

An Approach to Automated Construction of a General-Purpose Lexical Ontology Based on Wiktionary

A. A. Krizhanovsky and A. V. Smirnov

St. Petersburg Institute for Informatics and Automation, St. Petersburg, Russia

e-mail: Andrew.krizhanovsky@gmail.com

Received April 16, 2012

Abstract—An approach to the design of a system of automated construction of a general-purpose lexical ontology is proposed, and the architecture of such a system is described. Wiktionary is chosen as the online dictionary because it has a large database of words with translations into many languages. The structure of the dictionary entry is considered using the English Wiktionary as an example. This structure is used to design a database for storing the retrieved information. Ontologies are an important part of knowledge management systems. Ontologies require the development of approaches and algorithms for their construction. Lexical ontologies are constructed, and the main features of two ontology databases based on the Russian and English Wiktionaries are compared. The dynamics of numerical parameters of Wiktionaries and general-purpose lexical ontologies for 2010–2012 constructed by the authors is analyzed.

DOI: 10.1134/S1064230713020068

INTRODUCTION

In computational lexicology (a branch of computational linguistics) one can see a gradual transition (in terms of terminology and semantic content) from machine-readable dictionaries to lexical knowledge bases and then to lexical ontologies. A machine-readable dictionary [1] represents paper dictionary data in electronic form thus enabling these data to be processed on a computer. The lexical knowledge base differs from the machine-readable dictionary in that the word meanings are explicitly indicated and connections between the corresponding meanings are specified, which makes it possible to use these data for logical inference [2].

This paper describes an approach to building a general-purpose lexical ontology to integrate lexical and semantic information.

Lexical ontology contains structured information about the words and includes semantic relationships (e.g. synonymy, hypernymy, and holonymy) between the meanings of words [3]. The phrase *general purpose* in the name of the ontology implies that there is no attachment to a particular subject area; i.e., there is an attempt to include all the words of the language into the dictionary of the ontology. However, a significant part of applied ontologies is constructed for a specific subject area with the indication of relations between the concepts of this area [4]. There is a direction of the automated construction of “dedicated lexical ontologies” in which the argument for their creation is that such specialization significantly reduces the size of the ontology and thus reduces its processing time [5]. Currently, however, it is the insufficient size of dictionaries, thesauri, and ontologies that poses a great problem for applications [6].

Thus, a *general-purpose lexical ontology* contains structured information about words and semantic relations. At the same time, there is no attachment to a particular subject area. WordNet is considered one of the most successful projects of this kind.

WordNet is a machine-readable English dictionary and thesaurus. The concept of this dictionary is based on psycholinguistic theories, which were used to identify the meanings of words and relationships between words and meanings, as well as relationships between the meanings themselves [7]. WordNet data are used for many purposes, such as the definition of the word [8–10] and calculation of the self-consistency and coherence of sentences in the text [11, 12]. Many ontologies and knowledge bases include WordNet data or are connected with lists of WordNet synonyms, e.g., OpenCyc [13] and DBpedia [14]. There are several knowledge bases including not only WordNet but also Wiktionary. They are discussed below. Among them are the lexical-semantic resource UBY [15] for the English and German languages and the Lexvo.org system [16] containing relationships in the form of RDF triples between words of about 7000 languages.

Wiktionary was selected as a source of data for the construction of the general-purpose lexical ontology (hereinafter referred to as *ontology*) for several reasons.¹ A wiktionary is a freely editable multipurpose multilingual online dictionary and thesaurus that can be edited by users. It contains interpretations and translations of words, the description of phonetic and morphological properties, and semantic relations. In addition, it contains pronunciations of words (transcriptions and audio files), rules of word division into syllables, stresses in words, information about the etymology of words, as well as quotes from literary works that illustrate the use of words, and even video and photos illustrating the meaning of words. Wiktionary's advantages are a big volume and variety of lexicographic data. In [17–18], it is shown that the German Wiktionary is comparable to GermaNet and OpenThesaurus thesauri in terms of the amount of data, and the English Wiktionary even exceeds the amount of WordNet data.

The scientific importance of multifunctional online dictionaries (wiktionaries) is confirmed by the fact that the wiktionary and its sister project Wikipedia [19] is widely used in scientific experiments. The wiktionary is used for various purposes related to the processing of text and speech:

- (1) in machine translation between Dutch and Afrikaans [20];
- (2) for automated identification of the part of speech of words using a hidden Markov model for three languages—English, Vietnamese, and Korean [21];
- (3) in text processing by the NULEX parser, where a part of Wiktionary data (tenses) is integrated with WordNet and VerbNet databases [22];
- (4) in speech recognition and synthesis, where the wiktionary provides a basis for the rapid creation of pronunciation dictionaries [23];
- (5) for construction of ontologies [6];
- (6) in the representation of ontologies [26].

