

Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography

Christian M. Meyer and Iryna Gurevych
Ubiquitous Knowledge Processing Lab
Technische Universität Darmstadt
Hochschulstraße 10, 64289 Darmstadt, Germany

Abstract

With the rise of the Web 2.0, collaboratively constructed language resources are rivalling expert-built lexicons. The collaborative construction process of these resources is driven by what is called the “Wisdom of Crowds” phenomenon, which offers very promising research opportunities in the context of electronic lexicography. The vast number and broad diversity of authors yield, for instance, quickly growing and constantly updated resources. While expert-built lexicons have been extensively studied in the past, there is yet a gap in researching collaboratively constructed lexicons. We therefore provide a comprehensive description of Wiktionary – a freely available, collaborative online lexicon. We study the variety of encoded lexical, semantic, and cross-lingual knowledge of three different language editions of Wiktionary and compare the coverage of terms, lexemes, word senses, domains, and registers to multiple expert-built lexicons. We conclude our work by discussing several findings and pointing out Wiktionary’s future directions and impact on lexicography.

Keywords

Wiktionary, Collaboration of Web Communities, Lexical Resources, Online Dictionaries, Collaborative Lexicography, Comparative Study

Table of contents

1	Introduction	2
2	Describing Wiktionary: core features	3
2.1	Historical development	4
2.2	Wiktionary’s multilingual structure and language coverage	5
2.3	Wiktionary’s macrostructure and accessibility	8
2.4	Wiktionary’s microstructure and types of linguistic knowledge	9
2.5	Collaboration in Wiktionary	11
3	Analysing Wiktionary: a critical assessment	14
3.1	Coverage of terms	15
3.2	Coverage of lexemes	17
3.3	Coverage of word senses	21
3.4	Coverage of domains and registers	24
4	Conclusion	26
	Acknowledgements	28
	References	28

1 Introduction

Collaborative lexicography is a fundamentally new paradigm for compiling lexicons. Previously, lexicons have been the product of a small group of expert lexicographers specializing in a particular field. In contrast, collaborative lexicography is a bottom-up approach (Carr, 1997) which encourages lexicon readers to contribute to the writing of lexicon entries. The large-scale collaboration of many authors became possible with the rise of the Web 2.0 (i.e. the transition from static web pages to dynamic, user-generated content on the World Wide Web) and resulted in huge resources of very high quality (Giles, 2005). These resources represent the sum of the opinions of many authors, often known as their collective intelligence or as the ‘Wisdom of Crowds’ phenomenon (Surowiecki, 2005; Malone et al., 2010). The most prominent example of a collaboratively created resource is Wikipedia¹, which has emerged as the world’s largest encyclopedia.

Collaboratively constructed lexicons are continually updated by their community, and this yields a steeply increasing coverage of words and word senses. Each contributor has a certain field of expertise. This broad diversity of authors fosters the encoding of a vast amount of domain-specific knowledge. An important characteristic of collaborative lexicography is that the large number of authors has the ability to express the actual use of language in the spirit of Wittgenstein’s “meaning is use” rather than the often criticized record of “how people ‘ought to’ use language” (Atkins and Rundell, 2008: 2) in expert-built lexicons. Naturally, the entries in collaboratively constructed lexicons are repeatedly changed before a consensus is reached. Examining the complete edit history of an entry allows us to study the evolution of lexicon entries and their discussion within the community. This is not possible in expert-built lexicons, since neither the edit history nor the discussion of the lexicographers is usually publicly available.

In this chapter, we explore the possibilities of collaborative lexicography. The subject of our study is Wiktionary², which is the largest available collaboratively constructed lexicon for linguistic knowledge. Previous studies have compared Wiktionary with other online lexicons (Mann, 2010; Lew, 2011), but do not systematically study its collaborative construction process. Fuertes-Olivera (2009) provides a more focused qualitative study of English and Spanish entries in the English Wiktionary and compares them to expert-built lexicons. Hanks (this volume) assesses the quality and intelligibility of Wiktionary’s sense descriptions. Quantitative studies have also been undertaken on the German (Meyer and Gurevych, 2010a) and Russian Wiktionaries (Krizhanovsky, 2010).

Our work goes beyond these research efforts by providing, Section 2, a comprehensive description of Wiktionary’s macro- and microstructure, its community and collaboration mechanisms, as well as its various multilingual editions. In Section 3, we then compare the Wiktionaries of three different languages to multiple expert-built lexicons in a qualitative and quantitative manner and study the coverage of terms, lexemes, word senses, domains, and registers. In Section 4, we finally investigate the question of whether collaboratively created lexicons rival expert-built ones. To conclude, we discuss the new possibilities that collaborative lexicography has opened up for a range of lexicon users and what this implies for the future development of lexicography.

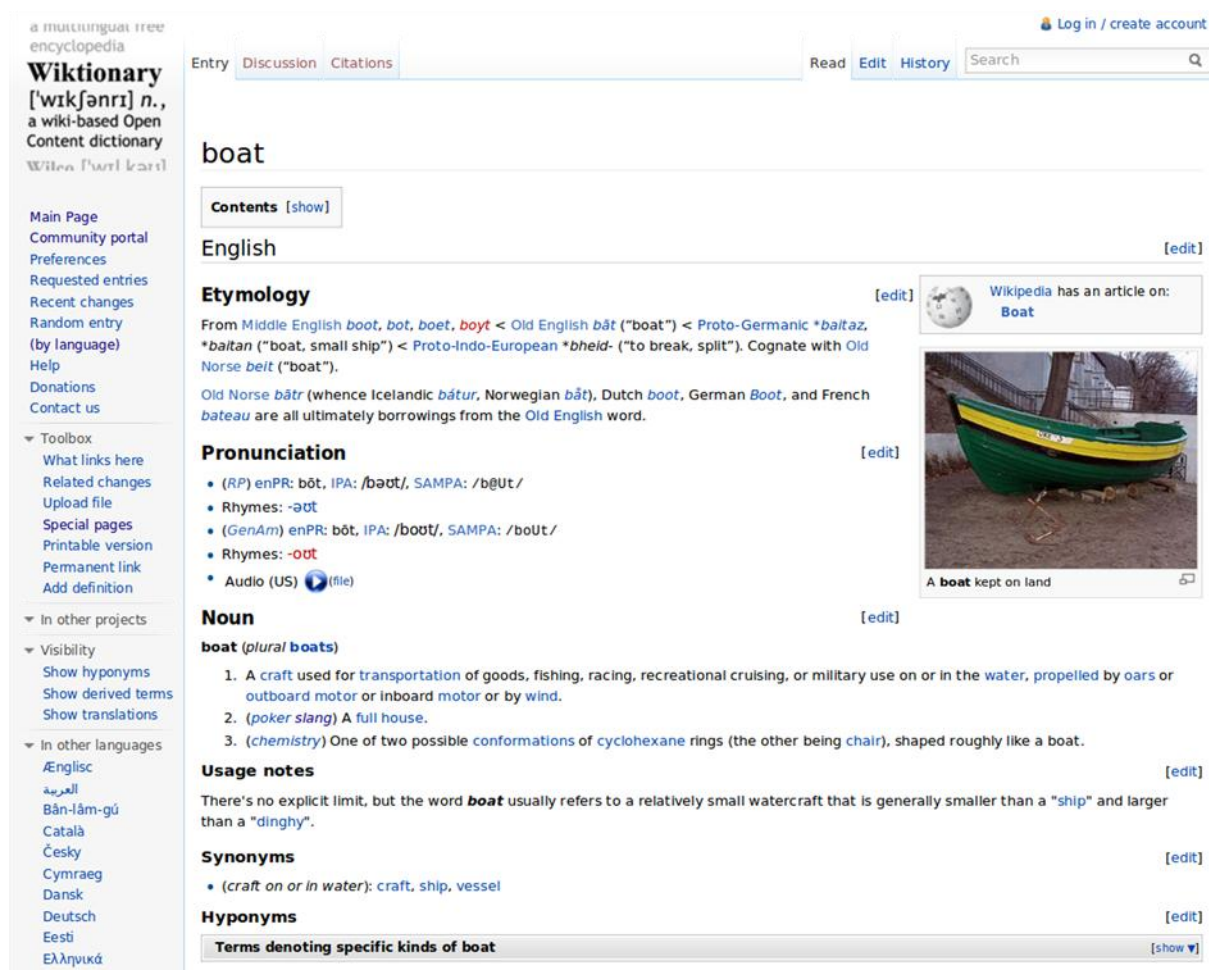
¹ <http://www.wikipedia.org/>

² <http://www.wiktionary.org/>

2 Describing Wiktionary: core features

Wiktionary is a multilingual online dictionary that is created and edited by volunteers and is freely available on the Web. The name “Wiktionary” is a portmanteau of the terms “wiki” and “dictionary”. A *wiki* is a web application allowing simple editing of hyperlinked web pages in a collaborative manner. Ward Cunningham set up the first system of this kind in 1995 and named it “wiki”, the Hawaiian word for “fast” (Leuf and Cunningham, 2001). Possibly the best-known example of a wiki-based resource is the online encyclopedia Wikipedia. A *dictionary* is a lexicon for human users that contains linguistic knowledge of how words are used (see Hirst, 2004). Wiktionary combines these two facets. Figure 1 shows the article “boat” from the English Wiktionary as an example lexicon entry.

This section first provides a short historical survey of Wiktionary. Then, we describe its multilingual aspects as well as its macro- and microstructure. We finally discuss how the Wiktionary community collaborates in order to compile the lexicon entries.



The image shows a screenshot of the English Wiktionary article for the word "boat". The page layout includes a sidebar on the left with navigation links such as "Main Page", "Community portal", and "Preferences". The main content area features a search bar at the top right, followed by the word "boat" in large font. Below the word, there are tabs for "Entry", "Discussion", and "Citations". The article is structured into sections: "Contents" (with a "show" link), "English" (with an "edit" link), "Etymology" (with an "edit" link), "Pronunciation" (with an "edit" link), "Noun" (with an "edit" link), "Usage notes", "Synonyms", and "Hyponyms". The "Etymology" section explains the word's origin from Middle English and Proto-Germanic. The "Pronunciation" section provides IPA and SAMPA notations. The "Noun" section lists three definitions: a craft used for transportation, a poker slang term for a full house, and a chemistry term for a cyclohexane ring conformation. The "Usage notes" section clarifies the word's relative size compared to "ship" and "dinghy". The "Synonyms" section lists "craft", "ship", and "vessel". The "Hyponyms" section includes a link to "Terms denoting specific kinds of boat". A small image of a green and yellow boat is included in the article.

Figure 1: The English Wiktionary’s article for “boat” (<http://en.wiktionary.org/wiki/boat>).

2.1 Historical development

Wiktionary was first launched for the English language on 12 December 2002 as a pure “companion volume” to Wikipedia. It originated from a long discussion within the Wikipedia community concerning the exclusion of linguistic knowledge from its encyclopaedic articles.³ Important initiators of this development were Daniel Alston, Brion Vibber, and Tim Starling.⁴ By 2004, Wiktionary had gradually turned into an independent project. Starting with the French and Polish Wiktionaries, a separate language edition had been created for all 143 active Wikipedia editions by May 2004, when Wiktionary also moved to its current URL <http://www.wiktionary.org/>. Since then, Wiktionary has rapidly increased and attracted a growing number of contributors. By the end of 2006, seven Wiktionaries exceeded 100,000 articles.

