Worth Its Weight in Gold or Yet Another Resource — A Comparative Study of Wiktionary, OpenThesaurus and GermaNet

Christian M. Meyer and Iryna Gurevych

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

Abstract. In this paper, we analyze the topology and the content of a range of lexical semantic resources for the German language constructed either in a controlled (GermaNet), semi-controlled (OpenThesaurus), or collaborative, i.e. community-based, manner (Wiktionary). For the first time, the comparison of the corresponding resources is performed at the word sense level. For this purpose, the word senses of terms are automatically disambiguated in Wiktionary and the content of all resources is converted to a uniform representation. We show that the resources' topology is well comparable as they share the small world property and contain a comparable number of entries, although differences in their connectivity exist. Our study of content related properties reveals that the German Wiktionary has a different distribution of word senses and contains more polysemous entries than both other resources. We identify that each resource contains the highest number of a particular type of semantic relation. We finally increase the number of relations in Wiktionary by considering symmetric and inverse relations that have been found to be usually absent in this resource.

1 Introduction

Large-scale acquisition of lexical semantic knowledge from unstructured corpora has become a hot research topic, since numerous natural language processing tasks like semantic search, automatic word sense disambiguation or calculating semantic relatedness require large lexical semantic resources as a source of background knowledge. *Expert-built lexical semantic resources* (ELSR) like WordNet [1] or GermaNet [2] are hand-crafted in a controlled manner by linguists and have been extensively used for such applications. Keeping ELSRs up-to-date is however a costly and time-consuming process, which leads to limited coverage and thus insufficiency for obtaining high quality results in above tasks. Especially for languages other than English, ELSRs suffer from their small size.

With the evolution of the socio-semantic web, a new type of resources has emerged: *collaboratively constructed lexical semantic resources* (CLSR) like Wikipedia or Wiktionary, which are created by a community of (mainly) non-experts

A. Gelbukh (Ed.): CICLing 2010, LNCS 6008, pp. 38–49, 2010.

on a voluntary basis. As CLSRs are constantly updated by their community, they benefit from the wisdom of crowds and avoid the costly maintenance process of ELSRs. Zesch et al. [3] found that Wiktionary outperforms ELSRs when used as a source of background knowledge for calculating semantic relatedness.

Our assumption is that a combination of ELSRs and CLSRs would lead to better results, since it profits from the high quality of ELSRs and the broad coverage of CLSRs. The structural and content related properties of the latter are however largely unknown. We therefore perform a comparative study of Wiktionary, GermaNet and OpenThesaurus, in order to learn about their content as well as the individual strengths and weaknesses.[1]

Previous studies regarded Wiktionary's lexical semantic relations at the term level, although they are generally marked with a certain word sense. For the first time, we analyze them at the word sense level, whereby an automatic word sense disambiguation algorithm is applied to relations without sense marker.

2 Description of Lexical Semantic Resources

We have chosen the three resources Wiktionary, GermaNet and OpenThesaurus for our study, because they cover well the range between ELSR and CLSR: GermaNet is fully expert-created, while Wiktionary is clearly a CLSR with a large community of volunteers. It is not controlled by an editorial board. OpenThesaurus is in between, as it is collaboratively constructed but has a much smaller community and is reviewed and maintained by an administrator [4]. In the following, we describe each individual resource and their representational units.

Our notation mainly follows [5]. A *term* is a word form that is characterized by a certain string, e.g., *bass* or *read*.[2] A *lexeme* is a term that is tagged with its part of speech, e.g., *bass (noun)* or *read (verb)*. Each lexeme can be used in one or more *word senses* that carry the meaning of a lexeme. For the lexeme *bass (noun)* there could e.g. be two word senses *bass*$\langle$*music*$\rangle$ and *bass*$\langle$*fish*$\rangle$. Note that in this definition, a word sense is bound to a certain lexeme rather than representing a concept. The latter will be called a *synset* (short for synonymy set) that combines word senses with the same meaning but represented by different lexemes. The set {*bass*$\langle$*fish*$\rangle$*, perch, Percidae*} is e.g. a synset for the meaning '*any of various marine and freshwater fish resembling the perch, all within the order of Perciformes*' that consists of three synonymous word senses. We use the notation $s \in S$ to indicate that word sense s is included in the synset S.

