# How to Tell a *Schneemann* from a *Milchmann*:
# An Annotation Scheme for Compound-Internal Relations

## Corina Dima, Verena Henrich, Erhard Hinrichs, Christina Hoppermann

Collaborative Research Center 833 and Department of Linguistics, University of Tübingen

Wilhelmstrasse 19, 72074 Tübingen, Germany

firstname.lastname@uni-tuebingen.de

## Abstract

This paper presents a language-independent annotation scheme for the semantic relations that link the constituents of noun-noun compounds, such as *Schneemann* 'snow man' or *Milchmann* 'milk man'. The annotation scheme is hybrid in the sense that it assigns each compound a two-place label consisting of a semantic property and a prepositional paraphrase. The resulting inventory combines the insights of previous annotation schemes that rely exclusively on either semantic properties or prepositions, thus avoiding the known weaknesses that result from using only one of the two label types. The proposed annotation scheme has been used to annotate a set of 5112 German noun-noun compounds. A release of the dataset is currently being prepared and will be made available via the CLARIN Center Tübingen. In addition to the presentation of the hybrid annotation scheme, the paper also reports on an inter-annotator agreement study that has resulted in a substantial agreement among annotators.

Keywords: noun compounds, annotation scheme, compound-internal semantic relations

## 1. Introduction

Compounding is a pervasive linguistic phenomenon across different a a a d . La (1995) points out that English noun compounds occur with high frequency in different text genres including fiction, press, and technical materials. This observation is not restricted to English. Baroni et al. (2002) report that almost half (47%) of the word types in the APA German news corpus are compounds. Moreover, Lemnitzer (2011) has shown that the overwhelming majority of German neologisms are compounds. This makes it all but impossible to exhaustively list their meanings in a dictionary. Instead, the interpretation of novel compounds has to be performed dynamically as new compounds are coined (Lauer, 1995).

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. This combination of indeterminacy of meaning and having to generalize from inherent underlying semantic patterns makes the interpretation of compounds a natural candidate for a machine learning approach. Due to the prevalent and dynamic nature of compounding, such a machine learning treatment of compounds is, in fact, crucial for a variety of natural language processing applications including text understanding, machine translation, information retrieval, and natural language generation.

The purpose of the present paper is to report on the construction of a language-independent, hybrid annotation scheme for compound-internal relations. The scheme has been instantiated for German, and used for annotating a dataset of German noun-noun compounds. The resulting annotations can be used as a resource in the automatic interpretation of German compounds. The paper discusses the principles underlying the annotation scheme and presents the results of an inter-annotator agreement study that evaluates the consistency of the annotations created using the annotation scheme.

## 2. Related Work

The interpretation of nominal compounds has received considerable attention in both theoretical and computational linguistics. 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) property-based inventories such as Girju et al. (2005) and Ó Séaghdha (2008), which classify the meaning of compounds by semantic category labels. Both approaches have not remained without criticism. The property-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. Moreover, prepositional paraphrases are often ambiguous themselves. This has the unwelcome consequence of blurring the underlying patterns of regularity of meaning that machine learning approaches require as the empirical basis for capturing suitable generalizations inherent in the data.

## 3. Annotation Scheme

The main contribution of this paper is to introduce a language-independent, hybrid annotation scheme that attempts to combine the relative strengths of both the property- and the paraphrase-based schemes. The motivation for combining semantic properties with prepositional paraphrases can be illustrated for German by considering the set of compounds involving the concrete noun *Haus* 'house' presented in Table 1. They illustrate the range of modifiers that this head can have and the diverse set of semantic properties and prepositional paraphrases that need to be assigned. For instance, *material*, *user*, *usage*, *location*, and *goods* represent a subset of the prototypical properties that a building, such as a *house*, can have.

