GjoSt-NN: A Representative Gold Standard of German Noun-Noun Compounds

Sabine Schulte im Walde, Anna Hätty, Stefan Bott, Nana Khvtisavrishvili
Institute for Natural Language Processing
University of Stuttgart, Germany
{schulte, anna.haetty, stefan.bott, nana.khvtisavrishvili}@ims.uni-stuttgart.de

Abstract
This paper presents a novel gold standard of German noun-noun compounds GjoSt-NN including 868 compounds annotated with corpus frequencies of the compounds and their constituents, productivity and ambiguity of the constituents, semantic relations between the constituents, and compositionality ratings of compound–constituent pairs. Moreover, a subset of the compounds containing 180 compounds is balanced for the productivity of the modifiers (distinguishing low/mid/high productivity) and the ambiguity of the heads (distinguishing between heads with 1, 2 and >2 senses).

Keywords: noun–noun compounds, compositionality ratings, semantic relations

1. Introduction
Compounds are combinations of two or more simplex words. They have been a recurrent focus of attention within theoretical, cognitive, and in the last decade also within computational linguistics. Our focus of interest is on German noun-noun compounds,1 such as Ahornblatt ‘maple leaf’, Feuerwerk ‘fireworks’, and Obstkuchen ‘fruit cake’, where both the grammatical head (in German, this is the rightmost constituent) and the modifier are nouns. More specifically, we are interested in the degrees of compositionality of German noun-noun compounds, i.e., the semantic relatedness between the meaning of a compound (e.g., Feuerwerk) and the meanings of its constituents (e.g., Feuer ‘fire’ and Werk ‘opus’).

The compositionality of noun-noun compounds has attracted much research across languages over the years. From a psycholinguistic point of view, researchers are interested in finding out how compound words are cognitively processed and represented in the mental lexicon. There is an ongoing debate about whether morphologically complex words are stored and processed as single units (full listing approach (Butterworth, 1983)), whether they are decomposed into their morphemes (Taft and Forster, 1975; Taft, 2004), or whether they can be accessed both ways: as whole forms and componentially (dual route models, cf. Caramazza et al. (1988), Baayen and Schreuder (1999)).

From a computational point of view, addressing the compositionality of noun compounds (and multi-word expressions in more general) is a crucial ingredient for lexicography and NLP applications, to know whether the expression should be treated as a whole, or through its constituents, and what the expression means. For example, studies such as Cholakov and Kordoni (2014); Weller et al. (2014); Cap et al. (2015); and Salehi et al. (2015) have integrated the prediction of compositionality into Statistical Machine Translation. Accordingly, research across languages has aimed for predicting the compositionality of noun compounds automatically. For example, Reddy et al. (2011) predicted the compositionality of 90 English noun–noun compounds via distributional information. Similarly, Schulte im Walde et al. (2013) assessed various types of distributional features to predict the compositionality of 244 German noun–noun compounds. Salehi and Cook (2013) and Salehi et al. (2014) explored multi-lingual dictionaries and distributional evidence to predict the compositionality of German and English noun–noun compounds.

Evaluating predictions of compositionality requires a gold standard of compositionality ratings, if the evaluation is not extrinsic. So in parallel to the emergence of computational systems predicting compositionality, there has also been an increase of gold standards to evaluate the predictions. Regarding noun compounds, Reddy et al. (2011) used heuristics about hypernymy and definitions in WordNet to induce intrinsic evidence to predict the compositionality of 90 English noun–noun compounds. Schulte im Walde et al. (2013) relied on an existing selection of noun compounds (von der Heide and Borgwaldt, 2009) and used a subset of concrete two-part noun–noun compounds. The work by Salehi et al. used both those datasets.

This paper presents a novel gold standard for the compositionality of German noun-noun compounds. In the next Section 2. we outline the desired properties of the gold standard. Sections 3. and 4. describe in detail how we created the gold standard, and what its resulting properties are.

2. Desired Properties of the Gold Standard
In previous work (Schulte im Walde et al., 2013), we used a gold standard of German noun-noun compositionality ratings that was based on a selection of noun compounds by von der Heide and Borgwaldt (2009). The original target set contained 450 concrete, depictable German noun compounds, with judgements on the compositionality of all compound–constituent pairs. From the compound set by von der Heide and Borgwaldt, we disregarded compounds with more than two constituents as well as compounds where the modifiers were not nouns. Our final set comprised a subset of their compounds including 244 two-part noun-noun compounds.
What is the motivation for creating a novel gold standard for the compositionality of German noun–noun compounds? We were interested in exploring factors that have been found to influence the cognitive processing and representation of compounds, such as

- frequency-based factors, i.e., the frequencies of the compounds and their constituents (e.g., van Jaarsveld and Rattink (1988), Janssen et al. (2008));
- the productivity (morphological family size), i.e., the number of compounds that share a constituent (de Jong et al., 2002); and
- semantic variables as the relationship between compound modifier and head: a teapot is a pot FOR tea, and a snowball is a ball MADE OF snow (Gagné and Spalding, 2009).
- In addition, we were interested in the effect of ambiguity (of both the modifiers and the heads) regarding the compositionality of the compounds.