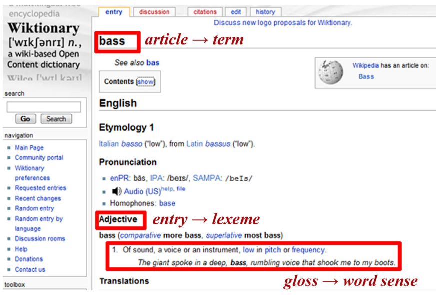
A *relation* is a pair (*source, target*), where *source* and *target* denote word senses that the relation connects. Relations are directed from source to target and have a certain relation type [5]. The term *bass* has e.g. a synonymy relation

[1] Although we focus on German resources, our methods are not language dependent and can also be applied to similar resources in other languages. Particularly, we conducted a study of the English Wiktionary, WordNet and Roget's Thesaurus and report our results at: `http://www.ukp.tu-darmstadt.de/data/lexical-resources`

[2] We provide English examples where possible to improve the understandability of the paper and choose words with similar ambiguities rather than translating literally.

(*bass⟨fish⟩, perch*) and a hypernymy relation (*bass⟨fish⟩, Perciformes*). For relations of type *synonymy* and *antonymy*, there can be a *symmetric* relation of the same type that connects the target with the source. Relations of the types *hypernymy*, *hyponymy*, *holonymy* and *meronymy* have however no symmetric but inverse relations that connect the target with the source. Instances of inverse relations are hypernymy–hyponymy and holonymy–meronymy. For the synonymy relation (*bass⟨fish⟩, perch*), there is e.g. a symmetric relation (*perch, bass⟨fish⟩*), while the hypernymy relation (*bass⟨fish⟩, Perciformes*) can have the inverse hyponymy relation (*Perciformes, bass⟨fish⟩*). A relation whose symmetric or inverse counterpart does not exist in a given resource will be called a *one-way relation*, otherwise a *two-way relation.*

Wiktionary[3] is a large online dictionary that is collaboratively constructed by a community. The resource is organized in article pages that represent a certain term and can consist of multiple lexemes. Each lexeme is tagged with its language and part of speech and can distinguish different word senses, which are represented by glosses. Figure 1 shows the Wiktionary entry *bass* as an example of this structure. Semantic relations are encoded as links to other articles. Wiktionary is available in more than 300 languages. Each language edition contains word entries from multiple languages. An entry about the English term *railway* can e.g. be found in both the German and the English Wiktionary. For our study, we focus solely on the German word entries in the German language version, which are parsed and accessed using the freely available Java-based Wiktionary Library[4] [6] and a Wiktionary dump of June 18, 2009.



bass

English

Etymology 1

Italian *basso* ("low"), from Latin *bassus* ("low").

Pronunciation

- enPR: bās, IPA: /beɪs/, SAMPA: /beIs/
- Audio (US)help, file
- Homophones: base

Adjective

bass (*comparative* more bass, *superlative* most bass)

1. Of sound, a voice or an instrument, low in pitch or frequency.
   The giant spoke in a deep, **bass**, *rumbling voice that shook me to my boots.*

Translations

Fig. 1. Wiktionary article *bass* with highlighted term, lexeme and word sense sections

GermaNet[5] [2] is an ELSR for the German language that is similar to the well-known Princeton WordNet [1]. GermaNet consists of a set of synsets that contain one or more word senses. While lexical relations such as antonymy are defined between lexemes, taxonomic relations like hypernymy can only exist between synsets. We use GermaNet 5.0 that is available upon a license.

