How Web Communities Analyze Human Language: Word Senses in Wiktionary

Christian M. Meyer
Ubiquitous Knowledge Processing Lab
Technische Universität Darmstadt
Hochschulstraße 10
D-64289 Darmstadt, Germany
http://www.ukp.tu-darmstadt.de

Iryna Gurevych
Ubiquitous Knowledge Processing Lab
Technische Universität Darmstadt
Hochschulstraße 10
D-64289 Darmstadt, Germany
http://www.ukp.tu-darmstadt.de

ABSTRACT

The emergence of Web 2.0 enables new insights in many research areas. In this study, we examine how communities of Web users define word senses. Although a fundamental notion in human language analysis, it is a major challenge to define what a word sense is. The Collective Intelligence of Web communities has the potential to provide fundamental insights into our understanding of word senses.

We focus on two human language resources from Computational Linguistics, namely WordNet and Wiktionary, and analyze their coverage and their word sense distribution. Then, we systematically study the nature of word sense definitions in both resources based on manually chosen representatives. We conclude our study by highlighting the potential of collaboratively defined word senses and suggest further analysis of collaborative resources, like Wiktionary.

Keywords

Collective Intelligence; Collaboration on the Web; Word Senses; Wiktionary; WordNet.

1. INTRODUCTION

The rise of the Socio-Semantic Web several years ago has vitally changed the nature of interaction and communication within the World Wide Web. Communities of Web users have started to create new resources of human knowledge, like Wikipedia, by means of Web 2.0 technologies. A crucial property of such resources is the collaborative construction process backed up by the phenomenon of Collective Intelligence as opposed to expert-formulated theories. This enables fundamentally new insights in many areas of research unthought of before and has the potential to radically influence previously existing research paradigms.

One specific example of how this can be done is a study presented in this paper, which investigates how communities on the Web collaboratively construct word senses. Word sense is a fundamental notion in human language analysis that has been a subject of intensive studies for many centuries in Philosophy, Linguistics, Lexicography, and Artificial Intelligence. While a key to successful understanding of the human language, it has turned out to be a major challenge to define word senses. Several approaches have been proposed to capture our intuition of a word's meaning, including both data-driven and theoretical considerations, which each fail to cover the whole variety of the intuitions of a native speaker [2]. As a result, our understanding of the nature of word senses is still limited today.

Kilgarriff particularly points out that *"The division of a word's meaning into senses is forced onto lexicographers by the economic and cultural setting within which they work"* [8, p. 100], which is especially a problem of expert-created word sense inventories, such as WordNet. In the field of Computational Linguistics, WordNet [3] has, however, become the de facto standard resource of word senses. This resource has been applied as a source of background knowledge for a long time, but has also been heavily criticized in the literature for inconsistently separating concepts and individuals, object- and meta-level, and for violations in its subsumption hierarchy [4]. In spite of numerous evaluation campaigns in the field of automatic sense disambiguation, e.g. SensEval[1], the progress made in this area of research has been rather moderate: At SensEval-4 [1], e.g., the best system exceeded the baseline approach by only 3% [13].

Ten years ago, the rapid development of the Web has put us in a position to use it as a source of lexical semantic knowledge [10]. Still, state-of-the-art methods for word sense discovery from the raw Web data are not yet mature enough to substitute for human-generated knowledge representations, since they are not able to model the large differences in the occurrence frequencies of the word senses and to identify the main senses of a word [16].

On the other hand, collaboratively constructed resources, such as Wikipedia and Wiktionary[2] emerged. While a lot of research effort has been spent on Wikipedia, the use of Wiktionary is yet poorly investigated. In spite of several successful methods that use collaborative resources as a source of background knowledge [5], e.g., for calculating the semantic relatedness of words [19], we, however, still lack a clear understanding of the information in these resources and the encoded collaboratively defined word senses in particular.

In previous work, we have studied the topology of the graph that is induced from the encoded word senses of both expert and collaborative resources [11]. We found that both resource types are governed by comparable topological properties. In this paper, we build upon these insights and study collaboratively defined word senses for the first time.

Web Science Conf. 2010, April 26-27, 2010, Raleigh, NC, USA.

[1] http://www.senseval.org/
[2] http://en.wikipedia.org/ resp. http://en.wiktionary.org/

In the following, we first review several definitions of word senses, before we introduce the expert built resource WordNet and the collaboratively constructed resource Wiktionary that both will be the subjects of our analysis. To relate the Wiktionary's word senses to those in WordNet, we first compare the coverage of the resources and analyze their word sense distributions in order to find differences in the encoded terms and word senses. We finally examine the nature of collaboratively defined word senses based on manually chosen representatives and draw our conclusions.

1.1 Word Sense – A Fundamental Notion

Speaking of *words* may cause confusion as it is unclear if they are regarded together with their part of speech or separated from them, if multi words, like *'plant life'*, are included or excluded, and if inflected word forms, such as *'planting'* are likewise meant as canonical word forms, e.g., *'(to) plant'*.

We therefore define the following terminology that will be used throughout the remaining article: A word form that is solely characterized by its representing string, will be called a *term*. This definition is independent from a part of speech and includes multi words and inflected forms. Examples are *'planted'*, *'plant life'*, or *'WebScience'*. The basic unit of a dictionary is called a *lexeme*. For the term *'plant'* there can, e.g., be the two lexemes *'plant (noun)'* and *'plant (verb)'*. Lexemes may be multi words. Although traditional dictionaries, like the American Heritage Dictionary[3], usually only contain canonical forms, we do not generally exclude inflected lexemes. We finally refer to a *word sense* as a discretized quantum of a lexeme's meaning [2]. There are, e.g., the two word senses *'plant⟨industry⟩'* and *'plant⟨botany⟩'* for the lexeme *'plant (noun)'*. While the former denotes a *'building for carrying on industrial labor'*, the latter represents *'a living organism lacking the power of locomotion'*.