A straightforward way of conveying the meaning of each of these compounds is to annotate them with the preposition that best characterizes the semantic relation between the head and the modifier. While such prepositional paraphrases are natural and intuitive, this type of annotation assigns in some cases the same preposition to compounds where the modifier serves different functions with respect to the same head. For example, the compound *Gästehaus* 'guest house' refers to a building whose intended users are guests. In contrast, the compound *Autohaus*

'car dealership' does not refer to a building whose intended users are cars but rather to a building that is used for displaying and selling goods. Still, both compounds are annotated with the preposition *for*. The differences in interpretation are reflected in our annotation scheme by the use of different semantic properties, namely *user* for *guest house* and *goods* for *car dealership*. Examples like these justify the hybrid nature of our annotation scheme: semantic properties that name the relation between the head and the modifier are required to further specify the meaning of the prepositions. Thus, the correlation between a preposition and a property facilitates the pairwise disambiguation of these two aspects of meaning. This annotation mechanism also yields more consistent annotations. As seen in Table 1, compounds that are annotated with the same property will typically be associated with the same preposition or the same set of prepositions, in the context of the same head. For example, the properties *material* and *user* are each associated with one preposition (*of* and *for*, respectively), while the property *location* is associated with three different prepositions that further specify the special arrangement of the objects denoted by the head. 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. Conversely, prepositions can be associated with more than one property so that the property serves to further disambiguate the meaning of the preposition. As mentioned above, the preposition *for*, which is associated with three different properties (see Table 1), is a case in point. This one-to-many mapping can be explained by the fact that the set of prepositions is outnumbered by the number of possible properties.

| Compound | Translation | Property | Preposition |
|---|---|---|---|
| *Holzhaus* | 'wooden house' | | |
| *Schneehaus* | 'igloo', lit. 'snow house' | *material* | *aus* 'of' |
| *Steinhaus* | 'stone house' | | |
| *Armenhaus* | 'poor house' | | |
| *Gästehaus* | 'guest house' | *user* | *für* 'for' |
| *Waisenhaus* | 'orphanage', lit. 'orphan house' | | |
| *Auktionshaus* | 'auction house' | | |
| *Geburtshaus* | 'birth house' | *usage* | *für* 'for' |
| *Konzerthaus* | 'concert house' | | |
| *Autohaus* | 'car dealership', lit. 'car house' | | |
| *Möbelhaus* | 'furniture store', lit. 'furniture house' | *goods* | *für* 'for' |
| *Modehaus* | 'fashion house' | | |
| *Baumhaus* | 'tree house' | | *in* 'in' |
| *Eckhaus* | 'corner house' | *location* | *an* 'on' |
| *Landhaus* | 'country house' | | *auf* 'in' |

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

| Preposition | Translations | Examples |
|---|---|---|
| *als* | 'as' | *Hausboot* 'boat *as* a house', *Signalflagge* 'flag *as* a signal' |
| *an* | 'on', 'in', 'at' | *Deckenbild* 'painting *on* the ceiling', *Abendessen* 'meal *in* the evening' |
| *auf* | 'on' | *Berghütte* 'hut *on* the mountain' |
| *aus* | 'from', 'of' | *Meerwasser* 'water *from* the sea', *Blätterhaufen* 'pile *of* leaves' |
| *bei* | 'at', 'near' | *Javasee* 'sea *near* Java' |
| *durch* | 'by', 'through' | *Satellitenbild* 'picture taken *by* a satellite', *Dorfstraße* 'road *through* village' |
| *für* | 'for' | *Kinderbuch* 'book *for* children', *Tennisball* 'ball *for* tennis' |
| *in* | 'in' | *Grundwasser* 'water *in* the ground', *Herbstblatt* 'leaf *in* autumn' |
| keine Präposition | 'no preposition' | *Lachsfisch* 'salmon fish', *Humuserde* 'humus soil' |
| *mit* | 'with' | *Gurkensalat* 'salad *with* cucumbers', *Obstbaum* 'tree *with* fruits' |
| *nach* | 'after' | *Maßanzug* 'suit tailor-made *after* measurements' |
| *über* | 'above', 'over', 'across', 'about' | *Sportzeitung* 'magazine *about* sports', *Straßenbrücke* 'bridge *across* a street', *Internetradio* 'radio *over* internet' |
| *um* | 'around', 'at' | *Stadtmauer* 'wall *around* a city', *Mittagssonne* 'sun *at* midday' |
| *unter* | 'below', 'under' | *Dachboden* 'floor *under* the roof' |
| *von* | 'from', 'of' | *Bauernbrot* 'bread *from* farmers', *Hausbesitzer* 'owner *of* a house' |
| *wie* | 'like' | *Marmorkuchen* 'cake *like* marble', *Kirschtomate* 'tomato *like* a cherry' |
| *zu* | 'at', 'to' | *Gartentür* 'door *to* the garden', *Geburtstagskuchen* 'cake *at* a birthday' |

Table 2: Preposition-based labels (for the German language): 17 prepositions.