The success of Wiktionary has also drawn increasing attention from both the public and academia. Descy (2006: 4) introduced Wiktionary as a “neat” platform that is “really easy to use” and “designed to be a multilingual, international dictionary”. Lepore (2006: 87) raised a criticism about the large-scale import of lexicon entries from copyright-expired dictionaries such as Webster’s New International Dictionary of the English Language and the low quality of contributions: “‘Be your own lexicographer!’ might be Wiktionary’s motto. [...] Why pay good money for a dictionary written by lexicographers when we can cobble one together ourselves?”⁵ Similar issues have been discussed by Fuertes-Olivera (2009) and Hanks (this volume), who particularly criticize the low quality of some Wiktionary definitions. Notwithstanding, Wiktionary has been successfully employed in multiple natural language processing applications including information retrieval, semantic relatedness calculation, and speech synthesis (Etzioni et al., 2007; Müller and Gurevych, 2009; Zesch et al., 2008b; Schlippe et al., 2010).

As of August 2010, there were over 170 Wiktionaries, of which 145 were active.⁶ With over 1.6 million articles, the English and the French editions are by far the largest ones. Table 1 shows the number of articles in the largest Wiktionary editions (over 100,000 articles) as at January of each year. The column “Recent growth” shows the average number of new articles within the last six months. Wiktionaries are generally growing, although their speed differs markedly: the largest ones are also growing very fast. However, the Tamil Wiktionary is currently decreasing by an average of 38 articles per month, and the Greek Wiktionary was reduced in size by more than 10,000 articles between 2008 and 2009: this is an indicator of a consolidation process after importing data from other resources.

³ See <http://en.wikipedia.org/w/index.php?oldid=294531> (1 November 2001), <http://meta.wikimedia.org/w/index.php?oldid=403> (22 November 2001)

⁴ See <http://meta.wikimedia.org/w/index.php?oldid=3149> (25 November 2002), <http://en.wikipedia.org/w/index.php?oldid=378432551> (11 August 2010), <http://meta.wikimedia.org/w/index.php?oldid=1759149> (11 December 2009)

⁵ Although a large number of lexicon entries have been automatically imported to Wiktionary, these entries are not supposed to stay in their original state, but are to be revised by the community. This is why the imported entries are kept in a separate part of Wiktionary; see http://en.wiktionary.org/wiki/Wiktionary:Webster_1913.

⁶ See <http://stats.wikimedia.org/wiktionary/EN/TablesArticlesTotal.htm>

Table 1: The development of the seventeen largest Wiktionary editions and their recent growth (in articles per month).

Edition	2002	2003	2004	2005	2006	2007	2008	2009	2010	Recent growth
English	3	866	31k	51k	113k	321k	650k	1.1M	1.6M	+1,179
French	–	6	10	23k	123k	225k	698k	1.2M	1.6M	+1,697
Lithuanian	–	–	–	7	75	529	23k	108k	410k	+556
Turkish	–	–	–	707	1k	118k	185k	252k	266k	+33
Chinese	–	–	–	425	27k	113k	116k	117k	263k	+50
Russian	–	–	–	896	2k	106k	132k	189k	234k	+199
Vietnamese	–	–	–	42	605	209k	225k	228k	228k	+1
Ido	–	–	–	–	32k	104k	126k	146k	165k	+49
Polish	–	1	2	28k	37k	54k	82k	111k	147k	+115
Greek	–	4	32	87	4k	79k	121k	120k	145k	+101
Finnish	–	–	–	497	18k	44k	75k	105k	139k	+97
Hungarian	–	–	–	4k	9k	35k	45k	98k	136k	+121
Norwegian	–	–	–	39	5k	5k	6k	17k	123k	+59
Portuguese	–	–	–	1k	11k	28k	44k	53k	112k	+26
Tamil	–	–	–	27	1k	6k	7k	103k	105k	–38
German	–	2	6	4k	18k	46k	72k	88k	104k	+39
Italian	–	1	32	5k	32k	43k	66k	94k	103k	+21

2.2 Wiktionary’s multilingual structure and language coverage

Multilingual structure. Wiktionary provides two different approaches to encoding linguistic knowledge in multiple languages. First, there are independent Wiktionaries for each language—called *language editions*—that are accessible via a subdomain denoting the respective ISO 639 language code.⁷ The Russian Wiktionary can, for instance, be found at <http://ru.wiktionary.org/>. This language is the *native language* of a Wiktionary edition, since it is used for the graphical user interface and for describing the lexicon entries. Second, each Wiktionary edition may include lexicon entries from multiple languages. There is, for example, an article about the Russian term “лодка” (English “boat”) both within the English and the Russian edition (the latter is shown in Figure 2). The rationale behind this is to provide lexicographic descriptions in different languages: the Russian Wiktionary uses Russian to describe “лодка” which corresponds to the practice of monolingual dictionaries. The description texts of “лодка” in the English Wiktionary are, in contrast, written in English—similar to a bilingual dictionary.⁸ This makes Wiktionary useful for both native speakers and language learners. Consider for instance the definition “водное транспортное средство, небольшое судно, идущее на вёслах, под парусом или на моторной тяге” (English “a water-based means of transport, a small vessel powered by oars, sails, or a motor”) for the simple Russian term “лодка”. To understand this definition, a learner needs to have a certain level of Russian. However, if the learner is a native English speaker, he or she can easily find that “лодка” means “(nautical) boat, dinghy, gig, yawl”. The language of the

⁷ http://www.infoterm.info/standardization/iso_639_1_2002.php.

⁸ Note that this distinction has not always been clear in previous work. Fuertes-Olivera (2009), for instance, uses the name “Spanish Wiktionary” to refer to Spanish terms within the English language edition. This has tended to exaggerate the claim that Wiktionary is language dominated by English, since the actual Spanish language edition that uses Spanish to define its terms was not considered in this study.

user interface also plays an important role here, since a menu item labelled with “Полный индекс” (English “full index”) might not be easily comprehensible for a language learner of Russian. Using the index of his/her native language edition to browse the Russian entries is much more convenient.

The screenshot shows the Russian Wiktionary article for "лодка". The page includes a navigation menu on the left, a search bar at the top right, and a main content area. The word "лодка" is highlighted in blue. Below the word, there is a section for "Русский" (Russian) and "Морфологические и синтаксические свойства" (Morphological and syntactic properties). A table of inflection forms is provided, showing the word in various cases (Им., Р., Д., В., Тв., Пр.) and numbers (ед. ч. - singular, мн. ч. - plural). The table is as follows:

падеж	ед. ч.	мн. ч.
Им.	лодка	лодки
Р.	лодки	лодок
Д.	лодке	лодкам
В.	лодку	лодки
Тв.	лодкой, лодкою	лодками
Пр.	лодке	лодках

Below the table, there is a section for "Значение" (Meaning) with a definition: "водное транспортное средство, небольшое судно, идущее на вёслах, под парусом или на моторной тяге". There is also a section for "Синонимы" (Synonyms) and "Антонимы" (Antonyms). The page also features a small image of several colorful kayaks on a river.

Figure 2: The Russian Wiktionary’s article for “лодка” (<http://ru.wiktionary.org/wiki/лодка>).

Coverage of languages. We explored which languages have a separate Wiktionary edition in order to clarify whether Wiktionary covers the full variety of languages in the world or is dominated by certain countries, continents, or cultures. We grouped the Wiktionary language editions by their language family (based on Ruhlen, 1987) and the main geographical region the language is spoken in.⁹

Table 2 shows our geographical classification of the Wiktionary language editions. We found six editions for man-made languages, namely Ido, Esperanto, Volapük, Interlingua, Interlingue, and Lojban. The English language also features a Simple Wiktionary that uses only a controlled vocabulary to describe its entries. Both constructed and simple language editions are marked as “Other” in the table. For geographical region, we created a “Worldwide” group covering English, Spanish, Portuguese, and French, which are spoken as

⁹ We are aware that some language families are subject to discussion (e.g. for the Korean language), and that a clear allocation to certain geographic regions is very fuzzy and debatable. Nevertheless, we do not aim at a full ethnological study but at gaining insights into the type of languages for which a Wiktionary language edition exists.

a main language in several continents of the world. For the remaining languages, we grouped them by the continent where the language is mainly spoken. The group “Near/Middle East” is an exception, as it forms the borderline between Europe, Asia, and Africa, subsuming the Arabic, Turkish, and Persian languages, as well as Hebrew.

Table 2: Wiktionary editions and the total of articles by region.

Region	Wiktionaries	Articles
Africa	17	86,084
Americas	8	30,012
Asia	36	1,714,125
Europe	54	2,600,675
Near/Middle East	9	511,225
Oceania	10	63,655
Worldwide	4	3,809,000
Other	7	334,210
Total	145	9,148,986

All the main regions of the world are covered by a Wiktionary language edition. Most Wiktionaries exist for the languages spoken in Europe and Asia, while the Americas seem to be under-represented. This is partly due to the large share of the four worldwide languages spoken in these countries. Regarding the size of the Wiktionaries (i.e. the number of articles that are contained in the corresponding language edition), the worldwide languages contain the bulk of the articles (41%), followed by Europe (29%), and Asia (19%). Wiktionaries for the languages spoken in Africa and Oceania (including Australia) are still very small (less than 2% of the total number of articles).

Table 3 shows the language families covered by Wiktionary. It is not surprising that the Indo-European language families (Germanic, Romance, and Slavic) represent the largest number of both Wiktionary editions and articles, since Indo-European languages are the most widespread, including English, Spanish, French, German, Russian, etc. There are, however, many less-common languages, and we also studied under-represented and missing language families. In particular, African languages from the Nilo-Saharan family (including, for example, the Masai language), the Berber family (e.g. Tarifit), and the Khoisan family are under-represented. Khoisan languages are known for their click sounds which may have impeded the creation of a corresponding Wiktionary edition due to the complicated script of the words (e.g. “!Xóǝ” or “ǀHǝǎ”). Another reason may be the lack of technical infrastructure in African countries. Besides African languages, Paleosiberian and Tungusic languages spoken in Siberia, Mongolia, and other regions of northern Asia are also missing. These languages are endangered because of their small number of speakers, which might explain the lack of Wiktionary editions. Many languages of Native American and Australian aborigines are also not yet represented in Wiktionary, probably because of the lack of written knowledge about them or technical infrastructure in cultures that live very close to nature.

Although the vast majority of encoded knowledge in Wiktionary relates to the most widespread languages, our analysis shows that Wiktionary also offers the rare chance to obtain linguistic resources for smaller languages (see also Prinsloo, this volume). Since Wiktionary is constantly growing, we expect the number of Wiktionaries for minor languages to reach a considerable size in the future. In this context, Wiktionary can become an important, easy-to-use platform for linguists who study endangered languages and want to share their research.

