Exploring Vector Space Models to Predict the Compositionality of German Noun-Noun Compounds

Sabine Schulte im Walde and Stefan Müller and Stephen Roller
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Pfaffenwaldring 5b, 70569 Stuttgart, Germany
{schulte,muellesn,roller}@ims.uni-stuttgart.de

Abstract

This paper explores two hypotheses regarding vector space models that predict the compositionality of German noun-noun compounds: (1) Against our intuition, we demonstrate that window-based rather than syntax-based distributional features perform better predictions, and that not adjectives or verbs but nouns represent the most salient part-of-speech. Our overall best result is state-of-the-art, reaching Spearman's $\rho = 0.65$ with a word-space model of nominal features from a 20-word window of a 1.5 billion word web corpus. (2) While there are no significant differences in predicting compound–modifier vs. compound–head ratings on compositionality, we show that the modifier (rather than the head) properties predominantly influence the degree of compositionality of the compound.

1 Introduction

Vector space models and distributional information have been a steadily increasing, integral part of lexical semantic research over the past 20 years. On the one hand, *vector space models* (see Turney and Pantel (2010) and Erk (2012) for two recent surveys) have been exploited in psycholinguistic (Lund and Burgess, 1996) and computational linguistic research (Schütze, 1992) to explore the notion of "similarity" between a set of target objects within a geometric setting. On the other hand, the *distributional hypothesis* (Firth, 1957; Harris, 1968) has been exploited to determine co-occurrence features for vector space models that best describe the words, phrases, sentences, etc. of interest.

While the emergence of vector space models is increasingly pervasive within data-intensive lexical semantics, and even though useful features have been identified in general terms:[1] when it comes to a specific semantic phenomenon, we need to explore the relevant distributional features in order to investigate the respective phenomenon. Our research is interested in the meaning of German compounds. More specifically, we aim to predict the degrees of compositionality of German noun-noun compounds (e.g., *Feuerwerk* 'fire works') with regard to the meanings of their constituents (e.g., *Feuer* 'fire' and *Werk* 'opus'). This prediction uses vector space models, and our goal is to identify salient features that determine the degree of compositionality of the compounds by relying on the distributional similarities between the compounds and their constituents.

In this vein, we systematically explore window-based and syntax-based contextual clues. Since the targets in our vector space models are all nouns (i.e., the compound nouns, the modifier nouns, and the head nouns), our hypothesis is that ***adjectives and verbs are expected to provide salient distributional properties***, as adjective/verb meaning and noun meaning are in a strong interdependent relationship. Even more, we expect adjectives and verbs that are syntactically bound to the nouns under consideration (syntax-based, i.e., attributive adjectives and subcategorising verbs) to outperform those that "just" appear in the window contexts of the nouns (window-based). In order to investigate this first

[1] See Agirre et al. (2009) and Bullinaria and Levy (2007; 2012), among others, for systematic comparisons of co-occurrence features on various semantic relatedness tasks.

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, pages 255–265, Atlanta, Georgia, June 13-14, 2013.

hypothesis, we compare window-based and syntax-based distributional features across parts-of-speech.

Concerning a more specific aspect of compound meaning, we are interested in the contributions of the *modifier* noun versus *head* noun properties with regard to the meaning of the noun-noun compounds. While there has been prior psycholinguistic research on the constituent contributions (e.g., Gagné and Spalding (2009; 2011)), computational linguistics has not yet paid much attention to this issue, as far as we know. Our hypothesis is that ***the distributional properties of the head constituents are more salient than the distributional properties of the modifier constituents in predicting the degree of compositionality of the compounds***. In order to assess this second hypothesis, we compare the vector space similarities between the *compounds and their modifier constituents* with those of the *compounds and their head constituents*, with regard to the overall most successful features.

The paper is organised as follows. Section 2 introduces the compound data that is relevant for this paper, i.e., the noun-noun compounds and the compositionality ratings. Section 3 performs and discusses the vector space experiments to explore our hypotheses, and Section 4 describes related work.

2 Data

2.1 German Noun-Noun Compounds

Compounds are combinations of two or more simplex words. Traditionally, a number of criteria (such as compounds being syntactically inseparable, and that compounds have a specific stress pattern) have been proposed, in order to establish a border between compounds and non-compounds. However, Lieber and Stekauer (2009a) demonstrated that none of these tests are universally reliable to distinguish compounds from other types of derived words.

Compounds have thus been a recurrent focus of attention within theoretical, cognitive, and in the last decade also within computational linguistics. Recent evidence of this strong interest are the Handbook of Compounding (Lieber and Stekauer, 2009b) on theoretical perspectives, and a series of workshops[2] and special journal issues with respect to multi-word expressions (including various types of compounds) and the computational perspective (Journal of Computer Speech and Language, 2005; Language Resources and Evaluation, 2010; ACM Transactions on Speech and Language Processing, to appear).

