Semi-Automatic Extension of GermaNet with Sense Definitions from Wiktionary

Verena Henrich, Erhard Hinrichs, Tatiana Vodolazova

University of Tübingen, Department of Linguistics
Wilhelmstr. 19, 72074 Tübingen, Germany
{verena.henrich,erhard.hinrichs,tatiana.vodolazova}@uni-tuebingen.de

Abstract

Comprehensive sense definitions enhance the usability of the German wordnet GermaNet for a wide variety of NLP applications. The purpose of this paper is to automatically add sense descriptions to GermaNet from the web-based dictionary Wiktionary. A customized version of the Lesk algorithm is developed that automatically harvests Wiktionary's sense descriptions and maps them to corresponding lexical units in GermaNet. Different setups of the algorithm are compared. This algorithm yields as the best result an accuracy of 93.8% and an F1-score of 84.3, which confirms the viability of the proposed method for automatically enriching GermaNet. This best result crucially involves the use of coordinated relations as a novel concept for calculating sense alignment.

Keywords: GermaNet, German wordnet, Wiktionary, word sense disambiguation, sense definitions, resource alignment

1. Introduction

Dictionaries that include information about the meaning of words provide definitions or descriptions that illustrate individual word senses. The Princeton WordNet for English provides comprehensive sense definitions for its synonymy sets (called *synsets*), whose members (called *lexical units*) are taken to have the same meaning. For example, the synset [pipe, tube] is defined by *a hollow cylindrical shape*. While the usefulness of such descriptions for identifying word senses is self-evident, many wordnets for languages other than English lack comprehensive coverage of such definitions. Rather, these wordnets rely on the implicit mutual disambiguation of word senses by the members of a synset. For example, for the synsets [pipe, tobacco pipe] and [pipe, tube], the contrast between the lexical units *tobacco pipe* and *tube* indicates which senses are documented by the two synsets. This type of implicit disambiguation has its limits for those words where the individual senses are synsets with only one member. For example, GermaNet contains two senses for the word *Pfeife,* which can either refer to a whistle or a tobacco pipe, with each sense represented by a single lexical unit as the only member of a synset. GermaNet's coverage of sense definitions is far from complete and does not include the synsets for *Pfeife*. Currently, only 10% of all synsets in GermaNet are accompanied by definitions. Given the considerable coverage of GermaNet, adding descriptions to the missing 62,582 synsets by purely manual, lexicographic work would be an arduous task. Therefore, the possibility of employing automatic or semi-automatic methods for adding sense descriptions would be extremely valuable. The purpose of this paper is to explore this possibility on the basis of Wiktionary, a freely available, web-based dictionary containing sense definitions. The idea is to automatically harvest Wiktionary's definitions by mapping word senses in GermaNet to the corresponding entries in Wiktionary. A survey of the overlaps of GermaNet and Wiktionary shows that, disregarding word sense disambiguation, about 30,488 terms (45.23%) in GermaNet are also present in the German Wiktionary (Zesch, 2010). There has been a considerable body of research for English that investigates the alignment of the Princeton WordNet with Wikipedia (including Niemann and Gurevych, 2011; Ponzetto and Navigli, 2009/2010; Ruiz-Casado et al., 2005; Suchanek et al., 2007; Toral et al., 2009) or with the Oxford Dictionary of English (Navigli, 2006). However, we are not aware of any other previous research that tries to align the German Wiktionary to GermaNet.

2. Resources

GermaNet (Kunze and Lemnitzer, 2002; Henrich and Hinrichs, 2010) is a lexical semantic network modeled after the Princeton WordNet for English (Fellbaum, 1998). It partitions adjectives, nouns, and verbs into a set of semantic concepts (called *synsets*) that are interlinked by semantic relations. A synset is a set of words (called *lexical units*) where all the words are taken to have (almost) the same meaning. There are two types of semantic relations in GermaNet. *Conceptual relations*, such as hypernymy, part-whole relations, entailment, or causation, hold between two synsets. *Lexical relations* hold between two individual lexical units. The current version of GermaNet (version 6.0 of April 2011) covers 93407 lexical units, which are grouped into 69594 synsets.