Consequently, we created a gold standard with a focus on two-part noun-noun compounds including

- compounds and constituents from various frequency ranges;
- compounds and constituents from various productivity ranges;
- compounds and constituents with various numbers of senses; and
- compounds with various semantic relations.

Optimally, the compound targets in the gold standard should be balanced according to all of the above criteria, to include a similar number of compounds and constituents across the conditions. The following section will describe details of the creation process, and to what extent we achieved such a balance.

3. Creation of the Gold Standard

We rely on one of the currently largest corpora for German to induce our new gold standard of German noun–noun compounds $G_{host\text{-}NN}$: the web corpus $DECOW14\text{AX}$$^2$ (Schäfer and Bildhauer, 2012; Schäfer, 2015), henceforth: $decow$, containing 11.7 billion words. The creation pipeline to acquire a balanced gold standard from the web corpus includes the following steps:

1. corpus-based induction of a noun-noun compound candidate list;
2. addition of empirical properties to the compound candidates;
3. random but balanced selection of a core set of noun-noun compounds;
4. systematic extension of the core set to the full gold standard; and
5. annotation of the gold standard.

3.1. Corpus-based induction of candidate list

Relying on the $decow$ corpus, we first extracted all words identified as common nouns by the Tree Tagger (Schmid, 1994), plus their lemmas. The noun lemmas were counted, resulting in a total of 365,786 lemma types and their corpus frequencies.

As we wanted to focus on two-part noun-noun compounds only, we applied the morphological analyser SMOR (Faab et al., 2010) to the set of corpus lemmas, providing us potentially ambiguous morphological analyses for all 365,786 noun lemmas. From these, we extracted only those lemmas where SMOR predicted an analysis with exactly two nominal constituents. This set of two-part noun-noun compounds contained 154,960 compound candidate types for our new gold standard.

3.2. Enrichment of empirical properties

The complete set of 154,960 N-N candidates was enriched with empirical properties relevant for the gold standard:

- corpus frequencies of the compounds and the constituents (i.e., modifiers and heads), relying on $decow$;
- productivity of the constituents (modifiers and heads), i.e., how many compound types contained a specific modifier or head constituent;
- number of senses of the constituents (modifiers and heads) and the compounds, relying on GermaNet (Hamp and Feldweg, 1997; Kunze, 2000).

Overall, the 154,960 compound candidates included 7,061 different modifiers and 6,903 different heads.

Figure 5 shows the productivity of the constituents, by plotting the logarithm of the productivity against the number of constituents. For example, 332/361 modifiers/heads appeared in more than 100 different compounds; in contrast, there are approx. 1,388/1,434 modifiers/heads that appeared in only one compound. Figure 2 shows the number of senses of the modifier and head constituents, as a proportion of those constituent types that were included in GermaNet. Overall, GermaNet covered 26,444 of our 154,960 compound candidate types (17%); 6,683 of the modifier types (95%) and 6,550 of the head types (95%). A large proportion of compound types (97%) has only one sense in GermaNet. In comparison, 65% and 62% of the modifiers/heads have only one sense in GermaNet; 23%/24% have two senses, and below 10% have more than two senses.

3.3. Random but balanced compound selection

From the set of compound candidates we wanted to extract a random subset that at the same time was balanced across frequency, productivity and ambiguity ranges of the compounds and their constituents. Since defining and combining several ranges for each of the three criteria and for compounds as well as constituents would have led to an explosion of factors to be taken into account, we focused on two main criteria instead:

\[^2\text{http://corporafromtheweb.org/decow14/}\]
We systematically extended the set of $20 \times 3.4$. Systematic extension of core set each mode.

randomly selecting 5 out of the 20 selected compounds in 5 noun-noun compounds for each criteria combination, by in the following subsection, we also created a subset of hair dress, hair care, floral glory, colour glory as Bl¨utenpracht, Farbenpracht, etc. we added Haarw¨asche, Haarkleid, Haarpflege, etc. Haar (the modifier is Haarpracht selected compounds. Taking either the same modifier or the same head as any of the inal set of compound candidates (cf. Section 3.1.) with selected compounds by adding all compounds from the orig-

The total of 9 criteria combinations is listed here:

| modifier prod. | low | mid | high |
|----------------|-----|-----|------|
| head senses    | 1   | 2   | >2   |
| modes          | 1   | 2   | 3   |

Table 1: Combined ranges for random extraction.

