231–253.
31. *Yoshida, Sho, Hiroaki Tsurumaru, and Tooru Hitaka*. «Man-assisted machine construction of a semantic dictionary for natural language processing.» In Proceedings of the 9th conference on Computational linguistics-Volume 1, pp. 419–424. Academia Praha, 1982.
32. *Zaliznyak A*. Russian Grammar Dictionary (Grammaticheskij slovar' russkogo jazyka). Moskva, 1977.