Wiktionary is a web-based dictionary that is available for many languages, including German. It is written collaboratively by volunteers and provides information such as part-of-speech, hyphenation, possible translations, inflection, etc. It covers, among others, the word categories of adjectives, nouns, and verbs (that are also available in GermaNet). Word senses are distinguished by sense descriptions and by example sentences. Further, Wiktionary provides relations to other words, e.g., in the form of synonyms, antonyms, hypernyms, hyponyms, holonyms, and meronyms. Different from GermaNet, the relations are (mostly) not disambiguated. For the present project, a dump of the German Wiktionary as of February 2, 2011 is utilized, consisting of 46457 German words comprising 70339 word senses.

3. Word Sense Disambiguation Using a Variation of Adapted Lesk (WSD-VAL)

Lesk (1986) introduces a word sense disambiguation algorithm that disambiguates two words by counting the overlaps between their respective sense definitions. Banerjee and Pedersen (2003) apply the Lesk algorithm to the task of computing the semantic relatedness between the synsets of WordNet. They extend the definitions overlap measure by adding the definition overlaps of the related synsets and by giving more weight to longer sequences of matching words. The underlying idea is that the more related two synsets are, the more overlaps their definitions and the definitions of corresponding related words have.

3.1. The Idea of WSD-VAL

In the current scenario of Wiktionary and GermaNet, the aim is to correctly map a GermaNet lexical unit to a Wiktionary sense. There can be more than one occurrence of a target word in GermaNet, thus a target word can correspond to a number of lexical units, each belonging to a distinct semantic concept, i.e., synset. The mapping of each sense definition in Wiktionary needs to consider all synsets containing the target word in GermaNet. Further, due to different sense granularities and distinct coverages of Wiktionary and GermaNet, senses in one resource may correspond to more than one, exactly one, or even no senses in the other resource. Even if there is exactly one sense in both resources, this does not necessarily mean that they match. For example, there is exactly one sense for *Angeln* 'fishing' in GermaNet and exactly one sense in Wiktionary described as *Landschaft im Nordosten Schleswig-Holsteins* 'region in the north-east of Schleswig-Holstein'; but these two senses are clearly distinct.

The other challenge for the mapping is the absence of sense definitions in GermaNet, which prohibits, e.g., simply applying a Lesk-based disambiguation out-of-the-box. We hence develop a word sense disambiguation algorithm (WSD-VAL), which accommodates auxiliary information from GermaNet and Wiktionary to enable a word overlaps approach as introduced by Lesk.

Lexical fields: Therefore, we introduce the notion of *lexical fields*, which substitute sense descriptions in GermaNet by encapsulating relations and semantic field information.[1] Fig. 1 visualizes the extraction of the lexical field information for one sense of the word *Eisen* 'iron' in GermaNet. All lexical units (the items in the boxes with a white background in Fig. 1) related to the target word – either directly by a lexical relation or indirectly by a conceptual relation – are extracted[2]. For example, the synonym *Ferrum* 'ferrum', the hypernyms *Schwermetall* 'heavy metal', *Mineralstoff* 'mineral', etc., the holonyms *Eisenerz* 'iron ore', *Stahl* 'steel', etc., the hyponyms *Magnet* 'magnet', *Gusseisen* 'cast iron', etc. – to name only a few. In addition to the terms obtained via lexical and conceptual relations, the lexical fields are further enriched by the labels of the top-most category (in WordNet terminology also called *unique beginner*) that the target word belongs to. Each part of speech in GermaNet is partitioned into such unique beginners. The word *Eisen*, for example, belongs to the unique beginner *Substanz* 'substance'. This is why the term *Substanz* is added to the lexical field.[3] All words that have been added to the lexical field in the example are listed on the right in Fig. 1.