Unfortunately, there are no clear borders to discretize the continuum of meaning. The word sense *'plant⟨botany⟩'* could, e.g., be split into a word sense *'plant⟨Plantae⟩'* that represents exactly the members from the biological kingdom *Plantae*, and a word sense *'plant⟨colloquial⟩'*, which represents a definition of plant that also covers organisms from other biological kingdoms, like algae, but sometimes also called plants in everyday life. Ide and Wilks note: *"That there is no absolutely right number of senses for a word is conceded by the fact that a publisher like Oxford University Press produces its major English dictionary in at least four sizes [..]"* [7, p. 49]. This impedes a strict definition of word senses and has led to several – partly contrary – points of view regarding meaning and word senses. It would exceed the scope of this paper to provide a full overview on this topic. We rather introduce selected theoretical and practical considerations about word senses in the following and refer to pertinent literature on the topic [2, 9, 7].

Theoretical Considerations

The study of meaning is very old and can be traced back to classical philosophers, like Plato and Aristotle. It has especially been a core topic in the philosophy at the beginning of the 20th century. Two of several main philosophical streams are associated with H. P. Grice and Gottlob Frege, respectively, and will be introduced in the following.

Grice argues that meaning is the act of communication. A speaker expects some kind of reaction when uttering something to a hearer. The meaning is induced by the context, the two subjects, and the way an utterance is communicated. The idea is summarized as *"meaning is something you do"*. In this framework, the words only *"have a meaning insofar as there are stable aspects to the role that the word plays in those utterances"* [9, p. 32].

Frege's philosophy, in contrast, assigns meanings to the building blocks of a sentence. Words bear a certain meaning and are combined by grammatical rules, which introduce additional meanings. The meaning of a full sentence is then composed from the meanings of the individual words and rules. This framework allows to apply operations from mathematical logic to model and evaluate meaning.

Both considerations have been criticized for being not comprehensive enough to model the full variety of meaning [14, 9], which is still an open research question.

Practical Considerations

Apart from the theoretical considerations about word meanings, there are more practical approaches that are commonly used in lexicography to define the individual word senses for a lexeme. On the one hand, lexicography aims at covering a word's meaning in a comprehensive and consistent way. On the other hand, the lexicographers are restricted by the intended dictionary size, audience, and editorial deadlines. Hence, their definition of meaning is more tangible, although maybe not as comprehensive as in philosophy.

The basic idea when compiling a dictionary is to find evidence about words and their meanings within the language. Since about twenty years, large corpora of written and spoken text are used for this purpose, whereas a KWIC (Key Word In Context) concordance is examined to define word senses and prove their evidence at the same time. A KWIC concordance lists every occurrence of a target word and the corresponding contexts within a large corpus. The lexicographer then groups occurrences with similar contexts and meanings. Those groups that cannot be merged any further form the word senses of a word. For each group, the lexicographer needs to identify core features that distinguish the given word sense from the others; these features are then used to compose a word sense definition.

It is obvious that the number and coverage of the defined word senses depend on the available corpora and the strategy that is applied to merge the occurrences. Hanks [6] distinguishes two types of lexicographers: *lumpers* that prefer a more coarse-grained or general word sense, and *splitters* that prefer a more fine-grained distinction of meaning. We will examine if this also holds for Web communities.

1.2 Human Language Resources

We aim at studying collaboratively defined word senses as opposed to the expert built word senses of traditional dictionaries or linguistic resources. We have chosen to focus our study on WordNet and Wiktionary, since WordNet is the standard resource for computational tasks involving word senses, and Wiktionary is possibly the only structured and collaborative resource for word senses, whose coverage is mature enough to compete with WordNet. Both resources will be introduced in the following.

[3] http://www.houghtonmifflinbooks.com/ahd/

WordNet

In 1985, George A. Miller and his group at the Cognitive Science Laboratory of the Princeton University started the development of WordNet[4] [3]. The current version 3.0 encodes 117,659 synsets, i.e. sets of synonymous lexemes that share a common meaning. For the lexeme *'plant (noun)'* there are, e.g., the synsets {*plant, works, industrial plant*} and {*plant, flora, plant life*}, which represent different word senses. WordNet encodes a short definition text, called a gloss, for each word sense, which is *'buildings for carrying on industrial labor'* and *'a living organism lacking the power of locomotion'* for the two synsets above. Example sentences can additionally be found for some of the synsets. Only the first of the synsets for *'plant'* has an example sentence: *'they built a large plant to manufacture automobiles'*. Due to its machine-readable structure and its open license policy, WordNet quickly became a standard resource for a wide range of tasks in the field of Computational Linguistics.

Wiktionary

The collaborative online project Wiktionary started in December 2002 with the goal of creating a large, Wiki-based, and multilingual dictionary that is both freely available and editable by volunteers. As a primer, the 1913 edition of the Webster's New International Dictionary has been imported to the English Wiktionary, of which still 994 entries remain unedited.[5] Since there are no special requirements for participating, the community of Wiktionary contributors grew, however, very fast: by the end of 2009, about 300,000 users have created over 1,500,000 articles in the English edition.

Currently, there are 172 language editions of Wiktionary. Each language edition contains lexemes from multiple languages – there is, e.g., an entry about the English noun *'plant'* in both the English and the German language edition. Since there are substantial differences in the syntax and the guidelines of the different language editions, we will solely focus on English entries of the English language edition that have been extracted by JWKTL[6] from a Wiktionary dump of January 19, 2010.

