The Role of Modifier and Head Properties in Predicting the Compositionality of English and German Noun-Noun Compounds: A Vector-Space Perspective

Sabine Schulte im Walde and **Anna Hätty** and **Stefan Bott**
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
Pfaffenwaldring 5B, 70569 Stuttgart, Germany
{schulte,haettyaa,bottsn}@ims.uni-stuttgart.de

Abstract

In this paper, we explore the role of constituent properties in English and German noun-noun compounds (*corpus frequencies* of the compounds and their constituents; *productivity* and *ambiguity* of the constituents; and *semantic relations* between the constituents), when predicting the degrees of compositionality of the compounds within a vector space model. The results demonstrate that the empirical and semantic properties of the compounds and the head nouns play a significant role.

1 Introduction

The past 20+ years have witnessed an enormous amount of discussions on whether and how the modifiers and the heads of noun-noun compounds such as *butterfly, snowball* and *teaspoon* influence the compositionality of the compounds, i.e., the degree of transparency vs. opaqueness of the compounds. The discussions took place mostly in psycholinguistic research, typically relying on reading time and priming experiments. For example, Sandra (1990) demonstrated in three priming experiments that both modifier and head constituents were accessed in semantically transparent English noun-noun compounds (such as *teaspoon*), but there were no effects for semantically opaque compounds (such as *buttercup*), when primed either on their modifier or head constituent. In contrast, Zwitserlood (1994) provided evidence that the lexical processing system is sensitive to morphological complexity independent of semantic transparency. Libben and his colleagues (Libben et al. (1997), Libben et al. (2003)) were the first who systematically categorised noun-noun compounds with nominal modifiers and heads into four groups representing all possible combinations of modifier and head transparency (T) vs. opaqueness (O) within a compound. Examples for these categories were *car-wash* (TT), *strawberry* (OT), *jailbird* (TO), and *hogwash* (OO). Libben et al. confirmed Zwitserlood's analyses that both semantically transparent and semantically opaque compounds show morphological constituency; in addition, the semantic transparency of the head constituent was found to play a significant role.

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. (2015b) have integrated the prediction of multi-word compositionality into statistical machine translation.

Computational approaches to automatically predict the compositionality of noun compounds have mostly been realised as vector space models, and can be subdivided into two subfields: (i) *approaches that aim to predict the **meaning** of a compound by composite functions*, relying on the vectors of the constituents (e.g., Mitchell and Lapata (2010), Coecke et al. (2011), Baroni et al. (2014), and Hermann (2014)); and (ii) *approaches that aim to predict the **degree of compositionality** of a compound*, typically by comparing the compound vectors with the constituent vectors (e.g., Reddy et al. (2011), Salehi and Cook (2013), Schulte im Walde et al. (2013), Salehi et al. (2014; 2015a)). In line with subfield (ii), this paper aims to distinguish the contributions of modifier and head properties when predicting the compositionality of English and German noun-noun compounds in a vector space model.

*Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*SEM 2016)*, pages 148–158,
Berlin, Germany, August 11-12, 2016.

Up to date, computational research on noun compounds has largely ignored the influence of constituent properties on the prediction of compositionality. Individual pieces of research noticed differences in the contributions of modifier and head constituents towards the composite functions predicting compositionality (Reddy et al., 2011; Schulte im Walde et al., 2013), but so far the roles of modifiers and heads have not been distinguished. We use a new gold standard of German noun-noun compounds annotated with *corpus frequencies* of the compounds and their constituents; *productivity* and *ambiguity* of the constituents; and *semantic relations* between the constituents; and we extend three existing gold standards of German and English noun-noun compounds (Ó Séaghdha, 2007; von der Heide and Borgwaldt, 2009; Reddy et al., 2011) to include approximately the same compound and constituent properties. Relying on a standard vector space model of compositionality, we then predict the degrees of compositionality of the English and German noun-noun compounds, and explore the influences of the compound and constituent properties. Our empirical computational analyses reveal that the empirical and semantic properties of the compounds and the head nouns play a significant role in determining the compositionality of noun compounds.

2 Related Work

Regarding relevant psycholinguistic research on the representation and processing of noun compounds, Sandra (1990) hypothesised that an associative prime should facilitate access and recognition of a noun compound, if a compound constituent is accessed during processing. His three priming experiments revealed that in transparent noun-noun compounds, both constituents are accessed, but he did not find priming effects for the constituents in opaque noun-noun compounds.

Zwitserlood (1994) performed an immediate partial repetition experiment and a priming experiment to explore and to distinguish morphological and semantic structures in noun-noun compounds. On the one hand, she confirmed Sandra's results that there is no semantic facilitation of any constituent in opaque compounds. In contrast, she found evidence for morphological complexity, independent of semantic transparency, and that both transparent and also partially opaque compounds (i.e., compounds with one transparent and one opaque constituent) produce semantic priming of their constituents. For the heads of semantically transparent compounds, a larger amount of facilitation was found than for the modifiers. Differences in the results by Sandra (1990) and Zwitserlood (1994) were supposedly due to different definitions of partial opacity, and different prime–target SOAs.

