Feature Norms of German Noun Compounds

Stephen Roller
Department of Computer Science
The University of Texas at Austin
roller@cs.utexas.edu

Sabine Schulte im Walde
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart, Germany
schulte@ims.uni-stuttgart.de

Abstract

This paper presents a new data collection of feature norms for 572 German noun-noun compounds. The feature norms complement existing data sets for the same targets, including compositionality ratings, association norms, and images. We demonstrate that the feature norms are potentially useful for research on the noun-noun compounds and their semantic transparency: The feature overlap of the compounds and their constituents correlates with human ratings on the compound–constituent degrees of compositionality, $\rho = 0.46$.

1 Introduction

Feature norms are short descriptions of typical attributes for a set of objects. They often describe the visual appearance (a firetruck *is red*), function or purpose (a cup *holds liquid*), location (mushrooms grow *in forests*), and relationships between objects (a cheetah *is a cat*). The underlying features are usually elicited by asking a subject to carefully describe a cue object, and recording their responses.

Feature norms have been widely used in psycholinguistic research on conceptual representations in semantic memory. Prominent collections have been pursued by McRae et al. (2005) for living vs. non-living basic-level concepts; by Vinson and Vigliocco (2008) for objects and events; and by Wu and Barsalou (2009) for noun and noun phrase objects. In recent years, feature norms have also acted as a loose proxy for perceptual information in data-intensive computational models of semantic tasks, in order to bridge the gap between language and the real world (Andrews et al., 2009; Silberer and Lapata, 2012; Roller and Schulte im Walde, 2013).

In this paper, we present a new resource of feature norms for a set of 572 concrete, depictable German nouns. More specifically, these nouns include 244 noun-noun compounds and their corresponding constituents. For example, we include features for *'Schneeball'* ('snowball'), *'Schnee'* ('snow'), and *'Ball'* ('ball'). Table 1 presents the most prominent features of this example compound and its constituents. Our collection complements existing data sets for the same targets, including compositionality ratings (von der Heide and Borgwaldt, 2009); associations (Schulte im Walde et al., 2012; Schulte im Walde and Borgwaldt, 2014); and images (Roller and Schulte im Walde, 2013).

The remainder of this paper details the collection process of the feature norms, discusses two forms of cleansing and normalization we employed, and performs quantitative and qualitative analyses. We find that the normalization procedures improve quality in terms of feature tokens per feature type, that the normalized feature norms have a desirable distribution of features per cue, and that the feature norms are useful in semantic models to predict compositionality.

2 Feature Norm Collection

We employ Amazon Mechanical Turk (AMT)[1] for data collection. AMT is an online crowdsourcing platform where *requesters* post small, atomic tasks which require manual completion by humans. Workers can complete these tasks, called *HITs*, in order to earn a small bounty.

2.1 Setup and Data

Workers were presented with a simple page asking them to describe the typical attributes of a given noun. They were explicitly informed in English that only native German speakers should complete

[1] http://www.mturk.com

Proceedings of the 10th Workshop on Multiword Expressions (MWE 2014), pages 104–108,
Gothenburg, Sweden, 26-27 April 2014.

Schneeball ‘snowball’			*Schnee* ‘snow’			*Ball* ‘ball’		
ist kalt	‘is cold’	8	*ist kalt*	‘is cold’	13	*ist rund*	‘is round’	14
ist rund	‘is round’	7	*ist weiß*	‘is white’	13	*zum Spielen*	‘for playing’	4
aus Schnee	‘made from snow’	7	*im Winter*	‘in the winter’	6	*rollt*	‘rolls’	2
ist weiß	‘is white’	7	*fällt*	‘falls’	3	*wird geworfen*	‘is thrown’	2
formt man	‘is formed’	2	*schmilzt*	‘melts’	2	*ist bunt*	‘is colorful’	2
wirft man	‘is thrown’	2	*hat Flocken*	‘has flakes’	2	*Fußball*	‘football’	2
mit den Händen	‘with hands’	2	*ist wässrig*	‘is watery’	1	*Basketball*	‘basketball’	2

Table 1: Most frequent features for example compound *Schneeball* and its constituents.

the tasks. All other instructions were given in German. Workers were given 7 example features for the nouns *‘Tisch‘* (‘table’) and *‘Katze‘* (‘cat’), and instructed to provide typical attributes per noun. Initially, workers were required to provide 6-10 features per cue and were only paid $0.02 per hit, but very few workers completed the hits. After lowering the requirements and increasing the reward, we received many more workers and collected the data more quickly. Workers could also mark a word as unfamiliar or provide additional commentary if desired.