Our focus of interest is on German noun-noun compounds (see Fleischer and Barz (2012) for a detailed overview and Klos (2011) for a recent detailed exploration), 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').

Our work is based on a selection of noun compounds by von der Heide and Borgwaldt (2009), who created a set of 450 concrete, depictable German noun compounds according to four compositionality classes: compounds that are transparent with regard to both constituents (e.g., *Ahornblatt* 'maple leaf'); compounds that are opaque with regard to both constituents (e.g., *Löwenzahn* 'lion+tooth → dandelion'); compounds that are transparent with regard to the modifier but opaque with regard to the head (e.g., *Feuerzeug* 'fire+stuff → lighter'); and compounds that are opaque with regard to the modifier but transparent with regard to the head (e.g., *Fliegenpilz* 'fly+mushroom → toadstool').

From the compound set by von der Heide and Borgwaldt, we disregarded noun compounds with more than two constituents (in some cases, the modifier or the head was complex itself) as well as compounds where the modifiers were not nouns. Our final set comprises a subset of their compounds including 244 two-part noun-noun compounds.

2.2 Compositionality Ratings

von der Heide and Borgwaldt (2009) collected human ratings on compositionality for all their 450 compounds. The compounds were distributed over 5 lists, and 270 participants judged the degree of compositionality of the compounds with respect to their first as well as their second constituent, on

[2] www.multiword.sourceforge.net

Compounds			Mean Ratings and Standard Deviations			
whole	literal meanings of constituents		whole	modifier	head	mean range
Ahornblatt 'maple leaf'	maple	leaf	6.03 ± 1.49	**5.64** ± 1.63	**5.71** ± 1.70	(1) high/high
Postbote 'post man'	mail	messenger	6.33 ± 0.96	**5.87** ± 1.55	**5.10** ± 1.99	
Seezunge 'sole'	sea	tongue	1.85 ± 1.28	**3.57** ± 2.42	**3.27** ± 2.32	(2) mid/mid
Windlicht 'storm lamp'	wind	light	3.52 ± 2.08	**3.07** ± 2.12	**4.27** ± 2.36	
Löwenzahn 'dandelion'	lion	tooth	1.66 ± 1.54	**2.10** ± 1.84	**2.23** ± 1.92	(3) low/low
Maulwurf 'mole'	mouth	throw	1.58 ± 1.43	**2.21** ± 1.68	**2.76** ± 2.10	
Fliegenpilz 'toadstool'	fly/bow tie	mushroom	2.00 ± 1.20	**1.93** ± 1.28	**6.55** ± 0.63	(4) low/high
Flohmarkt 'flea market'	flea	market	2.31 ± 1.65	**1.50** ± 1.22	**6.03** ± 1.50	
Feuerzeug 'lighter'	fire	stuff	4.58 ± 1.75	**5.87** ± 1.01	**1.90** ± 1.03	(5) high/low
Fleischwolf 'meat chopper'	meat	wolf	1.70 ± 1.05	**6.00** ± 1.44	**1.90** ± 1.42	

Table 1: Examples of compound ratings.

a scale between 1 (definitely opaque) and 7 (definitely transparent). For each compound–constituent pair, they collected judgements from 30 participants, and calculated the rating mean and the standard deviation. We refer to this set as our ***compound–constituent ratings***.

A second experiment collected human ratings on compositionality for our subset of 244 noun-noun compounds. In this case, we asked the participants to provide a unique score for each compound as a whole, again on a scale between 1 and 7. The collection was performed via Amazon Mechanical Turk (AMT)[3]. We randomly distributed our subset of 244 compounds over 21 batches, with 12 compounds each, in random order. In order to control for spammers, we also included two German fake compound nouns into each of the batches, in random positions of the lists. If participants did not recognise the fake words, all of their ratings were rejected. We collected between 27 and 34 ratings per target compound. For each of the compounds we calculated the rating mean and the standard deviation. We refer to this second set as our ***compound whole ratings***.

Table 1 presents example mean ratings for the compound–constituent ratings as well as for the compound whole ratings, accompanied by the standard deviations. We selected two examples each for five categories of mean ratings: the compound–constituent ratings were (1) high or (2) mid or (3) low with regard to both constituents; the compound–constituent ratings were (4) low with regard to the modifier but high with regard to the head; (5) vice versa. Roller et al. (2013) performed a thorough

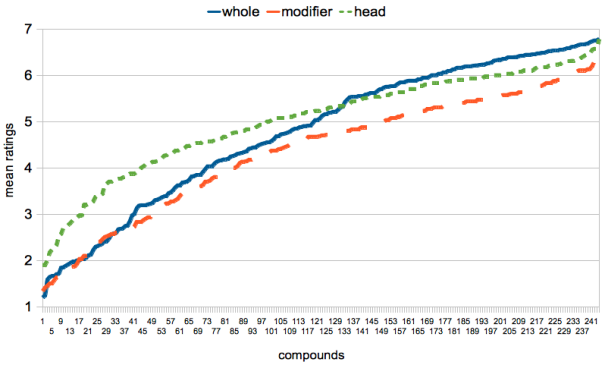

Figure 1: Distribution of compound ratings.

analysis of the two sets of ratings, and assessed their reliability from several perspectives.

