Determining Immediate Constituents of Compounds in GermaNet

Verena Henrich
University of Tübingen
verena.henrich@uni-tuebingen.de

Erhard Hinrichs
University of Tübingen
erhard.hinrichs@uni-tuebingen.de

Abstract

In order to be able to systematically link compounds in GermaNet to their constituent parts, compound splitting needs to be applied recursively and has to identify the immediate constituents at each level of analysis. Existing tools for compound splitting for German only offer an analysis of all component parts of a compound at once without any grouping of subconstituents. Thus, existing tools for splitting compounds were adapted to overcome this issue. Algorithms combining three heterogeneous kinds of compound splitters are developed to achieve better results. The best overall result with an accuracy of 92.42% is achieved by a hybrid combined compound splitter that takes into account all knowledge provided by the individual compound splitters, and in addition some domain knowledge about German derivation morphology and compounding.

1 Introduction

The present paper presents a compound splitter for German that is tailored to the needs of systematically enriching the set of lexical relations of *GermaNet* (Kunze and Lemnitzer, 2002; Henrich and Hinrichs, 2010), the German version of the Princeton WordNet for English (Fellbaum, 1998). Compounding is a highly productive word formation process resulting in complex words with two or more constituent parts. Baroni et al. (2002) report that almost half (47%) of the word types in the APA German news corpus are compounds.

For GermaNet, the numbers are comparable: The morphological analyzer *SMOR* (Schmid et al., 2004) for German classifies 46.89% of all lexical units contained in release 6.0 of GermaNet as compounds. Among those, nominal compounds make up 95% and are thus by far the largest class of compounds. It is for this reason that we concentrate exclusively on the treatment of nominal compounds in the present study.

Given the prevalence of compounds in GermaNet and its current coverage of 84586 lexical units, a systematic treatment of compounds is badly needed in order to enhance the usability of GermaNet for a wide variety of NLP applications, including machine translation, natural language generation, information extraction, etc. The size of GermaNet and the high frequency of compounds clearly prohibit a purely manual solution and mandate an automatic treatment. The treatment of compounds for GermaNet needs to be systematic along at least three dimensions: (i) it should cover all combinations of word classes present in GermaNet which can enter into noun compounding, (ii) it should apply to all lexical units already entered into GermaNet, and (iii) it should be extendable to all compounds which are candidates for inclusion in GermaNet in future data releases.

2 Nominal Compounds in German

Peter Eisenberg (Eisenberg, 2006) defines four major subclasses for compounds, where the rightmost head constituent is a noun.

1. Noun + Noun: *Apfelbaum* ‘apple tree’.
2. Adjective + Noun: *Weißbrot* ‘white bread’.
3. Verb + Noun: *Esstisch* ‘eating table’.
4. Preposition + Noun: *Oberarm* ‘upper arm’.

In addition to these four major classes, there is a small class of bound morphemes (i.e., morphemes that cannot appear as an independent word), such as *Him-*[1], that can also serve as the initial constituent of a nominal compound:

[1] In the German linguistics literature such bound morphemes are referred to as *unikale Elemente*.

Proceedings of Recent Advances in Natural Language Processing, pages 420–426,
Hissar, Bulgaria, 12-14 September 2011.

5. Bound Morpheme + Noun: *Himbeere* 'raspberry'.

What makes compound splitting for German a challenging task is the fact that compounding is not always simple string concatenation, but often involves the presence of intervening linking elements or the elision of word-final characters in the non-head constituent of a compound[2]. Word-final *e*, for example, is absent in compounds such as *Hüftschwung* 'hip swing', whose non-head constituent is *Hüfte* 'hip'. While such elision cases are relatively rare, the presence of linking morphemes in nominal compounds is a much more frequent phenomenon. Eisenberg (2006) distinguishes between the following linking elements: *n* (*Blumenvase*: *Blume* + *n* + *Vase*; 'flower vase'), *s* (*Zweifelsfall*: *Zweifel* + *s* + *Fall*; 'case of doubt'), *ns* (*Glaubensfrage*: *Glaube* + *ns* + *Frage*; 'question of believe'), *e* (*Pferdewagen: Pferd* + *e* + *Wagen*; 'horse carriage'), *er* (*Kindergarten*: *Kind* + *er* + *Garten*), *en* (*Heldenmut*: *Held* + *en* + *Mut*; 'hero's courage'), *es* (*Siegeswille*: *Sieg* + *es* + *Wille*; 'will to win'), and *ens* (*Schmerzensschrei*: *Schmerz* + *ens* + *Schrei;* 'scream of pain').