Table 3: Wiktionary editions and the total of articles by language family.

Language family	Wiktionaries	Articles
Afro-Asiatic	3	390
Austro-Asiatic	2	229,426
Austronesian	10	78,527
Baltic	2	574,000
Celtic	6	65,460
Creole	2	229
Dravidian	4	282,000
Finno-Ugric	3	346,000
Germanic	10	2,116,314
Indo-Aryan	2	38,151
Indo-Iranian	14	148,309
Inuit	3	12,319
Niger-Congo	9	54,893
Nordic	6	317,323
Romance	14	2,350,689
Semitic	5	56,847
Sino-Tibetan	3	404,000
Slavic	15	589,967
Tai	3	75,109
Turkic	8	463,923
Other	21	945,110
Total	145	9,148,986

2.3 Wiktionary’s macrostructure and accessibility

The content of Wiktionary is organized into *pages*, each of which consists of a formatted text body and a unique title describing the contents of the page. There are four main types of page:

1. *Article pages* constitute the heart of each Wiktionary edition, as they contain the actual linguistic information (see Section 2.4).
2. *Redirect pages* navigate the user to a certain target article. This is useful for terms with several typographic variants, such as “you’ve” (Unicode character U+2019) and “you've” (Unicode character U+0027) or varying capitalization (e.g. “pdf” and “PDF”) that should be described on only one page.¹⁰
3. *Talk pages* are available for each Wiktionary page to discuss its contents, collect ideas for extension, express criticism, and ask questions (Section 2.5).
4. *Internal pages* describe the motivation, goals, statistics, indices, and appendices of Wiktionary as well as its guidelines for contributors (Section 2.5).

In order to gain access to individual pages, Wiktionary’s user interface offers four different access paths (Bergenholtz and Tarp, 1995):

¹⁰ In the Russian Wiktionary, redirects are often used for denoting inflected word forms or misspellings. The plural form ‘красные’ redirects, for example, to ‘красный’ (English ‘red’). The other Wiktionary editions use separate article pages to encode inflected forms and misspellings, since they can be ambiguous. For instance, the misspelling “liar” of “liar” can also refer to a (rarely used) nominalization of “(to) lie” (i.e. a person or thing that is lying on a bed, for example).

1. Each page can be directly accessed by typing its title. In addition to jumping quickly to the article, this also allows it to be bookmarked or linked.
2. An internal search engine can be used to find specific articles. No automatic lemmatization is performed, since inflected word forms are included as separate entries (see Section 2.4). The search results in a list of articles which contain the search terms.
3. Index, category, and list pages are available for browsing through Wiktionary's contents. These pages organize the lexicon entries alphabetically, or by language, domain, register, style, etc. The full alphabetical index corresponds to the organization of most printed lexicons. It is particularly useful if the spelling of a term is unclear. A currently emerging part of Wiktionary is called *Wikisaurus*, which aims to organize the article pages in an onomasiological, thesaurus-like manner using synonyms, hyponyms, hypernyms, and the like.
4. Pages are connected via hyperlinks, i.e. cross-references to other pages. This can be used to point the user to related articles with further explanation or terms that are unclear to a reader.

2.4 Wiktionary's microstructure and types of linguistic knowledge

The actual linguistic information is found on Wiktionary's article pages. The title of the article represents a certain term described in the article, for example the noun "boat", the verb "sleep" or the proverb "Rome wasn't built in a day". Article titles are case-sensitive and can distinguish diacritic variations. The terms "cafe", "café" and "Café" are thus described in different articles. The text body of an article is divided into one or more lexicon entries containing a variety of linguistic information. Wiktionary has no specific target user profile (such as translators or language learners). All readers always have access to all the information available. This is usually organized in separate sections covering knowledge from all major fields of linguistics and is outlined briefly below.

It should, however, be noted that Wiktionary has no fixed structure for its entries. Rather, it allows flexible encoding schemas, which are necessary to describe culture- or language-specific information. The German language, for instance, distinguishes occupational titles of females and males (e.g. "Mechanikerin" and "Mechaniker"), while in English there is usually only one form ("mechanic"). Another example is the different word formation of Chinese based on radicals, which need to be encoded fundamentally differently from English entries.

Language. As shown in Figure 1, each lexicon entry starts with the language of the term being described. This is necessary, as entries in multiple languages can be encoded within the same article page. There is, for example, both an English and a French entry within the article "sensible" that need to be separated due to different meanings. The language sections are usually ordered alphabetically. An exception is made for entries in the Wiktionary's native language, which are always the first ones. These entries are expected to be looked up most frequently and are usually the most detailed ones. Information that cannot be associated with a certain language, such as the letters of an alphabet, internationally used abbreviations (e.g. chemical symbols or the ISO language codes), or scientific names within the biological taxonomy are encoded in a separate section entitled 'Translingual'.

Etymology. The 'Etymology' section describes the origin of a term—e.g. "from Middle English boot, bot, boet, boyd, from Old English bāt [...]" for the English "boat". Etymologies

can help readers to understand the similarities and dissimilarities of terms, and are also useful to explore the distinction between homonymy and polysemy, which is often not clearly defined. The English term “bass” distinguishes, for instance, two homonymous meanings originating from the Latin “bassus” for its musical meaning and from the Proto-Indo-European “*bhors-” for its biological meaning.

Phonetic knowledge. Wiktionary entries often encode a term’s pronunciation using the IPA (International Phonetic Alphabet) or SAMPA (Speech Assessment Methods Phonetic Alphabet) notation. Different variants can be distinguished using labels such as ‘Received Pronunciation’ (Standard English of England), ‘General American’, ‘Standard German’, ‘Swiss German’, etc. In addition to a textual representation using a phonetic alphabet, sound files may be added to help readers learn a certain pronunciation. Finally, the phonetic suffix for finding rhymes is sometimes included in a Wiktionary article. The term “boat” is, for example, represented by the IPA string “bəʊt” and has the rhyming suffix “-əʊt”.

Morphological knowledge. Most lexicons focus on entries for canonical word forms (e.g. the verb “(to) go”). Since Wiktionary has practically no space limitations, it can also include inflected forms (like the past tense form “went”) as separate entries. While canonical word forms are described in a fully fledged article, inflected forms usually explain the type of inflection and link to the canonical word form. In addition, Wiktionary entries often contain inflection tables explaining either the declension of a noun, adjective, etc. by noting the form for each combination of case, gender, and number (where appropriate) or the conjugation of a verb, by noting the forms for the respective combinations of person, number, gender, tense, aspect, mood, or voice. The Russian article “лодка” from Figure 2 contains, for instance, an inflection table with thirteen forms.

Syntactic knowledge. Each lexicon entry is marked with a syntactic category. For single words, the part of speech tags ‘noun’, ‘verb’, ‘adjective’, etc. are used, while encoded multi-word expressions are marked as ‘idiom’, ‘proverb’, ‘saying’, etc. The coverage of syntactic categories is explored in detail in Section 3.2. Apart from the syntactic categories, basic syntactic information is often found in Wiktionary articles. Nouns are, for example, tagged as ‘countable’/‘uncountable’ or as taking only the singular or only the plural form and verbs are labelled as ‘transitive’ or ‘intransitive’. However, Wiktionary does not provide deep lexical-syntactic knowledge, such as subcategorization frames.

Semantic knowledge. The core of each Wiktionary entry is its meaning section. Following the notation of traditional lexicons, the meaning of a term is described in an enumeration of discrete word senses. A word sense is explained by a gloss, example sentences illustrating its usage, quotations, and linguistic labels. The gloss of the third word sense of the article “boat” is, for instance, “(chemistry) One of two possible conformers of cyclohexane rings (the other being chair), shaped roughly like a boat”. The prefix “chemistry” in parentheses is a *linguistic label* (Atkins and Rundell, 2008), which associates word senses with a certain domain, register, style, etc. The underlined words in the gloss are hyperlinks to other Wiktionary articles. This is a useful feature for readers who have problems understanding the definition and wish to quickly look up the definition of one of these words.

The composition of glosses is known to be one of the most contentious aspects of lexicography (Johnson, 1755: Preface). Wiktionary’s collaborative construction process provides a fundamentally new perspective on this challenge, since readers can easily change formulations they do not understand or ask for a reformulation on the talk pages. We will analyse this further in Sections 2.5 and 3.3 below.

In addition to the hyperlinks in the glosses, there are separate sections for gathering hyperlinks to related articles. The section title denotes the type of relation between the linked terms (e.g. synonymy, hypernymy, hyponymy, antonymy, holonymy). More loosely defined relations are gathered within sections labelled ‘Derived terms’ and ‘See also’.

Cross-lingual knowledge. Another way of interconnecting articles is translation. Due to the multilingual nature of Wiktionary discussed in section 2.2, a translation can be defined both as a link to the term within the same Wiktionary edition and to the Wiktionary edition of the target language. For the German translation “Boot” of the term “boat” within the English Wiktionary, there is, for instance, a link to “Boot” in both the English and in the German Wiktionary. For each term and language, multiple translations can be encoded.

A third type of cross-language linking in Wiktionary makes use of inter-wiki links. These links are shown within the navigation pane and allow users to switch from one language edition to another without changing the term. The English article “boat” contains, for example, inter-wiki links to the article “boat” within the German, French, and Russian Wiktionaries (as opposed to linking the translated terms).

Pictorial knowledge. A picture is worth a thousand words, as the old adage goes. Since there are usually no size restrictions in electronic lexicons, the use of drawings, photographs, etc. is becoming increasingly popular to illustrate meanings (Lew, 2010). The Wiktionary community includes pictures in the lexicon entries as an additional description of meaning (see Fig. 1 for an example). The English Wiktionary has also set up a *picture dictionary*¹¹ that can be used to browse the entries graphically. This is a particularly useful feature for non-native speakers to gain a quick idea of a term’s meaning. Hanks (this volume) also envisages the inclusion of other multimedia (such as sound and video) to illustrate meanings in Wiktionary.

References. To include a new term in Wiktionary, the proposed term needs to be ‘attested’ (see the guidelines in Section 2.5 below). This attestation can be done by providing references to external sources. The article “boat” contains, for instance, a reference to “Weisenberg, Michael (2000): *The Official Dictionary of Poker*. MGI/Mike Caro University. ISBN 978-1880069523.” to attest the poker-related word sense of “boat”. Besides references to published books or articles, references to publicly available online lexicons are frequently used by the Wiktionary community.

2.5 Collaboration in Wiktionary

In contrast to traditional lexicons built by individual expert lexicographers, Wiktionary is collaboratively constructed by a large community of ordinary web users. To overcome the lack of lexicographic experience in such a community, Wiktionary relies on the collective intelligence of many different authors—the “Wisdom of Crowds” phenomenon (Surowiecki, 2005). In this section, we take a closer look at this community and its workflows and habits in compiling the lexicon entries.