Tables 2 and 3 present an overview of the inventory utilized in the annotations presented in this paper. The semantic properties in the inventory denote prototypical properties that are associated with a particular head and are assumed to be language-independent. The prepositional paraphrases, on the other hand, are language-specific and therefore have to be instantiated each time the annotation scheme is applied to a new language. With 37 semantic properties and, in its instantiation for German, 17 prepositional paraphrases[1], the proposed inventory is comparable in terms of size with the taxonomy described by Tratz and Hovy (2010), which uses 43 fine-grained relations. We expect, however, that the automatic prediction task will benefit from the disambiguation potential of the combined annotation scheme. The properties in Table 3 were derived in a data-driven fashion by considering a specific set of head nouns[2], rather than making use of an existing inventory of semantic categories.

Using an existing inventory of semantic categories runs the danger of having to fit the data to semantic distinctions that may not be suitable for the range of data considered. As Tratz and Hovy (2010) have pointed out, the heterogeneity of past approaches to the semantics of compounding poses the additional challenge of having to choose an inventory that was based on rather different basic assumptions. Moreover, our hybrid approach of pairing prepositional paraphrases with semantic properties requires a particular set of semantic properties. Their purpose is to further disambiguate the prepositional paraphrases within the scope of each head noun.

## 4. Dataset

This section introduces the dataset that was used to derive the annotation scheme described in Section 3. The compound dataset was obtained by extracting compounds headed by concrete nouns from the German wordnet *GermaNet* (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010). The added benefit of extracting the compounds from GermaNet is two-fold: (i) The compounds contained in GermaNet have already been split into their constituent parts (Henrich and Hinrichs, 2011). Each constituent is labeled with its parts of speech and with its function, i.e., whether it is a modifier or a head. (ii) Since the words in GermaNet are interconnected by lexical and semantic relations, such information can be harvested for the automatic analysis of compounds.

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. The motivation for starting

---

[1] It is debatable whether *als* 'as' and *wie* 'like' have the grammatical status of a preposition. Many reference grammars and dictionaries classify *als* as a conjunction and *wie* as a comparative particle only. Their inclusion in the preposition-based annotation set is motivated by the fact that *als* 'as' provides the most natural paraphrase for the property *function* and *wie* 'like' plays the same role for the properties *appearance* and *shape*.

[2] Further details about the particular dataset that was used to derive the present annotation scheme are given in Section 4.