3 Modeling Compounds in GermaNet

GermaNet is a lexical semantic network that is modeled after the Princeton WordNet for English. It partitions the lexical space into a set of semantic concepts (modeled by *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 the same meaning. There are two types of semantic relations in GermaNet. *Conceptual relations* hold between two synsets, including hypernymy, part-whole relations, entailment, or causation. *Lexical relations* hold between two individual lexical units.

To the best of our knowledge, a systematic treatment of compounds is largely absent from monolingual wordnets presently available. The only programmatic approach for how to treat compounds is documented in the final report of the *EuroWordNet* project (Vossen, 2002) from which the following illustrative example is taken:

```
guitar player
  HAS_HYPERONYM player
  CO_AGENT_INSTRUMENT guitar
```

In this EuroWordNet proposal, compounds such as *guitar player* are linked via conceptual relations to their component parts. The compound as a whole is related via the hypernymy relation to its head constituent (*player*) and via the bidirectional CO_ROLE relation to its modifier constituent (*guitar*). This CO_ROLE relation is then further specified by the particular thematic role realized by the modifier constituent. In short, the EuroWordNet treatment focuses on the semantics of compounds.

The current proposal of how to treat compounds in GermaNet is to some extent more modest in that it focuses on the morphosyntactic structure of compounds and leaves a semantic treatment to future work. A strong requirement for a compounding analysis for GermaNet is that it has to reflect the recursive nature of compounding in the case of compounds that have more than two constituent parts such as *Kraftfahrzeugsteuer* 'motor vehicle tax'. The immediate constituents of this compound are *Kraftfahrzeug* and *steuer*, with the first constituent then splitting further into *Kraft* and *fahrzeug*, etc. (see Figure 1). In order to be able to systematically link compounds in GermaNet to their constituent parts, compound splitting needs to be applied recursively and has to identify only the immediate constituents at each level of analysis.

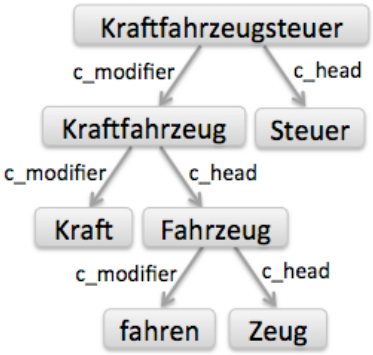

Figure 1. Compounds in GermaNet.

4 Related Work on Compound Splitting

For German, there are a number of morphological tools available that include compound splitting, such as GERTWOL (Haapalainen and Majorin, 1994), SMOR (Schmid et al., 2004), ASV Toolbox (Witschel and Biemann, 2005), BananaSplit[3], and Morfessor (Creutz and Lagus, 2005). After an initial evaluation of all publicly available tools, SMOR and ASV Toolbox are used as baseline tools for the present project.

[2] Langer (1998) presents a frequency table for German linking morphemes and elisions, according to which approximately half of the compounds he investigated contain some kind of linking morpheme or elision.

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

SMOR is a morphological analyzer for German inflection and productive word formation including composition, which has been developed at the University of Stuttgart. It provides analyses consisting of sequences of morphemes enriched with morphological information, however without grouping them into immediate constituents. Furthermore, although SMOR disambiguates its results to a certain extend, for many compounds there are still several distinct sequences of morphemes provided.