We collected responses from September 21, 2012 until January 31, 2013. Workers who were obvious spammers were rejected and not rewarded payment. Typically spammers pasted text from Google, Wikipedia, or the task instructions and were easy to spot. Users who failed to follow instructions (responded in English, did not provide the minimum number of features, or gave nonsensical responses) were also rejected without payment. Users who put in a good faith effort and consistently gave reasonable responses had all of their responses accepted and rewarded.

In total, 98 different workers completed at least one accepted hit, but the top 25 workers accounted for nearly 90% of the responses. We accepted 28,404 different response tokens over 18,996 response types for 572 different cues, or roughly 50 features per cue.

3 Cleansing and Normalization

We provide two cleaned and normalized versions of our feature norms.[2] In the first version, we correct primarily orthographic mistakes such as inconsistent capitalization, spelling errors, and surface usage, but feature norms remain otherwise unchanged. This version will likely be more useful to researchers interested in more subtle variations and distinctions made by the workers.

The second version of our feature norms are more aggressively normalized, to reduce the quantity of unique and low frequency responses while maintaining the spirit of the original response. The resulting data is considerably less sparse than the orthographically normalized version. This version is likely to be more useful for research that is highly affected by sparse data, such as multimodal experiments (Andrews et al., 2009; Silberer and Lapata, 2012; Roller and Schulte im Walde, 2013).

3.1 Orthographic Normalization

Orthographic normalization is performed in four automatic passes and one manual pass in the following order:

Letter Case Normalization: Many workers inconsistently capitalize the first word of feature norms as though they are writing a complete sentence. For example, *‘ist rund‘* and *‘Ist rund‘* (‘is round’) were both provided for the cue *‘Ball‘*. We cannot normalize capitalization by simply using lowercase everywhere, as the first letter of German nouns should always be capitalized. To handle the most common instances, we lowercase the first letter of features that began with articles, modal verbs, prepositions, conjunctions, or the high-frequency verbs *‘kommt‘*, *‘wird‘*, and *‘ist‘*.

Umlaut Normalization: The same German word may sometimes be spelled differently because some workers use German keyboards (which have the letters ä, ö, ü, and ß), and others use English keyboards (which do not). We automatically normalize to the umlaut form (i.e. *‘gruen‘* to *‘grün‘*, *‘weiss‘* to *‘weiß‘*) whenever two workers gave *both versions for the same cue*.

Spelling Correction: We automatically correct common misspellings (such as *errecihen* → *erreichen*), using a list from previous collection experiments (Schulte im Walde et al., 2008; Schulte im Walde et al., 2012). The list was created semi-automatically, and manually corrected.

[2] The norms can be downloaded from `www.ims.uni-stuttgart.de/forschung/ressourcen/experiment-daten/feature-norms.en.html`.

Usage of *‘ist‘* and *‘hat‘*: Workers sometimes drop the verbs *‘ist‘* (‘is’) and *‘hat‘* (‘has’), e.g. the worker writes only *‘rund‘* (‘round’) instead of *‘ist rund‘*, or *‘Obst‘* (‘fruit’) instead of *‘hat Obst‘*. We normalize to the *‘ist‘* and *‘hat‘* forms when two workers gave *both versions for the same cue*. Note that we cannot automatically do this across separate cues, as the relationship may change: a tree *has* fruit, but a banana *is* fruit.

Manual correction: Following the above automatic normalizations, we manually review all non-unique responses. In this pass, responses are normalized and corrected with respect to punctuation, capitalization, spelling, and orthography. Roughly 170 response types are modified in this phase.