Figure 1 shows how the mean ratings for the compounds as a whole, for the compound–modifier pairs as well as for the compound–head pairs are distributed over the range $[1, 7]$: For each set, we independently sorted the 244 values and plotted them. The purpose of the figure is to illustrate that the ratings for our 244 noun-noun compounds are not particularly skewed to any area within the range.[4]

Figure 2 again shows the mean ratings for the compounds as a whole as well as for the compound–constituent pairs, but in this case only the compound whole ratings were sorted, and the compound–constituent ratings were plotted against the compound whole ratings. According to the plot, the compound–modifier ratings (red) seem to correlate better with the compound whole ratings than the compound–head ratings (yellow) do. This intuition will be confirmed in Section 3.1.

[3] www.mturk.com

[4] The illustration idea was taken from Reddy et al. (2011b).

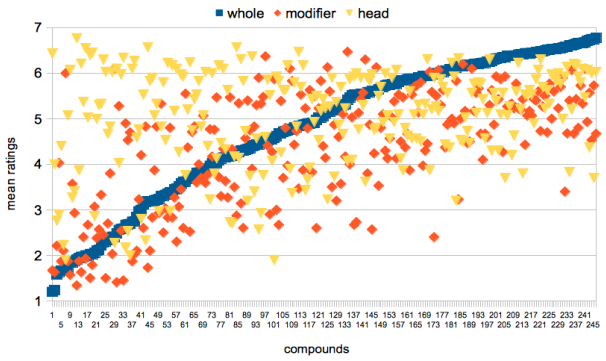

Figure 2: Compounds ratings sorted by *whole* ratings.

3 Vector Space Models (VSMs)

The goal of our vector space models is to identify distributional features that are salient to predict the degree of compositionality of the compounds, by relying on the similarities between the compound and constituent properties.

In all our vector space experiments, we used co-occurrence frequency counts as induced from German web corpora, and calculated *local mutual information (LMI)*[5] values (Evert, 2005), to instantiate the empirical properties of our target nouns with regard to the various corpus-based features. LMI is a measure from information theory that compares the observed frequencies O with expected frequencies E, taking marginal frequencies into account:

$$LMI = O \times log \frac{O}{E},$$

with E representing the product of the marginal frequencies over the sample size.[6] In comparison to (pointwise) mutual information (Church and Hanks, 1990), LMI improves the problem of propagating low-frequent events, by multiplying mutual information by the observed frequency.

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 compositionality between the compound and the constituents, assuming that the stronger the distributional similarity (i.e., the *cosine* values), the larger the degree of compositionality.

[5]Alternatively, we also used the raw frequencies in all experiments below. The insights into the various features were identical to those based on LMI, but the predictions were worse.

[6]See `http://www.collocations.de/AM/` for a more detailed illustration of association measures (incl. LMI).

The vector space predictions were evaluated against the human ratings on the degree of compositionality, using the Spearman Rank-Order Correlation Coefficient ρ (Siegel and Castellan, 1988). The ρ correlation is a non-parametric statistical test that measures the association between two variables that are ranked in two ordered series. In Section 3.3 we will compare the overall effect of the various feature types and correlate all 488 compound–modifier and compound–head predictions against the ratings at the same time; in Section 3.4 we will compare the different effects of the features for compound–modifier pairs vs. compound–head pairs and thus correlate 244 predictions in both cases.

After introducing a baseline and an upper bound for our vector space experiments in Section 3.1 as well as our web corpora in Section 3.2, Section 3.3 presents window-based in comparison to syntax-based vector space models (distinguishing various part-of-speech features). In Section 3.4 we then focus on the contribution of modifiers vs. heads in the vector space models, with regard to the overall most successful features.

3.1 Baseline and Upper Bound

Table 2 presents the baseline and the upper bound values for the vector space experiments. The *baseline* in the first two lines follows a procedure performed by Reddy et al. (2011b), and relies on a *random assignment* of rating values $[1, 7]$ to the compound–modifier and the compound–head pairs. The 244 random values for the compound–constituent pairs were then each correlated against the compound whole ratings. The random compound–modifier ratings show a baseline correlation of $\rho = 0.0959$ with the compound whole ratings, and the random compound–head ratings show a baseline correlation of $\rho = 0.1019$ with the compound whole ratings.

The *upper bound* in the first two lines shows the correlations between the human ratings from the two experiments, i.e., between the 244 compound whole ratings and the respective compound–modifier and compound–head ratings. The compound–modifier ratings exhibit a strong correlation with the compound whole ratings ($\rho = 0.6002$), while the correlation between the compound–head ratings and the compound whole ratings is not even moderate

Function	ρ	
	Baseline	Upper Bound
modifier only	.0959	.6002
head only	.1019	.1385
addition	.1168	.7687
multiplication	.1079	.7829

Table 2: Baseline/Upper bound ρ correlations.