ASV Toolbox has been developed at the University of Leipzig. It comprises several tools for linguistic classification and clustering, amongst them compound splitting, which is included in the tool described as *ASV Toolbox Baseforms*[4]. The result of the compound analysis identifies all constituent parts of the compound without internal bracketing. It reduces inflected word forms of constituents to their base forms.

5 Compound Splitting Algorithms

Three individual compound splitters are used in the present project: a compound splitter incorporating GermaNet (GN-CS) developed by the authors of this paper, a modified version of SMOR (SMOR-CS), and a modified version of the ASV Toolbox compound splitter (ASV-CS).

5.1 Compound Splitter Incorporating GermaNet (GN-CS)

This compound splitter is especially tailored for determining compounds in GermaNet and their immediate constituents. It uses pattern matching for gathering all potential modifiers and heads of a compound, considering intervening linking morphemes and the elision of word-final characters (as described in section 2). In case the pattern matching yields more than one potential modifier-head composition, the correct constituents are verified incorporating the semantic resource GermaNet and its graph structure. For example, compositions having both constituents in GermaNet are preferred over compositions where only one constituent is an existing entry in GermaNet. Further, more probability is assigned to compositions of simple string concatenation than to compositions showing a linking morpheme or the elision of word-final characters.

The availability of semantic relations, such as part-whole relations, direct or indirect hypernymy, or synonymy, is employed as well. Thus, a modifier or head that is semantically related to the compound determines the correct splitting of compounds into its immediate constituents with high probability. The following example illustrates this. For the compound *Flughafengelände* 'airport area', all relevant parts of the two candidate parses *Flug* + *Hafengelände* and *Flughafen* + *Gelände* are existing entries in GermaNet, i.e., existing words. Further, both potential analyses show neither linking morphemes nor the elision of word-final characters. In this case, the usage of GermaNet's semantic relations determines, that *Flughafen* is a holonym of the compound *Flughafengelände*, and thus clearly and correctly determines the modifier, resulting in the correct parse *Flughafen* + *Gelände*.

If there are two different modifier-head combinations having both their heads as hypernyms of the compound, GN-CS disambiguates the correct splitting by taking into account the hypernym's distances[5]. The splitting belonging to the head with the larger hypernym distance is preferred.[6] For example, *Nachttischlampe* 'bedside lamp' has both hypernyms *Tischlampe* 'table lamp' (hypernym distance is 1, i.e., direct hypernym) and *Lampe* 'lamp' (hypernym distance is 2, i.e., indirect hypernym). Thus, *Nachttischlampe* is correctly split into *Nachttisch* + *Lampe*.

5.2 Modified SMOR Compound Splitter (SMOR-CS)

To achieve better results in the specific task of determining compounds in GermaNet and their immediate constituents, SMOR's output has been adapted. Some steps, such as the denominalization of the head constituents or the splitting of all affixes, need to be reverted. Other results, such as the splitting into more than two constituents or the indication of more than one splitting possibility, require further processing. For example, SMOR splits *Änderungsanforderung* 'change request' into *ändern* + *ung* + *an* + *fordern* + *ung*. After reverting the denominalization and the separation of prefixes and suffixes, SMOR-CS returns: *Änderung* + *Anforderung*.

For those compounds with several distinct results, it is not trivial to disambiguate the correct splitting. Furthermore, the splitting of compounds having more than two constituents, such

[4] See http://wortschatz.uni-leipzig.de/~cbiemann/software/toolbox/Baseforms%20Tool.htm

[5] Here, *hypernym distance* describes the path length between the compound and a direct or indirect hypernym, i.e., a direct hypernym has a hypernym distance of one.

[6] Preference of longer hypernym distance may seem counterintuitive, but surprisingly turns out to be the correct heuristic.

as *Brennstofflagerungsbehälter* ('fuel storage container'), which is split into *brennen* + *Stoff* + *lagern* + *Behälter* cannot be used in this form for determining immediate constituents, since the constituents are not grouped.