Below, we give a brief overview of the structure of entry of the English Wiktionary (using the entries for the word @@ and @@ as examples). The approach and architecture of the system for ontology construction are considered. Ontologies are constructed on the basis of lexicographic data of wiktionaries, which made it possible to analyze and compare the vocabulary of the English language in multilingual dictionaries (English and Russian Wiktionaries) with WordNet data.

1. WIKTIONARY AND THE STRUCTURE OF ITS ENTRY

The wiktionary contains not only interpretations and translations of words. Entries also describe phonetic and morphological properties of words and indicate semantic relationships. Several complementary information structures are used in the wiktionary in order to specify semantic properties, such as semantic categories and usage labels (they specify style, subject area, and language).

The structure of the entry of the wiktionary is rather strict and is clearly defined by rules. There are such rules in the English Wiktionary,² in the Russian Wiktionary,³ and probably in the other 170 wiktionaries.⁴ The presence of the structure and formatting rules of entries make it possible to look at the entry as an interesting object from the point of view of automated data retrieval, e.g., using regular expressions [25]. Automatic retrieval should convert the “implicit” structure, which can be understood only by a human reader, into an explicit form suitable for computer processing thus enabling the use of wiktionary data in various projects associated with text processing in the future.

Let us consider the structure of the wiktionary using examples from the English Wiktionary. In the entry, one can distinguish the following sections: etymological; pronunciation; semantic; inflection, conjugation, or declension; semantic relations; related terms; and translation. Let us illustrate all these sections with fragments of entries.

Etymological section. This section (Fig. 1) contains information about the history of the word; i.e., it describes phonetic and semantic changes that the word has undergone. Different views with references to the relevant literature may be given. If possible, etimologies should be accompanied with references to the sources of information. There are hundreds of special templates for frequently cited sources (see http://en.wiktionary.org/wiki/Category:Reference_templates) Lexicographers (editors of the wiktionary) link the etymology text with the related entries.

¹ Hereinafter names of specific projects (English Wiktionary, Russian Wiktionary) are written with a capital letter and names of all dictionaries of this type, i.e., wiktionaries, are written in lower case.

² See <http://en.wiktionary.org/wiki/Wiktionary:ELE>.

³ See http://ru.wiktionary.org/wiki/@: @@_@@_@@_@@.

⁴ See <http://meta.wikimedia.org/wiki/Wiktionary/Table>.

Etymology

First attested in 1951; from the Czech *háček* (“háček”, literally “little hook”), the diminutive form of *hák* (“hook”, from the Middle High German *hāken*, from the Old High German *hāko*, “hook”, from the Proto-Germanic **hakō*, “hook”, from the Proto-Indo-European **keg-*, **keng-*, “peg”, “hook”) + the diminutive suffix *-ek*; parallel to the formation of the English *hooklet* and the German *Häkchen*; cognate with the German *Haken* (“hook”), the Old English *haca* (“hook”, “door-fastening”), and the Modern English *hook* and *hake* (more information *sub verbis*).

Fig. 1. Etymology section of the entry “háček.”

Pronunciation

■ IPA: /'n'ezɲij/


■ Audio 

Fig. 2. Pronunciation section of the Russian entry @@.

Adjective

нежный (néžnyj)

1. gentle, tender, fond [quotations ▲]

■ 1847, Ivan Turgenev, *It Tears Where It is Thin*:

Я слышу за человека насмешливого и холодного, и очень этому рад: с такой репутацией легко жить... но вчера мне пришлось прикинуться озабоченным и нежным.

Ja slyvú za nasmešlivogo i xolódnogo čelovéka, i óčen' rad étomu: s takóej reputácijej legkó žít'... no včerá mne prišlós' prikínut'sja ozabóčennym i néžnym.

I have a reputation as a mocking and cold person, and I'm very happy about it. It is easy to live with such a reputation... but yesterday I had to pretend to be concerned and gentle.

2. delicate, soft, fine [quotations ▼]

3. gentle, delicate, tender (sensitive or painful) [quotations ▼]

Fig. 3. Semantic section of the Russian adjective @@ with three definitions, a quotation, its transcription, and its translation.

Pronunciation section contains pronunciation in the International Phonetic Alphabet and an audio file voiced by a native speaker (Fig. 2).

