# WebCAGe – A Web-Harvested Corpus Annotated with GermaNet Senses

**Verena Henrich, Erhard Hinrichs, and Tatiana Vodolazova**
University of Tübingen
Department of Linguistics
{firstname.lastname}@uni-tuebingen.de

## Abstract

This paper describes an automatic method for creating a domain-independent sense-annotated corpus harvested from the web. As a proof of concept, this method has been applied to German, a language for which sense-annotated corpora are still in short supply. The sense inventory is taken from the German wordnet GermaNet. The web-harvesting relies on an existing mapping of GermaNet to the German version of the web-based dictionary Wiktionary. The data obtained by this method constitute WebCAGe (short for: *Web-Harvested Corpus Annotated with GermaNet Senses*), a resource which currently represents the largest sense-annotated corpus available for German. While the present paper focuses on one particular language, the method as such is language-independent.

## 1 Motivation

The availability of large sense-annotated corpora is a necessary prerequisite for any supervised and many semi-supervised approaches to word sense disambiguation (WSD). There has been steady progress in the development and in the performance of WSD algorithms for languages such as English for which hand-crafted sense-annotated corpora have been available (Agirre et al., 2007; Erk and Strapparava, 2012; Mihalcea et al., 2004), while WSD research for languages that lack these corpora has lagged behind considerably or has been impossible altogether.

Thus far, sense-annotated corpora have typically been constructed manually, making the creation of such resources expensive and the compilation of larger data sets difficult, if not completely infeasible. It is therefore timely and appropriate to explore alternatives to manual annotation and to investigate automatic means of creating sense-annotated corpora. Ideally, any automatic method should satisfy the following criteria:

(1) The method used should be language independent and should be applicable to as many languages as possible for which the necessary input resources are available.

(2) The quality of the automatically generated data should be extremely high so as to be usable as is or with minimal amount of manual post-correction.

(3) The resulting sense-annotated materials (i) should be non-trivial in size and should be dynamically expandable, (ii) should not be restricted to a narrow subject domain, but be as domain-independent as possible, and (iii) should be freely available for other researchers.

The method presented below satisfies all of the above criteria and relies on the following resources as input: (i) a sense inventory and (ii) a mapping between the sense inventory in question and a web-based resource such as Wiktionary[1] or

[1] http://www.wiktionary.org/



*Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 387–396,
Avignon, France, April 23 - 27 2012. 

Wikipedia[2].

As a proof of concept, this automatic method has been applied to German, a language for which sense-annotated corpora are still in short supply and fail to satisfy most if not all of the criteria under (3) above. While the present paper focuses on one particular language, the method as such is language-independent. In the case of German, the sense inventory is taken from the German wordnet GermaNet[3] (Henrich and Hinrichs, 2010; Kunze and Lemnitzer, 2002). The web-harvesting relies on an existing mapping of GermaNet to the German version of the web-based dictionary Wiktionary. This mapping is described in Henrich et al. (2011). The resulting resource consists of a web-harvested corpus WebCAGe (short for: *Web-Harvested Corpus Annotated with GermaNet Senses*), which is freely available at: http://www.sfs.uni-tuebingen.de/en/webcage.shtml

The remainder of this paper is structured as follows: Section 2 provides a brief overview of the resources GermaNet and Wiktionary. Section 3 introduces the mapping of GermaNet to Wiktionary and how this mapping can be used to automatically harvest sense-annotated materials from the web. The algorithm for identifying the target words in the harvested texts is described in Section 4. In Section 5, the approach of automatically creating a web-harvested corpus annotated with GermaNet senses is evaluated and compared to existing sense-annotated corpora for German. Related work is discussed in Section 6, together with concluding remarks and an outlook on future work.

## 2 Resources

### 2.1 GermaNet

GermaNet (Henrich and Hinrichs, 2010; Kunze and Lemnitzer, 2002) is a lexical semantic network that is modeled after the Princeton WordNet for English (Fellbaum, 1998). It partitions the lexical space into a set of concepts that are interlinked by semantic relations. A semantic concept is represented as a *synset*, i.e., as a set of words whose individual members (referred to as *lexical units*) are taken to be (near) synonyms. Thus, a synset is a set-representation of the semantic relation of synonymy.