($\rho = 0.1385$). Obviously, the semantics of the modifiers had a much stronger impact on the semantic judgements of the compounds, thus confirming our intuition from Section 2.2.

The lower part of the table shows the respective baseline and upper bound values when the compound–modifier ratings and the compound–head ratings were combined by standard arithmetic operations, cf. Widdows (2008) and Mitchell and Lapata (2010), among others: the compound–modifier and compound–head ratings were treated as vectors, and the vector features (i.e., the compound–constituent ratings) were added/multiplied to predict the compound whole ratings. As in the related work, the arithmetic operations strengthen the predictions, and multiplication reached an upper bound of $\rho = 0.7829$, thus outperforming not only the head-only but also the modifier-only upper bound.

3.2 German Web Corpora

Most of our experiments rely on the *sdeWaC* corpus (Faaß et al., 2010), a cleaned version of the German web corpus *deWaC* created by the *WaCky* group (Baroni et al., 2009). The corpus cleaning had focused mainly on removing duplicates from the *deWaC*, and on disregarding sentences that were syntactically ill-formed (relying on a parsability index provided by a standard dependency parser (Schiehlen, 2003)). The sdeWaC contains approx. 880 million words and can be downloaded from `http://wacky.sslmit.unibo.it/`.

While the sdeWaC is an attractive corpus choice because it is a web corpus with a reasonable size, and yet has been cleaned and parsed (so that we can induce syntax-based distributional features), it has one serious drawback for a window-based approach (and, in general, for corpus work going beyond the sentence border): The sentences in the corpus have been sorted alphabetically, so going beyond the sentence border is likely to entering a sentence that did not originally precede or follow the sentence of interest. So window co-occurrence in the sdeWaC actually refers to x words to the left and right BUT within the same sentence. Thus, enlarging the window size does not effectively change the co-occurrence information any more at some point. For this reason, we additionally use *WebKo*, a predecessor version of the sdeWaC, which comprises more data (approx. 1.5 billion words in comparison to 880 million words) and is not alphabetically sorted, but is less clean and had not been parsed (because it was not clean enough).

3.3 Window-based vs. Syntax-based VSMs

Window-based Co-Occurrence When applying window-based co-occurrence features to our vector space models, we specified a corpus, a part-of-speech and a window size, and then determined the co-occurrence strengths of our compound nouns and their constituents with regard to the respective context words. For example, when restricting the part-of-speech to adjectives and the window size to 5, we counted how often our targets appeared with any adjectives in a window of five words to the left and to the right. We looked at lemmas, and deleted any kind of sentence punctuation. In general, we checked windows of sizes 1, 2, 5, 10, and 20. In one case we extended the window up to 100 words.

The window-based models compared the effect of varying the parts-of-speech of the co-occurring words, motivated by the hypothesis that adjectives and verbs were expected to provide salient distributional properties. So we checked which parts-of-speech provided specific insight into the distributional similarity between nominal compounds and nominal constituents: We used common nouns vs. adjectives vs. main verbs that co-occurred with the target nouns in the corpora. Figure 3 illustrates the behaviour of the Spearman Rank-Order Correlation Coefficient values ρ over the window sizes 1, 2, 5, 10, and 20 within sdeWaC (sentence-internal) and WebKo (beyond sentence borders), when restricting and combining the co-occurring parts-of-speech. It is clear from the figure that relying on nouns was the best choice, even better than combining nouns with adjectives and verbs. The differences for nouns vs. adjectives or verbs in the 20-word windows were

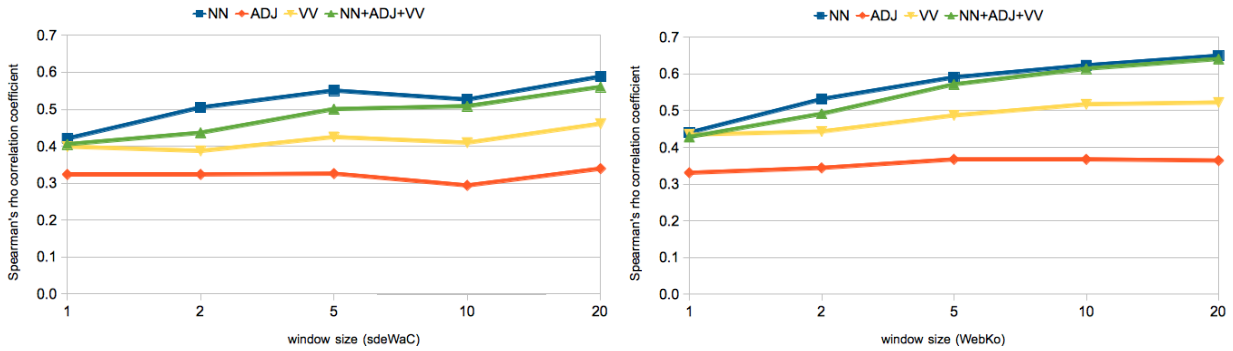

Figure 3: Window-based sdeWaC and WebKo ρ correlations across part-of-speech features.

significant.[7] Furthermore, the larger WebKo data outperformed the cleaned sdeWaC data, reaching an optimal prediction of $\rho = 0.6497$.[8] The corpus differences for NN and NN+ADJ+VV were significant.

