# $G_h$oSt-NN: A Representative Gold Standard of German Noun-Noun Compounds

Sabine Schulte im Walde, Anna Hätty, Stefan Bott, Nana Khvtisavrishvili
*{schulte, anna.haetty, stefan.bott, nana.khvtisavrishvili}@ims.uni-stuttgart.de*

Institut für Maschinelle Sprachverarbeitung

## Motivation for a Representative Gold Standard

Interest in systematically 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 & 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);
- the *relationship between compound modifier and head*: a teapot is a pot FOR tea, and a snowball is a ball MADE OF snow (Gagné & 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.

## Creation of the Gold Standard G$_h$ost-NN

### 1. Corpus-based induction of candidate list

- basis: German web corpus *DECOW14AX* (Schäfer & Bildhauer, 2012), with 11.7 billion words
- extraction of 365,786 common nouns and their lemmas, according to the *Tree Tagger* (Schmid, 1994)
- selection of 154,960 two-part noun-noun compounds, according to the morphological analyser *SMOR* (Faaß et al., 2010)

### 2. Enrichment of empirical properties

- *corpus frequencies* of compounds and constituents (i.e., modifiers and heads)
- *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 & Feldweg, 1997)

### 3. Random but balanced compound selection

- goal (optimum): random subset of compound candidates balanced across frequency, productivity and ambiguity ranges
- goal (compromise): two main criteria *productivity of the modifiers* (low/mid/high) and *ambiguity of the heads* (1, 2 and >2 senses)
- random selections of 20/5 compounds across 9 categories

### 4. Systematic extension of core set

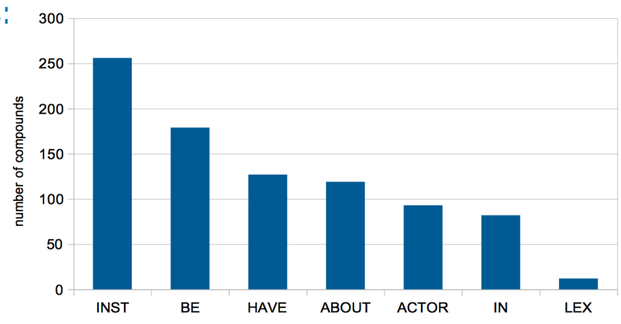
- systematic extension of the 20×9 and 5×9 randomly selected compounds by adding all compounds from the original set of compound candidates with either the same modifier or the same head as any of the selected compounds
- extension procedure destroyed the coherent balance of criteria underlying our random extraction, but ensured a variety of compounds with either the same modifiers or the same heads

### 5. Semantic annotation of gold standard

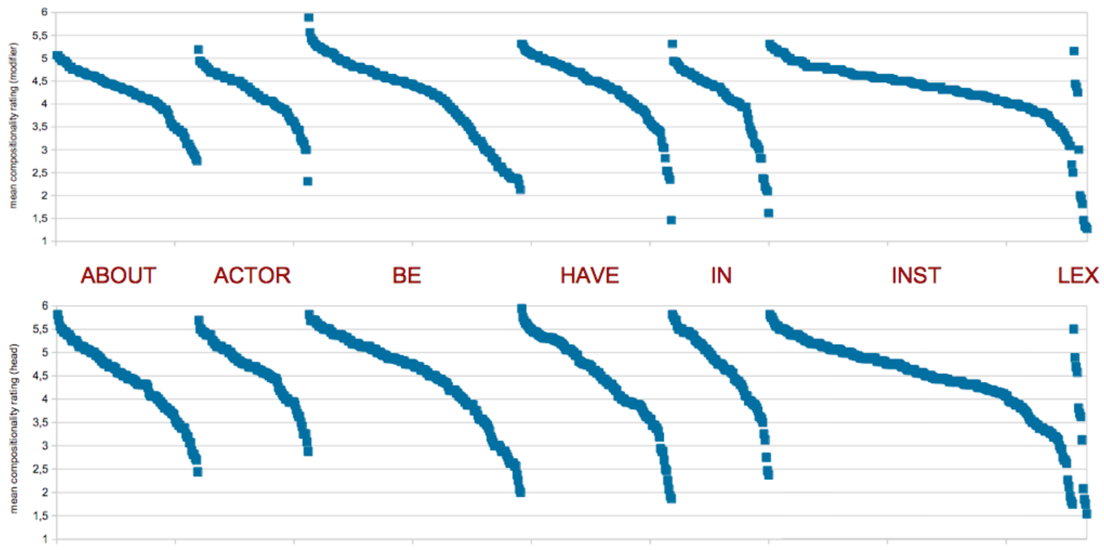
- size of novel gold standard G$_h$ost: 868 noun-noun compounds
- semantic relations according to Ó Séaghdha (2007):
  `BE, HAVE, IN, ABOUT, ACTOR, INST(rument), LEX`
- compositionality ratings from experts and via AMT