There are two types of semantic relations in GermaNet. *Conceptual relations* hold between two semantic concepts, i.e. synsets. They include relations such as hypernymy, part-whole relations, entailment, or causation. *Lexical relations* hold between two individual lexical units. Antonymy, a pair of opposites, is an example of a lexical relation.

GermaNet covers the three word categories of adjectives, nouns, and verbs, each of which is hierarchically structured in terms of the hypernymy relation of synsets. The development of GermaNet started in 1997, and is still in progress. GermaNet's version 6.0 (release of April 2011) contains 93407 lexical units, which are grouped into 69594 synsets.

### 2.2 Wiktionary

Wiktionary is a web-based dictionary that is available for many languages, including German. As is the case for its sister project Wikipedia, it is written collaboratively by volunteers and is freely available[4]. The dictionary provides information such as part-of-speech, hyphenation, possible translations, inflection, etc. for each word. It includes, among others, the same three word classes of adjectives, nouns, and verbs that are also available in GermaNet. Distinct word senses are distinguished by sense descriptions and accompanied with example sentences illustrating the sense in question.

Further, Wiktionary provides relations to other words, e.g., in the form of synonyms, antonyms, hypernyms, hyponyms, holonyms, and meronyms. In contrast to GermaNet, the relations are (mostly) not disambiguated.

For the present project, a dump of the German Wiktionary as of February 2, 2011 is uti-

[2] http://www.wikipedia.org/

[3] Using a wordnet as the gold standard for the sense inventory is fully in line with standard practice for English where the Princeton WordNet (Fellbaum, 1998) is typically taken as the gold standard.

[4] Wiktionary is available under the Creative Commons Attribution/Share-Alike license http://creativecommons.org/licenses/by-sa/3.0/deed.en