| Property | Examples |
|---|---|
| *access*[-1] | *Kellertreppe* 'basement stairs', *Gartentür* 'garden gate' |
| *activity* | *Ausflugsboot* 'excursion boat', *Laufschuhe* 'running shoes' |
| *appearance* | 'marble cake', *Silberfuchs* 'silver fox' |
| *cause* | *Regenbogen* 'rainbow' |
| *cause*[-1] | *Schimmelpilz* 'mold fungus', *Gallmücke* 'gall gnat' |
| *component* | *Duftöl* 'fragrance oil', *Chlorwasser* 'chlorine water' |
| *consistency* | *Panzerglas* 'bullet-proof glass' |
| *construction method* | *Blockhaus* 'log house' (lit. 'block house'), *Reihenhaus* 'row house' |
| *content* | *Sportzeitung* 'sports magazine' |
| *diet* | *Ameisenbär* 'anteater' (lit. 'ant bear'), *Mäusebussard* '(mice) buzzard' |
| *eponym* | *Sachertorte* 'Sacher cake', *Waldorfsalat* 'Waldorf salad' |
| *function* | *Hausboot* 'house boat', *Signalflagge* 'signal flag' |
| *goods* | *Blumenladen* 'flower shop', *Schuhfabrik* 'shoe factory' |
| *hyponym* | *Lachsfisch* 'salmon fish', *Humuserde* 'humus soil' |
| *ingredient* | *Gurkensalat* 'cucumber salad', *Orangensaft* 'orange juice' |
| *location* | *Berghütte* 'mountain hut', *Stadtmauer* 'city wall' |
| *manner of functioning* | *Atombombe* 'atomic bomb', *Gasherd* 'gas stove' |
| *material* | *Holzlöffel* 'wooden spoon', *Hornbrille* 'horn-rimmed glasses' |
| *measure* | *Literflasche* 'litre bottle' |
| *measure*[-1] | *Cholesterinspiegel* 'cholesterol level' |
| no property | *Eselsbrücke* 'mnemonic', *Meerkatze* 'subspecies of monkeys' (lit. 'sea cat') |
| *occasion* | *Geburtstagsgeschenk* 'birthday present', *Abendkleid* 'evening dress' |
| *origin* | *Bauernbrot* 'farmer's bread', *Meerwasser* 'sea water' |
| *origin*[-1] | *Erdbebenherd* 'epicentre' (lit. 'earthquake focus') |
| *owner* | *Bäckerladen* 'baker's shop', *Staatsbank* 'state bank' |
| *owner*[-1] | *Hundebesitzer* 'dog owner', *Wohnungseigentümer* 'flat owner' |
| *part* | *Obstbaum* 'fruit tree', *Propellerflugzeug* 'propeller plane' |
| *part*[-1] | *Kinderhand* 'child's hand', *Ordensritter* 'knight of an order' |
| *place of use* | *Gartenstuhl* 'garden chair', *Wandkalender* 'wall calendar' |
| *production method* | *Pfannkuchen* 'pancake', *Satellitenbild* 'satellite picture' |
| *shape* | *Kopfsalat* '(head) lettuce', *Kirschtomate* 'cherry tomato' |
| *shape*[-1] | *Erdball* 'Earth planet' (lit. 'earth ball'), *Eisberg* 'iceberg' (lit. 'ice mountain') |
| *storage* | *Taschentuch* '(pocket) handkerchief', *Taschenlampe* 'flashlight' (lit. 'pocket lamp') |
| *time point* | *Abendessen* 'evening meal', *Mittagssonne* 'midday sun' |
| *time span* | *Saisonarbeiter* 'seasonal worker', *Wochenzeitung* 'weekly newspaper' |
| *usage* | *Motoröl* 'motor oil', *Tennisball* 'tennis ball' |
| *user* | *Kinderbuch* 'children's book', *Maurerhammer* 'stonemason hammer' |

Table 3: Property-based labels: 37 properties. Note that the label *no property* is typically assigned for strongly lexicalized compounds – see Section 6 for more details.

with the class of concrete head nouns is that their associated properties are relatively easy to identify and therefore also to annotate.

At the time of writing this paper, 5112 German noun-noun compounds have been manually annotated. The annotation was performed on a per head basis: Each set of compounds with the same head was analyzed and grouped semantically as illustrated for the head *Haus* 'house' in Section 3. All annotated compounds correspond to 251 head nouns, resulting in an average of 20 compounds per head noun.

Table 4 serves as an overview of the distribution of properties used to annotate the dataset, with the most frequent property being *usage* (22.9%). The two-place properties in our inventory typically connect the modifier as the first argument of the relation with the head of the compound, which occupies the second argument of the relation. In the case of *Apfelbaum* 'apple tree', the property *part* encodes the fact that the apple is part of the tree. However, in some cases it is more natural to invert such a relation and switch the order of the arguments. Such inverse relations are marked by the *-1* superscript as in *Fingernagel* 'finger nail', annotated as *part*[-1], meaning that the nail is part of the finger.

