Annotating Semantic Relations in German Noun-Noun Compounds

Corina Dima, Verena Henrich, Erhard Hinrichs, Christina Hoppermann, Yannick Versley

Department of Linguistics and SFB 833, University of Tübingen, Germany

{firstname.lastname}@uni-tuebingen.de

1 Introduction

The semantic interpretation of individual words is an essential ingredient in text understanding. In many languages, the inventory of simplex words is relatively stable, which allows their meanings to be listed in a dictionary. This stability is reflected, among other things, by the fact that most novel words entering, for example, the German language, are complex words such as the compound *Herzbrötchen* 'heart roll'. This finding has been corroborated by the long-term corpus study Wortwarte (Lemnitzer, 2011), which has been recording all German neologisms for more than a decade.

Compounding is a word formation process that is ongoing and pervasive: Baroni et al. (2002) report that almost half (47%) of the word types in the APA German news corpus are compounds. The meaning of novel compounds is often not entirely predictable from the meanings of their constituent parts (i.e., modifier and head). At the same time, the construction and interpretation of new compounds often rely on semantic similarities with existing compounds involving either a similar head or a similar modifier.

The purpose of the present paper is to report on the construction of a data set containing German noun-noun compounds annotated with internal semantic relations. It will discuss the principles underlying the annotation scheme and will present the results of an inter-annotator agreement study that validates the reliability of the annotation scheme.

2 Related Work

The interpretation of nominal compounds has received considerable attention in both theoretical and computational linguistics. Nominal compounds are an intriguing linguistic phenomenon: syntactically, they are a conglomerate of simple tokens; semantically, they often express more than just the separate meanings of their constituents. Interpreting noun compounds has often been deemed to be a difficult task in the literature (Spärck Jones, 1983; Ryder, 1994). The difficulty stems mainly

from the lack of a generally accepted linguistic theory that clarifies the way compounds are created and interpreted, while accounting for all the possible cases. Nevertheless, significant efforts were made towards creating annotation schemes that use labels to encode common formation patterns. Broadly speaking, two types of annotation schemes have been used in the literature: (i) paraphrase-based inventories such as Levi (1978) and Lauer (1995), which try to capture the meaning of compounds in terms of prepositional or verbal paraphrases, and (ii) ontology-based inventories such as Girju et al. (2005) and Ó Séaghdha (2008), which classify the meaning of compounds by ontological category labels. Both approaches have not remained without criticism. The ontology-based approaches often rely on an intuitive, pre-theoretical understanding of the category labels involved and do not provide necessary and sufficient conditions for choosing one category over the other. The strength of the paraphrase-based approaches lies in the naturalness of the paraphrase task for native speakers. However, this strength is also a weakness because one compound can have multiple paraphrases with unclear criteria for choosing the best one.

3 Annotating German Compounds

This section introduces an annotation scheme that attempts to combine the relative strengths of both the ontology- and the paraphrase-based approaches. It makes use of a set of currently 37 properties and 17 prepositions for the interpretation of German noun-noun compounds. Annotated compounds are assigned a combined label, typically one property and one preposition. The combined approach is motivated by the observation that the semantic relation that holds between the constituents of a compound is better identified by the correlation between a property and a preposition. This observation is detailed in the next paragraphs.

The annotation was performed on a per head basis: each set of compounds with the same head was analyzed and grouped semantically. By applying this method to a pilot set of compounds, an initial set of properties was obtained. This set was further refined and enlarged iteratively through the subsequent annotation of additional compounds, and contains to date 37 properties.

Table 1 illustrates the semantic groupings for the head *Haus* ‘house’. It displays four semantic properties that prototypically connect the head of the compound with its modifier. *Material*, *user*, *use*, and *location* represent a subset of properties that a building, such as a house, can have.

In parallel with the property-based annotation illustrated above, a preposition-based annotation was performed that for each head typically associates a given preposition with one property. This association is illustrated in Table 1 where three of the four properties are annotated with exactly one preposition while the remaining property (*Lokation* ‘location’) is associated with multiple prepositions. For instance, *location* is used with the following prepositions: *in* ‘in’, *an* ‘on’, and *auf* ‘in’. These prepositions serve to further specify the spatial arrangement of the objects denoted by the modifier and by the head of a compound: For example, *Baumhaus* ‘tree house’

refers to a house that is located *in* (German *in*) a tree whereas *Eckhaus* 'corner house' signifies a house that is located *on* (German *an*) the corner of a street.