Libben and his colleagues (Libben et al. (1997), Libben (1998), and Libben et al. (2003)) were the first who systematically categorised noun-noun compounds with nominal modifiers and heads into four groups representing all possible combinations of a constituent's transparency (T) vs. opaqueness (O) within a compound: TT, OT, TO, OO. Libben's examples for these categories were *car-wash* (TT), *strawberry* (OT), *jailbird* (TO), and *hogwash* (OO). They confirmed Zwitserlood's analyses that both semantically transparent and semantically opaque compounds show morphological constituency, and also that the semantic transparency of the head constituent was found to play a significant role. Studies such as Jarema et al. (1999) and Kehayia et al. (1999) to a large extent confirmed the insights by Libben and his colleagues for French, Bulgarian, Greek and Polish.

Regarding related computational work, prominent approaches to model the meaning of a compound or a phrase by a composite function include Mitchell and Lapata (2010), Coecke et al. (2011), Baroni et al. (2014), and Hermann (2014)). In this area, researchers combine the vectors of the compound/phrase constituents by mathematical functions such that the resulting vector optimally represents the meaning of the compound/phrase. This research is only marginally related to ours, since we are interested in the degree of compositionality of a compound, rather than its actual meaning.

Most closely related computational work includes distributional approaches that predict the degree of compositionality of a compound regarding a specific constituent, by comparing the compound vector to the respective constituent vector. Most importantly, Reddy et al. (2011) used a standard distributional model to predict the compositionality of compound-constituent pairs for 90 English compounds. They extended their predictions by applying composite functions (see above). In a similar vein, Schulte im Walde et al. (2013) predicted the compositionality for 244 German compounds. Salehi et al. (2014) defined a cross-

lingual distributional model that used translations into multiple languages and distributional similarities in the respective languages, to predict the compositionality for the two datasets from Reddy et al. (2011) and Schulte im Walde et al. (2013).

3 Noun-Noun Compounds

Our focus of interest is on noun-noun compounds, such as *butterfly, snowball* and *teaspoon* as well as *car park, zebra crossing* and *couch potato* in English, and *Ahornblatt* 'maple leaf', *Feuerwerk* 'fireworks', and *Löwenzahn* 'dandelion' in German, where both the grammatical head (in English and German, this is typically the rightmost constituent) and the modifier are nouns. We are interested in the degrees of compositionality of noun-noun compounds, i.e., the semantic relatedness between the meaning of a compound (e.g., *snowball*) and the meanings of its constituents (e.g., *snow* and *ball*). More specifically, this paper aims to explore factors that have been found to influence compound processing and representation, such as

- *frequency-based factors*, i.e., the frequencies of the compounds and their constituents (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; a snowball is a ball MADE OF snow (Gagné and Spalding, 2009; Ji et al., 2011).

In addition, we were interested in the effect of *ambiguity* (of both the modifiers and the heads) regarding the compositionality of the compounds.

Our explorations required gold standards of compounds that were annotated with all these compound and constituent properties. Since most previous work on computational predictions of compositionality has been performed for English and for German, we decided to re-use existing datasets for both languages, which however required extensions to provide all properties we wanted to take into account. We also created a novel gold standard. In the following, we describe the datasets.[1]

[1]The datasets are available from `http://www.ims.uni-stuttgart.de/data/ghost-nn/`.

German Noun-Noun Compound Datasets As basis for this work, we created a novel gold standard of German noun-noun compounds: G_hOST-NN (Schulte im Walde et al., 2016). The new gold standard was built such that it includes a representative choice of compounds and constituents from various frequency ranges, various productivity ranges, with various numbers of senses, and with various semantic relations. In the following, we describe the creation process in some detail, because the properties of the gold standard are highly relevant for the distributional models.

Relying on the 11.7 billion words in the web corpus *DECOW14AX*[2] (Schäfer and Bildhauer, 2012; Schäfer, 2015), we extracted all words that were identified as common nouns by the *Tree Tagger* (Schmid, 1994) and analysed as noun compounds with exactly two nominal constituents by the morphological analyser *SMOR* (Faaß et al., 2010). This set of 154,960 two-part noun-noun compound 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 *DECOW14AX*;
- *productivity* of the constituents i.e., how many compound types contained a specific modifier/head constituent;
- *number of senses* of the compounds and the constituents, relying on *GermaNet* (Hamp and Feldweg, 1997; Kunze, 2000).

From the set of compound candidates we extracted a random subset that was balanced[3] for

- the *productivity of the modifiers*: we calculated tertiles to identify modifiers with low/mid/high productivity;
- the *ambiguity of the heads*: we distinguished between heads with 1, 2 and >2 senses.