5.3 Modified ASV Toolbox Compound Splitter (ASV-CS)

The output of the ASV Toolbox compound splitter is further processed in order to better fit the needs of the present project. To enhance the reliability of the determined constituents, the enhanced compound splitter ASV-CS searches for entries in GermaNet. If a result consists of more than two constituents, the different bracketing alternatives need to be verified. This is done by incorporating GermaNet's graph structure in the same way as for GN-CS (see section 5.1).

6 Combination of Compound Splitters

It has been shown for various NLP tasks, such as part-of-speech tagging (van Halteren et al., 2001) or word sense disambiguation (Florian and Yarowsky, 2002), that multiple classifier systems outperform single decision systems. Further, the performance of such methods is usually better the more diverse the individual systems are (Polikar, 2006). Thus, having three classifiers[7] (compound splitters) available that produce diverse results, the application of a combined method seems reasonable. As the compound splitters in the present project each return exactly one decision, the range of applicable combination algorithms is restricted. In the following subsection, the application of *majority voting* and *weighted majority voting* is described. Further, a combined algorithm, which is developed by the authors of this paper, is presented.

6.1 Majority Voting (MV) and Weighted Majority Voting (WMV)

In majority voting, equal weight is given to all compound splitters when voting for a result (i.e., a splitting of a compound into its immediate constituents). The votes from all compound splitters are summed up and the result with the highest number of votes is selected. In case a compound splitter does not return an analysis, it is disregarded, while the other two compound splitters vote for the final result.[8] In weighted majority voting, individual compound splitters are assigned different weights in such a way that the combination of weights minimizes errors.[9]

6.2 Combined Hybrid Compound Splitter (CH-CS)

In order to further increase performance, we created a hybrid combined compound splitter that takes into account all knowledge provided by the individual compound splitters, but that also takes into account some domain knowledge about German derivation morphology and compounding. One of the frequent mistakes made is to treat words like *Gutherzigkeit*[10] 'kindheartedness' or *Teilhaberschaft*[11] 'partnership' as compounds, while in reality these are complex nouns formed by derivation morphology. The hybrid model therefore incorporates knowledge about derivation morphology and filters out such erroneously marked compounds. As will be shown in the evaluation section, the hybrid model outperforms all individual compound splitters as well as the other combined compound splitters in all tasks described in section 7.

7 Evaluation

The automatic predictions of compounds and their immediate constituents are manually verified. The order of the manual verification is in the order of the IDs of the lexical units, which is actually randomly concerning the nouns themselves. For the purpose of evaluation, 68743[12] nouns were chosen, of which 42191 (61.37%) are compounds and 26552 (38.63%) are not. The evaluation is fourfold: (i) section 7.1 evaluates how many compounds are correctly identified, (ii) section 7.2 evaluates how many predicted compounds are split at the correct position, (iii) how many compounds are correctly predicted

[7] The task of compound splitting is, in a strict sense, not a classification task, because there is no predefined result set, such as a tagset for part-of-speech tagging. The results of the compound splitters are rather variable and, from a technical point of view, describe arbitrary content (although describing the splitting of a compound into its immediate constituents).

[8] In case of a tie, giving priority to SMOR-CS turned out to be the best strategy.

[9] Experimenting with several weighting combinations resulted in giving weight 2.0 to SMOR-CS, 0.9 to GN-CS, and 0.8 to ASV-CS. This adjustment helps in cases where both GN-CS and ASV-CS agree on an erroneous analysis.

[10] SMOR-CS treats *Gutherzigkeit* erroneously as a compound, although it is derived from the adjective *gutherzig* with the derivation suffix *-keit*.

[11] *Teilhaberschaft* is derived from the noun *Teilhaber* with the derivation suffix *-schaft*.

[12] Altogether, there are 93407 nouns in GermaNet. Note that all foreign words and named entities are disregarded in this evaluation.

regarding the word forms of their immediate constituents is evaluated in section 7.3, and, finally, (iv) there is an error analysis in section 7.4.