Wiktionary is organized in article pages that each represents a certain term and distinguishes one or more lexemes. Besides linguistic information, such as language, etymology, part of speech, or translations, different word senses are encoded for each lexeme entry. The word senses are represented by a short definition text that may come with example sentences or quotations. Figure 1 shows the article *'boat'* as an example of this representation.

2. RESOURCE COVERAGE

Before we approach the question of collaboratively defined word senses, we study the coverage of the resources in order to find out which kind of terms are encoded in expert created resources and which are merely found in collaborative resources. The term *'plant'* is, e.g., found in both resources, while the term *'exactor'* has only been found in Wiktionary and thus indicates a missing term within WordNet. Missing terms induce missing word senses and may reveal certain aspects of the resources and their encoded information.

[4]http://wordnet.princeton.edu/
[5]http://en.wiktionary.org/wiki/Category:Webster_1913, accessed on March 16, 2010
[6]http://www.ukp.tu-darmstadt.de/software/jwktl/

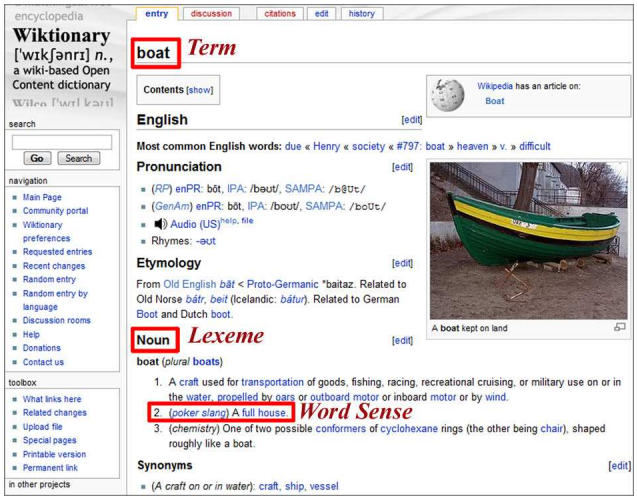

Figure 1: Wiktionary article *'boat'*

	Wiktionary	WordNet	Overlap
Number of Lexemes:	323,264	156,584	75,750
...only Nouns:	200,217	119,034	48,681
...only Verbs:	55,483	11,531	8,967
...only Adjectives:	46,636	21,538	14,484
...only Adverbs:	9,660	4,481	3,618
...other Parts of Speech:	11,268	0	0
Inflected Word Forms:	102,476	–	–
Latin Terms:	–	7,082	–
Abbreviations:	7,051	1,014	624
Proper Nouns:	13,494	14,236	3,110

Table 1: Coverage of Wiktionary and WordNet

We have counted 156,584 lexemes in WordNet and 323,264 lexemes in Wiktionary. Table 1 shows these numbers and the results of the following coverage analysis. At first sight, Wiktionary seems to encode twice as many lexemes as WordNet. WordNet, however, encodes only nouns, verbs, adjectives, and adverbs, while Wiktionary does not restrict the encoded parts of speech. In fact, 11,268 of Wiktionary's lexemes belong not to a part of speech encoded by WordNet. Those lexemes are marked as abbreviations, interjections, phrases, prepositions, affixes, numerals, or symbols.

Wiktionary also contains inflected terms, like the simple past form *'went'* or the plural form *'houses'* that are encoded as separate entries. Such word forms have not been included in WordNet. We found 102,476 Wiktionary lexemes that have been marked as an inflected word form. Verbs have the most inflected word forms, which leads to a significantly higher percentage of encoded verbs in Wiktionary: 17% of the lexemes in Wiktionary are verbs, compared to only 5% in WordNet. Without the inflected forms, the percentage drops to 6% in Wiktionary and thus leads to a similar part of speech distribution for verbs.

To relate the remaining lexemes, we aligned the resources at the lexeme level by matching lexemes with the same string representation and the same part of speech. We have counted 75,750 lexemes shared by both resources. This overlap is surprisingly low, which caused us to further analyze the differences of the resources. In WordNet, we found 7,082

Year/Dataset	Size	Wiktionary		WordNet	
2005 – BLOG	23	21	91.3%	10	43.5%
1994 – BCU	614	79	12.9%	7	1.1%
1997 – BCU	70	12	17.1%	0	0.0%
1998 – BCU	71	10	14.1%	0	0.0%
1999 – BCU	72	15	20.8%	3	4.2%
2000 – BCU	30	1	3.3%	0	0.0%
2001 – BCU	27	3	11.1%	0	0.0%
2002 – BCU	31	2	6.5%	0	0.0%
2003 – BCU	51	3	5.9%	0	0.0%
2004 – BCU	27	4	14.8%	1	3.7%
2005 – BCU	37	2	5.4%	0	0.0%
2006 – BCU	27	0	0.0%	0	0.0%
2007 – BCU	37	1	2.7%	0	0.0%
2008 – BCU	75	3	4.0%	0	0.0%
Total	**1,192**	**156**	**13.1%**	**21**	**1.8%**

Table 2: Coverage of Neologisms

Latin terms, like *'Sterculia rupestris'*, or *'genus Gastrolobium'*, that are scientific names within the biological taxonomy. Since in Latin, these terms are generally not encoded as English entries in Wiktionary and have thus not been extracted. Our analysis also showed a high number of abbreviations, like *'NFL'* or *'MD5'*, in Wiktionary, which are not present in WordNet. An abbreviation is thereby defined as a term that either ends with a period (like *'a.k.a.'*) or only consists of words with at least two upper case letters (like *'SWbW'*). WordNet encodes 1,014 abbreviations, Wiktionary contains 7,051. A total of 624 is found in both resources, thus indicating that the bulk of WordNet's abbreviations is covered by Wiktionary.