[3] http://www.wiktionary.org
[4] http://www.ukp.tu-darmstadt.de/software/jwktl
[5] http://www.sfs.uni-tuebingen.de/lsd

OpenThesaurus[6] [4] is a thesaurus for the German language. Its main focus is collecting synonyms, but also some taxonomic relations can be found in the resource. OpenThesaurus consists of a list of *meanings* (synsets) that can be represented by one or more *words* (terms). The resource is released as a full database dump from the project homepage. We use a dump of July 27, 2009.

3 Related Work

To our knowledge, there is no other comparative study of the three resources Wiktionary, GermaNet and OpenThesaurus that analyzes both topological and content related properties. The latter issue has been addressed for single resources, but without any comparison [6,2,4]. Garoufi et al. [7] compared the topological properties of Wiktionary with GermaNet and both the Wikipedia category and article graphs. They however do not convert the resources into a uniform representation. Topological properties are also analyzed by Navarro et al. [8], who built a graph of synonymy links from the French, English, German and Polish Wiktionaries. They found similar properties for the different language versions. Both the studies regard Wiktionary relations between terms rather than word senses. The two hypernymy relations (*smallmouth bass, bass*⟨*fish*⟩) and (*bass*⟨*music*⟩*, pitch*) then share the vertex *bass*, which leads to a path length of only 2 between *smallmouth bass* and *pitch*. This is different from ELSRs like WordNet or GermaNet that encode such relations between word senses or synsets and may result in a biased comparison of the resources. We solve this problem by applying automatic word sense disambiguation to the Wiktionary relations.

4 Representing Lexical Semantic Resources as Graphs

In order to allow a systematic and fair comparison, all resources need to be converted into a uniform representation. We therefore introduce a directed graph $G = (V, E)$ of all word senses V and the corresponding set of relations $E \subseteq V^2$. Each resource has however its unique representation and thus requires an individual approach to the graph construction described below.

Wiktionary. The source of a Wiktionary relation is usually associated with a certain word sense. The syntax *[2] fish* within the article *bass*, e.g., indicates that the second sense of *bass* (the fish within the order of Perciformes) is the source of a (hypernymy) relation to the target term *fish*. Unfortunately, the target of a relation is not sense disambiguated in general, as it is only given by a link to a certain article. For the term *fish* in the relation above, it is not clear whether the maritime animal, a part of a ship's mast or a card game is meant. Automatic word sense disambiguation is required to determine the correct sense of the target. To our knowledge, this issue has not been addressed in any of the works based on Wiktionary.

[6] `http://www.openthesaurus.de`

Let (u, v) be a Wiktionary relation with the source word sense u and a target term v. We first determine the set of candidate word senses, i.e. all word senses that are defined for term v. Then, the semantic relatedness between the source and each candidate is calculated, based on the sense gloss and usage examples that will be called *extended gloss* in the following. The candidate with the highest score is chosen as the relation target. Figure 2 outlines this approach formally.

```
function RelationTargetWSD($(u, v)$)
    $g1$ := gloss($u$) + examples($u$);
    $Candidates$ := {};
    $score : Candidates \longrightarrow \mathbb{R}$;
    for each Wiktionary word sense $c$ of term $v$ do
        $Candidates$ := $Candidates \cup \{c\}$;
        $g2$ := gloss($c$) + examples($c$);
        $score$(c) := calcESA($g1, g2$);
    end;
    return $\arg\max_{c \in \text{Candidates}} score(\text{c})$;
end.
```

Fig. 2. Automatic word sense disambiguation method for Wiktionary's relation targets

The semantic relatedness is computed using Explicit Semantic Analysis based on Wikipedia, which has been introduced to be capable of solving word sense disambiguation tasks [9]. It forms a vector space from all Wikipedia articles and creates a concept vector c for two input terms consisting of the *tfidf* scores [10] between the term and each Wikipedia article. The cosine of the concept vectors is then calculated as their semantic relatedness. Since we need to compare extended glosses, i.e. short texts, rather than single words, we use an extension of this method [3]: The concept vectors $c(t)$ of all non-stopword tokens $t \in g$ of the extended gloss g are calculated with the above method and combined by computing the normalized sum of the vectors, leading to:

$$\text{calcESA}(g1, g2) = \frac{c(g1) \cdot c(g2)}{|c(g1)| \cdot |c(g2)|} \quad \text{with} \quad c(g) = \frac{1}{|g|} \sum_{t \in g} c(t)$$