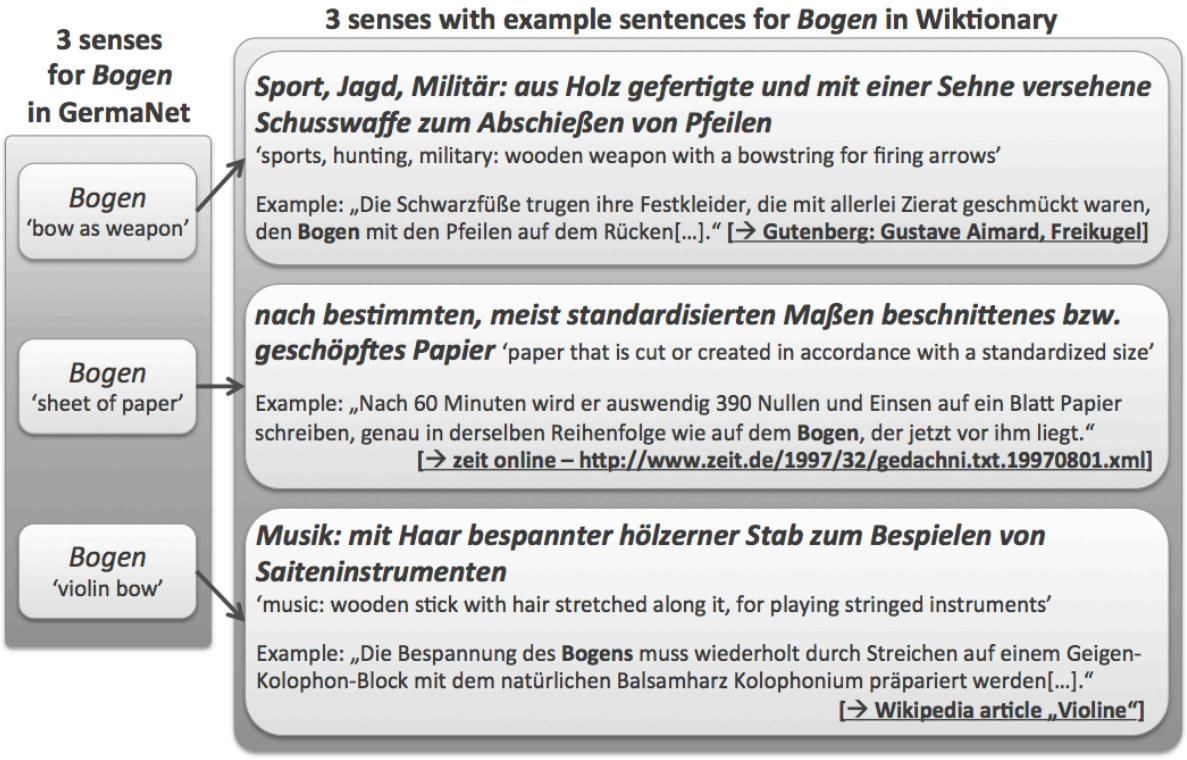


Figure 1: Sense mapping of GermaNet and Wiktionary using the example of *Bogen*.

lized, consisting of 46457 German words comprising 70339 word senses. The Wiktionary data was extracted by the freely available Java-based library JWKTL[5].

## 3 Creation of a Web-Harvested Corpus

The starting point for creating WebCAGe is an existing mapping of GermaNet senses with Wiktionary sense definitions as described in Henrich et al. (2011). This mapping is the result of a two-stage process: i) an automatic word overlap alignment algorithm in order to match GermaNet senses with Wiktionary sense descriptions, and ii) a manual post-correction step of the automatic alignment. Manual post-correction can be kept at a reasonable level of effort due to the high accuracy (93.8%) of the automatic alignment.

The original purpose of this mapping was to automatically add Wiktionary sense descriptions to GermaNet. However, the alignment of these two resources opens up a much wider range of possibilities for data mining community-driven resources such as Wikipedia and web-generated content more generally. It is precisely this potential that is fully exploited for the creation of the WebCAGe sense-annotated corpus.

Fig. 1 illustrates the existing GermaNet-Wiktionary mapping using the example word *Bogen*. The polysemous word *Bogen* has three distinct senses in GermaNet which directly correspond to three separate senses in Wiktionary[6]. Each Wiktionary sense entry contains a definition and one or more example sentences illustrating the sense in question. The examples in turn are often linked to external references, including sentences contained in the German Gutenberg text archive[7] (see link in the topmost Wiktionary sense entry in Fig. 1), Wikipedia articles (see link for the third Wiktionary sense entry in Fig. 1), and other textual sources (see the second sense entry in Fig. 1). It is precisely this collection of

[5]http://www.ukp.tu-darmstadt.de/software/jwktl

[6]Note that there are further senses in both resources not displayed here for reasons of space.

[7]http://gutenberg.spiegel.de/


Figure 2: Sense mapping of GermaNet and Wiktionary using the example of *Archiv*.

heterogeneous material that can be harvested for the purpose of compiling a sense-annotated corpus. Since the target word (rendered in Fig. 1 in bold face) in the example sentences for a particular Wiktionary sense is linked to a GermaNet sense via the sense mapping of GermaNet with Wiktionary, the example sentences are automatically sense-annotated and can be included as part of WebCAGe.

Additional material for WebCAGe is harvested by following the links to Wikipedia, the Gutenberg archive, and other web-based materials. The external webpages and the Gutenberg texts are obtained from the web by a web-crawler that takes some URLs as input and outputs the texts of the corresponding web sites. The Wikipedia articles are obtained by the open-source Java Wikipedia Library JWPL [8]. Since the links to Wikipedia, the Gutenberg archive, and other web-based materials also belong to particular Wiktionary sense entries that in turn are mapped to GermaNet senses, the target words contained in these materials are automatically sense-annotated.

Notice that the target word often occurs more than once in a given text. In keeping with the widely used heuristic of "one sense per discourse", multiple occurrences of a target word in a given text are all assigned to the same GermaNet sense. An inspection of the annotated data shows that this heuristic has proven to be highly reliable in practice. It is correct in 99.96% of all target word occurrences in the Wiktionary example sentences, in 96.75% of all occurrences in the external webpages, and in 95.62% of the Wikipedia files.

