Optimizing a Distributional Semantic Model for the Prediction of German Particle Verb Compositionality

Stefan Bott and Sabine Schulte im Walde

Institut für Maschinelle Sprachverabeitung
Univeristät Stuttgart
Pfaffenwaldring 5b, 70569 Stuttgart, Germany
{stefan.bott,schulte}@ims.uni-stuttgart.de

Abstract

In the work presented here we assess the degree of compositionality of German Particle Verbs with a Distributional Semantics Model which only relies on word window information and has no access to syntactic information as such. Our method only takes the lexical distributional distance between the Particle Verb to its Base Verb as a predictor for compositionality. We show that the ranking of distributional similarity correlates significantly with the ranking of human judgements on semantic compositionality for a series of Particle Verbs and the Base Verbs to which they correspond. We also investigate the influence of further linguistic factors, such as the ambiguity and the overall frequency of the verbs and a syntactically separate occurrences of verbs and particles that causes difficulties for the correct lemmatization of Particle Verbs. We analyse in how far these factors may influence the success with which the compositionality of the Particle Verbs may be predicted.

Keywords: Particle Verbs, Distributional Semantics, Multi Word Expressions

1. Introduction

Particle Verbs, such as the English "*to point out*" or "*to beat up*", are a special type of Phrasal Verbs and, with that, they are a type of Multi Word Expression. Similar to other Multi Word Expressions, their meaning may show various degrees of compositionality, i.e. they may be more or less opaque with respect to the meaning of their individual components. In German, particle verbs are very frequent and represent a highly productive paradigm. German Particle Verbs (PVs in what follows) are a challenge for computational lexicography, as well as for many NLP applications since, besides the mentioned problem of semantic composition, they are easily confounded with the base verbs from which they are composed. In addition, German syntax allows discontinuous realization with a potentially very large distance between the base verb (BV) and the verb particle, along with an unseparated variant. The verb particle may either occur together with the base verb as one word, as in (1-a), or separated with the verb in the second position of the sentence and the particle in clause-final position, as can be seen in (1-b).

(1) a. Peter isst das Eis auf.
Peter eats the ice cream PRT-up.
'Peter (completely) eats the ice cream.'

b. Peter möchte das Eis aufessen.
Peter wants-to the ice cream eat+PRT-up.

(2) Peter bringt den Eisverkäufer um.
Peter brings the ice cream vendor PRT-around.
'Peter kills the ice cream vendor.'

The formation of particle verb neologisms is possible and productive. Springorum et al. (2013) have shown that test subjects are perfectly able to associate a meaning to artificially created, previously unattested particle verbs and to construct example sentences for them. Different test subjects also agree to a large degree on the meaning they attribute to the newly formed lexical items. This shows that the creation of novel PVs is a regular process which must be governed by certain rules. There are two possible sources for the meaning of neologisms: the meaning of the BV or a transfer of the meaning of existing PVs to novel PVs which share the same particle which then creates a new PV meaning by analogy. The productive process appears to combine both. Here we are especially interested in the first case, the derivation of the lexical semantics of PVs from the semantics of BVs. As already mentioned, the relation between PVs and the BVs they are composed from can be more or less opaque. The meaning of some PVs if fully compositional, i.e. derived nearly directly from the meaning of the BV. Example (1) falls into this category: both *essen* and *aufessen* mean *to eat*. In other cases, like in (2), the meaning of the PV is fully opaque and not predicable at all from the meaning of the BV: *bringen* means to bring and *umbringen* means *to kill*. Most PVs can be found in a continuum between the two extremes, being neither fully opaque nor fully compositional.[1]

[1] It should be noted that the particles also may have a "meaning", although their semantics may be harder to define. (Springorum et al., 2012) survey various readings of the particle *an* and try to predict the semantics of PVs with respect to semantics of the particle, rather than with respect to BVs. A nice example is the verb *anpflaumen* (*to scold*), which is composed from the noun *Pflaume* (*plum*, i.e. highly opaque with respect to it's head), but nearly synonymous to other *an*-based PVs like *anschimpfen* or *anmotzen*, which are nearly perfectly transparent with respect to their base verb (both meaning *to scold*). One has to assume that a large part of the semantics of this PV comes from the particle, hence. For this reason we prefer the term *compositional* over the term *transparent*, although this is not relevant for the present