Wiktionarians. Wiktionary contributors are called *Wiktionarians*. They can be divided into three different types:

¹¹ http://en.wiktionary.org/wiki/Wiktionary:Picture_dictionary

1. The smallest group are the ninety-eight *administrators*, who must be nominated and elected by a majority. Administrators have the right to delete pages, change user permissions, and block articles or users.
2. *Registered users* are all the contributors who have created a personal account. This allows them to sign their edits with their name and make use of, for example, a watchlist to keep track of certain articles. There are currently 401,198 registered users for the English Wiktionary, 40,005 for the French edition, 36,900 for the German, and 32,692 for the Russian. In accordance with other collaboratively constructed resources, the number of edits per user follows a Zipf law. Therefore, most registered users perform only a few or perhaps not even a single article edit. When counting only users with at least ten edits, the number of actively contributing users drops to 3,958 for the English, 965 for the French, 794 for the German, and 277 for the Russian language edition.
3. The third type of contributors is *unregistered users*. They are also called *IPs*, because of their anonymous edits that are solely distinguishable by their Internet Protocol (IP) address. It is impossible to say how many people actually contribute to the project, since an IP address can be shared by many. Unregistered users perform about 5% of the article edits.

Automatic processing. In addition to human users, there are also so-called *bots*, i.e. computer programs that automatically crawl through the wiki pages and make changes according to certain patterns or rules. Currently, there are twenty-two active and seventeen inactive bots within the English Wiktionary. They have different responsibilities, which include automatic data imports, reformatting certain sections, and finding inter-wiki links to other Wiktionary language editions.

Discussion culture. The Wiktionary community has a lively discussion culture including both content (i.e. lexicographic) and technology (i.e. Wiki software) related concerns. As mentioned in Section 2.3 above, each article page has a *talk page* attached that can be used to discuss its content. It is good practice to sign a comment with one's own user name and the current date. The comments can address criticism and questions about the current state of the article or discuss possible extensions or modification of it. Figure 3 shows the talk page of the English Wiktionary article "colour". To date, the talk pages in Wiktionary have not been systematically studied. Similar works exist, however, for Wikipedia talk pages (Stegbauer, 2009; Stvilia et al., 2008) that might serve as a good starting point.

Besides the talk pages of individual articles, Wiktionary also offers general pages for discussing its organization and development as a whole. These pages are entitled "tea room", "etymology scriptorium", "beer parlour" and "grease pit". In general, the conversation is of an informal and colloquial style; a consensus is usually reached by voting.¹² Most questions and suggestions are quickly responded to. However, there are also topics that have been under discussion for a long period or have got completely stuck.

¹² <http://en.wiktionary.org/wiki/Wiktionary:Votes>

The image shows a screenshot of the Wiktionary website, specifically the talk page for the English article "colour". The page is titled "Talk:colour" and features a navigation bar with tabs for "Entry", "Discussion", "Citations", "Read", "Edit", and "History". A search box is visible in the top right corner. The main content area contains a discussion about the spelling "colour" versus "color". The discussion starts with a question: "Shouldn't this redirect to color (or vice versa)?" and includes several replies from users like Ortonmc, Hippletrail, and dmh, discussing the merits of each spelling and the impact of Americanization on the dictionary. The page also has a sidebar on the left with various navigation options like "Main Page", "Community portal", and "Feedback".

Fig. 3: Talk page for the English Wiktionary article “colour”.

Policies and guidelines. The Wiktionary community has developed a set of guidelines, mostly about the format of lexicon entries and the inclusion of new terms. Although there are slight differences in the guidelines of each language edition, they are largely similar. The main guideline for the inclusion of a new term is “if it’s likely that someone would run across it and want to know what it means”.¹³ Each encoded term needs to be “attested” within the language, which means “verified through (1) clearly widespread use, or (2) usage in a well-known work, or (3) usage in permanently recorded media, conveying meaning, in at least three independent instances spanning at least a year”.¹⁴ Unlike printed dictionaries, Wiktionary has practically no size restrictions. The guidelines therefore permit partial words, multi-word expressions, etc. that are often only partially considered or completely excluded from the headword list of other lexicons. In addition, Wiktionary also encodes inflected word forms (e.g. “went”) and common misspellings (like “aweful”) as separate lexicon entries, which is not done in most other lexicons, although it provides very interesting information for language learners.

Revision history. Every edit operation within Wiktionary is recorded and archived. In this way, a previous revision of an article can be reviewed at any time in order to inspect how the article has changed, and which users made particular changes. For each edit, the user can

¹³ <http://en.wiktionary.org/w/index.php?oldid=13078056> (10 May 2011).

¹⁴ <http://en.wiktionary.org/w/index.php?oldid=13078056> (10 May 2011).

provide a short note describing the modifications made and their reasons. Wiktionary contributors often use the revision history to revert vandalism, i.e. changes that introduced spam or deleted important parts of an article. The revision history also allows citations of a specific version of a Wiktionary article that does not change over time. This is an important feature within the World Wide Web which is constantly changing. For lexicographers, the revision history offers the unique possibility of recording how an article evolves (e.g. for exploring the semantic shift of a certain term). An example is the term “hand-held” which in 2005, when hand-helds were usually used as personal digital assistants, was described in Wiktionary as “a computing device (e.g. organiser, Internet-enabled cell phone) that is operated while held in the hands”. Today, these devices often contain portable video games, which has led the Wiktionary community to change the gloss to “a personal digital assistant or video game console that is small enough to be held in the hands”. Together with the talk pages described above, Wiktionary’s revision history provides us with the opportunity to study the lexicographic construction process as a whole (i.e. all decisions made on a certain entry). In a lexicographic publishing company, this information is either undocumented or private.

3 Analysing Wiktionary: a critical assessment

Having described Wiktionary in isolation, we now turn towards assessing its linguistic information in comparison to expert-built lexicons. Although the aim is to introduce Wiktionary in its full variety of language editions, we need to restrict our analysis to a selection of languages, due to the language skills of the authors and the limited availability of software libraries to analyse the encoded information quantitatively. As expert-built lexicons, we have chosen commonly used computational lexicons, since they allow their data to be automatically accessed in a similar way to Wiktionary. This is necessary for a fair comparison between the different types of lexicons. Traditional dictionaries are usually not intended for automatic processing and are therefore less suitable for a quantitative comparison; as Hirst (2004: 270–271) put it: “An ordinary dictionary is an example of a lexicon. However, a dictionary is intended for use by humans, and its style and format are unsuitable for computational use in a text or natural language processing system without substantial revision. [...] Nonetheless, a dictionary in a machine-readable format can serve as the basis for a computational lexicon, [...] Perhaps the best-known and most widely used computational lexicon of English is WordNet [...]”.

In our study, we analysed the English Wiktionary in comparison to the Princeton WordNet 3.0 (Fellbaum, 1998) and the electronic version of Roget’s thesaurus (Jarmasz and Szpakowicz, 2003), the German Wiktionary in comparison to GermaNet 6.0 (Kunze and Lemnitzer, 2002), and OpenThesaurus (Naber, 2005), as well as the Russian Wiktionary in comparison to the Russian WordNet 3.0 (Гельфейнбейн et al., 2003).¹⁵ For each lexicon, we studied the coverage of terms, lexemes, word senses, domains, and registers in both a qualitative and quantitative manner.

¹⁵ We use JWKTL (Zesch et al., 2008a) and Wikokit (Krizhanovsky, 2010) for parsing the Wiktionary data of 2 April 2011 (English edition), 6 April 2011 (German edition) and 4 April 2011 (Russian edition). For OpenThesaurus, we use a database dump from 8 September 2010.

3.1 Coverage of terms

A lexical form that consists of either a single word (e.g. “plant”), or multiple words (e.g. “freedom of speech”) will be called a *term* throughout this chapter. Note that, according to this definition, the noun and the verb “plant” are represented by the same term. Previously, terms (according to our definition) have also been referred to as *words* (Wilks et al., 1996) or *lexical items* (Atkins and Rundell, 2008). The number of terms encoded in Wiktionary is equivalent to its number of article pages (i.e. 2,402,397 in the English Wiktionary, 182,688 in the German Wiktionary, and 507,100 in the Russian Wiktionary). As described in Section 2.2 above, each Wiktionary edition can describe terms from multiple languages, which is not true for our expert-built lexicons. Therefore, we will focus solely on the terms in the native language of the respective Wiktionary edition (English terms in the English edition, etc.). The English and the Russian WordNet contain a large number of Latin terms that are part of the biological taxonomy of organisms. These terms are excluded from our study, since they are not encoded as native terms in Wiktionary. Expert-built lexicons usually do not list inflected word forms as separate terms. Wiktionary, however, also fosters the inclusion of inflected word forms, which is particularly useful for terms with irregular inflection. For our comparison of terms, we removed all inflected word forms.¹⁶

Table 4 shows the number of comparable terms that we used. We find that the German and the Russian Wiktionaries are of similar size to the expert-built lexicons. However, the English language Wiktionary exceeds the size of the Princeton WordNet by about 1.5 times and that of the Roget’s thesaurus by more than 3.5 times. In comparison to expert-built lexicons, Wiktionary is hence able to compete in terms of coverage, which makes it a valuable resource. Besides comparing the absolute sizes, we also analyse which types of terms are predominantly found in one of the lexicons by assessing the coverage of basic vocabulary and neologisms as well as the usage frequencies of the covered terms.

Basic vocabulary. The basic vocabulary of a language is known to change very slowly and should be well represented in a (general-purpose) lexicon. We compared the coverage of several word lists of basic vocabulary and report the results in Table 4. We used the *Swadesh* lists (Dyen et al., 1992) for English, German, and Russian; Ogden’s (1938) *Basic English* word list, West’s (1953) *General Service List* (GSL), and Nation’s (2006) *BNC 1–4* lists based on the British National Corpus for English; the *GUT1 Wortschatz*¹⁷ 100 and 500 for German; and Штейнфельдт’s (1963) list of common terms in modern Russian. Each Wiktionary edition covers the basic vocabulary very well. The English Wiktionary seems to be the most thorough, as it is the only lexicon that covers the full Swadesh list and over 99% of the other word lists. While the expert-built lexicons of English also have a good coverage of the basic vocabulary, the problem of coverage can be more acute for German and Russian. Wiktionary can be of great help here, as it retains a high coverage of over 96%.

¹⁶ Apart from the 363 inflected word forms noted for the Russian Wiktionary, there are a large number of inflected word forms within the 103,597 redirects. Redirects are not contained in the number of native terms and thus not included in this study.

¹⁷ http://www.gut1.de/grundwortschatz/grundwortschatz_500.html

Table 4: Number of terms in Wiktionary and comparable resources.