For each of the resulting nine categories (three productivity ranges $\times$ three ambiguity ranges), we randomly selected 20 noun-noun compounds

[2]`http://corporafromtheweb.org/decow14/`

[3]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, but 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, so we focused on two main criteria instead.

from our candidate set, disregarding compounds with a corpus frequency $< 2{,}000$, and disregarding compounds containing modifiers or heads with a corpus-frequency < 100. We refer to this dataset of 180 compounds balanced for modifier productivity and head ambiguity as **G_hOST-NN/S**.

We also created a subset of 5 noun-noun compounds for each of the 9 criteria combinations, by randomly selecting 5 out of the 20 selected compounds in each mode. This small, balanced subset was then systematically extended 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. Taking *Haarpracht* as an example (the modifier is *Haar* 'hair', the head is *Pracht* 'glory'), we added *Haarwäsche, Haarkleid, Haarpflege, etc.* as well as *Blütenpracht, Farbenpracht, etc.*[4] We refer to this dataset of 868 compounds that destroyed the coherent balance of criteria underlying our random extraction, but instead ensured a variety of compounds with either the same modifiers or the same heads, as **G_hOST-NN/XL**.

The two sets of compounds (G_hOST-NN/S and G_hOST-NN/XL) were annotated with the semantic relations between the modifiers and the heads, and compositionality ratings. Regarding *semantic relations*, we applied the relation set suggested by Ó Séaghdha (2007), because (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 languages. Ó Séaghdha (2007) himself had 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 referring to event participants (`ACTOR`, `INST(rument)`) that replaced the relations `MAKE`, `CAUSE`, `FOR`, `FROM`, `USE`. An additional relation `LEX` refers to lexicalised compounds where no relation can be assigned. Three native speakers of German annotated the compounds with these seven semantic relations.[5] Regarding *compositionality ratings*, eight native speakers of German annotated all 868 gold-standard compounds with compound–constituent compositionality ratings on a scale from 1 (definitely semantically opaque) to 6 (definitely semantically transparent). Another five native speakers provided additional annotation for our small core subset of 180 compounds on the same scale. As final compositionality ratings, we use the mean compound–constituent ratings across the 13 annotators.

As alternative gold standard for German noun-noun compounds, we used a dataset based on a selection of noun compounds by von der Heide and Borgwaldt (2009), that was previously used in computational models predicting compositionality (Schulte im Walde et al., 2013; Salehi et al., 2014). The dataset contains a subset of their compounds including 244 two-part noun-noun compounds, annotated by compositionality ratings on a scale between 1 and 7. We enriched the existing dataset with frequencies, and productivity and ambiguity scores, also based on *DECOW14AX* and *GermaNet*, to provide the same empirical information as for the G_hOST-NN datasets. We refer to this alternative German dataset as **VDHB**.

English Noun-Noun Compound Datasets Reddy et al. (2011) created a gold standard for English noun-noun compounds. Assuming that compounds whose constituents appeared either as their hypernyms or in their definitions tend to be compositional, they induced a candidate compound set with various degrees of compound–constituent relatedness from *WordNet* (Miller et al., 1990; Fellbaum, 1998) and *Wiktionary*. A random choice of 90 compounds that appeared with a corpus frequency > 50 in the *ukWaC* corpus (Baroni et al., 2009) constituted their gold-standard dataset and was annotated by compositionality ratings. Bell and Schäfer (2013) annotated the compounds with semantic relations using all of Levi's original nine relation types: `CAUSE`, `HAVE`, `MAKE`, `USE`, `BE`, `IN`, `FOR`, `FROM`, `ABOUT`. We refer to this dataset as **REDDY**.

Ó Séaghdha developed computational models to predict the semantic relations between modifiers and heads in English noun compounds (Ó Séaghdha, 2008; Ó Séaghdha and Copestake, 2013; Ó Séaghdha and Korhonen, 2014). As gold-standard basis for his models, he created a dataset of compounds, and annotated the compounds with semantic relations: He tagged and parsed the written part of the British National Cor-

[4] The translations of the example compounds are *hair washing, hair dress, hair care, floral glory*, and *colour glory*.

[5] In fact, the annotation was performed for a superset of 1,208 compounds, but we only took into account 868 compounds with perfect agreement, i.e. IAA=1.

Language	*Dataset*	*#Compounds*	*Annotation*		
			Frequency/Productivity	*Ambiguity*	*Relations*
DE	G_hOST-NN/S	180	DECOW	GermaNet	Levi (7)
	G_hOST-NN/XL	868	DECOW	GermaNet	Levi (7)
	VDHB	244	DECOW	GermaNet	–
EN	REDDY	90	ENCOW	WordNet	Levi (9)
	OS	396	ENCOW	WordNet	Levi (6)

Table 1: Noun-noun compound datasets.

pus using *RASP* (Briscoe and Carroll, 2002), and applied a simple heuristics to induce compound candidates: He used all sequences of two or more common nouns that were preceded or followed by sentence boundaries or by words not representing common nouns. Of these compound candidates, a random selection of 2,000 instances was used for relation annotation (Ó Séaghdha, 2007) and classification experiments. The final gold standard is a subset of these compounds, containing 1,443 noun-noun compounds. We refer to this dataset as **OS**.