WebCAGe is developed primarily for the purpose of the word sense disambiguation task. Therefore, only those target words that are genuinely ambiguous are included in this resource. Since WebCAGe uses GermaNet as its sense inventory, this means that each target word has at least two GermaNet senses, i.e., belongs to at least two distinct synsets.

The GermaNet-Wiktionary mapping is not always one-to-one. Sometimes one GermaNet sense is mapped to more than one sense in Wiktionary. Fig. 2 illustrates such a case. For the word *Archiv* each resource records three distinct senses. The first sense ('data repository')

[8] http://www.ukp.tu-darmstadt.de/software/jwpl/

in GermaNet corresponds to the first sense in Wiktionary, and the second sense in GermaNet ('archive') corresponds to both the second and third senses in Wiktionary. The third sense in GermaNet ('archived file') does not map onto any sense in Wiktionary at all. As a result, the word *Archiv* is included in the WebCAGe resource with precisely the sense mappings connected by the arrows shown in Fig. 2. The fact that the second GermaNet sense corresponds to two sense descriptions in Wiktionary simply means that the target words in the example are both annotated by the same sense. Furthermore, note that the word *Archiv* is still genuinely ambiguous since there is a second (one-to-one) mapping between the first senses recorded in GermaNet and Wiktionary, respectively. However, since the third GermaNet sense is not mapped onto any Wiktionary sense at all, WebCAGe will not contain any example sentences for this particular GermaNet sense.

The following section describes how the target words within these textual materials can be automatically identified.

## 4 Automatic Detection of Target Words

For highly inflected languages such as German, target word identification is more complex compared to languages with an impoverished inflectional morphology, such as English, and thus requires automatic lemmatization. Moreover, the target word in a text to be sense-annotated is not always a simplex word but can also appear as subpart of a complex word such as a compound. Since the constituent parts of a compound are not usually separated by blank spaces or hyphens, German compounding poses a particular challenge for target word identification. Another challenging case for automatic target word detection in German concerns particle verbs such as *ankündigen* 'announce'. Here, the difficulty arises when the verbal stem (e.g., *kündigen*) is separated from its particle (e.g., *an*) in German verb-initial and verb-second clause types.

As a preprocessing step for target word identification, the text is split into individual sentences, tokenized, and lemmatized. For this purpose, the sentence detector and the tokenizer of the suite of Apache OpenNLP tools[9] and the TreeTagger (Schmid, 1994) are used. Further, compounds are split by using BananaSplit[10]. Since the automatic lemmatization obtained by the tagger and the compound splitter are not 100% accurate, target word identification also utilizes the full set of inflected forms for a target word whenever such information is available. As it turns out, Wiktionary can often be used for this purpose as well since the German version of Wiktionary often contains the full set of word forms in tables[11] such as the one shown in Fig. 3 for the word *Bogen*.

| Kasus | Singular | Plural 1 | Plural 2 |
|---|---|---|---|
| Nominativ | **der Bogen** | **die Bogen** | **die Bögen** |
| Genitiv | des Bogens | der Bogen | der Bögen |
| Dativ | dem Bogen | den Bogen | den Bögen |
| Akkusativ | den Bogen | die Bogen | die Bögen |

Figure 3: Wiktionary inflection table for *Bogen*.

Fig. 4 shows an example of such a sense-annotated text for the target word *Bogen* 'violin bow'. The text is an excerpt from the Wikipedia article *Violine* 'violin', where the target word (rendered in bold face) appears many times. Only the second occurrence shown in the figure (marked with a `2` on the left) exactly matches the word *Bogen* as is. All other occurrences are either the plural form *Bögen* (`4` and `7`), the genitive form *Bogens* (`8`), part of a compound such as *Bogenstange* (`3`), or the plural form as part of a compound such as in *Fernambukbögen* and *Schülerbögen* (`5` and `6`). The first occurrence of the target word in Fig. 4 is also part of a compound. Here, the target word occurs in the singular as part of the adjectival compound *bogengestrichenen*.