As none of the window lines had reached an optimal correlation with a window size of 20 yet (i.e., the correlation values were still increasing), we enlarged the window size up to 100 words, in order to check on the most successful window size. We restricted the experiment to nominal features (with nouns representing the overall most successful features). The correlations did not increase with larger windows: the optimal prediction was still performed at a window size of 20.

Syntax-based Co-Occurrence When applying syntax-based co-occurrence features to our vector space models, we relied only on the sdeWaC corpus because WebKo was not parsed and thus did not provide syntactic information. We specified a syntax-based feature type and then determined the co-occurrence strengths of our compounds and constituents with regard to the respective context words.

In order to test our hypothesis that syntax-based information is more salient than window-based information to predict the compositionality of our compound nouns, we compared a number of potentially salient syntactic features for noun similarity: the syntactic functions of nouns in verb subcategorisation (intransitive and transitive subjects; direct and PP objects), and those categories that frequently modify nouns or are modified by nouns (adjectives and prepositions). With regard to subcategorisation functions, verbs subcategorising our target nouns represented the dimensions in the vector space models. For example, we used all verbs as vector dimensions that took our targets as direct objects, and vector values were based on these syntactic co-occurrences. For a noun like *Buch* 'book', the strongest verb dimensions were *lesen* 'read', *schreiben* 'write', and *kaufen* 'buy'. With regard to modification, we considered the adjectives and prepositions that modified our target nouns, as well as the prepositions that were modified by our target nouns. For the noun *Buch*, strong modifying adjective dimensions were *neu* 'new', *erschienen* 'published', and *heilig* 'holy'; strong modifying preposition dimensions were *in* 'in', *mit* 'with', and *zu* 'on'; and strong modified preposition dimensions were *von* 'by', *über* 'about', and *für* 'for'.

Figure 4 demonstrates that the potentially salient syntactic functions had different effects on predicting compositionality. The top part of the figure shows the modification-based correlations, the middle part shows the subcategorisation-based correlations, and at the bottom of the figure we repeat the ρ correlation values for window-based adjectives and verbs (within a window of 20 words) from the sdeWaC. The syntax-based predictions by modification and subcategorisation were all significantly worse than the predictions by the respective window-based parts-of-speech. Furthermore, the figure shows that there are strong differences with regard to the types of syntactic functions, when predicting compositionality: Relying on our target nouns as transitive subjects of verbs is al-

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

[8] For a fair corpus comparison, we repeated the experiments with WebKo on sentence-internal data. It still outperformed the sdeWaC corpus.

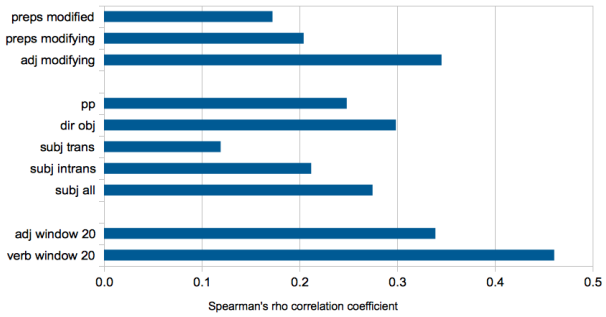

Figure 4: Syntax-based correlations.

most useless ($\rho = 0.1194$); using the intransitive subject function improves the prediction ($\rho = 0.2121$); interestingly, when abstracting over subject (in)transitivity, i.e., when we use all verbs as vector space features that appeared with our target nouns as subjects –independently whether this was an intransitive or a transitive subject– was again more successful ($\rho = 0.2749$). Relying on our noun targets as direct objects is again slightly better ($\rho = 0.2988$); as pp objects it is again slightly worse ($\rho = 0.2485$). None of these differences were significant, though.

Last but not least, we concatenated all syntax-based features to a large syntactic VSM (and we also considered variations of syntax-based feature set concatenations), but the results of any unified combinations were clearly below the best individual predictions. So the best syntax-based predictors were adjectives that modified our compound and constituent nouns, with $\rho = 0.3455$, which however just (non-significantly) outperformed the best adjective setting in our window-based vector space ($\rho = 0.3394$). Modification by prepositions did not provide salient distributional information, with $\rho = 0.2044/0.1725$ relying on modifying/modified prepositions.