Both English compound datasets were enriched with frequencies and productivities, based on the *ENCOW14AX*[6] containing 9.6 billion words. We also added the number of senses of the constituents to both datasets, using *WordNet*. And we collected compositionality ratings for a random choice of 396 compounds from the OS dataset relying on eight experts, in the same way as the G_hOST-NN ratings were collected.

Resulting Noun-Noun Compound Datasets Table 1 summarises the gold-standard datasets. They are of different sizes, but their empirical and semantic annotations have been aligned to a large extent, using similar corpora, relying on WordNets and similar semantic relation inventories based on Levi (1978).

4 VSMs Predicting Compositionality

Vector space models (VSMs) and distributional information have been a steadily increasing, integral part of lexical semantic research over the past 20 years (Turney and Pantel, 2010): They explore the notion of "similarity" between a set of target objects, typically relying on the *distributional hypothesis* (Harris, 1954; Firth, 1957) to determine co-occurrence features that best describe the words, phrases, sentences, etc. of interest.

[6] http://corporafromtheweb.org/encow14/

In this paper, we use VSMs in order to model compounds as well as constituents by distributional vectors, and we determine the semantic relatedness between the compounds and their modifier and head constituents by measuring the distance between the vectors. We assume that the closer a compound vector and a constituent vector are to each other, the more compositional (i.e., the more transparent) the compound is, regarding that constituent. Correspondingly, the more distant a compound vector and a constituent vector are to each other, the less compositional (i.e., the more opaque) the compound is, regarding that constituent.

Our main questions regarding the VSMs are concerned with the influence of constituent properties on the prediction of compositionality. I.e., how do the *corpus frequencies* of the compounds and their constituents, the *productivity* and the *ambiguity* of the constituents, and the *semantic relations* between the constituents influence the quality of the predictions?

4.1 Vector Space Models (VSMs)

We created a standard vector space model for all our compounds and constituents in the various datasets, using co-occurrence frequencies of nouns within a sentence-internal window of 20 words to the left and 20 words to the right of the targets.[7] The frequencies were induced from the German and English *COW* corpora, and transformed to *local mutual information (LMI)* values (Evert, 2005).

Relying on the LMI vector space models, the *cosine* determined the distributional similarity between the compounds and their constituents, which was in turn used to predict the degree

[7] In previous work, we systematically compared window-based and syntax-based co-occurrence variants for predicting compositionality (Schulte im Walde et al., 2013). The current work adopted the best choice of co-occurrence dimensions.

of compositionality between the compounds and their constituents, assuming that the stronger the distributional similarity (i.e., the cosine values), the larger the degree of compositionality. The vector space predictions were evaluated against the mean human ratings on the degree of compositionality, using the Spearman Rank-Order Correlation Coefficient ρ (Siegel and Castellan, 1988).

4.2 Overall VSM Prediction Results

Table 2 presents the overall prediction results across languages and datasets. The *mod* column shows the ρ correlations for predicting only the degree of compositionality of compound–modifier pairs; the *head* column shows the ρ correlations for predicting only the degree of compositionality of compound–head pairs; and the *both* column shows the ρ correlations for predicting the degree of compositionality of compound–modifier and compound–head pairs at the same time.

Dataset		mod	head	both
DE	G_hOST-NN/S	0.48	0.57	0.46
	G_hOST-NN/XL	0.49	0.59	0.47
	VDHB	0.65	0.60	0.61
EN	REDDY	0.48	0.60	0.56
	OS	0.46	0.39	0.35

Table 2: Overall prediction results (ρ).

The models for VDHB and REDDY represent replications of similar models in Schulte im Walde et al. (2013) and Reddy et al. (2011), respectively, but using the much larger *COW* corpora.

Overall, the *both* prediction results on VDHB are significantly[8] better than all others but REDDY; and the prediction results on OS compounds are significantly worse than all others. We can also compare within-dataset results: Regarding the two G_hOST-NN datasets and the REDDY dataset, the VSM predictions for the compound–head pairs are better than for the compound–modifier pairs. Regarding the VDHB and the OS datasets, the VSM predictions for the compound–modifier pairs are better than for the compound–head pairs. These differences do not depend on the language (according to our datasets), and are probably due to properties of the specific gold standards that we did not control. They are, however, also not the main point of this paper.

[8]All significance tests in this paper were performed by Fisher r-to-z transformation.

4.3 Influence of Compound Properties on VSM Prediction Results

Figures 1 to 5 present the core results of this paper: They explore the influence of compound and constituent properties on predicting compositionality. Since we wanted to optimise insight into the influence of the properties, we selected the 60 maximum instances and the 60 minimum instances for each property.[9] For example, to explore the influence of head frequency on the prediction quality, we selected the 60 most frequent and the 60 most infrequent compound heads from each gold-standard resource, and calculated Spearman's ρ for each set of 60 compounds with these heads.

Figure 1 shows that the distributional model predicts high-frequency compounds (red bars) better than low-frequency compounds (blue bars), across datasets. The differences are significant for G_hOST-NN/XL.