In WordNet, "no special attempt has been made to include proper nouns" [3, p. 23], while in Wiktionary, there are clear inclusion guidelines[7] for different types of proper nouns: Person or company names are not generally included in Wiktionary, only if they are commonly used to refer to a broader range of entities or events – *'kleenex'*, for example, refers nowadays to any disposable tissue, *'Google'* to any search engine on the Web, while *'Abraham Lincoln'* is defined as *'an emancipator or analogous reformer'*, which is derived from the quotation *'Clinton became the Abraham Lincoln of our movement'*. Since version 2.1 of WordNet, there is a distinction between instances and types [12] that can be used to count the number of proper nouns. The proper noun *'Cologne'* is, for example, an instance of *'city'*. We counted 14,236 proper nouns that occur as the target of an instance relation. Most of the proper nouns in Wiktionary are explicitly marked with the part of speech tag *'proper noun'*, e.g., the surname *'Meyer'*. We counted 13,494 terms with this part of speech. 3,110 of the identified proper nouns have been found in both resources, again indicating a low overlap. Manual inspection showed that person names, like *'Albert Camus'* or *'Johannes Kepler'*, have been predominantly found in WordNet, while Wiktionary showed more given names (e.g. *'Alice'* or *'Nadine'*) as well as named entities from non-US culture, like the Arabic broadcaster *'Al Jazeera'* or the Swiss canton *'Aargau'*.

We finally noticed numerous neologisms, i.e. newly coined terms, in Wiktionary. To quantify this observation, we evaluated the coverage of a list of modern words found on the Web[8] (referred to as the BLOG dataset) and of the neologism corpus of the Birmingham City University[9] (BCU), which contains neologisms from English newspapers published between 1994 and 2008. The manually compiled BLOG dataset contains the following 23 neologisms:

> *blog, RSS, podcast, WWW, weblog, mozilla, thunderbird, firefox, netscape, perl, usenet, CGI, HTTP, dotcom, flickr, technorati, google, iPod, doh, balti, ladette, mullet, alcopop*

With the exception of *'perl'* and *'flickr'*, all terms have been found in Wiktionary. WordNet, on the contrary, contains only the 10 underlined terms. As can be seen from Table 2, the coverage is generally low for the BCU dataset, since the corpus contains also terms that are only used once and thus not likely to be found in a dictionary (yet). Examples are *'Googleware'*, *'e-record'*, or *'schoolmaster-ish'*. The coverage of neologisms is significantly higher in Wiktionary, which confirms our intuition. On the one hand, Web communities are known to quickly reflect recent developments and trends [18]. On the other hand, Wiktionary allows constant changes rather than relying on editorial or release deadlines, which usually cause expert created resources to have a longer incubation time for newly established terms. Examples for terms of the BCU dataset only covered by Wiktionary are *'fashiony'*, *'bellgirl'*, and *'shockvertising'*. Only 2 of the 10 BCU terms covered by WordNet are not represented in Wiktionary: *'pawnee'* and *'up-tick'*.

We conclude this section by summarizing that the overlap between WordNet and Wiktionary at the lexeme level is rather low. This is due to different word inclusion policies of the resources. We additionally observed that none of the resources subsumes the other, which leads us to the conclusion that collaborative resources encode a large number of word senses for terms not even found in expert created resources. In Wiktionary, we particularly found abbreviations, given names, neologisms, and terms from non-US culture, while WordNet encodes especially person names as well as a large number of Latin terms from the biological taxonomy.

3. WORD SENSE DISTRIBUTION

Having studied the coverage of the resources at the lexeme level, we now turn to the word senses that are encoded for each lexeme. Therefore, we first extend our resource alignment such that each lexeme ℓ is associated with its corresponding number of word senses $\omega_r(\ell)$ encoded in resource r. The noun *'plant'*, e.g., distinguishes $\omega_{\text{wn}}(\ell) = 4$ word senses in WordNet and $\omega_{\text{wkt}}(\ell) = 9$ word senses in Wiktionary. Note that we solely focus on the 75,750 lexemes that are found in both resources as we have already studied terms encoded in only one resource within the previous section.

The *word sense distribution* $d_r(k)$ of resource r is the number of lexemes for a given number of word senses k:

$$d_r(k) = |\{\ell \mid \omega_r(\ell) = k\}|$$

Table 3 shows the number of lexemes for different values of k as well as the average $\overline{\omega}_r$ and maximum number of word senses $\hat{\omega}_r$. The word sense distribution d is similar in

[7] http://en.wiktionary.org/wiki/Wiktionary:CFI#Names

[8] http://softwareas.com/including-modern-words-in-modern-dictionaries, accessed on March 2, 2010

[9] http://rdues.bcu.ac.uk/neologisms.shtml

	All Lexemes		Nouns		Verbs		Adjectives		Adverbs		Other	
r	WKT	WN	WKT	WN	WKT	WN	WKT	WN	WKT	WN	WKT	WN
$d_r(1)$	85.2%	83.2%	85.8%	87.0%	87.3%	54.5%	82.2%	77.0%	88.4%	83.6%	75.4%	0.0%
$d_r(2)$	9.4%	10.5%	9.1%	8.4%	7.2%	22.0%	12.6%	15.2%	9.1%	12.0%	14.2%	0.0%
$d_r(3)$	2.8%	3.2%	2.7%	2.4%	2.7%	9.5%	3.3%	4.5%	1.7%	2.8%	4.3%	0.0%
$d_r(4)$	1.1%	1.3%	1.1%	0.9%	1.2%	5.3%	1.1%	1.6%	0.4%	0.7%	1.9%	0.0%
$d_r(\geq 5)$	1.3%	1.8%	1.3%	1.2%	1.7%	8.8%	0.9%	1.7%	0.4%	0.8%	3.2%	0.0%
$\overline{\omega}_r$	1.26	1.32	1.26	1.23	1.26	2.17	1.27	1.39	1.16	1.25	1.47	0.00
$\hat{\omega}_r$	58	59	57	33	58	59	22	27	12	13	47	0