In sum, attributive adjectives and verbs that subcategorised our target nouns as direct objects were the most salient syntax-based distributional features but nevertheless predicted worse than "just" window-based adjectives and verbs, respectively.

3.4 Role of Modifiers vs. Heads

This section tests our hypothesis that *the distributional properties of the head constituents are more salient than the distributional properties of the modifier constituents in predicting the degree of compositionality of the compounds*. Our rating data enables us to explore the modifier/head distinction with regard to two perspectives.

Perspective (i): Salient Features for Compound–Modifier vs. Compound–Head Pairs Instead of correlating all 488 compound–constituent predictions against the ratings, we distinguished between the 244 compound–modifier predictions and the 244 compound–head predictions. This perspective allowed us to distinguish between the salience of the various feature types with regard to the semantic relatedness between compound–modifier pairs vs. compound–head pairs.

Figure 5 presents the correlation values when predicting the degrees of compositionality of compound–modifier ($\boldsymbol{M}_{-}$ in the left panel) vs. compound–head ($\boldsymbol{H}_{-}$ in the right panel) pairs, as based on the window features and the various parts-of-speech. The prediction of the parts-of-speech is

NN > NN+ADJ+VV > VV > ADJ

and –with few exceptions– the predictions are improving with increasing window sizes, as the overall predictions in the previous section did. But while in smaller window sizes the predictions of the compound–head ratings are better than those of the compound–modifier ratings, this difference vanishes with larger windows. With regard to a window size of 20 there is no significant difference between predicting the semantic relatedness between compound–modifier vs. compound–head pairs.

When using the syntactic features to predict the degrees of compositionality of compound–modifier vs. head–compound pairs, in all but one of the syntactic feature types the verb subcategorisation as well as the modification functions allowed a stronger prediction of compound–head ratings in comparison to compound–modifier ratings. The only syntactic feature that was significantly better to predict compound–modifier ratings was relying on transitive subjects. In sum, the predictions based on syntactic features in most but not all cases behaved in accordance with our hypothesis.

As in our original experiments in Section 3.3, the syntax-based features were significantly outperformed by the window-based features. The syntactic features reached an optimum of $\rho = 0.2224$ and $\rho = 0.3502$ for predicting modifier–compound

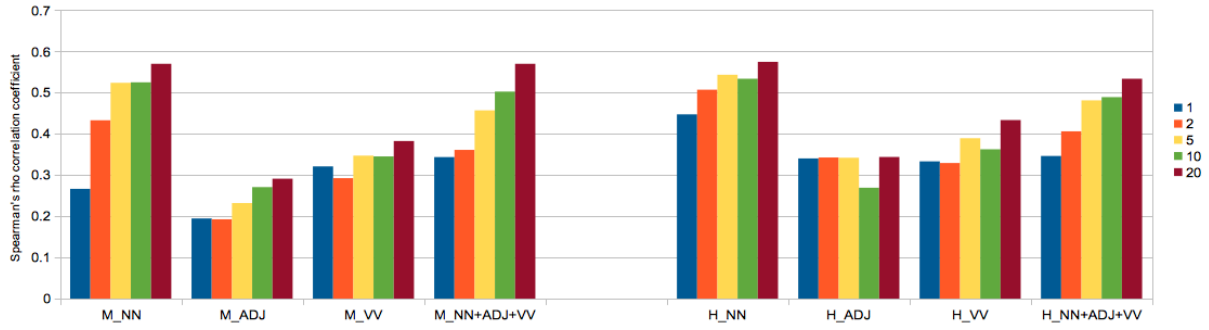

Figure 5: Window-based correlations (modifiers vs. heads).

vs. head–compound degrees of compositionality (in both cases relying on attributive adjectives), in comparison to $\rho = 0.5698$ and $\rho = 0.5745$ when relying on nouns in a window of 20 words.

Perspective (ii): Contribution of Modifiers/Heads to Compound Meaning This final analysis explores the contributions of the modifiers and of the heads with regard to the compound meaning, by correlating only one type of compound–constituent predictions with the compound whole ratings. I.e., we predicted the compositionality of the compound by the distributional similarity between the compound and only one of its constituents, checking if the meaning of the compound is determined more by the meaning of the modifier or the head. This analysis is in accordance with the upper bound in Section 3.1, where the compound–constituent ratings were correlated with the compound whole ratings.

Figure 6 presents the correlation values when determining the compound whole ratings by only compound–modifier predictions, or only compound–head predictions, or by adding or multiplying the modifier and head predictions. The underlying features rely on a 20-word window (adjectives, verbs, nouns, and across parts-of-speech). It is striking that in three out of four cases the predictions of the compound whole ratings were performed similarly well (i) by only the compound–modifier predictions, and (ii) by multiplying the compound–modifier and the compound–head predictions. So, as in the calculation of the upper bound, the distributional semantics of the modifiers had a much stronger impact on the semantics of the compound than the distributional semantics of the heads did.