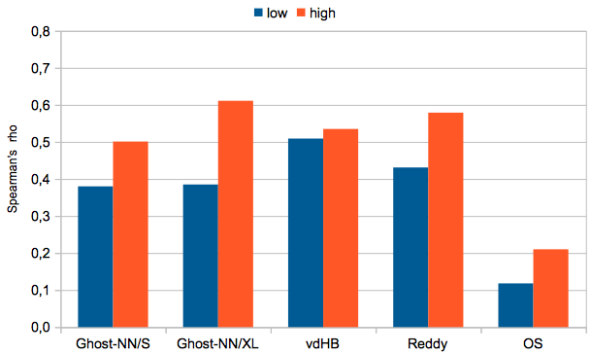

Figure 1: Effect of ***compound frequency***.

Figure 2 shows that the distributional model predicts compounds with low-frequency heads better than compounds with high-frequency heads (right panel), while there is no tendency regarding the modifier frequencies (left panel). The differences regarding the head frequencies are significant ($p = 0.1$) for both G_hOST-NN datasets.

Figure 3 shows that the distributional model also predicts compounds with low-productivity heads better than compounds with high-productivity heads (right panel), while there is no tendency regarding the productivities of modifiers (left panel). The prediction differences regarding the head productivities are significant for G_hOST-NN/S ($p < 0.05$).

[9]For REDDY, we could only use 45 maximum/minimum instances, since the dataset only contains 90 compounds.

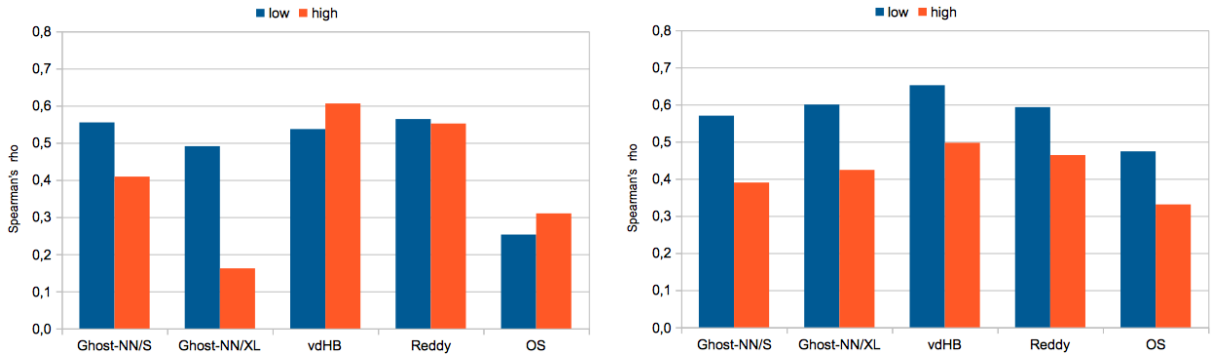

Figure 2: Effect of ***modifier/head frequency***.

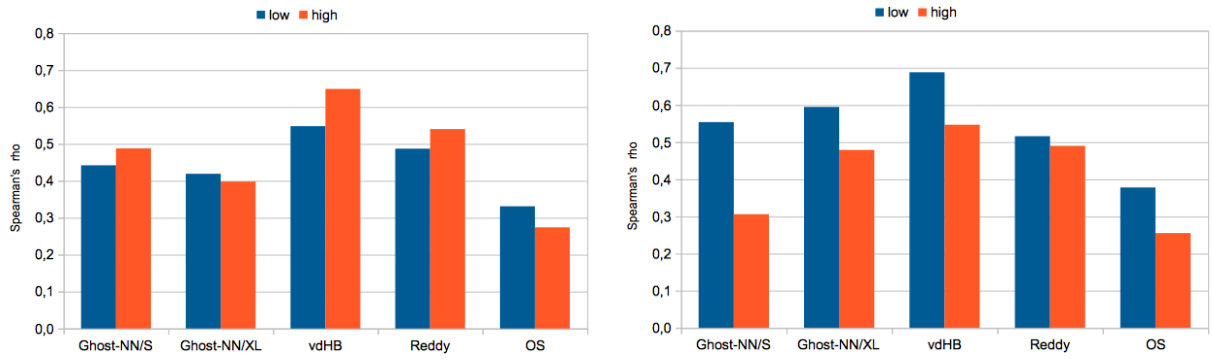

Figure 3: Effect of ***modifier/head productivity***.