English language	Wiktionary	WordNet	Roget's Thesaurus
Native terms:	352,865	148,730	59,391
– Latin terms:	160	7,080	22
– Inflected forms:	115,635	0	0
= Comparable terms:	237,070	141,650	59,369
Coverage of			
Swadesh list:	100.00%	92.68%	96.10%
Nation's BNC 1–4:	99.92%	97.75%	90.52%
West's GSL:	99.91%	97.24%	96.50%
Ogden's Basic English:	99.41%	96.94%	97.53%
Neologisms:	11.35%	0.72%	0.18%

German language	Wiktionary	GermaNet	OpenThesaurus
Native terms:	83,399	85,211	58,208
– Latin terms:	13	17	17
– Inflected forms:	34,146	0	0
= Comparable terms:	49,240	85,194	58,191
Coverage of			
Swadesh list:	97.70%	87.10%	91.71%
Gut1 Wortschatz 100:	98.99%	76.77%	89.90%
Gut1 Wortschatz 500:	99.20%	72.31%	83.47%
Neologisms:	0.37%	1.92%	0.44%

Russian language	Wiktionary	Russian WordNet
Native terms:	133,435	130,062
– Latin terms:	0	5,025
– Inflected forms:	363	0
= Comparable terms:	133,072	125,037
Coverage of		
Swadesh list:	96.97%	84.42%
Штейнфельдт:	97.05%	67.88%
Neologisms:	48.52%	4.72%

Neologisms. Meyer and Gurevych (2010b) have noticed a high number of neologisms within the English Wiktionary. In order to quantify this observation, we compared the coverage of these newly coined terms using a list of 555 English neologisms¹⁸ from 1997 to 2008 provided by Birmingham City University, a list of 36,220 German neologisms¹⁹ taken from the Wortwarte project for the years 2000–2010, and 7,482 Russian neologisms²⁰ provided by the Russian Academy of Sciences. The results can be found in Table 4. Note that, due to the different size of the neologism lists and the different language characteristics (such as the extensive use of compounding in German), the numbers are not comparable across the three languages. Both the English and the Russian Wiktionary editions encode significantly more neologisms than their respective expert-built lexicons. This can be explained by the

¹⁸ <http://rdues.bcu.ac.uk/neologisms.shtml>

¹⁹ <http://www.wortwarte.de>

²⁰ <http://dict.ruslang.ru/gram.php?act=search&orderby=word>

collaborative construction process of Wiktionary, which allows updating of the lexicons immediately, without being restricted to certain release cycles as is the case for expert-built lexicons. The German language lexicons cover only between 0.37% and 1.92% of the neologisms. GermaNet contains about 550 neologisms more than Wiktionary. Despite that, Wiktionary is a promising resource for neologisms, as it has the ability to encode neologisms not yet found in expert-built lexicons. We will study this in more detail in the subsequent section when measuring the overlap between the different lexicons.

3.2 Coverage of lexemes

A *lexeme* is a combination of a term and its part of speech that is used as a headword for a lexicon entry. The English Wiktionary encodes, for instance, three lexemes for the term “bass”: one for the adjective describing a sound and two for the noun distinguishing the music-related etymology from the biological organism. The latter distinction denotes homonymy—as opposed to polysemy, which is represented in Wiktionary by providing multiple word senses (describe in Section 3.3 below).

It is surprising that the Wiktionary community distinguishes between homonymy and polysemy, since homonymy “is gradually being abandoned as an organizing principle in many types of dictionary” (Atkins and Rundell, 2008: 281). The reason is that the distinction can cause confusion when looking up a term without knowing its etymology and is hence not very helpful for dictionary readers (see also Moon, 1987). The layout of lexicon entries in Wiktionary has been discussed for a long time and many different ways of organizing the article pages have been proposed.²¹ An early idea was to create a separate article page for each word sense. This suggestion was abandoned in 2003 because the different senses could not be easily compared. The Wiktionary community then used only a single article page per term and created separate lexicon entries whenever a term had multiple etymologies or pronunciations (e.g. two entries for the two possible pronunciations of “read”). However, this idea was soon abandoned as it was found to be too unstable. The basic principle of describing homonymous terms in separate entries was formulated as a guideline in 2004 and is still used today. By 2006, the distinction had been questioned mostly for usability reasons. The article pages should start with the list of word senses, since they represent the most important knowledge for the lexicon users. Etymologies and pronunciations should no longer be used to distinguish different lexicon entries but merely become additional information attached to the word senses. This suggestion was, however, rejected by the community, since etymologies were seen to play a major role in distinguishing word meanings.

Table 5 shows the number of lexemes in each lexicon and their part-of-speech distribution. As described above, we separated out lexemes that were not directly comparable, i.e. Latin terms and inflected word forms. The English Wiktionary was again the largest lexicon. It encoded the most nouns and verbs and more than twice as many adjectives and adverbs as WordNet and Roget’s thesaurus. The German Wiktionary shows a different picture: it is the smallest lexicon compared to GermaNet and OpenThesaurus. Verbs seem to be particularly under-represented, as GermaNet encodes more than twice and OpenThesaurus more than three times as many verbs as the German Wiktionary. However, it also encodes a lower number of adjectives and nouns. The Russian Wiktionary contains more verbs, adjectives, and adverbs than the Russian WordNet but, in turn, contains a lower number of nouns.

²¹ See http://en.wiktionary.org/wiki/Wiktionary_talk:Entry_layout_explained/ and the corresponding archive pages for a full discussion on this topic.

Table 5: Number of lexemes in Wiktionary and comparable resources.

English language	Wiktionary	WordNet	Roget’s Thesaurus
Lexemes:	379,694	156,584	62,819
Comparable lexemes:	247,192	149,502	60,760
<i>Nouns:</i>	154,452	111,954	29,854
<i>Verbs:</i>	23,172	11,531	15,150
<i>Adjectives:</i>	58,502	21,536	12,739
<i>Adverbs:</i>	11,066	4,481	3,017
<i>Other parts of speech:</i>	13,206	0	2,037
Not comparable:	119,296	7,082	22

German language	Wiktionary	GermaNet	OpenThesaurus
Lexemes:	85,574	85,257	58,213
Comparable lexemes:	43,843	85,240	57,916
<i>Nouns:</i>	33,841	68,211	38,281
<i>Verbs:</i>	4,280	8,812	10,667
<i>Adjectives/Adverbs:</i>	5,722	8,217	8,968
<i>Other parts of speech:</i>	7,455	0	280
Not comparable:	34,276	17	17

Russian language	Wiktionary	Russian WordNet
Lexemes:	134,994	131,251
Comparable lexemes:	115,001	126,224
<i>Nouns:</i>	64,190	97,257
<i>Verbs:</i>	18,508	8,995
<i>Adjectives:</i>	26,714	16,087
<i>Adverbs:</i>	5,589	3,885
<i>Other parts of speech:</i>	19,452	0
Not comparable:	541	5,027

Parts of speech. In total, we found sixty-nine different part-of-speech tags within the three Wiktionary editions. Table 6 shows the number of lexemes per part-of-speech tag. Since many tags are very fine-grained, for brevity we grouped them into the fourteen general categories shown in the table. The Wiktionary community uses, for instance, three different tags for abbreviations: initialisms (pronounced letter by letter; e.g. “CD” for “Compact Disc”), acronyms (pronounced like a regular word, e.g. “ROM” for “read only memory”), and abbreviations terminated by a full stop (such as “Apr.” for “April”). A similar distinction is made for pronouns (e.g. demonstrative, reflexive, or possessive pronouns), particles (e.g. comparative, intensifying, and answering particles), affixes (e.g. prefixes and suffixes), and phrases. The latter are tagged as proverbs (e.g. “love is blind”), interjections (e.g. “good God”), idioms (e.g. “in the same boat”), or collocations (like “strong tea”). Wiktionary encodes a high number of phrases which are particularly useful in combination with their translations into other languages, since idioms and proverbs are usually very hard to translate. This opens up very valuable opportunities for cross-lingual lexicography. The high number of named entities in the English Wiktionary is also notable. In comparison to the English WordNet, we predominantly find given names (e.g. “Alice” or “Nadine”), and toponyms (e.g. “Berlin” or “Ohio”), as well as named entities from the non-US culture (such as the Arabic broadcaster “Al Jazeera” or the Swiss canton “Aargau”). Interestingly, phrasal verbs (like

“turn off”) and compounds (like “toothpaste”) do not receive a special tag, but are considered as verbs, nouns, etc.

Table 6: The part-of-speech tags used in Wiktionary.

Part of speech	English Wiktionary	German Wiktionary	Russian Wiktionary
Noun	218,629	32,808	62,861
Verb	62,202	4,269	18,524
Adjective	58,872	5,015	26,717
Adverb	11,079	669	5,602
Named entity	15,635	1,062	15,063
Abbreviation	6,763	3,050	234
Phrase	3,217	1,915	930
Particle	8	36	93
Pronoun	364	132	106
Preposition	463	108	135
Numeral	376	140	63
Determiner	93	17	15
Affix	1,474	472	198
Other	519	1,610	4,453

Overlap of lexemes. To examine whether the lexicons largely overlap or contain complementary information, we aligned lexemes that shared the same term and part of speech and measured their lexical overlap. Figure 4 shows a Venn diagram of the number of lexemes shared by each pair of lexicons. We find that the total overlap of the lexicons is very small. For the English language, only 11% of the lexemes in Wiktionary, 19% of the lexemes in WordNet, and 46% of the lexemes in Roget’s thesaurus are found within the respective other lexicons. The highest number of lexemes is shared by Wiktionary and WordNet, which is, however, still quite low compared to the number of lexemes found in only one of the resources.

This is a very surprising result, since one would expect two lexicons covering general language to basically encode the same list of lexemes. We therefore analysed which lexemes are encoded in only one of the lexicons and particularly found named entities (e.g. “Grammy”), multi-word expressions (e.g. “air sick”), and alternative spellings (e.g. “narcist”), as well as domain-specific lexemes. Wiktionary predominantly encodes lexemes from information sciences (e.g. “sound card”), natural sciences (e.g. “benzoyl”), and sports (e.g. “libero”), as well as informal (e.g. “ear candy”), and archaic lexemes (e.g. “abaculus”). In WordNet, we mainly found lexemes from the biological or medical domain (e.g. the “napa” plant, or the “axial muscle”), but encountered also numerous lexemes covering shades of colour (such as “reddish-pink”).

The overlap between lexemes was similarly small for the Russian lexicons, and although the number of shared lexemes was slightly higher for the German lexicons, there were still large differences in their coverage.