In terms of size, our dataset is comparable with English datasets surveyed by Tratz and Hovy (2010),

and is, to the best of our knowledge, the largest German noun-noun compound dataset annotated with compound-internal relations.

| Property | | Preposition | |
|---|---|---|---|
| *usage* | 22.9% | *für* | 100% |
| *part* | 10.6% | *mit* | 81% |
| | | *aus* | 19% |
| *part*[-1] | 6.8% | *von* | 100% |
| no property | 6.3% | no preposition | 96% |
| | | *von* | 3% |
| | | *für* | 1% |
| *location* | 6.0% | *in* | 53% |
| | | *an* | 22% |
| | | *auf* | 13% |
| | | *über* | 4% |
| | | *bei* | 4% |
| | | *um* | 2% |
| | | *unter* | 2% |
| *ingredient* | 5.9% | *mit* | 56% |
| | | *aus* | 43% |
| | | *von* | 1% |
| *material* | 5.7% | *aus* | 90% |
| | | *mit* | 10% |
| *user* | 4.0% | *für* | 100% |
| *function* | 3.4% | *als* | 100% |
| *origin* | 3.4% | *aus* | 66% |
| | | *von* | 34% |
| *place of use* | 3.3% | *für* | 53% |
| | | *in* | 27% |
| | | *an* | 10% |
| | | *auf* | 10% |
| *owner* | 3.2% | *von* | 100% |
| *manner of functioning* | 2.8% | *mit* | 66% |
| | | no preposition | 18% |
| | | *für* | 6% |
| | | *wie* | 5% |
| | | *über* | 5% |
| *shape* | 2.4% | *wie* | 100% |
| *content* | 2.2% | *über* | 57% |
| | | *von* | 40% |
| | | *mit* | 3% |
| *appearance* | 1.7% | *wie* | 100% |
| *eponym* | 1.2% | *von* | 96% |
| | | *nach* | 4% |
| *activity* | 1.0% | *für* | 100% |
| *time point* | 1.0% | *an* | 56% |
| | | *in* | 40% |
| | | *zu* | 2% |
| | | *um* | 2% |

Table 4: Properties and prepositions[3] per property in the gold standard.

---

[3] The table disregards properties with less than 1% occurrences. See Table 3 for a complete list.

This dataset has been designed to support the automatic prediction of the implicit compound-internal semantic relations by using machine learning approaches. These relations can be predicted with the help of distributional representations of the simplex words that constitute the compound. As illustrated in Section 3, the correlation between a prepositional paraphrase and a semantic property facilitates the pairwise disambiguation of these two aspects of meaning. This approach enables the reuse of correlations learned for a simplex word in the context of a compound when that particular word occurs in other compounds.

## 5. Handling Ambiguity

Apart from settling on a well-motivated annotation scheme, the question of how to treat possible ambiguities in the interpretation of compounds needs to be addressed in a principled way. In the process of constructing an annotated dataset of German nominal compounds, we have encountered different types of ambiguities that can be attributed to different origins. The subcases are as follows:

- Differences in the interpretation due to genuine ambiguities in the semantics of the modifier and/or the head. These are cases where either the modifier or the head has two distinct readings. For example, the head *Welle* in the compound *Wasserwelle* can either refer to a wave in a body of water or to curly hair. Accordingly, *Wasserwelle* can then refer to a wave that consists of water or to curly hair produced by the use of water. In the case of *Bogengitarre*, the ambiguity comes from the different readings of the modifier *Bogen*: either a bow used to play instruments or an arc as a shape. That is, a *Bogengitarre* can either refer to a guitar that is played with a bow or to a guitar that has its riff shaped like an arc.
- Differences in the interpretation due to indeterminacy of plausible relations. The compound *Fischfrau* (Heringer, 1984) is a good example of this kind. It can either refer to a woman selling fish, a woman that looks like a fish, or even a female fish. Similarly, *Geburtshaus* 'birth house' (see Table 1) can be interpreted either as the house a famous person was born in or as a clinical facility where women give birth.

In the current annotation task, compounds are analyzed without taking context into account. This results in annotators picking the interpretation that appears most salient to them. In some cases, however, annotators are aware of ambiguities of the kind just described and will comment on other plausible interpretations of a compound.

For the time being, we concentrated exclusively on

the identification of the appropriate property and preposition, but did not specify different word senses of either modifier or head in the annotated compounds. Consider, for instance, the contrast between *Milchmann* 'milk man' and *Schneemann* 'snow man' in the title of our paper: it is only for the compound *milk man* that the head refers to the main sense of the noun, which is a male adult.