Table 3: Word Sense Distribution in Wiktionary (WKT) and WordNet (WN)

both resources, with the exception of verbs: WordNet has a significantly higher average number of word senses $\overline{\omega}_r$ per verb. 8.8% of the verbs even have 5 or more word senses, which is clearly higher than 1.7% in Wiktionary. Another observation is that the maximum number of word senses $\hat{\omega}_r$ for nouns differs significantly: 57 in Wiktionary compared to 33 in WordNet. As the other parts of speech do not display such a large difference, we plan to further analyze the encoded noun word senses in Wiktionary.

By now, we have only considered $\omega_r(\ell)$ for each resource individually. We can, however, not conclude that the encoded word senses for a given lexeme are similar by solely looking at the word sense distribution. Both resources, e.g., encode exactly one verb with 42 senses, but it turns out that this verb is *'take'* in WordNet and *'catch'* in Wiktionary. Therefore, we calculated the *polysemic difference* $\Delta(\ell) = |\omega_{\text{wn}}(\ell) - \omega_{\text{wkt}}(\ell)|$ for each of the 75,750 encoded lexemes ℓ shared by both resources. The noun *'plant'* has, e.g., a polysemic difference of $\Delta(\ell) = |4 - 9| = 5$.

We found 45,955 lexemes (60%) with $\Delta(\ell) = 0$, indicating that the majority of lexemes encodes the same number of word senses. 71,677 lexemes (95%) have $\Delta(\ell) \leq 2$, which shows that the polysemic difference is not very high for almost the whole resource. We, however, found a few lexemes with a very high difference of up to 44.

In this section, we have studied the word sense distribution of the resources, i.e. the number of word senses per lexeme. We have found that the maximum number of word senses is clearly higher for nouns in Wiktionary, while WordNet shows a higher average number of word senses for verbs. We also considered the polysemic difference, i.e. the difference in the number of word senses for each lexeme, and found that 60% of the lexemes shared by both resources have the same number of word senses. From these observations, we can conclude that the number and the distribution of word senses follow similar patterns in both expert and collaborative resources, although some minor differences exist.

4. WORD SENSE COMPARISON

In the previous section, we have studied the word sense distribution and found a high agreement in the number of word senses a lexeme has in each resource. Even though a lexeme has, e.g., two word senses in both resources, it does, however, not necessarily mean that these word senses are equal or at least similar. The adjective *'buggy'* has, e.g., a word sense *'infested with bugs'* shared by both resources. The second word sense is, however, *'containing programming errors'* in Wiktionary and *'informal or slang terms for mentally irregular'* in WordNet, which does not denote the same meaning.

Thus, the logical next step of our analysis would be to create an alignment of the resources at the word sense level in order to identify missing word senses. While an automatic alignment algorithm based on string matching could be applied at the level of lexemes, there is, unfortunately, no such simple method for aligning the resources at the word sense level. This is due to the different representations and definitions of word senses based on human intuitions. Typically, even humans disagree on some instances of this task, which makes automatic word sense alignment hard. Such an automatic alignment algorithm would need to match a word sense *'A factory or other industrial or institutional building or facility'* with *'buildings for carrying on industrial labor'*, but not with *'a living organism lacking the power of locomotion'*. Although we believe that it is possible to automatically align the resources at the word sense level – and actually aim to develop such methods – we focus on a manual analysis of meaningful representative examples here. The insights we gain from these examples are not only crucial for creating alignment methods, but also have the potential to shed some light on the definition of word senses itself, which still is an open research question.

We refrain from simply choosing the representatives randomly, as we aim at covering as many aspects of the resources as possible. We therefore manually select a controlled sample by taking into account the observations of both previous sections. In particular, we examine lexemes of different parts of speech, polysemic differences, and usage frequencies. We restrict our selection to lexemes covered by both resources, as we have already systematically analyzed the resource coverage in section 2. Table 4 shows the number of word senses in our sample that has been found in both resources (the sense overlap o), that has only been found in Wiktionary (n_{wkt}) or only in WordNet (n_{wn}), as well as the number of word senses in Wiktionary (v_{wkt}) and WordNet (v_{wn}) that are variants of some other sense in the respective other resource. In the following, we will describe the invidiual dimensions and give examples for this categorization.[10]

Low Frequent Terms

As a first criterion for choosing representative lexemes, we consider how often the term is used within the English Language, i.e. its *language frequency* f. Thereby we want to study if there are differences in the word sense definition between rare, seldomly used terms and common, often used terms. As an approximation of the real language frequency, we use the term's frequency within a large corpus, namely, the Wortschatz corpus [15] that contains 1,000,000 English

[10]To improve the understanding of the paper, we provide a document describing the word senses used for our discussion at http://www.ukp.tu-darmstadt.de/data/lexical-resources/

	o	n_{wkt}	n_{wn}	v_{wkt}	v_{wn}
Frequency: $f \leq 5$	4	6	0	1	1
Frequency: $5 < f \leq 100$	9	2	3	3	0
Frequency: $f > 100$	16	7	5	16	9
Polysemy: $\Delta(\ell) = 0$	6	7	4	3	6
Polysemy: $\Delta(\ell) > 30$	8	32	1	5	0
Part of Speech: Nouns	30	46	7	14	7
Part of Speech: Verbs	9	6	4	8	7
Part of Speech: Adjectives	4	2	2	6	5

Table 4: Word Sense Comparison Results

sentences taken from different newspapers. From this corpus, we choose terms with a low ($f \leq 5$), medium ($5 < f \leq 100$), and high ($f > 100$) frequency.