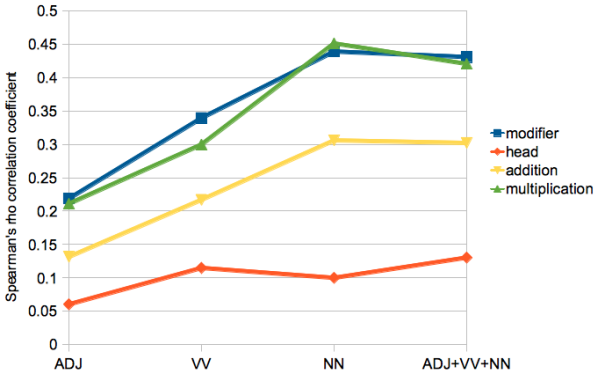

Figure 6: Predicting the compound whole ratings.

3.5 Discussion

The vector space models explored two hypotheses to predict the compositionality of German noun-noun compounds by distributional features. Regarding hypothesis 1, we demonstrated that –against our intuitions– not adjectives or verbs whose meanings are strongly interdependent with the meanings of nouns provided the most salient distributional information, but that relying on nouns was the best choice, in combination with a 20-word window, reaching state-of-the-art $\rho = 0.6497$. The larger but less clean web corpus *WebKo* outperformed the smaller but cleaner successor *sdeWaC*. Furthermore, the syntax-based predictions by adjective/preposition modification and by verb subcategorisation (as well as various concatenations of syntactic VSMs) were all worse than the predictions by the respective window-based parts-of-speech.

Regarding hypothesis 2, we distinguished the contributions of modifiers vs. heads to the compound meaning from two perspectives. (i) The predictions of the compound–modifier vs. compound–

head ratings did not differ significantly when using features from increasing window sizes, but with small window sizes the compound–head ratings were predicted better than the compound–modifier ratings. This insight fits well to the stronger impact of syntax-based features on compound–head in comparison to compound–modifier predictions because –even though German is a language with comparably free word order– we can expect many syntax-based features (especially attributive adjectives and prepositions) to appear in close vicinity to the nouns they depend on or subcategorise. We conclude that the features that are salient to predict similarities between the compound–modifier vs. the compound–head pairs are different, and that based on small windows the distributional similarity between compounds and heads is stronger than between compounds and modifiers, but based on larger contexts this difference vanishes. (ii) With regard to the overall meaning of the compound, the influence of the modifiers was not only much stronger in the human ratings (cf. Section 2) and in the upper bound (cf. Section 3.1), but also in the vector space models (cf. Figure 6). While this insight contradicts our second hypothesis (that the head properties are more salient than the modifier properties in predicting the compositionality of the compound), it fits into a larger picture that has primarily been discussed in psycholinguistic research on compound meaning, where various factors such as the semantic relation between the modifier and the head (Gagné and Spalding, 2009) and the modifier properties, inferential processing and world knowledge (Gagné and Spalding, 2011) were taken into account. However, also in psycholinguistic studies that explore the semantic role of modifiers and heads in noun compounds there is no agreement about which constituent properties are inherited by the compound.

4 Related Work

Most computational approaches to model the meaning or compositionality of compounds have been performed for English, including work on *particle verbs* (McCarthy et al., 2003; Bannard, 2005; Cook and Stevenson, 2006); *adjective-noun combinations* (Baroni and Zamparelli, 2010; Boleda et al., 2013); and *noun-noun compounds* (Reddy et al., 2011b; Reddy et al., 2011a). Most closely related to our work is Reddy et al. (2011b), who relied on window-based distributional models to predict the compositionality of English noun-noun compounds. Their gold standard also comprised compound–constituent ratings as well as compound whole ratings, but the resources had been cleaned more extensively, and they reached $\rho = 0.714$.

Concerning vector space explorations and semantic relatedness in more general terms, Bullinaria and Levy (2007; 2012) also systematically assessed a range of factors in VSMs (corpus type and size, window size, association measures, and corpus preprocessing, among others) against four semantic tasks, however not including compositionality ratings. Similarly, Agirre et al. (2009) compared and combined a WordNet-based and various distributional models to predict the pair similarity of the 65 Rubenstein and Goodenough word pairs and the 353 word pairs in WordSim353. They varied window sizes, dependency relations and raw words in the models. On WordSim353, they reached $\rho = 0.66$, which is slightly better than our best result, but at the same time the dataset is smaller.

Concerning computational models of German compounds, there is not much previous work. Our own work (Schulte im Walde, 2005; Kühner and Schulte im Walde, 2010) has addressed the degrees of compositionality of German particle verbs. Zinsmeister and Heid (2004) are most closely related to our current study. They suggested a distributional model to identify lexicalised German noun compounds by comparing the verbs that subcategorise the noun compound with those that subcategorise the head noun as direct objects.