For expository purposes, the data format shown in Fig. 4 is much simplified compared to the actual, XML-based format in WebCAGe. The infor-

[9] http://incubator.apache.org/opennlp/

[10] http://niels.drni.de/s9y/pages/bananasplit.html

[11] The inflection table cannot be extracted with the Java Wikipedia Library JWPL. It is rather extracted from the Wiktionary dump file.

```
[...] Das Wort Geige stammt aus dem deutschen Sprachraum und
umfasste im Mittelalter alle <tag lexUnit="19087" lemma="Bogen"
1 wcat="NN">bogen</tag>gestrichenen Saiteninstrumente. [...]

2 Der <tag lexUnit="19087" lemma="Bogen" wcat="NN">Bogen</tag>
besteht häufig aus dem Rotholz Pernambuk (Fernambuk). Gutes
Pernambuk ist gerade gewachsen und die Fasern verlaufen
parallel, die <tag lexUnit="19087" lemma="Bogen"
3 wcat="NN">Bogen</tag>stange kann besonders dünn gearbeitet
werden und weist eine ideale Elastizität auf. Das Holz eignet
sich somit besonders für qualitativ hochwertige <tag
4 lexUnit="19087" lemma="Bogen" wcat="NN">Bögen</tag>. Da das
Vorkommen der Holzart begrenzt ist, haben Fernambuk<tag
5 lexUnit="19087" lemma="Bogen" wcat="NN">bögen</tag> einen
entsprechen hohen Preis. Einfachere Schüler<tag lexUnit="19087"
6 lemma="Bogen" wcat="NN">bögen</tag> sind meist aus Brasilholz
gefertigt. Heute werden, auch von Berufsgeigern, zunehmend <tag
7 lexUnit="19087" lemma="Bogen" wcat="NN">Bögen</tag> aus
Kohlefaser (Karbonfiber) verwendet.

Am unteren Ende des <tag lexUnit="19087" lemma="Bogen"
8 wcat="NN">Bogens</tag> befindet sich der sogenannte Frosch aus
Ebenholz, meist verziert mit einer runden Perlmutt-Einlage.
[...]

                    Source: http://de.wikipedia.org/wiki/Violine
```

Figure 4: Excerpt from Wikipedia article *Violine* 'violin' tagged with target word *Bogen* 'violin bow'.

mation for each occurrence of a target word consists of the GermaNet sense, i.e., the lexical unit ID, the lemma of the target word, and the GermaNet word category information, i.e., *ADJ* for adjectives, *NN* for nouns, and *VB* for verbs.

## 5 Evaluation

In order to assess the effectiveness of the approach, we examine the overall size of WebCAGe and the relative size of the different text collections (see Table 1), compare WebCAGe to other sense-annotated corpora for German (see Table 2), and present a precision- and recall-based evaluation of the algorithm that is used for automatically identifying target words in the harvested texts (see Table 3).

Table 1 shows that Wiktionary (7644 tagged word tokens) and Wikipedia (1732) contribute by far the largest subsets of the total number of tagged word tokens (10750) compared with the external webpages (589) and the Gutenberg texts (785). These tokens belong to 2607 distinct polysemous words contained in GermaNet, among which there are 211 adjectives, 1499 nouns, and 897 verbs (see Table 2). On average, these words have 2.9 senses in GermaNet (2.4 for adjectives, 2.6 for nouns, and 3.6 for verbs).

Table 2 also shows that WebCAGe is considerably larger than the other two sense-annotated corpora available for German ((Broscheit et al., 2010) and (Raileanu et al., 2002)). It is important to keep in mind, though, that the other two resources were manually constructed, whereas WebCAGe is the result of an automatic harvesting method. Such an automatic method will only constitute a viable alternative to the labor-intensive manual method if the results are of sufficient quality so that the harvested data set can be used as is or can be further improved with a minimal amount of manual post-editing.