## 6. Inter-Annotator Agreement (IAA)

In order to validate the consistency of the combined annotation of property and preposition labeling, an IAA study was conducted. A total of 500 nominal compounds were chosen for this study. These compounds represent a balanced dataset whose head constituents are taken from the categories of concrete nouns identified by Melinger and Weber (2006) – see Section 4. To aid in the annotation of the dataset, 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') or *Meerkatze* (which refers to a subspecies of monkeys but whose literal meaning is 'sea cat'), it is impossible to capture the relationship between the head and the modifier with a property or a preposition. In such cases, annotators are instructed to mark the compound as lexicalized. Otherwise, annotators are asked to assign exactly one property. With preposition labels, annotators have three options depending on the particular compound under consideration: to assign exactly one, more than one, or no preposition at all. An experienced lexicographer inspected and in some cases post-corrected all candidate annotations and adjudicated cases of disagreement in order to arrive at a gold standard.

The IAA was computed separately for the property annotation and the preposition annotation. The property annotation resulted in a Cohen's Kappa (Cohen, 1960) score of 0.74[4], which corresponds to a *substantial* agreement according to the classification of Kappa coefficients proposed by Landis and Koch (1977). For the preposition annotation a Kappa score of 0.75 was obtained for all instances where exactly one preposition was assigned. Since in some cases the annotators chose more than one preposition, a Dice (1945) score of 0.79 was computed to measure both

---

[4] Cohen's Kappa measures the degree of agreement normalized by chance.

complete and partial agreement. The reported scores take into account the initial, uncorrected input of the two student annotators. 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. This finding is in stark contrast to the IAA results by Girju et al. (2005). 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.

It is also instructive to compare the percentages of agreement for the different property labels involved (see Table 5). The agreement ranges from perfect to low agreement and can be correlated with how easily a certain property can be distinguished from all the others.

| Property | IAA |
|---|---|
| *owner*[-1] | 100% |
| *time point* | 100% |
| *ingredient* | 92.86% |
| *material* | 90% |
| *user* | 85.71% |
| *location* | 85.19% |
| *origin* | 82.76% |
| *manner of functioning* | 79.17% |
| *part* | 78.38% |
| *usage* | 75.47% |
| *appearance* | 75% |
| *owner* | 75% |
| *part*[-1] | 74.42% |
| *hyponym* | 66.67% |
| *activity* | 64.29% |
| *shape* | 63.64% |
| *place of use* | 61.54% |
| *function* | 50% |
| no property | 25% |

Table 5: Inter-annotator agreement per property.

For example, the modifier of compounds that are assigned the property *material* refers usually to a substance like *Schnee* 'snow' in *Schneemann* 'snowman', thus making the identification of the property *material* reduce to observing that the modifier belongs to the class of substances. For other properties, the criteria for identifying a relation leave more space for interpretation and therefore result in a lower agreement. This is the case for *Kirschtomate* 'cherry tomato', where *appearance* and *shape* both seem plausible at first glance: The compound can be described as a tomato that has the color of a cherry (property *appearance*) or a tomato that has the shape of a cherry (property *shape*). However, since the head

*Tomate* 'tomato' already entails the red color, only the shape interpretation justifies the creation of a new compound.

In more than 60% of all disagreements (73 instances, see Table 6), the disagreements are genuine in the sense that 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. Typical examples include the compound *Bogengitarre* (see Section 5), which can either refer to a guitar that is played with a bow or to a guitar that has its riff shaped like an arc, *Frauenkalender*, which can either refer to a calendar produced for a female audience or to a calendar with pictures of women, and *Weihnachtsstern*, which can either refer to a plant blooming at Christmas time or the star of Bethlehem. Such examples thus do not point to a weakness in the annotation scheme but rather to a need to interpret compounds in their context of use (see Section 7 for more details).

| Disagreement in: | # of instances |
|---|---|
| Both property and preposition | 73 |
| Property only | 45 |
| Preposition only | 39 |
| Total # of property disagreements | 118 |
| Total # of preposition disagreements | 112 |

Table 6: Disagreement overview.

As shown in Table 6, the remaining disagreements are more or less equally divided between cases where both annotators agree on the property but disagree on the prepositional paraphrase (39) and cases where they disagree on the properties but agree on the preposition (45).