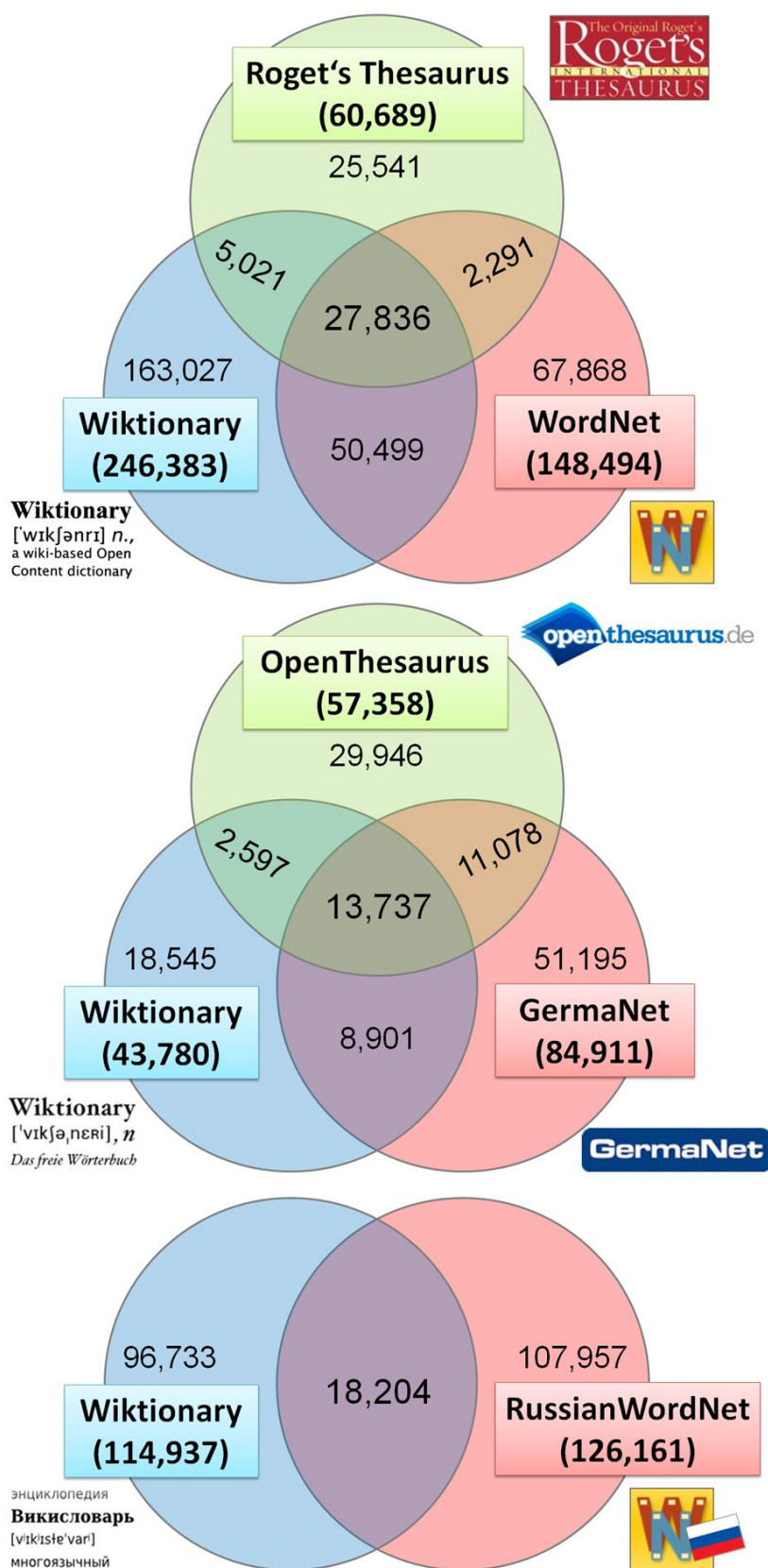


Fig. 4: Overlap of lexemes between the English (top), German (middle), and Russian (bottom) lexicons.

3.3 Coverage of word senses

In Wiktionary, the different meanings of a lexeme are enumerated in a list of *word senses*. Each word sense is described by a short gloss that is sometimes accompanied by usage examples or references to related word senses. There are, for instance, the following three word senses for the noun “boat”:

1. A craft used for transportation of goods, fishing, racing, recreational cruising, or military use on or in the water, propelled by oars or outboard motor or inboard motor or by wind.
2. (poker slang) A full house.
3. (chemistry) One of two possible conformers of cyclohexane rings (the other being chair), shaped roughly like a boat.

Table 7 shows the number of word senses encoded in our eight lexicons. The highest number of word senses is found within the English Wiktionary, which has more than twice the number of senses as WordNet and over four times as many as Roget’s thesaurus. The German lexicons are not so different, although GermaNet contains about 20,000 word senses more than the other two. However, the Russian Wiktionary encodes a much lower number of word senses than the Russian WordNet.

Table 7: Number of word senses in Wiktionary and comparable resources.

English language	Wiktionary	WordNet	Roget’s Thesaurus
Word senses:	474,128	206,978	98,464
Lexemes with 0 senses:	145	0	0
Lexemes with 1 sense:	327,274	130,207	44,317
Lexemes with 2 senses:	33,640	16,375	10,107
Lexemes with >2 senses:	18,635	10,002	8,395
Max. senses per lexeme:	59	59	18
Avg. senses per lexeme:	1.25	1.32	1.57

German language	Wiktionary	GermaNet	OpenThesaurus
Word senses:	73,500	95,715	72,897
Lexemes with 0 senses:	36,004	0	0
Lexemes with 1 sense:	36,242	77,335	49,219
Lexemes with 2 senses:	8,197	6,071	5,788
Lexemes with >2 senses:	5,131	1,851	3,206
Max. senses per lexeme:	52	26	14
Avg. senses per lexeme:	1.48	1.13	1.25

Russian language	Wiktionary	Russian WordNet
Word senses:	80,618	182,448
Lexemes with 0 senses:	82,261	0
Lexemes with 1 sense:	38,060	110,387
Lexemes with 2 senses:	8,266	11,830
Lexemes with >2 senses:	6,407	9,034
Max. senses per lexeme:	30	54
Avg. senses per lexeme:	1.53	1.39

Degree of polysemy. Comparing the absolute number of word senses only allows us to draw limited conclusions, since lexicographers can choose different sense granularities for their sense descriptions. A higher number of word senses hence does not necessarily imply a higher coverage of meanings per se. We therefore also compared the *degree of polysemy*, which we define as the number of word senses per lexeme. Table 7 shows the number of lexemes with 0, 1, 2, and more than 2 word senses as well as the maximum and the average number of word senses per lexeme in the different lexicons.

Expert-built lexicons do not contain lexemes without any word senses. This is different in Wiktionary, where users may encode entries without providing all the linguistic information at once or even as a *stub* (i.e. a skeleton of empty sections without any content-related information) that needs to be filled by other contributors over time. The low number of these lexicon entries in the English Wiktionary indicates that it is in a stable state and contains definitions for the vast majority of lexemes. This is different for the German Wiktionary, which lacks word sense definitions for 42% of its lexemes and for the Russian Wiktionary, lacking definitions for as many as 60% of its lexemes. Smaller Wiktionary editions obviously need more development time in order to fill their gaps.

Between 80% and 90% of the lexemes in expert-built lexicons have only one word sense. In the English Wiktionary, 81% of the lexemes are monosemous, which conforms to this range. The German Wiktionary, however, contains only 68% monosemous lexemes and hence encodes a higher number of polysemous lexemes. A possible explanation for this discrepancy might be that the Wiktionary community is more likely to create articles for polysemous lexemes, since they can cause confusion when understanding a text and are thus felt to be more important (Meyer and Gurevych, 2010a). This also applies to the Russian Wiktionary, in which 72% of the lexemes are monosemous.

The average number of encoded word senses is similar in all the lexicons, ranging between 1.14 and 1.57. The maximum number of word senses is, however, very different—ranging from eighteen word senses in Roget’s thesaurus to fifty-nine word senses in the English Wiktionary and WordNet. This can be used as an indicator for sense granularity. WordNet is known to be very fine-grained (Palmer et al., 2007); the English and German Wiktionary editions seem to be of similar granularity. Studying the number of word senses for a lexeme and the time when a word sense has been encoded by the Wiktionary community is an important strand of lexicographic research. This is because it characterizes how the lexicon is used and what types of entries have not yet been encoded because they have not yet been looked up.

Polysemic difference. The English Wiktionary and WordNet both have exactly one verb with fifty-nine word senses. This seems to show strong similarity. The verb is, however, “(to) break” in WordNet and “(to) go” in Wiktionary. To accommodate this issue in our analysis, we measured the *polysemic difference* between the lexicons, which we define as the difference in the number of word senses per lexeme (Meyer and Gurevych, 2010b). The verb “(to) break” from the example above has thirty-four word senses in Wiktionary and hence yields a polysemic difference of $|59 - 34| = 25$.

In the English Wiktionary, 60% of the lexemes shared with WordNet and 40% of the lexemes shared with Roget’s thesaurus have a polysemic difference of 0 (i.e. the same number of word senses). This is even higher for the German Wiktionary: 60% of the lexemes are shared with GermaNet and 51% of the lexemes are shared with OpenThesaurus. Over 86% of the lexemes in the English Wiktionary and over 91% in the German Wiktionary have a polysemic difference of less than or equal to 2, which means that the number of encoded word senses per lexeme is not dramatically different. As our example of “(to) break” shows, there are, however, also a few lexemes with a very high polysemic difference, which is again either

an indicator of different sense granularities or for a lack of sense coverage in one of the lexicons.

Overlap of word senses. The adjective “buggy” has two word senses both in Wiktionary and in WordNet, which yields a polysemic difference of 0 for this lexeme. However, it turns out that the two lexicons only share a single word sense “infested with bugs”. Wiktionary additionally encodes “containing programming errors”, while WordNet encodes “informal or slang terms for mentally irregular”. In order to gain a clearer insight into the coverage of word senses, we need to align the lexicons and quantify the number of shared word senses—similar to our study concerning the overlap of lexemes reported in Section 3.2. While an alignment of lexemes can be achieved using simple word-matching approaches, the alignment of word senses is a very complex task that is the subject of ongoing research (Navigli and Ponzetto, 2010; Niemann and Gurevych, 2011).

To our knowledge, the word sense alignment by Meyer and Gurevych (2011) between the English Wiktionary and WordNet is the only work integrating Wiktionary with other lexicons. According to this word sense alignment, Wiktionary and WordNet share 56,970 word senses. For 60,707 WordNet synsets²² there is no corresponding word sense in Wiktionary. Conversely, there are 371,329 word senses in Wiktionary that have no counterpart in WordNet. Similar to our observation regarding the overlap of lexemes in Section 3.2, the overlap of word senses is surprisingly small. Table 8 shows the number of senses per part of speech that are only found in Wiktionary or WordNet (but not vice versa) and the number of senses shared by both lexicons. The word senses of inflected word forms are naturally missing from WordNet. However, both Wiktionary and WordNet encode a large number of senses that are not found in the other lexicon. The collaboratively constructed Wiktionary is hence an important resource that should be considered by lexicographers when composing the word senses of a lexicon entry. In particular, newly coined word senses such as the computer-science-related word sense of “buggy” can be quickly included in Wiktionary due to its continual updatability.

Table 8: Number of word senses only found in Wiktionary or WordNet and shared by both lexicons.