Semantic section includes the interpretation and quotations that illustrate each meaning of the word (Fig. 3). A feature of the interpretations is the use of links to entries of the same dictionary. Quotations (example sentences) are accompanied by bibliographic information, such as the author, title, and year of publication. The quotations for the languages in non-Latin alphabets, a transcription for the quotations must be given; for the languages other than English, a translation is to be given (Fig. 3).

The following section can be titled differently and can contain different types of data depending on the part of speech of the entry: *Inflection* or *Conjugation* for verbs, or *Declension* for nouns and adjectives; this section is present only in non-English entries (Fig. 4).

declension of нежный [hide ▲]				
	singular			plural
	masculine	feminine	neuter	
nominative	нежный	нежная	нежное	нежные
genitive	нежного	нежной	нежного	нежных
dative	нежному	нежной	нежному	нежным
accusative	<small>inanimate</small>	нежный	нежное	нежные
	<small>animate</small>	нежного	нежную	нежных
instrumental	нежным	нежной (нежною)	нежным	нежными
prepositional	о нежном	о нежной	о нежном	о нежных
short form	нежен	нежна́	нежно	нежны́, нежны́

Fig. 4. Declension section of the Russian adjective @@.

Synonyms

- (*tender, fond*): ласковый, любящий
- (*delicate, soft, fine*): тонкий, лёгкий, изящный, приятный, мягкий
- (*tender, delicate*): хрупкий, мягкий, уязвимый, слабый, хилый, изнеженный

Antonyms

- (*tender, fond*): грубый
- (*tender, delicate*): выносливый, жёсткий

Hypernyms

- (*tender, fond*): чувство, любовь
- (*delicate, soft, fine*): толщина, вес, твёрдость
- (*tender, delicate*): жёсткость, слабость

Fig. 5. Semantic relations section of the Russian adjective @@.

Related terms

- нежность, неженка, нега, изнеженность
- изнеженный
- нежить
- нежно

Fig. 6. Related terms section of the Russian entry @@.

Semantic relations section describes semantic relationships (synonyms, hypernyms, etc.) for each meaning of the word (Fig. 5).

Related terms section groups paronyms classified by parts of speech (Fig. 6 shows nouns, adjectives, verbs, and adverbs). This section contains a set of words (family of words) in the same language that have strong etymological connections but are not derived terms. The word list includes hyperlinks to relevant entries.

Translation section contains translations of each meaning of the word into foreign languages (Fig. 7). Translations are given in the form of links to the relevant entries about foreign words. The translation section of the Russian Wiktionary includes a two- or three-letter language code at the end of the name of each

Translations



Fig. 7. Translation section of the entry “háček.”

language (Fig. 7). The set of these codes is a part of the base of lexicographic constants presented in detail in the next section.

2. APPROACH AND DESIGN OF THE SYSTEM OF CONSTRUCTION OF A GENERAL-PURPOSE LEXICAL ONTOLOGY

The proposed automated approach consists of two phases: (1) initial analysis of data of the online dictionary by experts to determine the features of its structure, and (2) the subsequent automatic construction of a general-purpose lexical ontology by means of the developed computer system.

In the wiktionary (www.wiktionary.org), special tags and other markers indicate semantic elements and define the hierarchical structure of the entry. However, without special processing of the dictionary, only full text searches in the entries or hypertext navigation can be used. For complex searches (for example, to obtain a list of all synonym sets that contain the word), the online dictionary should be converted into a format suitable for computer processing. Therefore, a team of linguists (experts in the online dictionary under examination), and engineers (database and IT experts) should work at the first phase of the approach.

The *first phase* includes the following subtasks:

- (1) analysis of online dictionary entries;
- (2) determination of the entry structure;
- (3) identification of “reference” elements of the entry (key words, elements of hypertext markup) using regular expressions [24];
- (4) design of a relational database based on the model of the entry structure to store the retrieved data;
- (5) configuration of the ontology construction system.

In designing the ontology relational database, sections of the online dictionary entry to be processed are determined. Presently, the system retrieves from the entry the following lexicographic data: language, part of speech, meaning of the word (interpretation), quotation (illustrating the meaning of the word), semantic relationships, and translation (Table 1). In the future, it is planned to take into account data from all sections.

Next, based on the structure of the entry and its processed sections, the computer system is adjusted to automatically construct an ontology using wiktionary data.

The *second phase* of the ontology construction is carried out in automatically (Fig. 8). The objectives of the second phase are the construction of the ontology (the main objective of the phase) and the automatic collection of debugging data.