7.1 Identification of Compounds

The first part of the evaluation concerns the prediction whether a noun in GermaNet is a compound or not. Table 1 lists all *true positives* (TP; correctly identified compounds), *false positives* (FP; erroneously identified as a compound), *true negatives* (TN; correctly identified as no compound), and *false negatives* (FN; erroneously not identified as a compound). The numbers are separately calculated for the individual algorithms and for the combined algorithms.

Algorithm	*TP*	*FP*	*TN*	*FN*
GN-CS	38489	1559	24993	3702
SMOR-CS	33765	544	26008	8426
ASV-CS	36356	555	25997	5835
MV & WMV	39675	1806	24746	2516
CH-CS	41894	1974	24578	297

Table 1: Identification of Compounds

The reason for MV and WMV performing alike in this task of identifying compounds is that in case a compound splitter does not return an analysis, it is disregarded. This means that, if at least one compound splitter returns a result, both MV and WMV decide that this noun is a compound regardless of any weighting.

There are remarkable improvements especially in the numbers of true positives and false negatives of the combined algorithms compared to the individual ones. The reason for these remarkable differences is obvious: the individual splitting algorithms are very heterogeneous, which leads to an improved overall coverage. Table 2 shows the calculated percentages for accuracy, precision, and recall of the task of identifying compounds.

Algorithm	*Accuracy*	*Precision*	*Recall*
GN-CS	92.34%	96.11%	91.23%
SMOR-CS	86.95%	98.41%	80.03%
ASV-CS	90.70%	98.50%	86.17%
MV & WMV	93.71%	95.65%	94.04%
CH-CS	96.70%	95.50%	99.30%

Table 2: Accuracy, Precision, and Recall of Identifying Compounds

Highest accuracy and best recall are achieved by CH-CS, whereas ASV-CS and SMOR-CS yield highest precision. The values in this section (Tables 1 and 2) are gathered with the aim of identifying if a noun in GermaNet is a compound or not. The correctness of the splitting into two constituents is considered in the following sections.

7.2 Predicting Immediate Splitting Position

This part of the evaluation regards the splitting position. It is evaluated for all 42191 compounds whether the predicted position at which the algorithms split the compounds into two constituents is correct. An obvious error is, e.g., the splitting of *Tiefkühltruhe* 'deep-freezer' into *tief* + *Kühltruhe* instead of *tiefkühlen* + *Truhe*. An example of an erroneous splitting that is not as obvious is the splitting of *Muskelshirt* 'muscle shirt' into *Muskel* + *Hirt* instead of *Muskel* + *Shirt*. In contrast, the predicted position of the splitting *Bundfaltenhose* 'pleated pants' into *Bundfalten* + *Hose* (instead of *Bundfalte* + *Hose*) is correct, although this example reveals a wrong inflection of the modifier. The evaluation results are presented in Table 3; where the accuracy specifies the number of correctly predicted splitting positions divided by the total number of compounds.

Algorithm	*Correct position*	*Erroneous position*	*Accuracy*
GN-CS	37779	4411	89.54%
SMOR-CS	32863	9326	77.89%
ASV-CS	35407	6783	83.92%
MV	38548	3636	91.38%
WMV	38688	3496	91.71%
CH-CS	40010	2181	94.83%

Table 3: Predicting Immediate Splitting Position

For the task of predicting the immediate splitting position again all combined algorithms outperform the individual compound splitters.

7.3 Prediction of Immediate Constituents

This section evaluates the correctness of the entire prediction of two immediate constituents, including word class and inflection. The predicted constituents for all 42191 compounds are analyzed and the results listed in Table 4. The evaluation takes into account, that for some compounds, there is more than one composition correct. For *Nachtspeicherheizung* 'night storage heater', e.g., two internal groupings are semantically correct: *Nacht* + *Speicherheizung* and *Nachtspeicher* + *Heizung*. In other compounds, two word classes are possible for the modifier. For example, *Spielecke* 'kid's corner' might be composed of *Spiel* + *Ecke* or *spielen* + *Ecke*.