Consider e.g. the hypernymy relation (*bass⟨fish⟩, fish*). There are three target candidates for the term v = *fish* with relatedness scores: *score*(*fish⟨maritime animal⟩*) = .35, *score*(*fish⟨part of a mast⟩*)= .13 and *score*(*fish⟨card game⟩*) = .16. The word sense with the maximum *score* is chosen, which is *fish⟨maritime animal⟩* in this case.

To evaluate this approach, we annotated 250 randomly sampled Wiktionary relations by marking each of the 920 possible target candidates with either + if the specified relation (u, v) holds, or with − otherwise. The annotators were allowed to assign multiple target senses of a relation with + if more than one relation holds, whereas also no + was possible. There is e.g. a hyponymy relation (*Antwerp⟨Province⟩, Essen*) about a Belgian municipality whose target has only the three word sense candidates *nutrition*, *meal* and *German city*, so none of them was selected.[7] The annotations were created independently by two annotators, who are both German native speakers. Table 1 shows the number of target candidates both annotators agreed on, namely $(+,+)$ and $(-,-)$, as well as the number of candidates that the annotators did not agree on: $(+,-)$ and $(-,+)$. We report these numbers separately for each level of ambiguity a, i.e. the number of possible targets for a given relation and note the relation count r_a for each level. There are e.g. $r_a = 2$ relations that both have $a = 15$ target candidates, of which 1 was considered correct and 27 incorrect by both annotators, while they disagreed on 2 of them. We observe a uniform disagreement D_a at each level of ambiguity, although it is slightly higher for $a = 4$ and $a = 10$.

Table 1. Agreement table for the word sense disambiguation of Wiktionary relations

a	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Sum
r_a	103	59	27	24	8	9	7	5	3	1	1	0	1	2	250
$(+,+)$	90	50	23	23	7	9	6	5	1	1	2	0	1	1	219
$(-,+)$	14	8	5	2	2	2	0	1	0	1	0	0	0	1	36
$(+,-)$	19	5	13	7	2	6	5	2	6	0	1	0	1	1	68
$(-,-)$	83	114	67	88	37	46	45	37	23	9	9	0	12	27	597
D_a	.16	.07	.17	.08	.08	.13	.09	.07	.20	.09	.08	.00	.07	.07	

We measured the inter-annotator agreement following the methods introduced in [11] to ensure the reliability of our data. Therefore, we first considered the data as 920 binary annotations that judge if a target candidate is valid for a given source word sense and measured an observed agreement $A_\mathrm{O} = .88$ and a chance-corrected agreement of $\kappa = .72$, which allows tentative conclusions [11]. We then interpreted our data as 250 set-valued annotations that provide a set of valid target word senses for a given relation. For measuring the agreement of set-valued data, we used MASI [12] as a distance function for Krippendorff's α, which resulted in $\alpha = .86$ and indicates good agreement. We refrained from removing items where no agreement was reached, since these are the most difficult instances whose removal would lead to biased results. We rather measured the agreement between our algorithm M and both the human annotators A and B. Besides the inter-annotator agreement A–B, which serves as an upper bound, we tried the naïve baseline approach *0* that always chooses the first target word sense. Table 2 summarizes the evaluation results. Our approach exceeds the baseline in each case. There is however room for improvements with respect to the upper bound A–B. We plan to compare several approaches in our future work.

The algorithm exclusively relies on the semantic relatedness of the word senses' extended glosses. Thus, the disambiguation is likely to fail if only a short or very general gloss is given, which has been found to be the most common source of errors. Besides cases, where the community did not provide a meaningful gloss, there are also minor errors in the extraction API that lead to truncated

[7] On June 27, 2009 the missing word sense has been added to the article *Essen*.