Typical disagreements of the former type include *Sahnejoghurt* 'cream yoghurt' and *Wandkalender* 'wall calendar'. In the case of *Sahnejoghurt*, both annotators chose the property *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. In the example of *Wandkalender*, the annotators agreed on the property *place of use*. One annotator chose the preposition *für* 'for', thus highlighting the functional aspect of the compound, while the other annotator chose the preposition *an* 'on', thus putting more emphasis on the location of the object. Such disagreements do not undermine the usage of the property *place of use* and therefore do not constitute sufficient ground for revising the annotation scheme. Rather, we consider such disagreements to be inevitable and motivated by different conceptualizations of the object denoted by the compound.

The remaining source of disagreement concerns cases where the annotators disagree on the properties but agree on the preposition. Such disagreements signal genuine annotation errors as in *Kokosnussmilch* 'coconut milk' where one of the annotators chose the property *origin* while the other one assigned the property *ingredient*. The correct label in this example is *ingredient* 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 *origin* and *ingredient*).

## 7. Conclusion and Future Work

This paper introduced a hybrid, language-independent annotation scheme for the internal semantic relations of noun-noun compounds. The annotation scheme was applied to a set of 5112 German noun-noun compounds derived from concrete head nouns. Starting with the class of concrete nouns is motivated by the fact that their associated properties are relatively easy to identify and therefore also 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. In effect, we have already started this process since some of the concrete head nouns in the present dataset also have senses that refer to abstract entities. Examples of this kind are *Staatsapparat* 'state apparatus' and *Machtapparat* 'power apparatus', where *Apparat* 'apparatus' does not refer to a physical device but rather to an abstract instrument. It will be an interesting question to what extent the current set of properties can cover the relations for non-concrete nouns and how it has to be extended.

Another set of compounds that will require further study are compounds such as *Nadelwald* 'conifer forest' (literally: 'needle forest') and *Metallgewerkschaft* 'metal worker union' (literally: 'metal union'). The semantic relation operable in such compounds no longer connects the two constituent parts of the compound but requires the insertion of an inferred or implied term that serves as an argument of the relation. In the case of *Metallgewerkschaft*, the part-whole relation (i.e., *part* in our inventory) holds between *metal worker* and *union* rather than *metal* and *union*. We refer to such examples as *metonymic*

compounds because their interpretation requires the insertion of a missing argument in order to assign the proper relation.

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. This makes the scheme suitable for the automatic interpretation of nominal compounds using machine learning experiments, which will be conducted in the scope of future work, using different setups.

The first experiments will assess how reliably the prepositions and the properties can be identified separately. These results will be compared with the outcome of an experiment that automatically assigns the combined property/preposition label to each compound in order to verify our conjecture that such joint labeling will yield better results.

As discussed in Section 5, the task of assigning semantic relations to compounds does not take into account the particular linguistic context in which a compound is used. This is a clear shortcoming of the present approach, albeit one shared with many previous annotation efforts. In future work, we plan to overcome this shortcoming by designing a task for collecting appropriate context examples for all annotated compounds. The collected contexts can then be used in machine learning experiments by mining the contexts for appropriate features that can aid in the disambiguation of compound-internal relations.

## 8. Acknowledgements

## 9. References

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

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), pp. 37–46.

Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3), pp. 297–302.

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

Hamp, B., Feldweg, H. (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., Hinrichs, E. (2010). GernEdiT - The GermaNet Editing Tool. In *Proceedings of Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, Valetta, Malta, pp. 2228–2235.

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

Heringer, H.-J. (1984). Wortbildung: Sinn aus dem Chaos. *Deutsche Sprache 12*, pp. 1–13.

Landis, J. R., Koch, G. (1977). The measurement of observer agreement for categorical data, *Biometrics,* 33 (1), pp. 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*, pp. 7–18.

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

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

Tratz, S., Hovy, E. H. (2010). A Taxonomy, Dataset, and Classifier for Automatic Noun Compound Interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, Uppsala, Sweden, pp. 678–687.

Ó Séaghdha, D. (2008). *Learning compound noun semantics*. Ph.D. thesis, Computer Laboratory, University of Cambridge.