For the purpose of the present evaluation, we conducted a precision- and recall-based analysis for the text types of Wiktionary examples, external webpages, and Wikipedia articles sep-

Table 1: Current size of WebCAGe.

| | | Wiktionary examples | External webpages | Wikipedia articles | Gutenberg texts | All texts |
|---|---|---|---|---|---|---|
| Number of tagged word tokens | adjectives | 575 | 31 | 79 | 28 | 713 |
| | nouns | 4103 | 446 | 1643 | 655 | 6847 |
| | verbs | 2966 | 112 | 10 | 102 | 3190 |
| | all word classes | 7644 | 589 | 1732 | 785 | 10750 |
| Number of tagged sentences | adjectives | 565 | 31 | 76 | 26 | 698 |
| | nouns | 3965 | 420 | 1404 | 624 | 6413 |
| | verbs | 2945 | 112 | 10 | 102 | 3169 |
| | all word classes | 7475 | 563 | 1490 | 752 | 10280 |
| Total number of sentences | adjectives | 623 | 1297 | 430 | 65030 | 67380 |
| | nouns | 4184 | 9630 | 6851 | 376159 | 396824 |
| | verbs | 3087 | 5285 | 263 | 146755 | 155390 |
| | all word classes | 7894 | 16212 | 7544 | 587944 | 619594 |

Table 2: Comparing WebCAGe to other sense-tagged corpora of German.

| | | WebCAGe | Broscheit et al., 2010 | Raileanu et al., 2002 |
|---|---|---|---|---|
| Sense tagged words | adjectives | 211 | 6 | 0 |
| | nouns | 1499 | 18 | 25 |
| | verbs | 897 | 16 | 0 |
| | all word classes | 2607 | 40 | 25 |
| Number of tagged word tokens | | 10750 | approx. 800 | 2421 |
| Domain independent | | yes | yes | medical domain |

arately for the three word classes of adjectives, nouns, and verbs. Table 3 shows that precision and recall for all three word classes that occur for Wiktionary examples, external webpages, and Wikipedia articles lies above 92%. The only sizeable deviations are the results for verbs that occur in the Gutenberg texts. Apart from this one exception, the results in Table 3 prove the viability of the proposed method for automatic harvesting of sense-annotated data. The average precision for all three word classes is of sufficient quality to be used as-is if approximately 2-5% noise in the annotated data is acceptable. In order to eliminate such noise, manual post-editing is required. However, such post-editing is within acceptable limits: it took an experienced research assistant a total of 25 hours to hand-correct all the occurrences of sense-annotated target words and to manually sense-tag any missing target words for the four text types.

## 6 Related Work and Future Directions

With relatively few exceptions to be discussed shortly, the construction of sense-annotated corpora has focussed on purely manual methods. This is true for SemCor, the WordNet Gloss Corpus, and for the training sets constructed for English as part of the SensEval and SemEval shared task competitions (Agirre et al., 2007; Erk and Strapparava, 2012; Mihalcea et al., 2004). Purely manual methods were also used for the German sense-annotated corpora constructed by Broscheit et al. (2010) and Raileanu et al. (2002) as well as for other languages including the Bulgarian and

Table 3: Evaluation of the algorithm of identifying the target words.

| | | **Wiktionary examples** | **External webpages** | **Wikipedia articles** | **Gutenberg texts** |
|---|---|---|---|---|---|
| Precision | adjectives | 97.70% | 95.83% | 99.34% | 100% |
| | nouns | 98.17% | 98.50% | 95.87% | 92.19% |
| | verbs | 97.38% | 92.26% | 100% | 69.87% |
| | all word classes | 97.32% | 96.19% | 96.26% | 87.43% |
| Recall | adjectives | 97.70% | 97.22% | 98.08% | 97.14% |
| | nouns | 98.30% | 96.03% | 92.70.% | 97.38% |
| | verbs | 97.51% | 99.60% | 100% | 89.20% |
| | all word classes | 97.94% | 97.32% | 93.36% | 95.42% |

the Chinese sense-tagged corpora (Koeva et al., 2006; Wu et al., 2006). The only previous attempts of harvesting corpus data for the purpose of constructing a sense-annotated corpus are the semi-supervised method developed by Yarowsky (1995), the knowledge-based approach of Leacock et al. (1998), later also used by Agirre and Lopez de Lacalle (2004), and the automatic association of Web directories (from the Open Directory Project, ODP) to WordNet senses by Santamaría et al. (2003).