Table 2. Evaluation results of our word sense disambiguation approach

	0–A	*0–B*	*M–A*	*M–B*	*A–B*
A_O	.791	.780	.820	.791	.886
κ	.498	.452	.567	.480	.728
α	.679	.620	.726	.649	.866

glosses. Other errors are caused by references to other word senses within a gloss; the second sense of *tomato*, e.g., refers to its first sense: *[2] the fruit of [1].*

GermaNet and OpenThesaurus. To obtain a uniform representation of the resources, the synsets in GermaNet and OpenThesaurus need to be decomposed into the individual word senses. We therefore add a node to V for each word sense $s \in S$ of any synset S. Accordingly, an edge (s_1, s_2) is added to E for each word sense $s_1 \in S_1$ and $s_2 \in S_2$ of a relation (S_1, S_2) between synsets. As synsets represent sets of synonyms, we also add a synonymy edge (s_1, s_2) for all $s_1, s_2 \in S$, which results in a fully connected subgraph for each synset. Consider e.g. the synset {*bass⟨fish⟩, perch, Perciformes*}. The three word senses are added to V and the synonymy edges (*bass⟨fish⟩, perch*), (*bass⟨fish⟩, Perciformes*) and (*perch, Perciformes*) as well as their symmetric counterparts are added to E.

5 Topological Analysis of Resources

We now use this uniform representation of the resources and study topological properties of their graphs. Table 3 shows the results of our study for Wiktionary (WKT), GermaNet (GN) and OpenThesaurus (OT).

For applications that aim to calculate semantic relatedness using a lexical semantic resource, it is often crucial that the resource graph is connected. As none of the resources is connected as a whole, we studied the number of connected components CC, the largest ($lcc1$) and the second largest ($lcc2$) connected components. GermaNet was found to contain the fewest connected components, only about 2% of the respective number in Wiktionary and OpenThesaurus. 98% of all vertices are within $lcc1$ in GermaNet, thus allowing to use almost the whole

Table 3. Comparison of topological properties

	$\lvert V\rvert$	$\lvert E\rvert$	CC	$\lvert V_{\mathrm{lcc1}}\rvert$	$\lvert E_{\mathrm{lcc1}}\rvert$	$\lvert V_{\mathrm{lcc2}}\rvert$	$\lvert E_{\mathrm{lcc2}}\rvert$
WKT	107,403	157,786	20,114	80,638	149,947	69	68
GN	76,864	394,856	471	75,848	393,072	49	149
OT	87,735	288,121	26,624	12,742	48,590	704	4,078

	γ_{lcc1}	R^2	ℓ_{lcc1}	ℓ_{rand}	c_{lcc1}	c_{rand}	o_{lcc1}	o_{rand}
WKT	-2.37	96.2%	1.3	8.5	0.13	<0.01	0.59	0.32
GN	-1.71	75.9%	10.8	9.1	0.24	<0.01	0.41	0.11
OT	-1.91	63.4%	<0.01	4.8	0.26	<0.01	0.48	0.15

resource for applications that require connected graphs. For Wiktionary, 75% of the vertices are in $lcc1$, which leads to a similar number of nodes compared to GermaNet — the difference in $|V_{\text{lcc1}}|$ is merely 4,950. In OpenThesaurus, only 14% of the vertices are contained in $lcc1$, which makes it less useful for such tasks. We also analyzed $lcc2$, as it reveals if the remaining graph forms a usable semantic network itself or only consists of mainly unconnected vertices. Each resource showed a very small $lcc2$, both in the number of vertices and edges. It is thus sufficient to focus on the $lcc1$ as it contains the bulk of semantic information.