## G$_h$ost-NN Properties

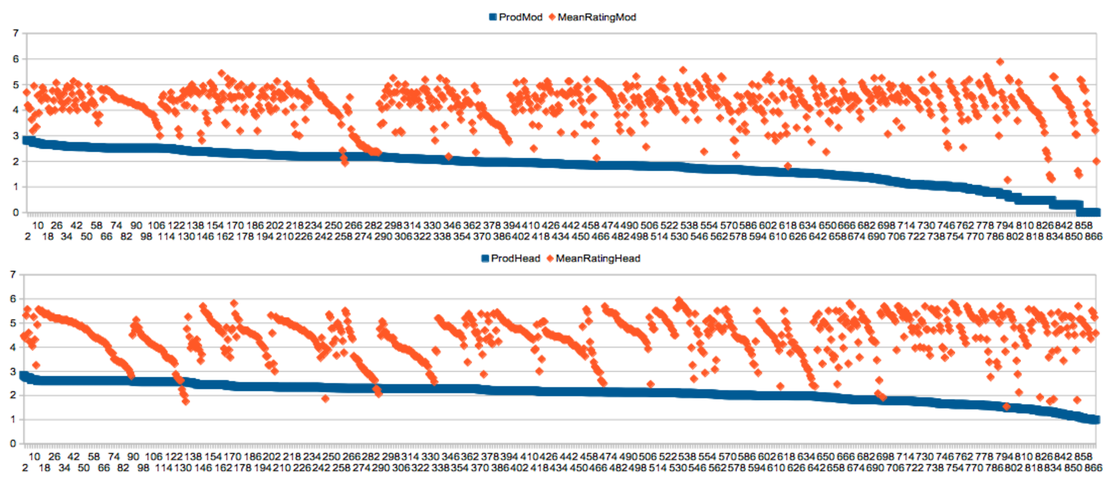
**Semantic relations in compounds:**




**Compositionality ratings across relation types:**




**Productivity and compositionality ratings of modifiers and heads:**




## Resource

1. The set of 154,960 noun-noun candidate compounds and their constituents, accompanied by corpus frequency, productivity and degree of ambiguity.
2. The final gold standard *G$_h$ost-NN* of 868 noun-noun compounds and their constituents, accompanied by corpus frequency, productivity, ambiguity, and annotated with semantic relations and compositionality ratings.
3. The carefully balanced *G$_h$ost-NN* subsets of 20×9 and 5×9 compounds and their constituents, categorised according to our 9 categories for modifier productivity and head ambiguity.