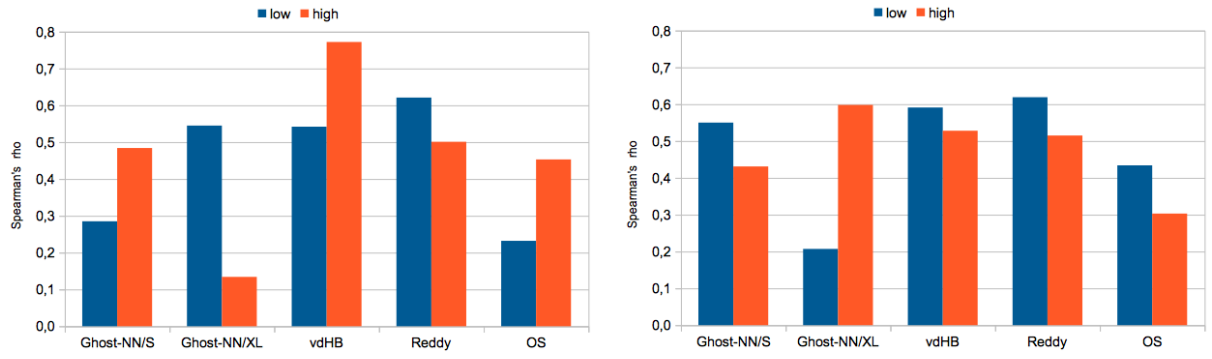

Figure 4: Effect of ***modifier/head ambiguity***.

Figure 4 shows that the distributional model also predicts compounds with low-ambiguity heads better than compounds with high-ambiguity heads (right panel) –with one exception (G_hOST-NN/XL)– while there is no tendency regarding the ambiguities of modifiers (left panel). The prediction differences regarding the head ambiguities are significant for G_hOST-NN/XL ($p < 0.01$).

Figure 5 compares the predictions of the distributional model regarding the semantic relations between modifiers and heads, focusing on G_hOST-NN/XL. The numbers in brackets refer to the number of compounds with the respective relation. The plot reveals differences between predictions of compounds with different relations.

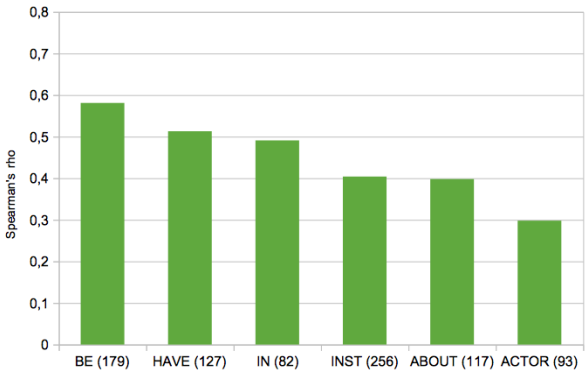

Figure 5: Effect of ***semantic relation***.

Table 3 summarises those differences across gold standards that are significant (where filled cells refer to rows significantly outperforming columns). Overall, the compositionality of BE compounds is predicted significantly better than the compositionality of HAVE compounds (in REDDY), INST and ABOUT compounds (in G_hOST-NN) and ACTOR compounds (in G_hOST-NN and OS). The compositionality of ACTOR compounds is predicted significantly worse than the compositionality of BE, HAVE, IN and INST compounds in both G_hOST-NN and OS.

	HAVE	INST	ABOUT	ACTOR
BE	REDDY	G_hOST	G_hOST	G_hOST, OS
HAVE			OS	G_hOST, OS
IN				G_hOST, OS
INST				G_hOST, OS

Table 3: Significant differences: relations.

5 Discussion

While modifier frequency, productivity and ambiguity did not show a consistent effect on the predictions, head frequency, productivity and ambiguity influenced the predictions such that the prediction quality for compounds with low-frequency, low-productivity and low-ambiguity heads was better than for compounds with high-frequency, high-productivity and high-ambiguity heads. The differences were significant only for our new G_hOST-NN datasets. In addition, the compound frequency also had an effect on the predictions, with high-frequency compounds receiving better prediction results than low-frequency compounds. Finally, the quality of predictions also differed for compound relation types, with BE compounds predicted best, and ACTOR compounds predicted worst. These differences were ascertained mostly in the G_hOST-NN and the OS datasets. Our results raise two main questions:

(1) What does it mean if a distributional model predicts a certain subset of compounds (with specific properties) "better" or "worse" than other subsets?

(2) What are the implications for (a) psycholinguistic and (b) computational models regarding the compositionality of noun compounds?

Regarding question (1), there are two options why a distributional model predicts a certain subset of compounds better or worse than other subsets. On the one hand, one of the underlying gold-standard datasets could contain compounds whose compositionality scores are easier to predict than the compositionality scores of compounds in a different dataset. On the other hand, even if there were differences in individual dataset pairs, this would not explain why we consistently find modelling differences for head constituent properties (and compound properties) but not for modifier constituent properties. We therefore conclude that the effects of compound and head properties are due to the compounds' morphological constituency, with specific emphasis on the influences of the heads.