Developers need debugging data at the first phase of the ontology construction for the next iteration (for example, to expand the base of lexicographic constants when a language code is added because the data are retrieved from a multilingual dictionary). In other words, the development of the system is carried out iteratively, modules are gradually added to the system to extract lexicographic data (semantic module,

Table 1. Sections and data of a dictionary entry included in the ontology

Entry Section	Section Data	General-purpose lexical ontology
@	@	Language, part of speech
Pronunciation Section	Transcription, audio file	–
Semantic Section	@	synonyms, antonyms, hyponyms, hypernyms, meronyms, holonyms
Etymological Section	Etymology of the word	–
Related terms	List of paronyms	–
@	@	–
Translation Section	Translations	Translations

translation module, etc.). The developed architecture of the system of the ontology construction is modular and expandable. It is planned to create modules for each section of the entry. Fig. 8 illustrates interactions between the main parts of the system.

The software system requires the specification of the following input parameters. Firstly, language of the wiktionary (Russian or English) since there are differences in the structure of entries in the Russian and English Wiktionaries. Secondly, the address of databases of online dictionaries and ontology, i.e., the parameters to connect to remote databases such as IP-address, database name, user name, and password. The source of data is the online dictionary database, i.e., a corpus of entries.

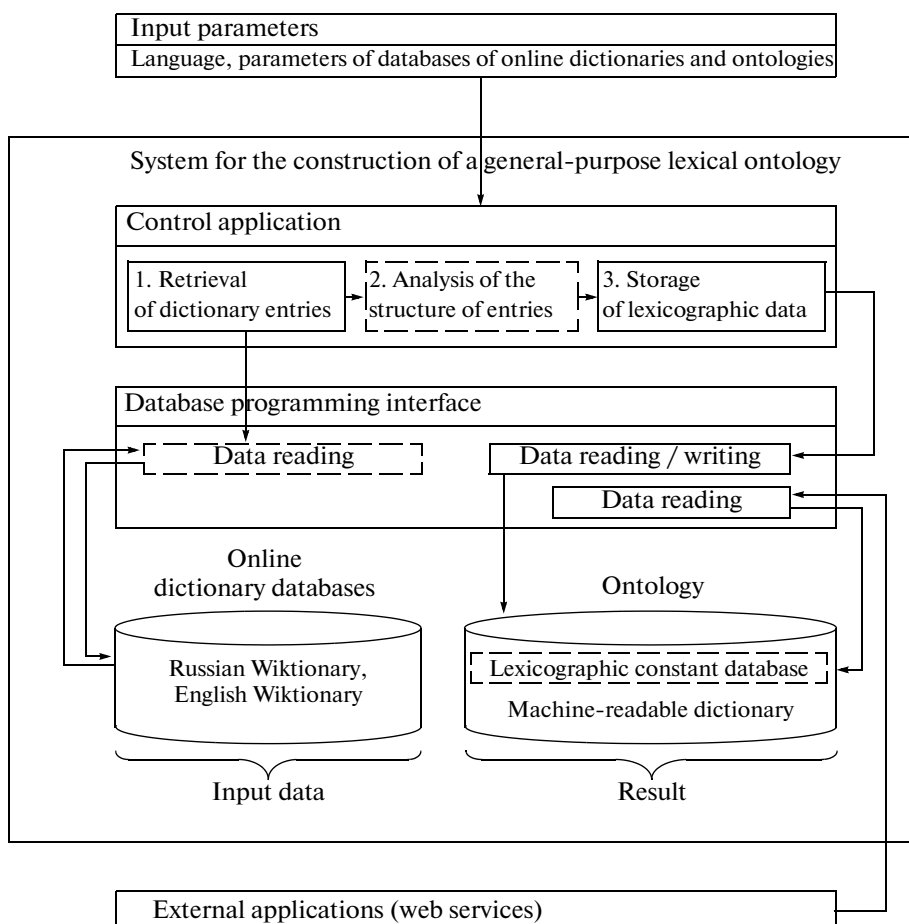


Fig. 8. Architecture of the system of automatic construction of a general-purpose lexical ontology.

The *control application* performs a sequence of three steps for each entry (wiki text) retrieved from the wiktioary database (which is the first step). At the second step the entry is analyzed; more precisely, reference elements of the entry (keywords and hypertext markup elements) indicating its subsections are sought using regular expressions. The reference elements are defined by experts based on the analysis of entry submission guidelines of the wiktioary. The analysis is performed from the general to the special. First, the entry is broken up into large parts, then these parts are scanned by the analyzer once more and broken up into smaller parts.