Albert and Barabási [13] studied the topology of several real world graphs and found governing organizational principles that significantly differ from those in random graphs. We applied their experimental approaches to our resource based graphs. The *degree distribution* of graph G is a function $D\colon \mathbb{N} \longrightarrow \mathbb{N}$ that maps each possible degree to its number of occurrences: $d \mapsto |\{v \in V \mid \deg(v) = d\}|$. While the function follows a normal distribution for random graphs, it shows a *power law distribution* for many real world graphs, which results from the way a graph grows over time and its organizational structures [14]. Such graphs are called scale-free, since their topology remains stable, regardless of their size. For a power law, the probability of each node $v \in V$ to have degree k is proportional to the γ-th power of k:

$$P(\deg(v) = k) \propto k^{-\gamma}$$

Garoufi et al. [7] studied if the degree distribution of Wiktionary and GermaNet follows a power law but did not provide any goodness-of-fit analysis to evaluate the quality of the fitted parameter γ. We use the coefficient of determination R^2 [15] for this purpose. The nearer R^2 is to 100%, the stronger is the evidence for a power law. In our setting, the Wiktionary graph shows a clear power law and can be considered scale-free, which was previously reported in [7,8]. For both other resources, R^2 is considerably lower. This is a surprising observation, since [7] found a power law in the degree distribution of the GermaNet graph. One explanation could be that their observed power law is not significant, as no goodness-of-fit analysis is provided. Another possibility is that our uniform representation of resources leads to different results. Further analyses need to be applied to learn about GermaNet's degree distribution. While the scale-free Wiktionary graph allows to project our topological insights to future (larger) versions of Wiktionary, this does not necessarily hold for GermaNet and OpenThesaurus.

Real world graphs tend to show a *small world property* [13]. Such graphs usually have a small average path length ℓ over each node pair $(u, v) \in V^2$. Besides that, they have a high fraction of transitive triplets, which can be measured by the clustering coefficient c, i.e. the average probability that two neighbors of a node are connected by an edge [16]. Both measures are required to clearly differ from the corresponding values of a random graph with similar vertex and edge count [13]. Table 3 contains the two measures for the resource's largest connected component (ℓ_{lcc1} and c_{lcc1}) together with the corresponding results of a random graph (ℓ_{rand} and c_{rand}). The clustering coefficient differs about an order of magnitude from a corresponding random graph. In Wiktionary and OpenThesaurus,

ℓ_{lcc1} is clearly lower than in the corresponding random graph. The small world property is thus clearly visible for these two resources.

The average path length of the GermaNet graph is slightly higher than ℓ_{rand}, which can also be seen in [7]. Especially terms from different parts of speech contribute path lengths of up to 39, which is the diameter of the graph. An average path length of 10.8 is still low for a graph of this size, we however aimed at comprehensibly verifying the existence or absence of the small world property. We therefore calculated the *topological overlap* o_{lcc1} for each resource graph as a third topological measure and compared it to the o_{rand} of the corresponding random graph. The topological overlap is the average $o(u, v)$ for each pair $(u, v) \in V^2$, which measures the number of vertices to which both u and v are linked. A high topological overlap characterizes hierarchical and small world graphs [17]. Our results in Table 3 show a considerably higher o_{lcc1} for the resource graphs compared to o_{rand} — in particular for the GermaNet graph, which hence reveals also a small world property for this resource.

Comparing lexical semantic resources requires a similar topology of their induced graphs. The small world property is a good indicator for that. It not only allows a fair and unbiased comparison but also promises that a combination of the resources is governed by the same structures that they show individually.

6 Content Analysis of Resources

After studying the resource topology, we focused on their content and examined the number of lexemes, word senses and relations. Table 4 shows the determined results. Each of the three resources contains a comparable number of lexemes and word senses. Wiktionary is however the largest resource with 23,857 lexemes more than OpenThesaurus, which is the smallest. GermaNet on the contrary contains the highest number of relations, 2.5 times more than Wiktionary and 1.3 times more than OpenThesaurus. This makes GermaNet the most densely connected resource. Wiktionary encodes a distinction between polysemy and homonymy: The former is expressed in word senses, while the latter is represented by different lexemes that arise from different etymology. None of the other resources explicitly encodes this type of information.