Given a target word, the customized algorithm WSD-VAL counts the overlaps between each of Wiktionary's sense definitions and each of the lexical fields representing lexical units in GermaNet. For example, the lexical field of *Eisen* 'iron' in GermaNet contains the hypernym *Chemisches Element* (see the box on the right in Fig. 1), which appears in the first sense definition of *Eisen* in Wiktionary described as *Chemie, ohne Plural: chemisches Element, silberweißes, bei Feuchtigkeit leicht oxidierendes Metall* 'Chemistry, without plural: chemical element, silver-white, on dampness easily rusting metal'.

Coordinated relations: Besides the application of lexical fields for counting word overlaps, WSD-VAL utilizes the occurrence of the same relations in GermaNet and Wiktionary – which we call *coordinated relations*. As mentioned in Section 2, Wiktionary and GermaNet have several relations in common, for example the hypernymy and the hyponymy relations. Thus, if a lexical unit in GermaNet and a sense in Wiktionary both show the same hypernyms, this is a strong indicator for their equality. For example, *Eisen* in the sense of 'iron' in GermaNet and the first sense of *Eisen* in Wiktionary both show, among others, the same hypernym *Schwermetall* and the same two hyponyms *Roheisen* and *Gusseisen*, which are a good indicator that these two entries express the same semantic concept.

Utilizing the overlap information of lexical fields and coordinated relations is the underlying idea of the WSD-VAL algorithm.

3.2. Implementation of WSD-VAL

Preprocessing: The Wiktionary sense descriptions are tokenized and stopwords, such as determiners, are withdrawn. All words can also be normalized using either stemming or tokenization.[4] As compounding is a highly productive word formation process in German (Eisenberg, 2006), it is necessary to split compounds occurring in the sense descriptions in Wiktionary and in the lexical fields in GermaNet to achieve a higher overlap rate. For example, the compound *Wasserwelle* 'waterwave' is – as a synonym – contained in the lexical field of one sense of *Welle* 'wave' in GermaNet (Henrich and Hinrichs, 2011). After splitting the compound into its two components *Wasser* 'water' and *Welle*, an overlap (of *Wasser*) with the first sense definition in Wiktionary, i.e. *Physik: Erhebung von Wasser* 'physics: elevation of water', can be captured. Duplicates, which arise due to compound splitting, are eliminated to avoid multiple

[1] A similar kind of technique using all related words for constructing *pseudo glosses* has been used by Gurevych (2005) for the purpose of computing semantic relatedness for any two words in GermaNet.

[2] As indicated by the panel (with the caption *key*) in the top left part of Fig. 1 conceptual relations connect synsets whereas lexical relations connect lexical units contained in synsets.

[3] Other unique beginners for nouns include *Nahrung* 'food', *Tier* 'animal', *Körper* 'body', *Pflanze* 'plant', etc.

[4] Several experiments with stemming and lemmatization yielded better results with stemming. Thus, all below described experiments use stemming (Snowball stemmer; Porter, 1980) in the preprocessing step.