Compound	Translation	Property	Preposition
Holzhaus	'wooden house'	*Material* 'material'	*aus* 'of'
Schneehaus	'igloo', lit. 'snow house'		
Steinhaus	'stone house'		
Armenhaus	'poor house'	*Nutzer* 'user'	*für* 'for'
Gästehaus	'guest house'		
Waisenhaus	'orphanage', lit. 'orphan house'		
Auktionshaus	'auction house'	*Verwendung* 'use'	*für* 'for'
Geburtshaus	'birth house'		
Konzerthaus	'concert house'		
Baumhaus	'tree house'	*Lokation* 'location'	*in* 'in'
Eckhaus	'corner house'		*an* 'on'
Landhaus	'country house'		*auf* 'in'

Table 1. Semantic grouping of compounds with the head Haus 'house'.

There are also cases when the same preposition combines with more than one property. For instance, the preposition *für* 'for' occurs with the properties *use* and *user*. This one-to-many mapping can be explained by the fact that the set of prepositions is outnumbered by the number of possible properties.

The joint annotation using prepositions and properties combines the relative strengths of the paraphrase- and the ontology-based approaches. The correlation between a preposition and a property facilitates the pairwise disambiguation of these two aspects of meaning, thus ensuring the consistency of the annotation. At the time of writing this paper, 4359 German noun-noun compounds have been manually annotated using the annotation scheme described in this section.

The compound data set was obtained by extracting compounds headed by concrete nouns from the German wordnet *GermaNet* (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010). All compounds in GermaNet have been split into their constituent parts (Henrich and Hinrichs, 2011), i.e., modifier and head. The particular choice of head nouns was based on an earlier list by Melinger and Weber (2006). This list is organized by semantic categories such as buildings, clothing, food, furniture, weapons, etc. and contains for each category a sample set of concrete nouns that fall under that category.

4 Agreement Study

In order to validate the reliability of the combined property and preposition annotation, an inter-annotator agreement (IAA) study was conducted. The data for this study was collected by randomly selecting nouns from GermaNet that belong to the categories of concrete nouns identified by Melinger and Weber (2006) – see Section 3. A total of 500 nominal compounds headed by these concrete nouns were

then extracted from GermaNet. This set of compounds was used for the IAA study presented in this section.

To aid in the annotation of the data set, written guidelines were given to two student annotators who are native Germans and who performed the annotation independently. They had been previously trained on the compound annotation task, but had never seen any of the compounds in the current study.

The annotation was performed on a per head basis. The task consisted of assigning a property and a preposition label to each compound whenever possible. For strongly lexicalized compounds such as *Eselsbrücke* (‘mnemonic‘, literally: ‘donkey bridge’), it is impossible to capture the relationship between the head and the modifier with a property or a preposition. In such cases, annotators were instructed to mark the compound as lexicalized. Otherwise, annotators were asked to assign exactly one property. With preposition labels, annotators had three options depending on the particular compound under consideration: to assign exactly one, more than one, or no preposition at all.

The IAA was computed separately for the property annotation and the preposition annotation. The reported numbers reflect uncorrected annotations, as produced by the student annotators. The property annotation resulted in a percentage of agreement of 76.4% and a Cohen’s Kappa score of 0.74, which corresponds to a *substantial* agreement according to the classification of Kappa coefficients proposed by Landis and Koch (1977). For the preposition annotation a percentage of agreement of 79.5% and a Kappa score of 0.75 were obtained for all instances where exactly one preposition was assigned (96.4% of the compounds). Since in some cases the annotators chose more than one preposition, a Dice score of 0.79 was computed to measure both complete and partial agreement. It is noteworthy that the amount of agreement is roughly the same for both property and preposition labeling. We conjecture that this similar agreement is due to the parallel annotation as the property labeling helped to disambiguate the preposition labeling and vice versa. Our findings regarding the agreement levels for the preposition and property labels are in stark contrast with the IAA results by Girju et al. (2005). In a similar two-label annotation experiment, they report a Kappa of 0.80 for annotation with the 8 prepositions proposed by Lauer (1995) and 0.58 for the annotation with their inventory of 35 semantic relations.

In order to understand the differences in the individually performed annotations, a disagreement analysis was performed. In more than 60% of all disagreements, the annotators disagree on both the property and the preposition. More often than not, these are cases where the compounds to be annotated are genuinely ambiguous and where the annotators annotated different senses of the compound. A typical example is the compound *Frauenkalender*, which can either refer to a calendar produced for a female audience or to a calendar with pictures of women.