At the third phase, the data are stored in the ontology database. As a result, the system creates the ontology consisting of two parts—a database of lexicographic constants and a machine readable dictionary. In Fig. 8, the dashed line shows the parts of the system that require preliminary adjustment. These are (1) the database of lexicographic constants, (2) the analysis module in the control application, and (3) the data read module in the database programming interface. The fact is that the community of editors of wiktioaries is making various improvements and refinements to the structure of the dictionary as the project evolves. These changes should be duly taken into account by developers of the system. The adjustment is carried out at the first phase of the ontology construction in the process of the joint work of lexicographers and software engineers.

Thus, the *lexicographic constant database* is created manually by experts; it establishes a correspondence between identifiers and some lexicographic information. The lexicographic constant database contains the following lists, which are required for the analysis of entries and search in the machine-readable dictionary database:

(1) Language codes in accordance with the international standard of abbreviation of the names of languages ISO 639, self-designation of languages, language names in English and in Russian (370 languages and their codes in the Russian Wiktionary, 274 corresponding items in the English Wiktionary).

(2) “Third-level headings” in the terminology of the wiktioary including the names of the parts of speech and headings such as proper noun, article, prefix, suffix, acronym, abbreviation, etc. (58 headings in the English Wiktionary, 25 headings in the Russian Wiktionary).

(3) Names of semantic relations (synonymy, antonymy, etc.).

The lexicographic constant database is filled by experts, and it operates in the read-only mode when the data from the wiktioary are retrieved.

The *database programming interface* is a collection of functions for reading data from the wiktioary and for reading and writing ontology data. The functions related to the processing of ontology data can be divided into *low-level functions* that search, delete, add, or update records in a single table, and *high-level functions* that can operate on several tables at once. For example, high-level software interface functions make it possible to obtain the following data for a given language and part of the speech: (i) the list of interpretations, (ii) the list of synonyms, antonyms, etc., and (iii) the list of translations from the source to the target language for the given word.

External applications (e.g., Web services) can work with the ontology remotely if the parameters of the ontology database and the data access programming interface are known. If there is need for high data processing speed, the ontology database and software that provide ontology’s functionality can be downloaded from the site of the authors of this paper (<http://code.google.com/p/wikokit/>) and installed locally.

3. RESULTS OF THE AUTOMATIC CONSTRUCTION OF A GENERAL-PURPOSE LEXICAL ONTOLOGY

Two databases of lexical ontologies were constructed; this allowed us to compare them and wiktioaries in terms of various parameters (Tables 2–4). Data from the English Wiktionary as of October 8, 2011 (the so-called database dump) and data from the Russian Wiktionary as of May 21, 2011 were used as initial databases. The ontology relational database consists of a set of related tables. The second column in Table 3 gives the name of each table in the ontology database and some comments (in parentheses). The other columns of Table 3 give the size of the corresponding tables (the number of rows). In the next to last column of these three table there is the ratio of the two previous columns, i.e., the ratio of the ontology parameters of the English Wiktionary to the parameters of the ontology of the Russian Wiktionary for 2011–2012. To analyze the dynamics of growth of wiktioaries and ontologies, the last column contains similar ratios

Table 2. Key indicators of the English and Russian Wiktionaries

No.	Property	Wiktionary		en/ru	
		English (en)	Russian (ru)	2011–2012	2010
	Wiktionary version (date)	08.10.2011	21.05.2011	2011	2010
1	Number of pages	2936016	325128	9.03	7.13
2	Number of edits since Wiktionary was set up	16971706	3183428	5.33	3.65
3	Average edits per page	5.35	5.12	1.04	1.03
4	Number of active editors of the dictionary	1060	176	6.02	7.17

Table 3. Key indicators of the ontology databases constructed on the basis of the English and Russian Wiktionaries

No.	The table name in the ontology database (and comment)*	Ontology		en/ru	
		English Wiktionary (en)	Russian Wiktionary (ru)	2011–2012	2010
1	page (number of entries)	1283011	532024	2.41	3.77
2	relation (number of semantic relationships)	227430	144675	1.57	1.57
3	lang_pos (number of parts of speech in all the languages)	1219090	427876	2.85	4.63
4	wiki_text (number of text fragments in interpretations, translations, and semantic relations)	1641186	375787	4.37	7.81
5	wiki_text_words (number of words-hyperlinks)	2304041	427116	5.39	10.81
6	meaning (number of meanings)	1634749	248497	6.58	12.68
7	inflection (number of word forms)	102322	28582	3.58	8.84
8	translation (number of blocks with translations, i.e., the number of translated meanings of the words)	88450	29856	2.96	1.55
9	translation_entry (total number of pairs of translations in all languages)	801943	228805	3.50	1.96
10	Number of entries (number of “language–part of speech” pairs) with semantic relations	144530	53554	2.70	3.89
11	Number of entries in the native language (counting “language–part of speech” pairs)	282281	135396	2.08	2.43
12	Number of semantic relations for words in the native language	60844	104513	0.58	0.52