3.2 Variant Normalization

The second manual pass consists of more aggressive normalization of expression variants. In this pass, features are manually edited to minimize the number of feature types while preserving as much semantic meaning as possible:

- Replacing plurals with singulars;
- Removing modal verbs, e.g. *‘kann Kunst sein‘* (‘can be art’) to *‘ist Kunst‘*;
- Removing quantifiers and hedges, e.g. *‘ist meistens blau‘* (‘is mostly blue’) to *‘ist blau‘*;
- Splitting into atomic norms, e.g. *‘ist weiß oder schwarz‘* (‘is white or black’) to *‘ist weiß‘* and *‘ist schwarz‘*, or *‘jagt im Wald‘* (‘hunts in forest’) to *‘jagt‘* and *‘im Wald‘*;
- Simplifying verbiage, e.g. *‘ist in der Farbe schwarz‘* (‘is in the color black’) to *‘ist schwarz‘*.

These selected normalizations are by no means comprehensive or exhaustive, but do handle a large portion of the cases. In total, we modify roughly 5400 tokens over 1300 types.

4 Quantitative Analysis

In the following two analyses, we explore the type and token counts of our feature norms across the steps in the cleansing process, and analyze the underlying distributions of the features per cues.

Type and Token counts Table 2 shows the token and type counts for all features in each step of the cleansing process. We also present the counts for *non-idiosyncratic* features, or features which are provided for at least two distinct cues. The orthographic normalizations generally lower the number of total and non-idiosyncratic types, and increase the number of non-idiosyncratic tokens. This indicates we are successfully identifying and correcting many simple orthographic errors, resulting in a less sparse matrix. The necessary amount of manual correction is relatively low, indicating we are able to catch the majority of mistakes using simple, automatic methods.

Data Version	Total		Non-idiosyncratic	
of Responses	**Types**	**Tokens**	**Types**	**Tokens**
Raw	18,996	28,404	2,029	10,675
Case	18,848	28,404	2,018	10,801
Umlaut	18,700	28,404	1,967	10,817
Spelling	18,469	28,404	1,981	11,072
ist/hat	18,317	28,404	1,924	11,075
Manual	18,261	28,404	1,889	11,106
Aggressive	17,503	28,739	1,374	11,848

Table 2: Counts in the cleansing process.

The more aggressively normalized norms are considerably different than the orthographically normalized norms. Notably, the number of total tokens increases from the atomic splits. The data is also less sparse and more robust, as indicated by the drops in both total and non-idiosyncratic types. Furthermore, the number of non-idiosyncratic tokens also increases considerably, indicating we were able to find numerous edge cases and place them in existing, frequently-used bins.

Number of Features per Cue Another important aspect of the data set is the number of features per cue. An ideal feature norm data set would contain a roughly equal number of (non-idiosyncratic) features for every cue; if most of the features are underrepresented, with a majority of the features lying in only a few cues, then our data set may only properly represent for these few, heavily represented cues.

Figure 1 shows the number of features per cue for (a) all features and (b) the non-idiosyncratic features, for the aggressively normalized data set. In the first histogram, we see a clear bimodal distribution around the number of features per cue. This is an artifact of the two parts of our collection process: the shorter, wider distribution corresponds to the first part of collection, where workers gave more responses for less reward. The taller, skinnier distribution corresponds to the second half of collection, when workers were rewarded more for less work. The second collection procedure was clearly effective in raising the number of hits completed, but resulted in fewer features per cue.

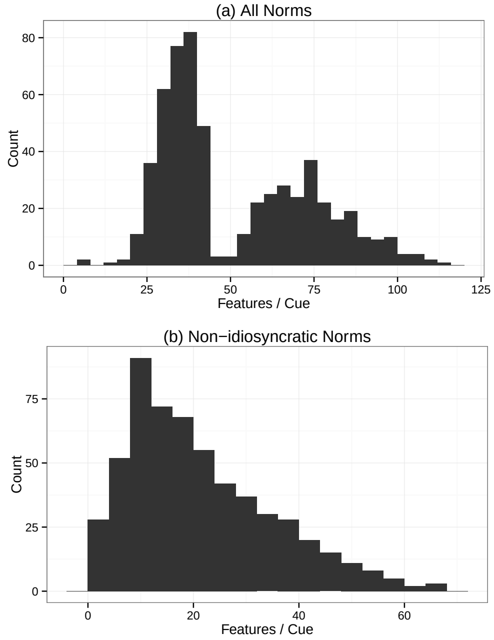

Figure 1: Distribution of features per cue.