The remaining disagreements are more or less equally divided between (i) cases where both annotators agree on the property but disagree on the prepositional paraphrase and (ii) cases where they disagree on the properties but agree on the preposition. A typical example for (i) is *Sahnejoghurt* ‘cream yoghurt’. Both

annotators chose the property *Zutat* ‘ingredient’ but two different prepositions: *aus* ‘from’ and *mit* ‘with’. The annotation guidelines state that both prepositions are candidate prepositions for this property and that annotators should choose *aus* if the modifier refers to the sole ingredient and that that they should choose *mit* if the modifier refers to one of several ingredients. In cases like *Sahnejoghurt*, it is very hard to decide this matter since in principle it could refer to a yoghurt consisting mainly of cream or of several ingredients.

The disagreements in case (ii) signal genuine annotation errors. For example, in *Kokosnussmilch* ‘coconut milk’ one of the annotators chose the property *Herkunft* ‘origin’ while the other one assigned the property *Zutat* ‘ingredient’. The correct label in this example is *Zutat* because the coconut milk is obtained by processing the grated coconut and does not refer to the liquid that is naturally contained in a coconut. The preposition agreement does not represent an erroneous preposition assignment made by the annotators but rather underlines the inherent ambiguity of the prepositions that can be used with more than one property (*aus* is the prototypical preposition for both *Herkunft* and *Zutat*).

5 Conclusion and Future Work

This paper introduced an annotation scheme for the internal semantic relations of concrete German noun-noun compounds. The substantial score of the inter-annotator agreement study shows that a combined approach using both property- and preposition-based annotations reliably disambiguates the compound-internal relation.

The motivation for starting with the class of concrete head nouns is that their associated properties are relatively easy to identify and therefore also easy to annotate. In future work, we plan to extend the coverage of the dataset by including head nouns that do not refer to concrete objects. It will be an interesting question to what extent the current set of properties can cover the relations for abstract nouns and in what ways it has to be extended.

Acknowledgements

We are very grateful to our student assistants Kathrin Adlung, Nadine Balbach, and Tabea Sanwald, who helped us with the annotations reported in this paper. Financial support was provided by the German Research Foundation (DFG) as part of the Collaborative Research Center ‘Emergence of Meaning’ (SFB 833) and by the German Ministry of Education and Technology (BMBF) as part of the research grant CLARIN-D.

References

Baroni, M., J. Matiasek, & H. Trost (2002) Predicting the Components of German Nominal Compounds. In F. van Harmelen, ed., *Proceedings of the 15th European Conference on Artificial Intelligence (ECAI)*, IOS Press, Amsterdam, 470–474.

Girju R., D. Moldovan, M. Tatu, & Daniel Antohe (2005) On the Semantics of Noun Compounds. In A. Villavicencio, F. Bond, and D. McCarthy, eds., *Journal of Computer Speech and Language – Special Issue on Multiword Expressions,* 19(4):479–496.

Hamp, B. & H. Feldweg (1997) GermaNet - a Lexical-Semantic Net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid.

Henrich, V. & E. Hinrichs (2010) GernEdiT – The GermaNet Editing Tool. In *Proceedings of LREC 2010*, Valetta, Malta, 2228–2235.

Henrich, V. & E. Hinrichs (2011) Determining Immediate Constituents of Compounds in GermaNet. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2011)*, Hissar, Bulgaria, 420-426.

Landis, J. R. & G. Koch (1977) The measurement of observer agreement for categorical data. *Biometrics*, 33 (1):159–174.

Lauer, M. (1995) *Designing Statistical Language Learners: Experiments on Compound Nouns*. Ph.D. thesis, Macquarie University.

Lemnitzer, L. (2011) Making sense of nonce words. In M. Heidemann Andersen & J. Nörby Jensen, eds., *Nye Ord*, 7–18.

Levi, J. N. (1978) *The Syntax and Semantics of Complex Nominals*. Academic Press, New York.

Melinger, A. & A. Weber (2006) Database of noun associations for German, http://www.coli.uni-saarland.de/projects/nag/ (accessed October 11, 2013).

Ryder, M. E. (1994) *Ordered Chaos: The Interpretation of English Noun-Noun Compounds*. University of California Press, Berkley, CA.

Ó Séaghdha, D. (2008) *Learning Compound Noun Semantics*. Ph.D. thesis, Computer Laboratory, University of Cambridge.

Späck Jones, K. (1983) Compound Noun Interpretation Problems. In F. Fallside & W. A. Woods, eds., *Computer Speech Processing*, Prentice-Hall, NJ.