* These and more recent data for the English Wiktionary are available at http://en.wiktionary.org/wiki/User:AKA_MBG/Statistics:Parameters_of_the_database_created_by_the_Wiktionary_parser, for the Russian Wiktionary, they are available at <http://ru.wiktionary.org/wiki/>

for the English Wiktionary as of January 6, 2010 and for the Russian Wiktionary as of April 5, 2010; these data are taken from the paper.⁵

The comparison of wiktionaries themselves (Table 2) shows that the English Wiktionary is approximately 9 times larger than the Russian Wiktionary in terms of the number of pages and 6 times larger in terms of the number of active participants. Among surprisingly stable parameters the values of which remain almost invariable in 2010–2011 is the average ratio of modifications, which is 1.03–1.04 (row 3 in Table 2), in 2010 the number of modifications was 4.8–4.96 per entry, and in 2011 it was 5.12–5.35 in Russian and English Wiktionaries, respectively.

Another stable parameter is the ratio of the number of semantic relations (row 2 in Table 3). The English Wiktionary contains by a factor of 1.57 more relations than the Russian Wiktionary; in 2010, there

⁵ A. A. Krizhanovsky, The Comparison of Wiktionary Thesauri Transformed into the Machine-Readable Format, *Preprint of St. Petersburg Inst. for Informatics and Automation, Russ. Acad. Sci.*, St. Petersburg, 2010. <http://arxiv.org/abs/1006.5040>.

Table 4. Dictionary statistics obtained by analyzing ontologies

Property	Wiktionary		en/ru	
	English (en)	Russian (ru)	2011–2012	2010
Number of entries in the native language (language & POS) to the total number of pages (ratio of the value in row 11 of Table 3 to the values in row 1 of Table 2), %	9.61	41.64	0.23	0.34
Number of entries with semantic relations to the total number of pages (ratio of value in row 10 of Table 3 to the value in row 1 in Table 2), %	4.92	16.47	0.30	0.55
Average number of semantic relations for entries about the words in the native language (the ratio of the value in row 12 to the value in row 11 in Table 3)	0.22	0.77	0.29	0.21

were 157 and 100 thousands of relations in these wiktionaries, respectively; and in 2011 there were 227 and 145 thousands of them, respectively.

The term “native language” (rows 1 and 3 in Table 4) means the basic or main language of the wiktionary, i.e., Russian for the Russian Wiktionary and English for the English Wiktionary. The main common feature (and, at the same time the difference) of all the 170 wiktionaries is as follows:

(1) The interpretation of all the words, including foreign words (row 6 in Table 3), are given only in the native language.

(2) Translations into all languages (rows 8 and 9 in Table 3) are given only in the entries for the words of the native language.

We compared the entries for the words in the native language, i.e., entries for English words in the English Wiktionary and entries for Russian words in the Russian Wiktionary.

1. The Russian Wiktionary contains almost 1.7 times (104.5 thousands) more semantic relations between the Russian words (row 12 in Table 3) than the semantic relations between the English words in the English Wiktionary (60.8 thousands). In 2010, this ratio was 2 (84 and 44 thousands, respectively).

2. In the English Wiktionary, the entries for the words of the native language (row 1 of Table 4) constitute less than a tenth (9.6%) of all the entries (in 2010 they accounted for a fifth, 18.3%). In the Russian Wiktionary, the percentage of the entries for the native words is much higher, although it reduced to 41% (in 2010, more than half the entries were for the Russian words, 53.7%). Thus, despite the common goal of both wiktionaries, i.e., the description of all lexical units of all languages, the Russian Wiktionary remains more monolingual.

3. The average number of semantic relations per entry (row 3 in Table 4) in the Russian Wiktionary is 3.5 times higher than in the English Wiktionary, 0.77 and 0.22, respectively (in 2010 these values were 0.65 and 0.14).

The construction of the ontology made it possible to carry out the numerical analysis and compare the vocabulary of the English language in the English Wiktionary, in the Russian Wiktionary, and in WordNet. In the comparison with the WordNet database, only entries for the English words were taken into account in multilingual wiktionaries because WordNet contains only entries of this kind. The comparison gave the following results:

(1) The number of English words and meanings in dictionaries. The English Wiktionary contains more English words (276470) and meanings (369778) as of 2011. This is by a factor of 1.78 more entries and by a factor of 1.79 more meanings than in WordNet.