Algorithm	*Correct constituents*	*Erroneous constituents*	*Accuracy*
GN-CS	32738	9449	77.60%
SMOR-CS	31757	10432	75.27%
ASV-CS	31621	10568	74.95%
MV	33349	8832	79.06%
WMV	33176	9005	78.65%
CH-CS	38994	3197	92.42%

Table 4: Prediction of Immediate Constituents

Table 4 reveals that all combined compound splitters outperform the individual compound splitters in the main task of the present project, i.e., in determining immediate constituents of compounds in GermaNet. The best overall result with an accuracy of 92.42% is achieved by the hybrid combined compound splitter CH-CS.

7.4 Error Analysis

To distinguish different cases that cause erroneous predictions of the immediate constituents, the following error types were identified. The occurrences of these error types – presented in Table 5 – are gathered for the combined algorithm CH-CS only as this error classification is done in a manual verification step.

- *Position*: The proposed splitting position is wrong, e.g., *Eislaufbahn* 'ice rink' is split into *Eis* + *Laufbahn* instead of *Eislauf* + *Bahn*.
- *Not parsed*: Some compounds are recognized but not parsed. For example, a compound such as *Kreuzschlitzschraubenzieher* ‚Philips screwdriver', consisting of four parts, is recognized as a compound, but not grouped into its immediate constituents.
- *Wrong lemma*: For some predictions, the lemmatization of the modifier is erroneous. For example, the immediate lemmatized constituents of *Hühnerleiter* 'chicken ladder' are *Huhn* and *Leiter*, but CH-CS splits the compound into *Hühner* + *Leiter* without lemmatizing the modifier.
- *Word class*: The modifier has been assigned a wrong word class. Two different subcases are distinguished:
  1. The proposed word does not exist. For example, *Mischanlage* 'mixing plant' is erroneously split into *Misch* + *Anlage*, but the modifier needs to be the verb *mischen*, because a noun like *Misch* does not exist.
  2. The proposed word (class) has a wrong reading, e.g., the splitting of *Allesschneider* 'slicing maschine' into *All* + *Schneider* instead of *alles* + *Schneider* reveals a wrong reading of the modifier.
- *False negatives*: Those compounds that are erroneously not identified as a compound.

Error type	*CH-CS*
Position	384 (12.01%)
Not parsed	1490 (46.60%)
Wrong lemma	207 (6.47%)
Word class 1	325 (10.17%)
Word class 2	311 (9.73%)
False negatives	297 (9.29%)
Other	183 (5.72%)
Total errors	3197

Table 5: Occurrences of Different Error Types

Two (obvious) causes of errors are identified in Table 6: bound morphemes and missing entries in GermaNet. Bound morphemes such as *Him-* in *Himbeere* 'raspberry' (cf. section 2 above) are a common source of error because the algorithm cannot reliably identify such words. Second, if either the modifier or the head is not in GermaNet, the algorithm may propose a wrong splitting. For example, the correct splitting of *Feincordhose* 'narrow wale corduroy pants' is *Feincord* + *Hose*, but as *Feincord* is not in GermaNet, the algorithm erroneously proposes *fein* + *Cordhose* as those two constituents are entries in GermaNet.

Error type	*Total*	*Bound morpheme*	*No entry in GermaNet*
Position	384	18 (4.7%)	280 (72.9%)
Not parsed	1490	98 (6.6%)	1061 (71.2%)
Wrong lemma	207	7 (3.4%)	150 (72.5%)
Word class 1	325	112 (34.5%)	226 (69.5%)
Word class 2	311	14 (4.5%)	87 (28.0%)
FN	297	15 (5.2%)	153 (51.5%)
Other	183	2 (1.1%)	23 (12.6%)
Total errors	3197	266 (8.3%)	1980 (61.9%)

Table 6: Causes of Errors

A third error type is identified for false positives – actually for 35.3% of all false positives (696 of 1974): Words like *Bausparen* 'building society savings' or *Zusammenprall* 'collision' are frequently treated as compounds, while these are nouns derived from compound verbs.