Looking at the individual effects of the compound and head properties that influence the distributional predictions, we hypothesise that high-frequent compounds are easier to predict because they have a better corpus coverage (and less

sparse data) than low-frequent compounds, and that they contain many clearly transparent compounds (such as *Zitronensaft* 'lemon juice'), and at the same time many clearly opaque compounds (such as *Eifersucht* 'jealousy', where the literal translations of the constituents are 'eagerness' and 'addiction'). Concerning the decrease in prediction quality for more frequent, more productive and more ambiguous heads, we hypothesise that all of these properties are indicators of ambiguity, and the more ambiguous a word is, the more difficult it is to provide a unique distributional prediction, as distributional co-occurrence in most cases (including our current work) subsumes the contexts of all word senses within one vector. For example, more than half of the compounds with the most frequent and also with the most productive heads have the head *Spiel*, which has six senses in GermaNet and covers six relations (`BE, IN, INST, ABOUT, ACTOR, LEX`).

Regarding question (2), the results of our distributional predictions confirm psycholinguistic research that identified morphological constituency in noun-noun compounds: Our models clearly distinguish between properties of the whole compounds, properties of the modifier constituents, and properties of the head constituents. Furthermore, our models reveal the need to carefully balance the frequencies and semantic relations of target compounds, and to carefully balance the frequencies, productivities and ambiguities of their head constituents, in order to optimise experiment interpretations, while a careful choice of empirical modifier properties seems to play a minor role.

For computational models, our work provides similar implications. We demonstrated the need to carefully balance gold-standard datasets for multi-word expressions according to the empirical and semantic properties of the multi-word expressions themselves, and also according to those of the constituents. In the case of noun-noun compounds, the properties of the nominal modifiers were of minor importance, but regarding other multi-word expressions, this might differ. If datasets are not balanced for compound and constituent properties, the qualities of model predictions are difficult to interpret, because it is not clear whether biases in empirical properties skewed the results. Our advice is strengthened by the fact that most significant differences in prediction results were demonstrated for our new gold standard, which includes compounds across various frequency, productivity and ambiguity ranges.

6 Conclusion

We explored the role of constituent properties in English and German noun-noun compounds, when predicting compositionality within a vector space model. The results demonstrated that the empirical and semantic properties of the compounds and the head nouns play a significant role. Therefore, psycholinguistic experiments as well as computational models are advised to carefully balance their selections of compound targets according to compound and constituent properties.

Acknowledgments

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).

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.

Marco Baroni, Raffaella Bernardi, and Roberto Zamparelli. 2014. Frege in Space: A Program for Compositional Distributional Semantics. *Linguistic Issues in Language Technologies*, 9(6):5–110.

Melanie J. Bell and Martin Schäfer. 2013. Semantic Transparency: Challenges for Distributional Semantics. In *Proceedings of the IWCS Workshop on Formal Distributional Semantics*, pages 1–10, Potsdam, Germany.

Ted Briscoe and John Carroll. 2002. Robust Accurate Statistical Annotation of General Text. In *Proceedings of the 3rd Conference on Language Resources and Evaluation*, pages 1499–1504, Las Palmas de Gran Canaria, Spain.

Fabienne Cap, Manju Nirmal, Marion Weller, and Sabine Schulte im Walde. 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.

Kostadin Cholakov and Valia Kordoni. 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.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2011. Mathematical Foundations for a Compositional Distributional Model of Meaning. *Linguistic Analysis*, 36(1-4):345–384.

Nicole H. de Jong, Laurie B. Feldman, Robert Schreuder, Michael Pastizzo, and Harald R. Baayen. 2002. The Processing and Representation of Dutch and English Compounds: Peripheral Morphological and Central Orthographic Effects. *Brain and Language*, 81:555–567.

Stefan Evert. 2005. *The Statistics of Word Co-Occurrences: Word Pairs and Collocations*. Ph.D. thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Gertrud Faaß, Ulrich Heid, and Helmut Schmid. 2010. Design and Application of a Gold Standard for Morphological Analysis: SMOR in Validation. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 803–810, Valletta, Malta.

Christiane Fellbaum, editor. 1998. *WordNet – An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

John R. Firth. 1957. *Papers in Linguistics 1934-51*. Longmans, London, UK.

Christina L. Gagné and Thomas L. Spalding. 2009. Constituent Integration during the Processing of Compound Words: Does it involve the Use of Relational Structures? *Journal of Memory and Language*, 60:20–35.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet – A Lexical-Semantic Net for German. In *Proceedings of the ACL Workshop on Automatic Information Extraction and Building Lexical Semantic Resources for NLP Applications*, pages 9–15, Madrid, Spain.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Karl Moritz Hermann. 2014. *Distributed Representations for Compositional Semantics*. Ph.D. thesis, University of Oxford.

Niels Janssen, Yanchao Bi, and Alfonso Caramazza. 2008. A Tale of Two Frequencies: Determining the Speed of Lexical Access for Mandarin Chinese and English Compounds. *Language and Cognitive Processes*, 23:1191–1223.

Gonia Jarema, Celine Busson, Rossitza Nikolova, Kyrana Tsapkini, and Gary Libben. 1999. Processing Compounds: A Cross-Linguistic Study. *Brain and Language*, 68:362–369.