(2) The distribution of English words in parts of speech (in the English Wiktionary, WordNet, and in the Russian Wiktionary) is as follows: nouns account for 52%, 71%, and 83% (the largest share in all the dictionaries); adjectives account for 20%, 14%, and 6%; verbs account for 15%, 12%, and 8%; and adverbs account for 4%, 3%, and 1%.

(3) The percentage of words with a single meaning is 81% in the English Wiktionary and 88% in WordNet. Thus, both dictionaries (WordNet and the English Wiktionary) contain more words with a single meaning than with words with several meanings. In the English Wiktionary, every fifth noun, verb, and adjective (17–22%) has several meanings. However, among the adverbs there are only 12% of words with multiple meanings.

(4) From the viewpoint of the average number of meanings for words belonging to different parts of speech, verbs are most multivocal in all the three dictionaries. The average number of meanings of verbs (excluding verbs with a single meaning) is more than 3 (in the range 3.08–3.57). Adverbs have the least number of meanings (excluding the words with a single meaning) (in the range 2.35–2.88). A more detailed analysis and comparison of dictionaries based on the constructed ontologies and WordNet can be found in [17].

CONCLUSIONS

An approach to the design of a system of automated construction of a general-purpose lexical ontology was proposed. The architecture of such a system was described. At the first stage, a team of linguists, lexicographers (experts in the online dictionary under examination) and software engineers (experts in databases and programmers) analyzes the entry of the online dictionary to design and develop the system. At the second stage, the system automatically checks all entries and stores the retrieved information in the general-purpose lexical ontology database.

The developed system can be used for the following purposes: (1) in research prototypes and applications (automation of ontology and knowledge base construction, recognition of meanings of words), (2) office applications (spell checkers and translation [25]), (3) in industrial applications (identifying users' interests, customer profile creation), and (4) in monitoring and text flow clustering systems, identifying of texts pertaining to a given topic.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (project nos. 11-01-00251, 12-01-00481, and 12-07-00070), the Russian Foundation for Humanities (project no.12-04-12062), project no. 213 of the Program of Fundamental Research of the Presidium of the Russian Academy of Sciences "Intelligent Information Technologies, Mathematical Modeling, System Analysis and Automation," and project no. 2.2 of the Program of the Department of Nano and Information Technologies of the Russian Academy of Sciences "Intelligent Information Technologies, System Analysis, and Automation."

REFERENCES

1. R. A. Amsler, "Machine Readable Dictionaries," in *Annual Review of Information Science and Technology*, Ed. by M. E. Williams (Knowledge Industry Publications, NY, 1984), Vol. 19, pp. 161–209.
2. N. Calzolari, "Lexical Databases and Textual Corpora: Perspectives of Integration for a Lexical Knowledge Base," in *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, Ed. by U. Zernik (Hillsdale, New Jersey, 1991), pp. 191–208.
3. T. Wandmacher, E. Ovchinnikova, U. Krumnack, et al., "Extraction, Evaluation and Integration of Lexical-Semantic Relations for the Automated Construction of a Lexical Ontology," in *Third Australasian Ontology Workshop (AOW)* (Gold Coast, Australia, 2007), Vol. 85 of CRPIT, pp. 61–69.
4. E. S. Bolotnikova, T. A. Gavrilova, and V. A. Gorovoi, "To a Method of Evaluating Ontologies," *J. Comput. Syst. Sci. Int.* **50**, 448–461 (2011).
5. O. Les'ko and Yu. Rogushina, "The Use of Specialized Lexical Ontology for Automating the Construction of a Subject Area Ontology from Texts in Natural Languages," in *Information Models of Knowledge*, Ed. by K. Markov, V. Velichko, and O. Voloshin (ITHEA, Kiev, 2010), pp. 93–100. http://www.foibg.com/ibs_isc/ibs-19/ibs-19-p10.pdf
6. C. M. Meyer and I. Gurevych, "OntoWiktionary—Constructing an Ontology from the Collaborative Dictionary Wiktionary," in *Semiautomatic Ontology Development: Processes and Resources*, Ed. by M. T. Paziienza and A. Stellato (IGI Global, Hershey, Pennsylvania, 2012), pp. 131–161. <http://www.ukp.tu-darmstadt.de/data/lexical-resources/ontowiktionary/>
7. R. Ferrer-i-Cancho, "The Structure of Syntactic Dependency Networks: Insights from Recent Advances in Network Theory," in *Problems of Quantitative Linguistics*, Ed. by V. Levickij and G. Altamann (Ruta, Chernivitsi, Ukraine, 2005), pp. 60–75.
8. A. Montoyo, M. Palomar, and G. Rigau, "Method for WordNet Enrichment Using WSD," in *Proc. 4th Int. Conf. on Text Speech and Dialogue TSD'2001, Zelezna Ruda, Spieak, Czech Republic*, Lect. Notes in Artif. Intell. **2166** (Springer, 2001).
9. P. Resnik and D. Yarowsky, "Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation," *Natural Language Eng.* **5** (2), 113–133 (2000).
10. D. Yarowsky, "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods," in *33rd Annual Meeting of the Association for Computational Linguistics* (Cambridge, MA, 1995), pp. 189–196.