Terms with a low frequency are both rare and known for having less word senses than frequent terms. Hence, three of our representatives encode only a single word sense in both resources: *'seagrass'* ($f = 1$), *'philology'* ($f = 2$) and *'prioress'* ($f = 4$). The encoded word senses of these terms can be considered equal, as there are only minor differences in the word sense definitions. Wiktionary, e.g., encodes *'a nun in charge of a priory; an abbess or mother superior'* for the term *'prioress'*, while WordNet defines *'the superior of a group of nuns'*, which clearly denotes the same meaning. At first glance, the Wiktionary definition seems to be more detailed as also synonyms are given. It should, however, be noted that the word *'prioress'* is contained within the WordNet synset {*abbess, mother superior, prioress*} and thus also contains these synonyms. Another difference between the word sense definitions is that Wiktionary restricts the *prioress* to be a *nun*, and WordNet to be a *superior*, which is defined by the resource as *'the head of a religious community'*. It thus lacks the restriction to females that is implicitly contained in Wiktionary's definition of *'nun'*.

For *'senescent'* ($f = 1$), both resources encode the word sense *'growing old'*. Wiktionary additionally encodes *'characteristic of old age'*, which is a very similar word sense, but defines a state rather than a process. We tried to verify this word sense by examining a KWIC concordance. Each examined sentence, however, uses either the word sense *'growing old'*, or is ambiguous such that both senses are appropriate. An example for the latter is *'My glass shall not persuade me I'm senescent'*[11]. It is thus unclear how to distinguish both word senses in their concrete language instantiations.

There is a large difference in the number of word senses for the term *'spec'* ($f = 2$): Wiktionary encodes 5 noun and 1 verb word sense, while WordNet only encodes the single noun word sense *'a detailed description of design criteria for a piece of work'*. It is interesting that Wiktionary has no exact counterpart for this meaning; it rather defines *'spec'* as a *'short form of specification'* and thus refers to another term that covers this word sense. It is arguable whether *'spec'* is, nowadays, used as a short form or as a synonym for *'specification'*. The verb word sense *'to specify, especially in a formal specification document'*, as well as the second noun word sense *'short form of speculation'* are both found in the American Heritage Dictionary but missing from WordNet. The remaining word senses include two other short forms denoting *'specialization'* and *'special'*, as well as a dialectal word sense denoting *'a special place (for hiding or viewing)'*.

All in all, we see a higher coverage of word senses for low frequent terms within Wiktionary. We, however, also found that these defined word senses are often informally defined, which makes a sense distinction hard.

Medium Frequent Terms

We have also chosen five terms from the Wortschatz corpus with a medium frequency. For the noun *'tortoise'* ($f = 14$) exactly one word sense is encoded in the resources, which denotes a land turtle. The word sense definitions are, however, slightly different: Wiktionary encodes *'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.'* and thus focuses on the animal's anatomy and unique behavior. WordNet stresses the habitat and nutrition of the animal: *'usually herbivorous land turtles having clawed elephant-like limbs; worldwide in arid area except Australia and Antarctica'*.

Both the terms *'intersect'* ($f = 8$) and *'freeway'* ($f = 100$) have one matching word sense, and one word sense that is only encoded in Wiktionary. While for *'intersect'*, the mathematical intersection operation is additionally defined (but missing from WordNet), there is a word sense about freeways especially in Australia, Canada, and the United States for the latter term. On the one hand, this word sense, *'(Australian/Canadian/US) A road designed for safe high-speed operation of motor vehicles [..]'*, is very similar to the generic word sense *'a toll-free highway'*, and thus makes a distinction very hard. On the other hand, there are indeed slightly different interpretations for the terms *'freeway'*, *'highway'*, *'expressway'*, etc., which, e.g., caused some discussion on Wikipedia about the corresponding articles.[12]

Until now, Wiktionary seems to encode a greater or equal number of word senses for our representatives than WordNet. This is different for *'alloy'* ($f = 29$), which has only one noun and one verb word sense in Wiktionary, but two for each part of speech in WordNet. Both Wiktionary word senses are fully covered by WordNet, which additionally encodes *'the state of impairing the quality or reducing the value of something'* as well as the corresponding verbalized form. As this word sense can be found in other printed dictionaries, it is clearly missing from Wiktionary.

The word senses in WordNet are ordered according to their frequency in the SemCor corpus [3, p. 41]. Manual inspection showed that the first word sense in Wiktionary often denotes the most frequently used one and thus should equal the WordNet's one. For the term *'tattoo'* ($f = 14$), the first word sense is, however, different: Wiktionary encodes *'An image made in the skin with ink and a needle'* and WordNet *'a drumbeat or bugle call that signals the military to return to their quarters'*. We examined two large corpora to find which word sense is more frequently used. It turned out that Wiktionary's sense is used in 23 of 30 sentences within the Wortschatz corpus and in 22 of 50 sentences within the British National Corpus. This confirms our intuition that *'tattoo'* in the sense of a skin image is more frequently used than the bugle call. Of the remaining 28 sentences in the British National Corpus, 18 could be

[11] Taken from the British National Corpus http://www.natcorp.ox.ac.uk/

[12] http://en.wikipedia.org/wiki/Talk:Freeway

assigned to the word sense *'a military display or pageant'* that is, e.g., used to refer to the Edinburgh Military Tattoo. This word sense has been found in Wiktionary, but is missing from WordNet. Apart from that, there are two further additional word senses from the military domain within Wiktionary that are not encoded in WordNet: *'(nautical) from taptoe, the time to close the taps'* and *'(nautical) a signal played five minutes before taps (lights out)'*.