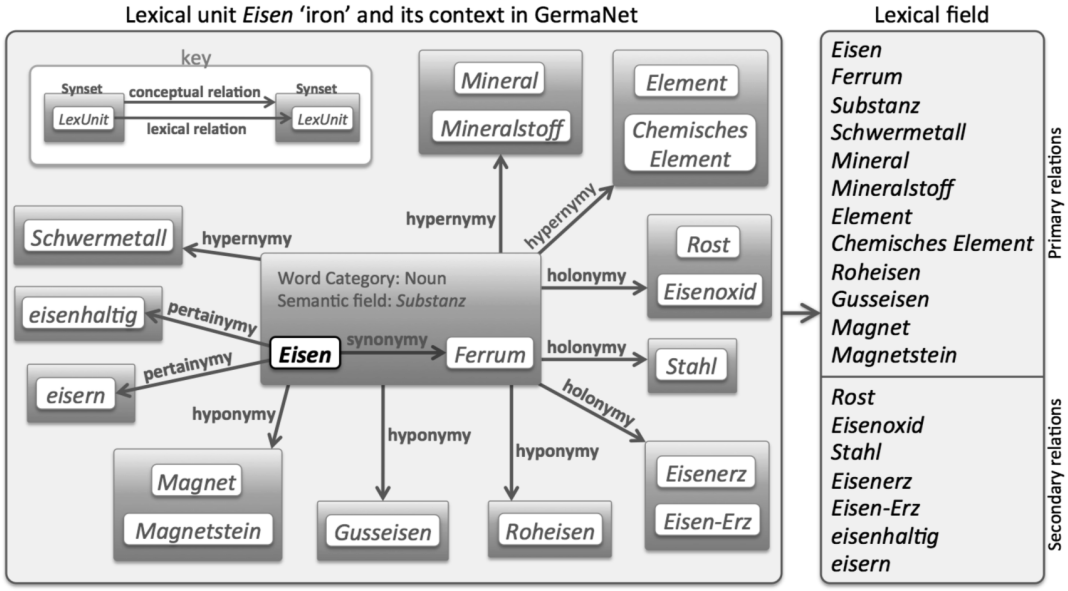

Fig. 1: Lexical field example using *Eisen* 'iron'

overlap counts for the same word. For example, after splitting *Wasserwelle* into its two components *Wasser* and *Welle*, the duplicate *Welle* (which is actually also the target word itself) is withdrawn.

Different versions: Basically, two versions of WSD-VAL are implemented, which can be run separately or in combination: The first variant utilizes the lexical fields in GermaNet as described above. All words are included into the lexical field that are directly connected[5] to the target word. For experimenting with different sets of words, two types of relations are distinguished: *Primary relations*, such as synonymy, hypernymy, and hyponymy, constituting the fundamental structure of a wordnet; and *secondary relations*, such as association, causation, entailment, holonymy, meronymy, and pertainymy with a subordinated importance. In the previous example of *Eisen* 'iron', all words that are connected by a primary relation are listed above the line in the box on the right in Fig. 1, and all words connected by a secondary relation are listed below the line. An overlap of single words is calculated between a tokenized Wiktionary sense description and a lexical field belonging to a target lexical unit in GermaNet.

The second variant of WSD-VAL counts the overlaps of coordinated relations between GermaNet and Wiktionary (as explained in Section 3.1). Therefore, all relations that occur in both resources, such as synonymy, antonymy, hypernymy, hyponymy, meronymy, and holonymy, are considered.

Overlap count: More precisely, each of the target lexical units in GermaNet is represented by a lexical field (as described in Section 3.1). An overlap is calculated between a lexical field in GermaNet and a sense description in Wiktionary. The overlap is a mere count of the number of words x_i for $x_i \in X$, where X is a set of words representing the lexical field of a target lexical unit in GermaNet found in the set of words of a Wiktionary sense description – optionally augmented by the coordinated relations overlap. Further, in case that there is exactly one sense in both resources for a given word, an initial count of 1 is given to the overall count of overlaps.[6]

WSD-VAL maps the Wiktionary sense definitions with the highest overlap counts to a given lexical unit in GermaNet and disregards all other overlap counts (even if those are above zero).[7] Notice that the overlap calculation of WSD-VAL can result in the same overlap score for several senses in Wiktionary. In these cases, more than one Wiktionary sense definition is mapped onto the lexical unit in question and is taken to mean that the lexical unit in GermaNet is jointly described by the Wiktionary sense descriptions in question.

Algorithm Setups: As the lexical fields do not consist of continuous word sequences but rather of single words, it is not possible to give more weight to longer sequences of word matches as it is done by Banerjee and Pedersen (2003). In order to be able to fine-tune the most reliable set of relations, the algorithm includes the possibility of specifying individual weights for different relations.[8]

A set of WSD-VAL experiments were conducted that differ from each other in the weight assigned to the terms that make up the lexical field of a given GermaNet lexical unit (see Table 1). For setups A – C, only the terms contributed by a single primary relation are considered (given a non-zero weight). Setup D considers only all

[5] Here, *directly connected* means that the path length between two words is exactly one – disregarding the type of relation (lexical or conceptual).