The latter study (Santamaría et al., 2003) is closest in spirit to the approach presented here. It also relies on an automatic mapping between wordnet senses and a second web resource. While our approach is based on automatic mappings between GermaNet and Wiktionary, their mapping algorithm maps WordNet senses to ODP subdirectories. Since these ODP subdirectories contain natural language descriptions of websites relevant to the subdirectory in question, this textual material can be used for harvesting sense-specific examples. The ODP project also covers German so that, in principle, this harvesting method could be applied to German in order to collect additional sense-tagged data for WebCAGe.

The approach of Yarowsky (1995) first collects all example sentences that contain a polysemous word from a very large corpus. In a second step, a small number of examples that are representative for each of the senses of the polysemous target word is selected from the large corpus from step 1. These representative examples are manually sense-annotated and then fed into a decision-list supervised WSD algorithm as a seed set for iteratively disambiguating the remaining examples collected in step 1. The selection and annotation of the representative examples in Yarowsky's approach is performed completely manually and is therefore limited to the amount of data that can reasonably be annotated by hand.

Leacock et al. (1998), Agirre and Lopez de Lacalle (2004), and Mihalcea and Moldovan (1999) propose a set of methods for automatic harvesting of web data for the purposes of creating sense-annotated corpora. By focusing on web-based data, their work resembles the research described in the present paper. However, the underlying harvesting methods differ. While our approach relies on a wordnet to Wiktionary mapping, their approaches all rely on the monosemous relative heuristic. Their heuristic works as follows: In order to harvest corpus examples for a polysemous word, the WordNet relations such as synonymy and hypernymy are inspected for the presence of unambiguous words, i.e., words that only appear in exactly one synset. The examples found for these monosemous relatives can then be sense-annotated with the particular sense of its ambiguous word relative. In order to increase coverage of the monosemous relatives approach, Mihalcea and Moldovan (1999) have developed a gloss-based extension, which relies on word overlap of the gloss and the WordNet sense in question for all those cases where a monosemous relative is not contained in the WordNet dataset.

The approaches of Leacock et al., Agirre and Lopez de Lacalle, and Mihalcea and Moldovan as

well as Yarowsky's approach provide interesting directions for further enhancing the WebCAGe resource. It would be worthwhile to use the automatically harvested sense-annotated examples as the seed set for Yarowsky's iterative method for creating a large sense-annotated corpus. Another fruitful direction for further automatic expansion of WebCAGe is to use the heuristic of monosemous relatives used by Leacock et al., by Agirre and Lopez de Lacalle, and by Mihalcea and Moldovan. However, we have to leave these matters for future research.

In order to validate the language independence of our approach, we plan to apply our method to sense inventories for languages other than German. A precondition for such an experiment is an existing mapping between the sense inventory in question and a web-based resource such as Wiktionary or Wikipedia. With BabelNet, Navigli and Ponzetto (2010) have created a multilingual resource that allows the testing of our approach to languages other than German. As a first step in this direction, we applied our approach to English using the mapping between the Princeton WordNet and the English version of Wiktionary provided by Meyer and Gurevych (2011). The results of these experiments, which are reported in Henrich et al. (2012), confirm the general applicability of our approach.

To conclude: This paper describes an automatic method for creating a domain-independent sense-annotated corpus harvested from the web. The data obtained by this method for German have resulted in the WebCAGe resource which currently represents the largest sense-annotated corpus available for this language. The publication of this paper is accompanied by making WebCAGe freely available.

## Acknowledgements

The research reported in this paper was jointly funded by the SFB 833 grant of the DFG and by the CLARIN-D grant of the BMBF. We would like to thank Christina Hoppermann, Marie Hinrichs as well as three anonymous EACL 2012 reviewers for their helpful comments on earlier versions of this paper. We are very grateful to Reinhild Barkey, Sarah Schulz, and Johannes Wahle for their help with the evaluation reported in Section 5. Special thanks go to Yana Panchenko and Yannick Versley for their support with the webcrawler and to Emanuel Dima and Klaus Suttner for helping us to obtain the Gutenberg and Wikipedia texts.