Although we found several missing word senses in both resources, a large number of word senses for terms with a medium frequency is shared by Wiktionary and WordNet. We additionally observed that the first word sense is usually the most frequent one within Wiktionary. Since in WordNet, the most frequent sense is determined from a sense tagged corpus of limited size, the collaborative building process in Wiktionary might yield a better definition of the most frequently used word sense, which is an important feature in automatic word sense disambiguation tasks. We aim at further studying the first word senses of collaborative language resources in the future.

High Frequent Terms

Finally, we studied words with a high frequency in the language. For the term *'million'* ($f = 52,440$), Wiktionary encodes only a single word sense *'(long and short scales) The cardinal number 1,000,000: 10^6'*. This sense is also covered by WordNet, which additionally encodes *'a very large indefinite number (usually hyperbole)'*. The latter is, e.g., used in the utterance *'there were millions of flies'* where the exact number is undefined. Wiktionary lacks this word sense.

For the term *'people'* ($f = 27,425$), both resources share the word senses *'any group of human beings'*, *'the body of citizens of a state or country'*, *'the members of a family line'*, and *'the commonality'*. Wiktionary additionally contains the word senses *'a group of persons regarded as being employees, followers, companions or subjects of a ruler'* and *'one's colleagues or employees'* that are variations of the shared word senses listed above, but not explicitly covered by WordNet. Consider the example sentences *'the new boss met his people'* and *'some people followed Hitler till his end'*. The word sense *'any group of human beings'* would be very general here, as there are no restrictions on the type of human beings, although the former addresses only employees and the latter a group of followers. It is, however, arguable if the distinction between *'one's colleagues or employees'* and *'a group of persons regarded as [..]'* is necessary (and clearly separable) as well as if the latter word sense should be split into 4 individual word senses concerning *employees*, *followers*, *companions of a ruler*, and *subjects of a ruler*.

A similar observation can be made for the verb *'(to) be'* ($f = 102,079$) that has 19 word senses in Wiktionary and 13 in WordNet. We found 6 of them denoting the same or highly similar meanings. Wiktionary distinguishes 5 word senses with a definition *'used to indicate age/height/time/weather/temperature'*, and example sentences, like *'He looks twelve, but is actually thirteen [..]'*. Although these word senses could be assigned to a more general word sense, like *'have the quality of being; (copula, used with an adjective or a predicate noun)'*, there is a slight difference between indicating an age (*He is thirteen*) and a copula for predicate nouns (*He is a boy*). The former would need a word sense *'a 13-year-old person'* for the term *thirteen* (and all other numerals) in order to be able to fully interpret the meaning of the sentence. Amongst the remaining word senses, 4 Wiktionary senses are syntactic usages of *'be'*, like *'used to form the passive voice'*. This kind of knowledge is usually found in grammatical resources rather than lexical semantic resources, as it does not define a word's meaning in the narrower sense. WordNet, on the other hand, contains the word senses *'have life, be alive'*, *'to remain unmolested, undisturbed, or uninterrupted [..]'* and *'be priced at'* that are missing from Wiktionary. Especially the last one is very similar to the *'used to indicate age'* word sense and thus underlines that none of the resources is fully comprehensive.

We observed a rather low overlap of the word senses of highly frequent lexemes. Both resources encode relevant word senses that are missing from the other resource. Wiktionary contains slightly more word senses that are mainly variations of WordNet's senses, but also include grammatical knowledge that is not traditionally encoded in dictionaries.

Polysemic Difference

In section 3, we have studied the polysemic difference, i.e. the difference $\Delta(\ell)$ in the number of word senses encoded in both resources for lexeme ℓ. We now study the word sense definitions for lexemes with different values of $\Delta(\ell)$.

The noun *'order'* has 14 word senses in each resource and thus a $\Delta(\ell)$ of 0. We found, however, only 6 of them to be similar or equal in both resources. Among the Wiktionary word senses, there are many technical or domain-specific definitions from mathematics (number of elements in a set), graph theory (number of vertices), order theory (partial order), electronics (polynomial order of a circuit block), chemistry (polynomial order of a rate law), and cricket (batting order). None of them has a counterpart in WordNet. Accordingly, there are word senses in WordNet that are not covered by Wiktionary. These are *'a degree in a continuum of size or quantity'* (order of magnitude), *'established customary state'* (law and order), *'a legally binding command or decision entered on the court record'*, *'a body of rules followed by an assembly'*, and *'(architecture) one of original three styles of Greek architecture [..]'*. We observe that the Wiktionary word senses of this example focus on natural sciences and technology, while WordNet covers more terms from law and humanities. The noun *'resident'* also shows $\Delta(\ell) = 0$, but encodes only 2 word senses in either resource. These word senses are identical – both the *'individual living at a location/area'* and the *'medical student assisting in a hospital'* are listed as distinct word senses. It is, however, noticeable that the Wiktionary word sense is more general, as the first word sense allows persons, animals, and plants as individuals, while WordNet restricts the word sense to persons. It is thus unable to cover the meaning of *'resident'* in the sentence *'the resident frog in our pond has died'*.