Figure 1: Productivity of constituents.

For each of the 9 categories, we randomly selected 20 noun-
noun compounds from our candidate set, disregarding com-

pounds containing modifiers or heads with a corpus-

frequency < 2,000, and disregarding compounds containing modifiers or heads with a corpus-

frequency < 100. For reasons which will become clear in the following subsection, we also created a subset of 5 noun-noun compounds for each criteria combination, by randomly selecting 5 out of the 20 selected compounds in each mode.

3.4. Systematic extension of core set

We systematically extended the set of $20 \times 9=180$ compounds without much overlap in modifiers or heads, or the complete $G_{10}$ containing 868 compounds out of the larger, less-balanced set of 1,208 compounds.\textsuperscript{4} These two sets of compounds were annotated with two kinds of semantic information, (i) the semantic relations between the modifiers and the heads, and (ii) compositionality ratings.

3.5. Annotation of gold standard

For computational experiments, researchers can either use the well-balanced set of $20 \times 9=180$ compounds without much overlap in modifiers or heads, or the complete $G_{10}$ containing 868 compounds out of the larger, less-balanced set of 1,208 compounds.\textsuperscript{4}

The total set of target compounds after the systematic exten-
tion contained $> 5,000$ types; relying only on the set of $5 \times 9$ randomly selected compounds, it contained 1,208 types. The latter selection appeared as a reasonable size for a gold standard to be annotated for semantic criteria (see below).

While the extension procedure destroyed the coherent bal-
ance of criteria that underlied our random extraction de-
scribed in the previous section, it ensured a variety of com-
pounds with either the same modifiers or the same heads.

3.5.1. Semantic relation annotation

In previous work, different kinds of annotation schemes have been used for compound relation annotation. Girju et al. (2005) annotated 282 English two-part noun compounds and 244 English three-part noun compounds with a list of eight prepositional paraphrases previously proposed by Lauer (1995), and also with a set of 35 semantic relations introduced by Moldovan and Girju (2003). Ó Séaghdha (2007) relied on a set of nine semantic relations suggested by Levi (1978), and designed and evaluated a set of relations that took over four of Levi’s relations (BE, HAVE, IN, ABOUT) and added two relations refering to event participants (ACTOR, INST(ument)) that replaced the relations MAKE, CAUSE, FOR, FROM, USE. The relation LEX refers to lexicalised compounds where no relation can be assigned. A set of 1,443 En-

glish noun compounds was annotated with his modified relation set. Dima et al. (2014) worked on German noun compounds and suggested to combine paraphrase- and property-based relation annotation.

We decided to apply the relation set suggested by Ó Séaghdha (2007) to our German noun compounds, for two reasons: (i) He had evaluated his annotation relations and annotation scheme, and (ii) his dataset had a similar size as ours, so we could aim for comparing results across lan-
guages. The three authors of this paper who are native speakers of German annotated the compounds with his semantic relations. We used three rounds for the annota-
tion, with discussions in between. If disagreement could not be resolved in the discussions, we disregarded the re-
spective compounds. In the end, we accepted 868 from the 1,208 compounds as gold standard target compounds. These compounds were annotated with the same relation by all three annotators. The distribution of the compounds over the semantic relations is shown in Figure 3.

\textsuperscript{3}The translations of the example compounds are hair washing, hair dress, hair care, floral glory, colour glory, respectively.

\textsuperscript{4}From the enlarged set of 1,208 compounds, the final dataset contains only 868 instances, to ensure a reliable agreement on relation annotation, see Section 3.5.1. below.
There is no tight correlation\(^6\) between compound and modifier frequencies \((\rho = 0.2345, p < 0.001)\) and compound and head frequencies \((\rho = 0.1451, p < 0.001)\). Figure 5 compares the log productivity scores of the modifiers and heads within the same compounds. While most compounds seem to include modifiers and heads with log frequencies between 1.5 and 2.8 (roughly corresponding to productivity scores between 30 and 650), other ranges are also covered sufficiently. The correlation between the modifier and head productivity scores confirms their independencies: \(\rho = 0.1271, p < 0.001\).

Figure 6 provides the same information regarding the ambiguity of modifiers and heads within compounds. We can see that we cover most combinations of modifiers and heads with ambiguities between 1 (monosemous) and 7. Again, their is no correlation \((\rho = 0.0193, p = 0.2840)\).