In the second histogram, we see only the non-idiosyncratic features for each cue. Unlike the first histogram, we see only one mode with a relatively long tail. This indicates that mandating more features per worker (as in the first collection process) often results in more idiosyncratic features, and not necessarily a stronger representation of each cue. We also see that roughly 85% of the cues have at least 9 non-idiosyncratic features each. In summary, our representations are nicely distributed for the majority of cues.

5 Qualitative Analysis

Our main motivation to collect the feature norms for the German noun compounds and their constituents was that the features provide insight into the semantic properties of the compounds and their constituents and should therefore represent a valuable resource for cognitive and computational linguistics research on compositionality. The following two case studies demonstrate that the feature norms indeed have that potential.

Predicting the Compositionality The first case study relies on a simple feature overlap measure to predict the degree of compositionality of the compound–constituent pairs of nouns: We use the proportion of shared features of the compound and a constituent with respect to the total number of features of the compound. The degree of compositionality of a compound noun is calculated with respect to each constituent of the compound.

For example, if a compound noun N_0 received a total of 30 features (tokens), out of which it shares 20 with the first constituent N_1 and 10 with the second constituent N_2, the predicted degrees of compositionality are $\frac{20}{30} = 0.67$ for N_0–N_1, and $\frac{10}{30} = 0.33$ for N_0–N_2. The predicted degrees of compositionality are compared against the mean compositionality judgments as collected by von der Heide and Borgwaldt (2009), using the Spearman rank-order correlation coefficient. The resulting correlations are $\rho = 0.45, p < .000001$ for the standard normalized norms, and $\rho = 0.46, p < .000001$ for the aggressively normalized norms, which we consider a surprisingly successful result concerning our simple measure. Focusing on the compound–head pairs, the feature norms reached $\rho = 0.57$ and $\rho = 0.59$, respectively.

Perceptual Model Information As mentioned in the Introduction, feature norms have also acted as a loose proxy for perceptual information in data-intensive computational models of semantic tasks. The second case study is taken from Roller and Schulte im Walde (2013), who integrated feature norms as one type of perceptual information into an extension and variations of the LDA model by Andrews et al. (2009). A bimodal LDA model integrating textual co-occurrence features and our feature norms significantly outperformed the LDA model that only relied on the textual co-occurrence. The evaluation of the LDA models was performed on the same compositionality ratings as described in the previous paragraph.

6 Conclusion

This paper presented a new collection of feature norms for 572 German noun-noun compounds. The feature norms complement existing data sets for the same targets, including compositionality ratings, association norms, and images.

We have described our collection process, and the cleaning and normalization, and we have shown both the orthographically normalized and more aggressively normalized feature norms to be of higher quality than the raw responses in terms of types per token, and that the normalized feature norms have a desirable distribution of features per cue. We also demonstrated by two case studies that the norms represent a valuable resource for research on compositionality.

References

Mark Andrews, Gabriella Vigliocco, and David Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, 116(3):463.

Ken McRae, George S. Cree, Mark S. Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559.

Stephen Roller and Stephen Schulte im Walde. 2013. A multimodal LDA model integrating textual, cognitive and visual modalities. In *Proceedings of the 2013 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1146–1157, Seattle, Washington, USA.

Sabine Schulte im Walde and Susanne Borgwaldt. 2014. Association norms for German noun compounds and their constituents. Under review.

Sabine Schulte im Walde, Alissa Melinger, Michael Roth, and Andrea Weber. 2008. An empirical characterisation of response types in German association norms. *Research on Language and Computation*, 6(2):205–238.

Sabine Schulte im Walde, Susanne Borgwaldt, and Ronny Jauch. 2012. Association norms of German noun compounds. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 632–639, Istanbul, Turkey.

Carina Silberer and Mirella Lapata. 2012. Grounded models of semantic representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1423–1433, Jeju Island, Korea.

David Vinson and Gabriella Vigliocco. 2008. Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, 40(1):183–190.

Claudia von der Heide and Susanne Borgwaldt. 2009. Assoziationen zu Unter-, Basis- und Oberbegriffen. Eine explorative Studie. In *Proceedings of the 9th Norddeutsches Linguistisches Kolloquium*, pages 51–74.

Ling-ling Wu and Lawrence W. Barsalou. 2009. Perceptual simulation in conceptual combination: Evidence from property generation. *Acta Psychologica*, 132:173–189.