5 Conclusion

This paper presented experiments to predict the compositionality of German noun-noun compounds. Our overall best result is state-of-the-art, reaching Spearman's $\rho = 0.65$ with a word-space model of nominal features from a 20-word window of a 1.5 billion word web corpus. Our experiments demonstrated that (1) window-based features outperformed syntax-based features, and nouns outperformed adjectives and verbs; (2) the modifier properties predominantly influenced the compositionality.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. 2009. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In *Proceedings of the North American Chapter of the Association for Computational Linguistics and Human Language Technologies Conference*, pages 19–27, Boulder, Colorado.

Collin Bannard. 2005. Learning about the Meaning of Verb–Particle Constructions from Corpora. *Computer Speech and Language*, 19:467–478.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are Vectors, Adjectives are Matrices: Representing Adjective-Noun Constructions in Semantic Space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193, Cambridge, MA, October.

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.

Gemma Boleda, Marco Baroni, Nghia The Pham, and Louise McNally. 2013. On Adjective-Noun Composition in Distributional Semantics. In *Proceedings of the 10th International Conference on Computational Semantics*, Potsdam, Germany.

John A. Bullinaria and Joseph P. Levy. 2007. Extracting Semantic Representations from Word Co-Occurrence Statistics: A Computational Study. *Behavior Research Methods*, 39(3):510–526.

John A. Bullinaria and Joseph P. Levy. 2012. Extracting Semantic Representations from Word Co-Occurrence Statistics: Stop-Lists, Stemming, and SVD. *Behavior Research Methods*, 44:890–907.

Kenneth W. Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29.

Paul Cook and Suzanne Stevenson. 2006. Classifying Particle Semantics in English Verb-Particle Constructions. In *Proceedings of the ACL/COLING Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, Sydney, Australia.

Katrin Erk. 2012. Vector Space Models of Word Meaning and Phrase Meaning: A Survey. *Language and Linguistics Compass*, 6(10):635–653.

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.

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

Wolfgang Fleischer and Irmhild Barz. 2012. *Wortbildung der deutschen Gegenwartssprache*. de Gruyter.

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.

Christina L. Gagné and Thomas L. Spalding. 2011. Inferential Processing and Meta-Knowledge as the Bases for Property Inclusion in Combined Concepts. *Journal of Memory and Language*, 65:176–192.

Zellig Harris. 1968. Distributional Structure. In Jerold J. Katz, editor, *The Philosophy of Linguistics*, Oxford Readings in Philosophy, pages 26–47. Oxford University Press.

Verena Klos. 2011. *Komposition und Kompositionalität*. Number 292 in Reihe Germanistische Linguistik. Walter de Gruyter, Berlin.

Natalie Kühner and Sabine Schulte im Walde. 2010. Determining the Degree of Compositionality of German Particle Verbs by Clustering Approaches. In *Proceedings of the 10th Conference on Natural Language Processing*, pages 47–56, Saarbrücken, Germany.

Rochelle Lieber and Pavol Stekauer. 2009a. Introduction: Status and Definition of Compounding. In *The Oxford Handbook on Compounding* (Lieber and Stekauer, 2009b), chapter 1, pages 3–18.

Rochelle Lieber and Pavol Stekauer, editors. 2009b. *The Oxford Handbook of Compounding*. Oxford University Press.

Kevin Lund and Curt Burgess. 1996. Producing High-Dimensional Semantic Spaces from Lexical Co-Occurrence. *Behavior Research Methods, Instruments, and Computers*, 28(2):203–208.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a Continuum of Compositionality in Phrasal Verbs. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan.

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

Siva Reddy, Ioannis P. Klapaftis, Diana McCarthy, and Suresh Manandhar. 2011a. Dynamic and Static Prototype Vectors for Semantic Composition. In *Proceedings of the 5th International Joint Conference on*

Natural Language Processing, pages 705–713, Chiang Mai, Thailand.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011b. 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.

Stephen Roller, Sabine Schulte im Walde, and Silke Scheible. 2013. The (Un)expected Effects of Applying Standard Cleansing Models to Human Ratings on Compositionality. In *Proceedings of the 9th Workshop on Multiword Expressions*, Atlanta, GA.

Michael Schiehlen. 2003. A Cascaded Finite-State Parser for German. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 163–166, Budapest, Hungary.

Sabine Schulte im Walde. 2005. Exploring Features to Identify Semantic Nearest Neighbours: A Case Study on German Particle Verbs. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 608–614, Borovets, Bulgaria.

Hinrich Schütze. 1992. Dimensions of Meaning. In *Proceedings of Supercomputing*, pages 787–796.

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.

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.

Dominic Widdows. 2008. Semantic Vector Products: Some Initial Investigations. In *Proceedings of the 2nd Conference on Quantum Interaction*, Oxford, UK.

Heike Zinsmeister and Ulrich Heid. 2004. Collocations of Complex Nouns: Evidence for Lexicalisation. In *Proceedings of Konvens*, Vienna, Austria.