| ***Nouns*** | | | | | | ***Frequencies*** | | | ***Productivities*** | | ***Ambiguities*** | | ***Relation*** | ***Ratings*** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Compound* | | *Modifier* | | *Head* | | *Compound* | *Modifier* | *Head* | *Modifier* | *Head* | *Modifier* | *Head* | | *Modifier* | *Head* |
| Stadthotel | *city hotel* | Stadt | *city* | Hotel | *hotel* | 3,405 | 4,053,206 | 1,199,856 | 543 | 59 | 1 | 1 | IN | 3.35 | 5.35 |
| Stadtrand | *suburb* | Stadt | *city* | Rand | *border* | 25,099 | 4,053,206 | 523,473 | 543 | 98 | 1 | 2 | HAVE | 4.94 | 4.25 |
| Stadtwerk | *public services* | Stadt | *city* | Werk | *plant* | 107,754 | 4,053,206 | 1,354,148 | 543 | 366 | 1 | 6 | ACTOR | 3.81 | 3.69 |
| Sonnenenergie | *solar energy* | Sonne | *sun* | Energie | *energy* | 25,398 | 832,636 | 1,191,333 | 155 | 30 | 3 | 2 | INST | 4.58 | 5.44 |
| Sonnenkönig | *Sun King* | Sonne | *sun* | König | *king* | 2,680 | 832,636 | 494,221 | 155 | 109 | 3 | 3 | LEX | 1.94 | 5.50 |
| Sonnenscheibe | *solar disc* | Sonne | *sun* | Scheibe | *slice* | 3,155 | 832,636 | 364,567 | 155 | 96 | 3 | 4 | BE | 4.56 | 3.75 |
| Sonnenseite | *sunny side* | Sonne | *sun* | Seite | *side* | 7,279 | 832,636 | 5,508,445 | 155 | 256 | 3 | 6 | IN | 4.00 | 4.31 |
| Sonnenstrahl | *sunbeam* | Sonne | *sun* | Strahl | *beam* | 44,612 | 832,636 | 32,182 | 155 | 27 | 3 | 3 | HAVE | 5.13 | 4.69 |
| Sonnenuhr | *sundial* | Sonne | *sun* | Uhr | *clock* | 8,407 | 832,636 | 4,507,590 | 155 | 63 | 3 | 2 | INST | 3.75 | 5.31 |
| Jeanshose | *jeans* | Jeans | *jeans* | Hose | *trousers* | 2,971 | 66,789 | 273,665 | 19 | 61 | 1 | 1 | BE | 5.25 | 5.44 |
| Latzhose | *overall* | Latz | *bib* | Hose | *trousers* | 3,296 | 5,324 | 273,665 | 1 | 61 | 2 | 1 | HAVE | 3.54 | 5.23 |
| Strumpfhose | *tights* | Strumpf | *stockings* | Hose | *trousers* | 20,535 | 26,331 | 273,665 | 13 | 61 | 1 | 1 | BE | 4.35 | 4.42 |
| Kirchspiel | *parish* | Kirche | *church* | Spiel | *game* | 6,583 | 1,761,187 | 4,122,168 | 319 | 403 | 3 | 6 | LEX | 4.44 | 3.13 |
| Machtspiel | *power game* | Macht | *power* | Spiel | *game* | 4,408 | 806,162 | 4,122,168 | 169 | 403 | 2 | 6 | ABOUT | 4.63 | 3.44 |
| Ritterspiel | *knights' tournament* | Ritter | *knight* | Spiel | *game* | 2,365 | 115,484 | 4,122,168 | 47 | 403 | 1 | 6 | ACTOR | 3.94 | 4.75 |
| Testspiel | *tryout* | Test | *test* | Spiel | *game* | 37,800 | 660,169 | 4,122,168 | 100 | 403 | 3 | 6 | BE | 4.25 | 5.19 |
| Windspiel | *wind chimes* | Wind | *wind* | Spiel | *game* | 2,284 | 551,317 | 4,122,168 | 88 | 403 | 3 | 6 | INST | 4.31 | 2.94 |
| Winterspiel | *winter games* | Winter | *winter* | Spiel | *game* | 16,067 | 721,552 | 4,122,168 | 207 | 403 | 1 | 6 | IN | 4.43 | 5.14 |
| Würfelspiel | *game of dice* | Würfel | *dice* | Spiel | *game* | 4,408 | 80,371 | 4,122,168 | 14 | 403 | 2 | 6 | INST | 4.94 | 5.56 |
| Bergkette | *mountain chain* | Berg | *mountain* | Kette | *chain* | 8,799 | 564,178 | 207,479 | 205 | 139 | 2 | 4 | BE | 5.13 | 2.56 |
| Halskette | *necklace* | Hals | *neck* | Kette | *chain* | 8,707 | 271,703 | 207,479 | 39 | 139 | 3 | 4 | IN | 3.94 | 5.44 |
| Handelskette | *trade chain* | Handel | *trade* | Kette | *chain* | 6,509 | 428,611 | 207,479 | 240 | 139 | 1 | 4 | INST | 4.75 | 3.38 |
| Hotelkette | *hotel chain* | Hotel | *hotel* | Kette | *chain* | 6,410 | 1,199,856 | 207,479 | 134 | 139 | 1 | 4 | BE | 5.00 | 3.13 |
| Produktionskette | *production chain* | Produktion | *production* | Kette | *chain* | 2,738 | 579,419 | 207,479 | 244 | 139 | 2 | 4 | HAVE | 4.69 | 3.19 |
| Schneekette | *snow chains* | Schnee | *snow* | Kette | *chain* | 5,167 | 324,839 | 207,479 | 95 | 139 | 1 | 4 | INST | 4.19 | 4.21 |