11. S. Harabagiu and D. Moldovan, "A Marker-Propagation Algorithm for Text Coherence," in *Working Notes of the Workshop on Parallel Processing at the 14th Int. Joint Conf. on Artificial Intelligence, Montreal, 1995*, pp. 76–86.
12. E. Teich and P. Fankhauser, "WordNet for Lexical Cohesion Analysis," in *Proc. of the Second Global WordNet Conf., Brno, Czech Republic, 2004*, pp. 326–331.
13. S. Reed and D. Lenat, "Mapping Ontologies Into Cyc," in *Proc. of AAAI 2002 Conf. Workshop on Ontologies for the Semantic Web, Edmonton, Canada, 2002*.
14. C. Bizer, J. Lehmann, G. Kobilarov, et al., "DBpedia—A Crystallization Point for the Web of Data," *Web Semantics: Science, Services, and Agents on the World Wide Web*, Vol. 7, No. 3, pp. 154–165. <http://www.wiwi.fu-berlin.de/en/institute/pwo/bizer/research/publications/Bizer-et-al-DBpedia-CrystallizationPoint-JWS-Preprint.pdf>
15. I. Gurevych, J. Eckle-Kohler, S. Hartmann, et al., "A Large-Scale Unified Lexical-Semantic Resource," in *Proc. 13th Conf. of the European Chapter of the Association for Computational Linguistics, Avignon, France, 2012*, pp. 580–590. http://www.ukp.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2012/uby_eacl2012_cameraready.pdf
16. G. Melo and G. Weikum, "Language as a Foundation of the Semantic Web," in *Proc. Poster and Demonstration Session of the 7th Int. Semantic Web Conf. (ISWC 2008), Karlsruhe, Germany, 2008*. http://www.informatik.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2011/oup-elex2012-meyer-wiktionary.pdf
17. A. A. Krizhanovskii, "A Quantitative Analysis of the English Lexicon in Wiktionaries and WordNet," in *Trudy SPIIRAN (Institut Avtomatiki i Avtomatizatsii, St. Petersburg, 2011)*, No. 19, pp. 87–101; *International Journal of Intelligent Information Technologies (IJIT)*. <http://sciepeople.ru/publication/108159/>
18. C. M. Meyer and I. Gurevych, "Wiktionary: A New Rival for Expert-built Lexicons? Exploring the Possibilities of Collaborative Lexicography" in *Electronic Lexicography* (Oxford Univ. Press, Oxford, 2012). http://www.informatik.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2011/oup-elex2012-meyer-wiktionary.pdf
19. A. A. Krizhanovsky and A. V. Smirnov, "On the Problem of Wiki Texts Indexing," *J. Comput. Syst. Sci. Int.*, **48**, 616–624 (2009).
20. P. Otte and F. M. Tyers, "Rapid Rule-based Machine Translation between Dutch and Afrikaans," in *16th Annual Con the European Association of Machine Translation, EAMT11, Leuven, Belgium, 2011*.
21. Kiem-Hieu Nguyen and Cheol-Young Ock, "Using Wiktionary to Improve Lexical Disambiguation in Multiple Languages," *CICLing Lect. Notes in Comput. Sci.* **7181**, 238–248 (21012).
22. C. McFate and K. Forbus, "NULEX: An Open-License Broad Coverage Lexicon," in *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, 2011*, pp. 363–367.
23. He. Qingyue, "Automatic Pronunciation Dictionary Generation from Wiktionary and Wikipedia," Thesis. Karlsruhe Institute of Technology (2009).
24. J. E. F. Friedl, *Mastering Regular Expressions* (O'Reilly, Cambridge, 1997; Piter, St. Petersburg, 2001).
25. F. Lin and A. Krizhanovsky, "Multilingual Ontology Matching Based on Wiktionary Data Accessible via SPARQL Endpoint," in *Proc. of the 13th Russian Conf. on Digital Libraries RCDL'2011, Voronezh, Russia, 2011*, pp. 19–26.

SPELL: OK