Part of speech	Wiktionary and WordNet	Only Wiktionary	Only WordNet
Nouns:	34,464	158,085	47,651
Verbs:	8,252	29,119	5,515
Adjectives/adverbs:	14,236	60,977	7,541
Other parts of speech:	0	16,778	0
Inflected word forms:	0	106,328	0

Composition of glosses. In a qualitative study on the composition of glosses in the English Wiktionary and WordNet (Meyer and Gurevych, 2010b), we often observed only minor differences in the wording of glosses for overlapping word senses. Wiktionary encodes, for example, “a nun in charge of a priory; an abbess or mother superior” to describe “prioress”. This meaning is described in WordNet as “the superior of a group of nuns”. The WordNet

²² Note that the alignment matches Wiktionary word senses with WordNet synsets—i.e. lists of synonymous word senses. This notion is not directly comparable to our definition of word senses. However, this only affects the scale of senses found only in WordNet, which we will not analyse any further, but rather focus on the word senses in Wiktionary.

gloss is broader as it does not restrict the prioress to a female (“superior” is defined there as “the head of a religious community”). The lexeme “tortoise” is described as “any of various land-dwelling reptiles of family *Testudinidae*, whose body is enclosed in a shell [...]. The animal can withdraw its head and four legs partially into the shell, providing some protection from predators” in Wiktionary, and as “usually herbivorous land turtles having clawed elephant-like limbs; worldwide in arid area except Australia and Antarctica” in WordNet. Although describing the same meaning, the two lexicons set a different focus: Wiktionary concentrates on the animal’s anatomy and unique behaviour, while WordNet stresses its habitat and nutrition. Comparing such small differences can be very helpful in the composition of glosses, which is one of the most challenging tasks of lexicographers.

Wiktionary is often criticized for providing unspecific or too-general glosses. As Fuertes-Olivera (2009: 123) points out, the noun “takeover” is, for instance, described as “the purchase of one company by another; a merger without the formation of a new company”, which does not really differentiate between the general purchase of a company and the specialized concepts of a takeover or a merger. Other issues are spelling errors in the lexicon entries, e.g. the use of “bootle feeding” in the article “bottle feed”. Hanks (this volume) observed many old-fashioned descriptions in Wiktionary, which stem from copying information from copyright-expired dictionaries.

Many of such errors are likely to be removed in a collaborative effort. In an experiment, Hanks (this volume) found that the Wiktionary community is very active and revises new entries within minutes. However, it is a serious problem to distinguish well-crafted entries from those that need substantial revision by the community. Although there are mechanisms to indicate a need for revision provided by the Wiki software, there is as yet no fixed review or release workflow for lexicon entries.

Sense ordering. The word senses in WordNet are ordered according to the SemCorpus frequencies (Fellbaum, 1998: 41). This promotes the most frequently used word sense to the first position, which is also a common strategy in practical lexicography (Atkins and Rundell, 2008: 250). However, using a corpus such as SemCorpus to obtain the sense frequencies might not yield very realistic data because sense-tagged corpora are usually very small and often limited to certain types of document or vocabulary (e.g. newspaper text).

Although there is no specific guideline for the sense ordering in Wiktionary, we observed that the first entry is often the most frequently used one. For the noun “tattoo”, the first word sense in Wiktionary is “an image made in the skin with ink and a needle” but “a drumbeat or bugle call that signals the military to return to their quarters” in WordNet. Intuitively, the Wiktionary word sense is the more frequently used one nowadays. The majority of the sentences in, for example, the British National Corpus refer to this meaning. Hence, the sum of subjective opinions on the usage of word senses that coins Wiktionary’s sense ordering can alleviate the limitations and sparseness of sense-tagged corpora and provide a viable resource for lexicographers when ordering word senses by usage frequencies.

3.4 Coverage of domains and registers

In Section 2.4, we introduced the notation of *linguistic labels*, which describe the domain, register, style, time period, etc. of a word sense. We identified 714 different linguistic labels within the English Wiktionary, 238 within the German, and 125 within the Russian edition.²³ About half of the word senses, 273,960 (58%) are tagged with at least one linguistic label in

²³ We only counted linguistic labels used at least ten times and removed labels describing grammatical properties, such as “not countable”.

the English Wiktionary. In the German and Russian Wiktionary, this percentage is lower: only 28,035 (38%) word senses in the German, and 34,937 (43%) in the Russian Wiktionary have a linguistic label.

Domain labels. A broad and balanced coverage of domain-specific vocabulary is a common challenge when compiling a lexicon (Pantel and Lin, 2002), since it largely depends on the individual expertise of the lexicographers. In Wiktionary, a large number of contributors work on the lexicon entries collaboratively and hence can develop domain-specific knowledge for practically every domain. Consequently, the majority of Wiktionary’s linguistic labels denote the thematic domain in which a word sense is used, e.g. the label “chemistry” of the third word sense of “boat”. These *domain labels* are very fine-grained. In order to identify over- or under-represented domains, we manually grouped similar domain labels by their general topic. The labels “cycling” and “weightlifting” are, for example, combined with the category ‘sports’. As a comparable resource, we use *WordNet Domains*²⁴ (Bentivogli et al., 2004), which encodes 157 different domain labels for 128,669 (62%) word senses of the English WordNet. Table 9 shows the distribution of domain labels in the three Wiktionary editions and WordNet Domains.

About a quarter of the domain labels in WordNet Domains are from biology, because WordNet covers the entire taxonomy of plants and animals, which is only partly encoded in Wiktionary. The Wiktionaries have a stronger focus on the other natural sciences—most prominently on chemistry (10,912 word senses) in the English Wiktionary. Other well-represented domains include information technology (IT), maths, medicine, and sports. The contributors in Wiktionary encode word senses on a voluntary basis, which might cause a focus on knowledge from their leisure time (such as sport) rather than from work-related topics. Clearly under-represented are the humanities and social sciences, although they are covered better within WordNet. While linguistics and engineering seem to be predominantly encoded by the German and the Russian Wiktionary communities, these domains are rarer in the English Wiktionary. The different focus of the Wiktionary language editions and WordNet Domains may help to close domain-specific coverage gaps in other lexicons in future.

²⁴ We use version 3.2, available from <http://wndomains.fbk.eu>, and ignore the label “factotum” as it does not represent a domain.

Table 9: Distribution of domain labels in the English, German, and Russian Wiktionaries and WordNet Domains.

Domain	English Wiktionary	German Wiktionary	Russian Wiktionary	WordNet Domains
Biology	9.5%	13.6%	11.3%	27.8%
Chemistry	15.4%	4.8%	4.7%	5.9%
Engineering	2.3%	6.9%	9.8%	2.7%
Geology	5.9%	3.9%	5.6%	6.1%
Humanities	6.0%	11.4%	13.6%	13.7%
IT	7.6%	2.9%	2.6%	1.2%
Linguistics	2.5%	18.9%	14.5%	3.7%
Maths	6.0%	3.5%	4.0%	1.1%
Medicine	10.0%	9.5%	9.1%	8.3%
Military	1.5%	1.4%	3.8%	1.4%
Physics	6.5%	3.7%	3.9%	2.9%
Religion	2.5%	3.4%	3.4%	1.9%
Social sciences	7.6%	6.9%	7.4%	10.6%
Sports	6.6%	4.2%	2.8%	1.8%
Other	10.1%	5.0%	3.5%	10.9%

Register and style labels. In addition to domain labels, there are also a large number of linguistic labels that denote a register of language, i.e. a variety of language used in a certain manner of speech or writing. The majority of these *register labels*—about 40%—comprise slang- or jargon-related word senses. This was also observed in Section 3.2. These labels comprise Internet jargon, argot, or young people’s language. Additionally, there are register labels denoting the degree of formality or marking offensive terms. *Style labels* are similar, as they mark word senses found in a certain type of text (e.g. newspaper or literary style). In total, we counted 14,266 word senses in the English Wiktionary, 3,237 in the German Wiktionary, and 2,573 in the Russian Wiktionary that encode at least one of these register or style labels.

Other labels. A third type of linguistic label used in Wiktionary comprises *temporal qualifiers*. The word sense “A sturdy merchant sailing vessel” of “cat” is, for example, marked as archaic. Sometimes, word senses are also associated with a particular period of time, such as “19th century”. Apart from that, we noticed a high number of dialect word senses marked as Scottish English, Swiss German, or Yorkshire English.

4 Conclusion

Collaborative lexicography is a novel paradigm for compiling dictionaries in which large communities, backed up by the phenomenon of collective intelligence, compete with expert lexicographers. In this chapter, we have studied the main principle of collaborative lexicography based on the collaboratively constructed, multilingual online lexicon Wiktionary.

First, we gave a comprehensive description of Wiktionary, including its historical development and its macro- and microstructure. We found that Wiktionary offers many different access paths to its knowledge and thus makes use of many innovative features of

online dictionaries that allow users to search, cross-reference, and browse through the lexicon entries by alphabet, topic, register, or in an onomasiological way. The lexicon entries often encode a large variety of heterogeneous linguistic knowledge, including etymological, phonetic, morphological, syntactic, semantic, crosslingual, and pictorial information. The flexible and easy-to-use Wiki software attracts a large number of contributors and enables the encoding of culture- or language-specific variations. This recipe for success has yielded over 170 language editions describing over nine million multilingual articles. In addition to the major languages (English and French), which represent the largest language editions, we also found smaller and endangered languages in Wiktionary. Finally, we examined different types of contributors and how they collaborate.

In the second part of our chapter, we compared Wiktionary to multiple expert-built lexicons in three languages. We found a very high coverage of terms in the English Wiktionary and competitive coverage in the German and the Russian Wiktionaries. In particular, neologisms and the basic vocabulary of a language are well covered by Wiktionary. The lexical overlap between the different lexicons is surprisingly small, which makes Wiktionary an important resource for additional linguistic information missing from other lexicons. We found more polysemous lexemes in Wiktionary, which might be looked up more frequently than monosemous ones, because they can cause confusion when understanding a text. The creation order of the lexicon entries can help to reveal the information needs of the dictionary users. We also inspected the formation of glosses and identified important additions in them, compared to expert-built lexicons. However, we also found erroneous and unspecific glosses which are not useful for a dictionary user. Wiktionary has as yet no reviewing or releasing workflow. Quality assurance and trustworthiness are hence important determinants when working with Wiktionary (and other collaboratively constructed resources). We studied the coverage of Wiktionary's linguistic labels that are used to further describe the domains, registers, styles, etc. of its word senses. We observed a general focus on domain-specific vocabulary from natural sciences, information sciences, and leisure, while social sciences and humanities were rather under-represented.

Our analysis emphasizes many different aspects of Wiktionary that rival expert-built lexicons. We believe that its unique structure and collaboratively constructed contents are particularly useful for a wide range of dictionary users, including:

1. Laypeople who want to quickly look up the definition of an unknown term or search for a forum to ask a question on a certain usage or meaning.
2. Language learners who benefit from the densely interlinked multilingual organization, the good coverage of basic vocabulary, and the use of graphics to illustrate word senses.
3. Professional translators who can exploit the translations of proverbs, interjections, idioms, or domain-specific vocabulary which have been collaboratively contributed by a multilingual community.
4. Journalists who take advantage of Wiktionary's up-to-dateness regarding neologisms or newly coined word senses.
5. Social scientists who study how a language is used, cultural peculiarities across languages, or the collaboration of web communities.
6. Linguists who wish to investigate semantic shifts or endangered languages.
7. Computational linguists who use Wiktionary data in natural language processing applications.
8. Lexicographers who can gain totally new insights about the users of their dictionaries. This includes questions of what it is important to include in a lexicon and what is comprehensible to the reader. In particular, the semantic knowledge of Wiktionary can be of great help for composing sense glosses (when used in addition to corpus-based

methods). Since every edit is archived, Wiktionary also allows the lexicographic process to be studied as a whole, in order to examine how a lexicon develops over time and what considerations and decisions are necessary. This is a new field of lexicographic research, since no corresponding data is available from expert-built dictionary manufacturers.