One of the highest polysemic differences was found for the noun *'stick'* ($\Delta(\ell) = 36$), which encodes 45 word senses in Wiktionary and 9 word senses in WordNet. It is surprising that the WordNet word sense *'informal terms for the leg'* has no direct counterpart amongst the 45 Wiktionary word senses. The remaining 8 word senses are, however, covered by Wiktionary. Most of the additional Wiktionary word senses are either related to slang or fairly domain-specific.

We conclude by finding that Wiktionary additionally encodes domain-specific word senses that are generally not found in WordNet. This is especially the case for lexemes with a high polysemic difference. We also found some evi-

dence that Wiktionary is more focused on word senses from natural sciences rather than social sciences or humanities, which have been predominantly found in WordNet. Such an observation could be explained by the current composition of Wiktionary's community, which might be rather technophile. We plan to further analyze this observation.

5. CONCLUSION

In our study, we analyzed collaboratively defined word senses from the collaborative language resource Wiktionary and compared them with expert defined word senses from WordNet, a standard resource for word senses in the field of Computational Linguistics.

We first aligned the resources at the level of lexemes and measured the overlap of the resources, i.e. the number of lexemes that are found in both resources. The overlap is surprisingly low, which means that collaborative resources contain many terms (and thus word senses) that are not found in traditional expert built resources and vice-versa. For the lexemes covered by both resources, we examined the encoded word senses. The word sense distribution in Wiktionary and WordNet is very similar, although there are on average more word senses for verbs in WordNet and a higher maximum number of word senses for nouns in Wiktionary. We measured the polysemic difference, i.e. the difference in the number of word senses for each lexeme, and found a difference of less or equal than 2 for 95% of the encoded lexemes. 60% even encoded the same number of word senses.

Finally, we selected several representative examples and compared the encoded word senses. We found that Wiktionary encodes additional word senses for seldomly used terms and has a better coverage of slang-related and domain-specific word senses. We especially found word senses from natural sciences, sports, and military that are badly covered by WordNet. Word senses from social sciences and humanities are, in contrast, better covered by WordNet within our example terms. While Wiktionary and WordNet share many word senses for terms with a medium language frequency, Wiktionary encodes a large number of word senses for terms with a high frequency or a high polysemic difference, which are either missing from WordNet or are variants of the encoded senses there. We argue that collaborative word sense inventories have a great potential and further analysis of these resources should be performed.

We particularly plan to study the resources by means of Social Network Analysis [17] and perform an automatic word sense alignment in order to combine expert and collaborative resources in the future.

6. ACKNOWLEDGMENTS

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806. We thank Elisabeth Wolf and Dr. György Szarvas for their helpful comments.

7. REFERENCES

[1] E. Agirre, L. Marquez, and R. Wicentowski, editors. *Proceedings of the 4th International Workshop on Semantic Evaluations*, Prague, Czech Republic, 2007.

[2] D. A. Cruse. *Lexical Semantics*. Cambridge Textbooks in Linguistics. Cambridge, UK: Cambridge University Press, 1986.

[3] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. Cambridge, MA: MIT Press, 1998.

[4] A. Gangemi, N. Guarino, and A. Oltramari. Conceptual Analysis of Lexical Taxonomies: The Case of WordNet Top-Level. In *Proceedings of the International Conference on Formal Ontology in Information Systems*, pages 285–296, Ogunquit, ME, USA, 2001.

[5] I. Gurevych and T. Zesch, editors. *Proceedings of the ACL '09 Workshop 'The People's Web Meets NLP: Collaboratively Constructed Semantic Resources'*, Suntec, Singapore, 2009.

[6] P. Hanks. Linguistic Norms and Pragmatic Exploitations, or Why Lexicographers need Prototype Theory, and Vice Versa. In *Papers in Computational Lexicography*, pages 89–113, Budapest, Hungary, 1994.

[7] N. Ide and Y. Wilks. Making Sense about Sense. In *Word Sense Disambiguation: Algorithms and Applications*, volume 33 of *Text, Speech and Language Technology*, pages 47–74. 2007.

[8] A. Kilgarriff. "I Don't Believe in Word Senses". *Computers and the Humanities*, 31(2):91–113, 1997.

[9] A. Kilgarriff. Word Senses. In *Word Sense Disambiguation: Algorithms and Applications*, volume 33 of *Text, Speech and Language Technology*, pages 29–46. 2007.

[10] A. Kilgarriff and G. Grefenstette. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, 29(3):333–347, 2003.

[11] C. M. Meyer and I. Gurevych. Worth its Weight in Gold or Yet Another Resource — A Comparative Study of Wiktionary, OpenThesaurus and GermaNet. In *Proceedings of the 11th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 38–49. Iaşi, Romania, 2010.

[12] G. A. Miller and F. Hristea. WordNet Nouns: Classes and Instances. *Computational Linguistics*, 32(1):1–3, 2006.

[13] R. Navigli. Word Sense Disambiguation: A Survey. *ACM Computing Surveys*, 41(2):1–69, 2009.

[14] H. Putnam. *Reason, Truth, and History*. Cambridge: Cambridge University Press, 1981.

[15] U. Quasthoff, M. Richter, and C. Biemann. Corpus Portal for Search in Monolingual Corpora. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 1799–1802, Genoa, Italy, 2006.

[16] R. Rapp. Word Sense Discovery Based on Sense Descriptor Dissimilarity. In *Proceedings of the Ninth Machine Translation Summit*, pages 315–322, New Orleans, LA, USA, 2003.

[17] J. Scott. *Social Network Analysis: A Handbook*. London, UK: Sage Publications, 2nd edition, 2000.

[18] J. Surowiecki. *The Wisdom of Crowds*. New York, NY: Anchor Books, 2005.

[19] T. Zesch, C. Müller, and I. Gurevych. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 1646–1652, Marrakech, Morocco, 2008.