The target of a Wiktionary relation is represented by a link to a certain article, which is sometimes yet missing due to the collaborative construction approach. Therefore, a large number of relations exist whose targets are fairly rare terms still not encoded in the resource in the form of a dedicated Wiktionary entry. The article *bass* e.g. links to an article *bass music*, which has not yet been created by the community. We will refer to such relation targets as *dangling lexemes*. 56% of the lexemes in Wiktionary are dangling, thus showing that the resource contains many gaps. As Wiktionary is constantly growing by 1–2% of its size each month,[8] these gaps are however likely to be filled in the future and yield a lexical semantic resource with high coverage.

[8] `http://stats.wikimedia.org/wiktionary/EN/TablesWikipediaDE.htm`

Table 4. Descriptive statistics about the resources' content

	WKT	**GN**	**OT**
Number of lexemes:	90,611	67,402	66,754
…Homonyms:	2,327	-	-
…Monosemous:	29,025	61,129	54,939
…Polysemous:	10,643	6,273	11,815
…Dangling lexemes:	50,943	0	0
Number of word senses:	107,403	76,864	87,735
Number of relations:	157,786	394,856	288,121
…One-way:	139,453	11,941	5,731
…Synonymy:	62,235	69,097	282,390
…Antonymy:	24,167	3,486	0
…Hypernymy:	37,569	155,385	5,731
…Hyponymy:	33,815	155,237	0
…Holonymy:	0	8,977	0
…Meronymy:	0	2,674	0
Number of two-way relations:	297,120	406,328	293,846
…Two-way synonymy:	117,318	69,134	282,384
…Two-way antonymy:	43,128	3,134	0
…Two-way hypernymy:	136,674	310,856	11,462
…Two-way holonymy:	0	23,204	0

9% of the lexemes in GermaNet and 17% of the lexemes in OpenThesaurus are polysemous, i.e. at least two word senses are encoded for a lexeme. Wiktionary however contains 26% polysemous lexemes, which is significantly higher than in both other resources. Different explanations are possible for this observation: Either Wiktionary contains mainly high-frequency words that are known to be more ambiguous, or the community more likely creates articles for polysemous terms, since they might be more interesting to create. Besides that, it is also possible that the coverage of senses for a lexeme is on average higher within Wiktionary, or that the Wiktionary word senses are more fine-grained than those of the other resources. This remains to be thoroughly studied in the future.

GermaNet is the only resource that contains holonymy and meronymy relations, while its number of hypernymy and hyponymy relations is also higher than in the other resources. OpenThesaurus contains the most synonyms as it was the major goal for its creation. It yet contains less hypernyms and neither antonymy nor hyponymy relations. Wiktionary shows the most antonyms and contains nearly as many synonyms as GermaNet. At first glance, Wiktionary seems to have less relations than GermaNet and OpenThesaurus. Especially the difference to GermaNet is very prominent. Further examination however shows that 88% of the Wiktionary relations are one-way relations. GermaNet and OpenThesaurus have only between 2–3% one-way relations, which can be explained by their creation guidelines. Since synonymy and antonymy relations are symmetric and taxonomic relations are invertible, the number of relations can be increased by generating the corresponding counterparts, thus converting each relation to a two-way relation. The results of this extension are included in Table 4 (hyponymy and meronymy are equal to their inverse counterpart and therefore omitted). Wiktionary benefits most from the extension and finally contains slightly more relations than OpenThesaurus. It still is the resource with the most antonyms and the second most synonymy and hypernymy relations.

7 Conclusions and Future Work

We analyzed the topological and content related properties of Wiktionary and compared them with GermaNet and OpenThesaurus. We have chosen the three resources, since they represent well the range between expert-built and collaboratively constructed lexical semantic resources. For the first time, we provide an analysis of lexical semantic relations in Wiktionary based on word senses. We applied word sense disambiguation to the relation targets in order to find the correct word sense of the relation target. We also transformed the synsets within GermaNet and OpenThesaurus into a set of synonymous word senses for each contained term, which allows a uniform representation of the three resources and thus a fair comparison of their encoded information. This setting is unique and has not been reported before to our knowledge.