[6] Needless to say, assigning an arbitrary count of at least 1 to the overlap score between words occurring exactly once in both resources will result in a positive mapping of these two senses which, in turn, will result in a prediction of false positives for all cases, where those senses do not match (see the example of *Angeln* in Section 3.1 above). However, such cases are rare and therefore the heuristic in question works well in practice (see the evaluation section below).

[7] Experiments with more sophisticated calculations, such as determining the average of all overlap counts and defining all counts that are above this average as a predicted mapping, did not show noticeable improvements.

[8] In computing semantic relatedness, for example, it has been shown that the hypernymy relation provides the best results (Gurevych, 2005).

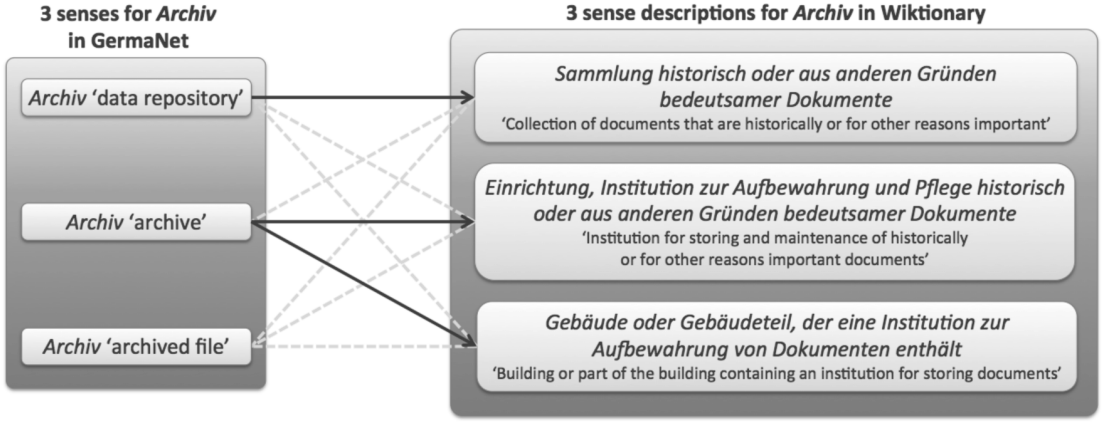

Fig. 2: Sense mapping example using *Archiv* 'data repository; archive; archived file'

secondary relations and setup E only all coordinated relations. In setup F, all terms obtained by the primary, secondary, and coordinated relations are given equal weight. In addition, a set of experiments was conducted where the terms obtained by the different relations were given different weights. Setup G shows the weight assignments that produced optimal results for a precision and recall evaluation (see Section 4 below).

Setup	*Lexical field overlap*				*Coordi-nated relations*
	Primary relations			*2nd-ary relations*	
	Hyper.	*Hypo.*	*Syno.*		
A	1	0	0	0	0
B	0	1	0	0	0
C	0	0	1	0	0
D	0	0	0	1	0
E	0	0	0	0	1
F	1	1	1	1	1
G	2	0.5	3	0.5	3

Table 1. Different algorithm setups

4. Evaluation

In order to be able to evaluate the automatic alignment of lexical units in GermaNet with senses in Wiktionary, the mappings produced by WSD-VAL were manually checked by two experienced lexicographers. Altogether, 20997 distinct words with an average of 1.3 senses (i.e., 27309 lexical units of which 3241 adjectives, 19423 nouns, and 4645 verbs) were manually checked.