Hongbo Ji, Christina L. Gagné, and Thomas L. Spalding. 2011. Benefits and Costs of Lexical Decomposition and Semantic Integration during the Processing of Transparent and Opaque English Compounds. *Journal of Memory and Language*, 65:406–430.

Eva Kehayia, Gonia Jarema, Kyrana Tsapkini, Danuta Perlak, Angela Ralli, and Danuta Kadzielawa. 1999. The Role of Morphological Structure in the Processing of Compounds: The Interface between Linguistics and Psycholinguistics. *Brain and Language*, 68:370–377.

Claudia Kunze. 2000. Extension and Use of GermaNet, a Lexical-Semantic Database. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pages 999–1002, Athens, Greece.

Judith N. Levi. 1978. *The Syntax and Semantics of Complex Nominals*. Academic Press, London.

Gary Libben, Martha Gibson, Yeo Bom Yoon, and Dominiek Sandra. 1997. Semantic Transparency and Compound Fracture. Technical Report 9, CLASNET Working Papers.

Gary Libben, Martha Gibson, Yeo Bom Yoon, and Dominiek Sandra. 2003. Compound Fracture: The Role of Semantic Transparency and Morphological Headedness. *Brain and Language*, 84:50–64.

Gary Libben. 1998. Semantic Transparency in the Processing of Compounds: Consequences for Representation, Processing, and Impairment. *Brain and Language*, 61:30–44.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to Wordnet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4):235–244.

Jeff Mitchell and Mirella Lapata. 2010. Composition in Distributional Models of Semantics. *Cognitive Science*, 34:1388–1429.

Diarmuid Ó Séaghdha and Ann Copestake. 2013. Interpreting Compound Nouns with Kernel Methods. *Journal of Natural Language Engineering*, 19(3):331–356.

Diarmuid Ó Séaghdha and Anna Korhonen. 2014. Probabilistic Distributional Semantics with Latent Variable Models. *Computational Linguistics*, 40(3):587–631.

Diarmuid Ó Séaghdha. 2007. Designing and Evaluating a Semantic Annotation Scheme for Compound Nouns. In *Proceedings of Corpus Linguistics*, Birmingham, UK.

Diarmuid Ó Séaghdha. 2008. *Learning Compound Noun Semantics*. Ph.D. thesis, University of Cambridge, Computer Laboratory. Technical Report UCAM-CL-TR-735.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An Empirical Study on Compositionality in Compound Nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand.

Bahar Salehi and Paul Cook. 2013. Predicting the Compositionality of Multiword Expressions Using Translations in Multiple Languages. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics*, pages 266–275, Atlanta, GA.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014. Using Distributional Similarity of Multi-way Translations to Predict Multiword Expression Compositionality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 472–481, Gothenburg, Sweden.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015a. A Word Embedding Approach to Predicting the Compositionality of Multiword Expressions. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies*, pages 977–983, Denver, Colorado, USA.

Bahar Salehi, Nitika Mathur, Paul Cook, and Timothy Baldwin. 2015b. The Impact of Multiword Expression Compositionality on Machine Translation Evaluation. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 54–59, Denver, Colorado, USA.

Dominiek Sandra. 1990. On the Representation and Processing of Compound Words: Automatic Access to Constituent Morphemes does not occur. *The Quarterly Journal of Experimental Psychology*, 42A:529–567.

Roland Schäfer and Felix Bildhauer. 2012. Building Large Corpora from the Web Using a New Efficient Tool Chain. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 486–493, Istanbul, Turkey.

Roland Schäfer. 2015. Processing and Querying Large Web Corpora with the COW14 Architecture. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora*, pages 28–34, Mannheim, Germany.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging using Decision Trees. In *Proceedings of the 1st International Conference on New Methods in Language Processing*.

Sabine Schulte im Walde, Stefan Müller, and Stephen Roller. 2013. Exploring Vector Space Models to Predict the Compositionality of German Noun-Noun Compounds. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics*, pages 255–265, Atlanta, GA.

Sabine Schulte im Walde, Anna Hätty, Stefan Bott, and Nana Khvtisavrishvili. 2016. G_hoSt-NN: A Representative Gold Standard of German Noun-Noun Compounds. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 2285–2292, Portoroz, Slovenia.

Sidney Siegel and N. John Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, Boston, MA.

Peter D. Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Henk J. van Jaarsveld and Gilbert E. Rattink. 1988. Frequency Effects in the Processing of Lexicalized and Novel Nominal Compounds. *Journal of Psycholinguistic Research*, 17:447–473.

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.

Marion Weller, Fabienne Cap, Stefan Müller, Sabine Schulte im Walde, and Alexander Fraser. 2014. Distinguishing Degrees of Compositionality in Compound Splitting for Statistical Machine Translation. In *Proceedings of the 1st Workshop on Computational Approaches to Compound Analysis*, pages 81–90, Dublin, Ireland.

Pienie Zwitserlood. 1994. The Role of Semantic Transparency in the Processing and Representation of Dutch Compounds. *Language and Cognitive Processes*, 9:341–368.