---

\(^6\)We rely on the Spearman rank-order correlation coefficient \(\rho\) (Siegel and Castellan, 1988) to calculate correlations.
1. the set of 154,960 noun-noun candidate compounds and their constituents (cf. Section 3.1.), accompanied by corpus frequency, productivity and degree of ambiguity (cf. Section 3.2.);
2. the final gold standard $G_{1,ost-NN}$ of 868 (out of 1,208) noun-noun compounds and their constituents, accompanied by corpus frequency, productivity, ambiguity, and annotated with semantic relations and compositionality ratings (cf. Section 3.5.); and
3. the carefully balanced $G_{1,ost-NN}$ subsets of $20 \times 9$ and $5 \times 9$ compounds and their constituents, categorised according to our 9 criteria combinations for modifier productivity and head ambiguity (cf. Section 3.3.).

For computational experiments, researchers can either use the well-balanced set of $20 \times 9=180$ compounds without much overlap in modifiers or heads, or a larger, but less-balanced set of 868 compounds. The datasets are available from http://www.ims.uni-stuttgart.de/data/ghost-NN.

Table 2 provides some example compounds and their properties. They were chosen to illustrate the variety of property combinations across items, while at the same time they include compounds with common modifiers or heads.

6. Acknowledgements
The research presented in this paper was funded by the DFG Heisenberg Fellowship SCHU 2580/1 (Sabine Schulte im Walde), the DFG Research Grant SCHU 2580/2 “Distributional Approaches to Semantic Relatedness” (Stefan Bott), and the DFG Collaborative Research Center SFB 732 (Anna Hätty, Nana Khvitsishvili).

7. Bibliographical References
Baayen, H. and Schreuder, R. (1999). War and Peace: Morphemes and Full Forms in a Noninteractive Activation Parallel Dual–Route Model. *Brain and Language*, 68:27–32.
Butterworth, B. (1983). Lexical Representation. In Brian Butterworth, editor, *Language Production*, pages 257–294. Academic Press, London.
Cap, F., Nirmal, M., Weller, M., and Schulte im Walde, S. (2015). How to Account for Idiomatic German Support Verb Constructions in Statistical Machine Translation. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 19–28, Denver, Colorado, USA.
Caramazza, A., Laudanna, A., and Romani, C. (1988). Lexical Access and Inflectional Morphology. *Cognition*, 28:297–332.
Cholakov, K. and Kordoni, V. (2014). Better Statistical Machine Translation through Linguistic Treatment of Phrasal Verbs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 196–201, Doha, Qatar.
de Jong, N. H., Feldman, L. B., Schreuder, R., Pastizzo, M., and Baayen, H. R. (2002). The Processing and Representation of Dutch and English Compounds: Peripheral Morphological and Central Orthographic Effects. *Brain and Language*, 81:555–567.
Dima, C., Henrich, V., Hinrichs, E., and Hoppermann, C. (2014). How to Tell a Schneemann from a Milchmann: An Annotation Scheme for Compound-Internal Relations. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 1194–1201, Reykjavik, Iceland.
Figure 7: Log frequencies of compounds across relations types.

Figure 8: Compositionality ratings across relation types.