Since the number of senses assigned to a word in GermaNet and Wiktionary may differ and since the lexical coverage of the two resources only partially coincides, a given word sense in GermaNet may have no counterpart in Wiktionary at all, or it may correspond to exactly one or more than one senses in Wiktionary. The same holds true in the inverse direction. Fig. 2 illustrates the range of possible mappings for the word *Archiv*, for which each resource records three distinct senses. Accordingly, three times three sense combinations need to be considered. This number rises exponentially with more available senses. The solid arrows in Fig. 2 denote the correct mappings: The first sense in GermaNet ('data repository') corresponds to the first sense in Wiktionary, the second sense in GermaNet ('archive') corresponds two both the second and third senses in Wiktionary, and the third sense in GermaNet ('archived file') does not map onto any sense in Wiktionary.

A truly meaningful evaluation has to reflect the nature of the task at hand: the semi-automatic enrichment of lexical units in GermaNet with appropriate sense descriptions from Wiktionary. A very crude approach would simply map all word senses recorded in GermaNet for a given word with all corresponding sense descriptions documented in Wiktionary. Such an approach would require a lot of human post-editing of eliminating all inappropriate mappings. Since the illicit mappings by far outnumber the number of correct mappings, the approach would be clearly inappropriate. In fact, the motivation behind WSD-VAL is precisely to map only plausible candidates for correct mappings in terms of word overlap between lexical fields in GermaNet and sense descriptions in Wiktionary.

The above considerations clearly show that the task at hand requires maximization of accuracy which amounts to minimization of the amount of human post-processing required. Accuracy of word sense mapping is computed as the ratio of true positives and true negatives compared to the overall number of possible mappings for a word. The overall accuracy is then computed as the average of all word accuracies. Table 2 shows the results for the described task separately for the different setups listed in Table 1.

Setup	*Accuracy*	*Recall*	*Precision*	*F1*
A	93.3%	72.1%	71.3%	71.7
B	93.1%	61.2%	60.8%	61.0
C	**93.8%**	63.8%	63.4%	63.6
D	93.2%	61.3%	60.8%	61.0
E	93.2%	73.6%	73.5%	73.5
F	92.3%	83.8%	82.8%	83.3
G	91.9%	**84.6%**	**84.1%**	**84.3**
Basel.	53.7%	50.7%	44.2%	47.2

Table 2. Accuracy, precision, recall, and F1-measure

As setups A – E each take into account only one property (see Table 1), a comparison of the results directly reflects the suitability of the different relation types when applied independently of each other. The use of hypernyms (setup A) and synonyms (setup C) outperform the use of hyponyms (setup B), secondary relations (setup D), and coordinated relations (setup E). The fact that hypernyms and synonyms outperform the other relations is not surprising since sense definitions often refer to a hypernym or synonym term which is then

described in more detail to fit the specific properties of the entity being described. For example, the English WordNet defines the noun *convertible* as 'a car [hypernym] that has a top that can be folded or removed'.

Apart from the accuracy of the mappings proposed by the algorithm, the recall behavior of WSD-VAL is also relevant. Poor recall would mean that many empirically correct mappings go undetected by the algorithm and therefore have to be manually added. We therefore also computed recall (see Table 2). Recall of the single application of hypernyms (setup A) and coordinated relations (setup E) is better than all other setups of single relations (setups B, C, and D). Again, this result is hardly surprising since coordinated relations are present when the same two terms are connected by the same lexical relations in both resources (see Section 3.1 for a more detailed discussion on coordinated relations). The best recall values are obtained by those settings where all relations are taken into account for the construction of the lexical field (setups F and G). Notice also that, compared to accuracy, there is a much wider spread in the results for recall, ranging from 61.2% (setup B) to 84.6% (setup G). This is hardly surprising since recall is bound to improve with the number of terms included in the lexical field as candidate for overlap. For completeness, Table 2 also contains the scores for precision and F1.

Note also that the accuracies for all setups are above 90%. Human correction is, thus, needed on average only for one out of ten mappings between GermaNet and Wiktionary suggested by the algorithm. The baseline of randomly mapping Wiktionary senses to lexical units in GermaNet (see row *Basel.* in Table 2) demonstrates that the mapping task as such is far from trivial. All setups A to G significantly outperform the baseline. This underscores the overall feasibility of the approach.