8 Conclusion and Future Work

Existing tools for splitting compounds were adapted to overcome issues with determining immediate constituents of compounds. Combinatory algorithms using three heterogeneous kinds of compound splitters are developed to achieve better results. As the combined compound split-

ting algorithms all outperform the individual compound splitters, the overall combined result should improve further, including even more individual compound splitters. The best overall result with an accuracy of 92.42% is achieved by a hybrid combined compound splitter that takes into account all knowledge provided by the individual compound splitters, and in addition some domain knowledge about German derivation morphology and compounding.

There are two obvious problems with the used individual compound splitters. First, lemmatized forms are never generated by GN-CS. Extending GN-CS with a lemmatizer to determine base forms can enhance this drawback. Second, the immediate constituents of compounds consisting of more than two or three constituents are not determined by SMOR-CS and ASV-CS, respectively. This issue can be improved through bracketing those compounds by ASV-CS and SMOR-CS.

In future work, we plan to automatically predict compound-internal relations between the now determined immediate constituents by using GermaNet's relations. This would also mean that the immediate compound constituents would have to be automatically disambiguated. Further, an automatic extension of GermaNet with compounds by using statistical information of existing compounds in GermaNet is envisioned.

Acknowledgments

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 our research assistant Sarah Schulz for her help with the evaluation reported in Section 7. Special thanks go to our GermaNet colleague Reinhild Barkey for extensive discussions on the syntax and semantics of compounds and on their modeling in GermaNet.

References

Marco Baroni, Johannes Matiasek, and Harald Trost. 2002. Predicting the Components of German Nominal Compounds. *Frank van Harmelen (eds.), Proceedings of the 15th European Conference on Artificial Intelligence (ECAI)*, Amsterdam: IOS Press, 470-474.

Mathias Creutz and Krista Lagus. 2005. Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0. *Publications in Computer and Information Science*, Report A81. Helsinki University of Technology Helsinki, Finland.

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

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

Radu Florian and David Yarowsky. 2002. Modeling consensus: classifier combination for word sense disambiguation. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing (EMNLP '02),* Vol. 10. Association for Computational Linguistics, Stroudsburg, PA, USA, 25-32.

Mariikka Haapalainen and Ari Majorin. 1994. *GERTWOL: Ein System zur automatischen Wortformerkennung Deutscher Wörter*. Technical report, Lingsoft Inc. https://files.ifi.uzh.ch/cl/volk/LexMorphVorl/Lexikon04.Gertwol.html

Verena Henrich and Erhard Hinrichs. 2010. GernEdiT – The GermaNet Editing Tool. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, main conference. Valletta, Malta.

Claudia Kunze and Lothar Lemnitzer. 2002. GermaNet – representation, visualization, application. In *Proceedings of LREC 2002*, main conference, Vol V. pp. 1485-1491.

Stefan Langer. 1998. Zur Morphologie und Semantik von Nominalkomposita. *Tagungsband der 4. Konferenz zur Verarbeitung natürlicher Sprache*, KONVENS, pp. 83–97.

Robi Polikar. 2006. Ensemble based systems in decision making. In *IEEE Circuits and Systems Magazine*, 16(3):21–45.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection. In *Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 2004)*, p. 1263-1266, Lisbon, Portugal.

Hans van Halteren, Walter Daelemans and Jakub Zavrel. 2001. Improving accuracy in word class tagging through the combination of machine learning systems. In *Compuational Linguistics*, 27, 2, 199-229.

Piek Vossen. 2002. EuroWordNet General Document. EuroWordNet Project LE2-4003 & LE4-8328 report, Version 3, Final, University of Amsterdam.

Hans F. Witschel and Chris Biemann. 2005. Rigorous dimensionality reduction through linguistically motivated feature selection for text categorisation. In *Proceedings of NODALIDA 2005*, Joensuu, Finland.