In the first part of our analysis, we created a word sense based graph for each resource and studied the graph topology. All graphs showed the small world property, which is important for being able to compare the analysis results. The Wiktionary graph is additionally scale-free and thus allows to project our observations to future (larger) Wiktionary versions. Studying content related properties revealed that although Wiktionary contains the lowest number of relations it has the highest number of word senses. It however contains lots of dangling word senses, i.e. word senses that are used as targets of semantic relations but are not yet described in an article. The number of Wiktionary's lexical semantic relations has been greatly increased by considering also the symmetric and inverse counterpart of each relation if not directly encoded in the resource. While GermaNet provides the highest number of taxonomic relations and OpenThesaurus the highest number of synonyms, Wiktionary contains the most antonyms and the second most synonymy and hypernymy relations.

Our future work will focus on an enhanced automatic word sense disambiguation of Wiktionary's relation targets in order to compare different approaches and give a comprehensive evaluation of our method. We also plan to study the information overlap of the resources in order to learn if the resources share a large common vocabulary or contain complementary information. Besides that, we aim at analyzing English resources in a similar manner.

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 Torsten Zesch from the UKP Lab for their contributions to this paper and Dr. Lothar Lemnitzer from BBAW for his helpful comments.

References

1. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database (Language, Speech, and Communication). MIT Press, Cambridge (1998)
2. Kunze, C., Lemnitzer, L.: GermaNet — representation, visualization, application. In: Proceedings of the Third International Conference on Language Resources and Evaluation, Las Palmas, Canary Islands, Spain, vol. 5, pp. 1485–1491 (2002)
3. Zesch, T., Müller, C., Gurevych, I.: Using Wiktionary for Computing Semantic Relatedness. In: Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, Chicago, IL, USA, pp. 861–867 (2008)
4. Naber, D.: OpenThesaurus: ein offenes deutsches Wortnetz. In: Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen: Beiträge zur GLDV-Tagung, Bonn, Germany, pp. 422–433 (2005)
5. Jurafsky, D., Martin, J.H.: Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall Series in Artificial Intelligence. Prentice Hall, Upper Saddle River (2000)
6. Zesch, T., Müller, C., Gurevych, I.: Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In: Proceedings of the 6th International Conference on Language Resources and Evaluation, Marrakech, Morocco, pp. 1646–1652 (2008)
7. Garoufi, K., Zesch, T., Gurevych, I.: Graph-Theoretic Analysis of Collaborative Knowledge Bases in Natural Language Processing. In: Proceedings of the Poster Session of the 7th International Semantic Web Conference, Karlsruhe, Germany (2008)
8. Navarro, E., Sajous, F., Gaume, B., Prévot, L., Hsieh, S., Kuo, I., Magistry, P., Huang, C.R.: Wiktionary for natural language processing: methodology and limitations. In: Proceedings of the ACL 2009 Workshop, The People's Web Meets NLP: Collaboratively Constructed Semantic Resources, Suntec, Singapore, pp. 19–27 (2009)
9. Gabrilovich, E., Markovitch, S.: Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, pp. 1606–1611 (2007)
10. Salton, G., McGill, M.J.: Introduction to modern information retrieval. McGraw-Hill, New York (1983)
11. Artstein, R., Poesio, M.: Inter-Coder Agreement for Computational Linguistics. Computational Linguistics 34(4), 555–596 (2008)
12. Passonneau, R.J.: Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa, Italy, pp. 831–836 (2006)
13. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Reviews of Modern Physics 74(1), 47–97 (2002)
14. Barabási, A.L., Albert, R.: Emergence of Scaling in Random Networks. Science 286(5439), 509–512 (1999)
15. Nagelkerke, N.J.D.: A note on a general definition of the coefficient of determination. Biometrika 78(3), 691–692 (1991)
16. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature 393(6684), 440–442 (1998)
17. Ravasz, E., Somera, A.L., Mongru, D., Oltvai, Z.N., Barabási, A.L.: Hierarchical Organization of Modularity in Metabolic Networks. Science 297(5586), 1551–1555 (2002)