5. Conclusion

Sense definitions are a crucial component for wordnets. However, as the current version of GermaNet rarely shows sense definitions, comprehensive sense definitions are badly needed in order to enhance its usability for a wide variety of NLP applications. The present paper has presented a method for semi-automatically enriching lexical units in GermaNet with appropriate sense descriptions from Wiktionary. An accuracy of more than 90% suggests that human correction is needed on average only for one out of ten mappings suggested by the algorithm. Moreover, the estimations of recall and precision result in 84.6% and 84.1%, respectively. These numbers underscore the overall feasibility of the approach and verify its usability for the task at hand.

The publication of this paper will be accompanied by making the extension of GermaNet with sense descriptions from Wiktionary freely available.

In future work, we plan to use this sense-mapping between GermaNet and Wiktionary to increase GermaNet's coverage with words and senses that are present in Wiktionary but not in GermaNet.

References

Banerjee, S. and Pedersen, T. (2003). Extended Gloss Overlaps as a Measure of Semantic Relatedness. *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pp. 805-810.

Eisenberg, P. (2006). *Das Wort – Grundriss der deutschen Grammatik*. 3rd edition, Verlag J. B. Melzer, Stuttgart/Weimar, Germany.

Fellbaum, C. (eds.) (1998). *WordNet – An Electronic Lexical Database*. The MIT Press.

Gurevych, I. (2005). Using the Structure of a Conceptual Network in Computing Semantic Relatedness. *Proceedings of Second International Joint Conference on Natural Language Processing*, pp. 767-778.

Henrich, V. and Hinrichs, E. (2010). GernEdiT – The GermaNet Editing Tool. *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, pp. 2228-2235.

Henrich, V. and Hinrichs, E. (2011). Determining Immediate Constituents of Compounds in GermaNet. *Proceedings of the International Conference RANLP-2011*, pp. 420-426.

Kunze, C. and Lemnitzer, L. (2002). GermaNet – representation, visualization, application. *Proceedings of the Third Conference on International Language Resources and Evaluation*, Vol V., pp. 1485-1491.

Lesk M. (1986). Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to tell a pine cone from an ice cream cone. *Proceedings of the Fifth Annual Conference on Systems Documentation*, pp. 24-26.

Niemann, E. and Gurevych, I. (2011). The People's Web meets Linguistic Knowledge: Automatic Sense Alignment of Wikipedia and WordNet. *Proceedings of the Ninth International Conference on Computational Semantics*, pp. 205-214.

Navigli, R. (2006). Meaningful clustering of senses helps boost word sense disambiguation performance. *Proceedings of COLING 2006 and ACL 2006*. Association for Computational Linguistics, pp. 105-112.

Ponzetto, S. P. and Navigli R. (2009). Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia. *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence*, pp. 2083-2088.

Ponzetto, S. P. and Navigli, R. (2010). Knowledge-rich Word Sense Disambiguation rivaling supervised system. *Proceedings of the 48th Annual Meeting of the ACL*, pp. 1522-1531.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3) :130-137.

Ruiz-Casado, M., Alfonseca, E. and Castells P. (2005). Automatic Assignment of Wikipedia Encyclopedic Entries to WordNet Synsets. *Advances in Web Intelligence*, Volume 3528 of LNCS, Springer Verlag, pp. 380-386.

Suchanek, F. M., Kasneci, G. and Weikum, G. (2007). Yago: A Core of Semantic Knowledge. *Proceedings of the 16th International World Wide Web Conference*, pp. 697-706.

Toral, A., Ferrandez, Ó., Agirre, E. and Muñoz, R. (2009). A study on Linking Wikipedia categories to Wordnet using text similarity. *Proceedings of the International Conference RANLP-2009*, pp. 449-454.

Zesch, T. (2010). What's the Difference? - Comparing Expert-Built and Collaboratively-Built Lexical Semantic Resources. *FLaReNet Forum 2010*, Barcelona, Spain.