In conclusion, as an exemplary product of collaborative lexicography, Wiktionary opens up a variety of interesting use cases and research opportunities. We believe that collaborative lexicography will not replace traditional lexicographic theories, but will provide a different viewpoint that can improve and contribute to the lexicography of the future. Thus, Wiktionary is a rival to expert-built lexicons—no more, no less.

Acknowledgements

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806. We thank Michael Matuschek and Yevgen Chebotar for their valuable contributions to this work, and Andrew A. Krizhanovsky at the Russian Academy of Sciences Saint Petersburg for sharing the Wikokit software for parsing the Russian Wiktionary edition.

References

- Atkins, B. T. Sue, and Michael Rundell (2008) *The Oxford Guide to Practical Lexicography*. Oxford, UK: Oxford University Press.
- Bentivogli, Luisa, Pamela Forner, Bernardo Magnini, and Emanuele Pianta (2004) Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing. In *Proceedings of the COLING '04 Workshop on 'Multilingual Linguistic Resources'*, 101–108, Geneva, Switzerland.
- Bergenholtz, Henning, and Sven Tarp (eds.) (1995) *Manual of Specialised Lexicography: The preparation of specialised dictionaries* (=Benjamins Translation Library 12). Amsterdam: John Benjamins Publishing.
- Carr, Michael (1997) Internet Dictionaries and Lexicography. *International Journal of Lexicography* 10(3): 209–230.
- Descy, Don E. (2006) The Wiki: True Web Democracy. *TechTrends* 50(1): 4–5.
- Dyen, Isidore, Joseph B. Kruskal, and Paul Black (1992) *An Indoeuropean Classification: A Lexicostatistical Experiment* (= Transactions of the American Philosophical Society 82), Philadelphia, PA: The American Philosophical Society.
- Etzioni, Oren, Kobi Reiter, Stephen Soderland, and Marcus Sammer (2007) Lexical Translation with Application to Image Search on the Web. In *Proceedings of Machine Translation Summit XI*, Copenhagen, Denmark.
- Fellbaum, Christiane (ed.) (1998) *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. Cambridge, MA: MIT Press.
- Fuertes-Olivera, Pedro A. (2009) The Function Theory of Lexicography and Electronic Dictionaries: Wiktionary as a Prototype of Collective Free Multiple-Language Internet Dictionary. In Henning Bergenholtz, Sandro Nielsen, and Sven Tarp (eds.), *Lexicography at a Crossroads: Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow* (=Linguistic Insights: Studies in Language and Communication 90), 99–134. Bern: Peter Lang.

- Giles, Jim (2005) Internet encyclopaedias go head to head. *Nature* 438(7070): 900–901.
- Hirst, Graeme (2004) Ontology and the Lexicon. In Steffen Staab, and Rudi Studer (eds.), *Handbook on Ontologies* (=International Handbooks on Information Sciences), 209–230. Berlin/Heidelberg: Springer.
- Jarmasz, Mario, and Stan Szpakowicz (2003) Roget's Thesaurus and Semantic Similarity. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, 212–219, Borovets, Bulgaria.
- Johnson, Samuel (1755) *A Dictionary of the English Language*. London: W. Strahan.
- Krizhanovsky, Andrew A. (2010) The comparison of Wiktionary thesauri transformed into the machine-readable format, *arXiv:1006.5040v1* [cs.IR].
- Kunze, Claudia, and Lothar Lemnitzer (2002) GermaNet — representation, visualization, application. In *Proceedings of the Third International Conference on Language Resources and Evaluation Vol. 5*, 1485–1491, Las Palmas, Spain.
- Lepore, Jill (2006) Noah's Mark: Webster and the original dictionary wars. *The New Yorker* LXXXII(36): 78–87. New York, NY: Condé Nast Publications.
- Leuf, Bo, and Ward Cunningham (2001) *The Wiki Way: Quick Collaboration on the Web*. Boston, MA: Addison-Wesley.
- Lew, Robert (2010) Multimodal Lexicography: The Representation of Meaning in Electronic Dictionaries. *Lexikos* 20: 290–306.
- Lew, Robert (2011) Online dictionaries of English. In Pedro A. Fuertes-Olivera, and Henning Bergenholtz (eds.), *E-Lexicography: The Internet, Digital Initiatives and Lexicography*. London/New York: Continuum.
- Malone, Thomas W., Robert Laubacher, and Chrysanthos Dellarocas (2010) Harnessing Crowds: Mapping the Genome of Collective Intelligence. MIT Sloan School Working Paper 4732-09.
- Mann, Michael (2010) Internet-Wörterbücher am Ende der „Nullerjahre“: Der Stand der Dinge. Eine vergleichende Untersuchung beliebter Angebote hinsichtlich formaler Kriterien. In Ulrich Heid, Stefan Schierholz, Wolfgang Schweickard, Herbert Ernst Wiegand, Rufus H. Gouws, and Werner Wolski (eds.), *Lexicographica* 26, 19–46. Berlin/New York: de Gruyter. [Mann, Michael (2010) Internet dictionaries at the end of the “00s”: The state of affairs. A comparative study of popular services in terms of formal criteria. In Ulrich Heid, Stefan Schierholz, Wolfgang Schweickard, Herbert Ernst Wiegand, Rufus H. Gouws, and Werner Wolski (eds.), *Lexicographica* 26, 19–46. Berlin/New York: de Gruyter.]
- Meyer, Christian M., and Iryna Gurevych (2010a) Worth its Weight in Gold or Yet Another Resource — A Comparative Study of Wiktionary, OpenThesaurus and GermaNet. In Alexander Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing: 11th International Conference* (=Lecture Notes in Computer Science 6008), 38–49. Berlin/Heidelberg: Springer.
- Meyer, Christian M., and Iryna Gurevych (2010b) How Web Communities Analyze Human Language: Word Senses in Wiktionary. In *Proceedings of the Second Web Science Conference*, Raleigh, NC, USA.
- Meyer, Christian M., and Iryna Gurevych (2011) What Psycholinguists Know About Chemistry: Aligning Wiktionary and WordNet for Increased Domain Coverage. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, 883–892, Chiang Mai, Thailand.
- Moon, Rosamund (1987) The Analysis of Meaning. In John M. Sinclair (Ed.), *Looking Up: An Account of the COBUILD Project in Lexical Computing*, 86–103. London: Collins.
- Müller, Christof, and Iryna Gurevych (2009) Using Wikipedia and Wiktionary in Domain-Specific Information Retrieval. In Carol Peters, Danilo Giampiccol, Nicola Ferro, Vivien Petras, Julio Gonzalo, Anselmo Penas, Thomas Deselaers, Thomas Mandl,

- Gareth Jones, and Mikko Kurimo (eds.), *Evaluating Systems for Multilingual and Multimodal Information Access: Proceedings of the 9th Workshop of the Cross-Language Evaluation Forum* (=Lecture Notes in Computer Science 5706), 219–226. Berlin/Heidelberg: Springer.
- Naber, Daniel (2005) OpenThesaurus: ein offenes deutsches Wortnetz. In *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen: Beiträge zur GLDV-Tagung*, 422–433. Frankfurt: Peter Lang. [Naber, Daniel (2005) OpenThesaurus: a free German wordnet. In *Language technology, mobile communication, and linguistic resources: Contributions of the GLDV conference*, 422–433. Frankfurt: Peter Lang.]
- Nation, I. S. P. (2006) How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review / La revue canadienne des langues vivantes* 63(1):59–82.
- Navigli, Roberto, and Simone Paolo Ponzetto (2010) BabelNet: Building a Very Large Multilingual Semantic Network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 216–225, Uppsala, Sweden.
- Niemann, Elisabeth, and Iryna Gurevych (2011) The People’s Web meets Linguistic Knowledge: Automatic Sense Alignment of Wikipedia and WordNet. In Johan Bos, and Stephen Pulman (eds.), *Proceedings of the Ninth International Conference on Computational Semantics*, 205–216, Oxford, UK.
- Ogden, Charles K. (1938) *Basic English: A General Introduction with Rules and Grammar*, 7th edition. London: Kegan Paul, Trench, Trubner & Co.
- Palmer, Martha, Hoa Trang Dang, and Christiane Fellbaum (2007) Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering* 13(2): 137–163.
- Pantel, Patrick, and Dekang Lin (2002) Discovering Word Senses from Text. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 613–619, Edmonton, AB, Canada.
- Ruhlen, Merritt (1987) *A Guide to the World's Languages. Vol. 1: Classification*. Stanford, CA: Stanford University Press.
- Schlippe, Tim, Sebastian Ochs, and Tanja Schultz (2010) Wiktionary as a Source for Automatic Pronunciation Extraction. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association*, 2290–2293, Makuhari, Japan.
- Stegbauer, Christian (2009) *Wikipedia: Das Rätsel der Kooperation*. Wiesbaden: VS Verlag für Sozialwissenschaften. [Stegbauer, Christian (2009) *Wikipedia: The mystery of cooperation*. Wiesbaden: VS Verlag für Sozialwissenschaften.]
- Stvilia, Besiki, Michael B. Twidale, Linda C. Smith, and Les Gasser (2008) Information Quality Work Organization in Wikipedia. *Journal of the American Society for Information Science and Technology* 59(6): 983–1001.
- Surowiecki, James (2005) *The Wisdom of Crowds*. New York, NY: Anchor Books.
- West, Michael (1953) *A General Service List of English Words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. London: Longman, Green & Co.
- Wilks, Yorick A., Brian M. Slator, and Louise M. Guthrie (1996) *Electric Words: Dictionaries, Computers, and Meanings* (=ACL-MIT Press series in natural-language processing). Cambridge MA: The MIT Press.
- Zesch, Torsten, Christof Müller, and Iryna Gurevych (2008a) Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias (eds.), *Proceedings of the 6th International Conference on Language Resources and Evaluation*, 1646–1652, Marrakech, Morocco.

- Zesch, Torsten, Christof Müller, and Iryna Gurevych (2008b) Using Wiktionary for Computing Semantic Relatedness. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, 861–867, Chicago, IL, USA.
- Гельфейнбейн, Илья Г., Артем В. Гончарук, Влад П. Лехельт, Антон А. Липатов, and Виктор В. Шило (2003) Автоматический перевод семантической сети WordNet на русский язык. *Труды Международного семинара Диалог по компьютерной лингвистике и её приложениям*, Протвино, Россия. [Gelfenbeyn, Ilya, Artem Goncharuk, Vladislav Lehelt, Anton Lipatov, and Victor Shilo (2003) Automatic translation of WordNet's semantic network into Russian. In *Proceedings of the International Dialog Conference*, Protvino, Russia.]
- Штейнфельдт, Э. (1963) *Частотный словарь современного русского литературного языка*. Москва: Прогресс. [Steinfeldt, E. (1963) *Frequency dictionary for the modern Russian literary language*. Moscow: Progress.]