Figure 9: Productivity and compositionality ratings of modifiers and heads.
| Compound                          | Nouns     | Modifier | Head        | Frequencies | Productivities | Ambiguities | Relation | Ratings |
|----------------------------------|-----------|----------|-------------|-------------|----------------|-------------|----------|---------|
| Stadthotel                       | city      | city     | Hotel       | 3,405       | 4,053,206      | 1,199,856   | 543      | 59      | 1       | 1      | 3.35  | 3.35  |
| Stadtrand                        | suburb    | city     | Rand        | 25,099      | 4,053,206     | 523,473     | 543      | 98      | 1       | 2      | HAVE | 4.94  | 4.25  |
| Stadtwerkhôtel                   | public    | city     | Werk        | 107,754     | 4,053,206     | 1,354,148   | 543      | 366     | 1       | 6      | ACTOR | 3.81  | 3.69  |
| Sonnenenergie                    | solar     | energy   | Energie     | 25,598      | 832,636       | 1,191,333   | 155      | 30      | 3       | 2      | INST  | 4.58  | 5.44  |
| Sonnenkönig                      | Sun King  | sun      | König       | 2,680       | 832,636       | 494,221     | 155      | 109     | 3       | 3      | LEX   | 1.94  | 5.50  |
| Sonnenmasse                      | sun mass  | sun      | Masse       | 3,433       | 832,636       | 468,284     | 155      | 108     | 3       | 3      | HAVE  | 4.56  | 4.75  |
| Sonnenscheibe                    | solar disc| sun      | Scheibe     | 3,155       | 832,636       | 364,567     | 155      | 96      | 3       | 4      | BE    | 4.56  | 3.75  |
| Sonnenseite                      | sunny side| sun      | Seite       | 7,279       | 832,636       | 5,508,445   | 155      | 256     | 3       | 6      | HAVE  | 4.00  | 4.31  |
| Sonnenstrahl                     | sunbeam   | Sonne    | Strahl      | 44,612      | 832,636       | 32,182      | 155      | 27      | 3       | 3      | HAVE  | 5.13  | 4.69  |
| Sonnenuhr                        | sundial   | Sonne    | Uhr         | 8,407       | 832,636       | 4,507,590   | 155      | 63      | 3       | 2      | 1ST   | 3.75  | 5.31  |
| Jeanshose                        | jeans     | Jeans    | Hose        | 2,971       | 66,789        | 273,665     | 19       | 61      | 1       | 1      | BE    | 5.25  | 5.44  |
| Latzhose                         | overall   | Latz     | Hose        | 3,296       | 5,324         | 273,665     | 1        | 61      | 2       | 1      | HAVE  | 3.54  | 5.23  |
| Strumpfhose                      | tight     | Strumpf  | Hose        | 20,535      | 26,331        | 273,665     | 13       | 61      | 1       | 1      | BE    | 4.35  | 4.42  |
| Kirchspiel                       | parish    | Kirche   | Spiel       | 6,583       | 1,761,187     | 4,122,168   | 319      | 403     | 3       | 6      | LEX   | 4.44  | 3.13  |
| Machtspiel                       | power game| Macht    | Spiel       | 4,408       | 806,162       | 4,122,168   | 169      | 403     | 2       | 6      | ABOUT | 4.63  | 3.44  |
| Ritterspiel                      | knights’ tournament | Ritter | Spiel   | 2,365       | 115,484       | 4,122,168   | 47       | 403     | 1       | 6      | ACTOR | 3.94  | 4.75  |
| Testspiel                        | tryout    | Test     | Spiel       | 37,800      | 660,169       | 4,122,168   | 100      | 403     | 3       | 6      | BE    | 4.25  | 5.19  |
| Trauerspiel                      | fiasco    | Trauer   | Spiel       | 10,763      | 134,379       | 4,122,168   | 77       | 403     | 3       | 6      | ABOUT | 3.06  | 2.81  |
| Windspiel                        | wind chimes| Wind    | Spiel       | 2,284       | 551,317       | 4,122,168   | 88       | 403     | 3       | 6      | INST  | 4.31  | 2.94  |
| Winterspiel                      | winter games| Winter | Spiel      | 16,067      | 721,552       | 4,122,168   | 207      | 403     | 1       | 6      | IN    | 4.43  | 5.14  |
| Würtelspiel                      | game of dice| Würfel dice   | Spiel       | 4,408       | 80,371        | 4,122,168   | 14       | 403     | 2       | 6      | INST  | 4.94  | 5.56  |
| Bergkette                        | mountain chain | Berg   | Kette      | 8,790       | 564,178       | 207,479     | 205      | 139     | 2       | 4      | BE    | 5.13  | 2.56  |
| Halbkette                       | necklace  | Hals     | Kette      | 8,707       | 271,703       | 207,479     | 39       | 139     | 3       | 4      | IN    | 3.94  | 5.44  |
| Handelskette                     | trade chain | Handel | Kette      | 6,509       | 428,611       | 207,479     | 240      | 139     | 1       | 4      | INST  | 4.75  | 3.38  |
| Hotelkette                       | hotel chain | Hotel | Kette      | 6,410       | 1,199,856     | 207,479     | 134      | 139     | 1       | 4      | BE    | 5.00  | 3.13  |
| Menschekette                     | human chain | Mensch | Kette      | 6,583       | 8,884,087     | 207,479     | 191      | 139     | 1       | 4      | BE    | 4.94  | 3.75  |
| Produktionkette                  | production chain | Produktion | Kette       | 2,738       | 579,419       | 207,479     | 244      | 139     | 2       | 4      | HAVE  | 4.69  | 3.19  |
| Schneckkette                     | snow chains | Schnee | Kette      | 5,167       | 324,839       | 207,479     | 95       | 139     | 1       | 4      | INST  | 4.19  | 4.21  |
| Zeichenkette                     | string    | Zeichen  | Kette      | 8,836       | 749,903       | 207,479     | 62       | 139     | 3       | 4      | BE    | 4.34  | 2.95  |

Table 2: Example compounds in $G_{host-NN}$ and their properties.