## References

Agirre, E., Lopez de Lacalle, O. 2004. Publicly available topic signatures for all WordNet nominal senses. *Proceedings of the 4th International Conference on Languages Resources and Evaluations* (LREC'04), Lisbon, Portugal, pp. 1123–1126

Agirre, E., Marquez, L., Wicentowski, R. 2007. *Proceedings of the 4th International Workshop on Semantic Evaluations*. Assoc. for Computational Linguistics, Stroudsburg, PA, USA

Broscheit, S., Frank, A., Jehle, D., Ponzetto, S. P., Rehl, D., Summa, A., Suttner, K., Vola, S. 2010. Rapid bootstrapping of Word Sense Disambiguation resources for German. *Proceedings of the 10. Konferenz zur Verarbeitung Natürlicher Sprache*, Saarbrücken, Germany, pp. 19–27

Erk, K., Strapparava, C. 2010. *Proceedings of the 5th International Workshop on Semantic Evaluation*. Assoc. for Computational Linguistics, Stroudsburg, PA, USA

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

Henrich, V., Hinrichs, E. 2010. GernEdiT – The GermaNet Editing Tool. *Proceedings of the Seventh Conference on International Language Resources and Evaluation* (LREC'10), Valletta, Malta, pp. 2228–2235

Henrich, V., Hinrichs, E., Vodolazova, T. 2011. Semi-Automatic Extension of GermaNet with Sense Definitions from Wiktionary. *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics* (LTC'11), Poznan, Poland, pp. 126–130

Henrich, V., Hinrichs, E., Vodolazova, T. 2012. An Automatic Method for Creating a Sense-Annotated Corpus Harvested from the Web. Poster presented at *13th International Conference on Intelligent Text Processing and Computational Linguistics* (CICLing-2012), New Delhi, India, March 2012

Koeva, S., Leseva, S., Todorova, M. 2006. Bulgarian Sense Tagged Corpus. *Proceedings of the 5th SALTMIL Workshop on Minority Languages:*

*Strategies for Developing Machine Translation for Minority Languages*, Genoa, Italy, pp. 79–87

Kunze, C., Lemnitzer, L. 2002. GermaNet representation, visualization, application. *Proceedings of the 3rd International Language Resources and Evaluation* (LREC'02), Las Palmas, Canary Islands, pp. 1485–1491

Leacock, C., Chodorow, M., Miller, G. A. 1998. Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics*, 24(1):147–165

Meyer, C. M., Gurevych, I. 2011. What Psycholinguists Know About Chemistry: Aligning Wiktionary and WordNet for Increased Domain Coverage. *Proceedings of the 5th International Joint Conference on Natural Language Processing* (IJCNLP), Chiang Mai, Thailand, pp. 883–892

Mihalcea, R., Moldovan, D. 1999. An Automatic Method for Generating Sense Tagged Corpora. *Proceedings of the American Association for Artificial Intelligence* (AAAI'99), Orlando, Florida, pp. 461–466

Mihalcea, R., Chklovski, T., Kilgarriff, A. 2004. *Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Barcelona, Spain

Navigli, R., Ponzetto, S. P. 2010. BabelNet: Building a Very Large Multilingual Semantic Network. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (ACL'10), Uppsala, Sweden, pp. 216–225

Raileanu, D., Buitelaar, P., Vintar, S., Bay, J. 2002. Evaluation Corpora for Sense Disambiguation in the Medical Domain. *Proceedings of the 3rd International Language Resources and Evaluation* (LREC'02), Las Palmas, Canary Islands, pp. 609–612

Santamaría, C., Gonzalo, J., Verdejo, F. 2003. Automatic Association of Web Directories to Word Senses. *Computational Linguistics* 29 (3), MIT Press, PP. 485–502

Schmid, H. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK

Wu, Y., Jin, P., Zhang, Y., Yu, S. 2006. A Chinese Corpus with Word Sense Annotation. *Proceedings of 21st International Conference on Computer Processing of Oriental Languages* (ICCPOL'06), Singapore, pp. 414–421

Yarowsky, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics* (ACL'95), Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 189–196