Since PVs are a productive open class, it would be highly desirable to be able to predict the degree of compositionality of unseen PVs. This would be a first step towards predicting the actual meaning of the PV itself from the meaning of the BV and the particle. Apart from the potential benefits for computational lexicography, many NLP applications could benefit from these predictions, for example machine translation, where highly compositional PVs may not require special treatment, while the highly opaque ones would need such special attention. Additionally, from a theoretical point of view a proper model of compositionality prediction promises to provide insights into the semantics of PVs and the nature of the productive process which allows the creation of novel PVs on the basis of given BVs and particles.

A complicating factor one has to bear in mind is that both BVs and the PVs may be semantically ambiguous and different word senses of PVs may show different degrees of compositionality. This can be seen in example (3):

(3) a. Mutti rührt den Teig an.
Mummy stirs the dough PRT.
Mummy mixes the dough

b. Pappi rührt keinen Schnaps mehr an.
Daddy stirs no booze more PRT.
Daddy doesn't touch/drink booze any more.

Both sentences in (3) share the same BV *rühren* (literally *to stir*), but while *anrühren* (3-a) has a highly transparent meaning, the meaning of the same PV in (3-b) is highly opaque with respect to its BV. The ambiguity here can be traced to two different readings of the particle *an*, one expressing a partitive and the other a topological reading (Springorum et al., 2012).[2] Further on, the BV *rühren* is ambiguous itself and can also mean *to touch emotionally*. One of our goals is to examine in how far these ambiguities influence in the success with which the degree of compositionality can be predicted automatically. We expect that ambiguity is a complicating factor for this task and that the prediction of the compositionality of a set of unseen PVs will be more accurate if only the unambiguous ones are taken into consideration.

Previous approaches to the assessment of the compositionality of German PVs have either used information on syntactic subcategorization (Hartmann, 2008) or lexical cooccurrences which are filtered by syntactic functions (Kühner and Schulte im Walde, 2010). The current approach investigates in how far syntactic information can be omitted entirely in the representation of distributional context while still capturing a substantial part of the information which is necessary to predict degrees of compositionality.

The already mentioned fact that German PVs may occur either separately or together - as one word - also gives rise to some technical problems. Most notably, in the separate paradigm, illustrated by (1-a), morphologic taggers or lemmatizers typically do not correct the lemma of the PV and output the lemma of the BV, instead. Syntactic parsers resolve the connection between the PV and its separated particle, which allows the reconstruction of the PV lemma, but the long distance dependency involved in these constructions may often induce parsing errors. Such technical problems can be expected to have an important influence on any kind of automatic treatment of PVs, including the assessment of compositionality, which we tackle here. This is also a problem for the indexing of BVs: in this case, BV lemmas given by a lemmatizer may not correspond to the actual BV itself, but rather to a PV which is formed from the BV with some verb particle.

In this paper we investigate in how far the degree of compositionality of PVs can be predicted on the basis of windows based lexical distributional information. We also investigate in how far a series of additional factors influence the degree of success by which compositionality can be predicted on the basis of such information, namely the choice of cooccurrent parts of speech, the application of term weighting, the ambiguity of the PV and the correction of PV lemmas in cases where the parser outputs the lemma of the corresponding BV. The rest of the paper is organized as follows: section 2. describes previous works which are relevant here, section 3. spells out the research questions we are interested in. Sections 4. and 5. are dedicated to the data used, the evaluation measures and the experimental setup. Sections 6. and 7. present the results and discuss them. Finally, section 8. draws some conclusions and gives an outlook to future works we plan.

2. State of the Art

The problem of predicting degrees of PV compositionality is not new and has been previously addressed in the literature. For English, there is work on the automatic extraction of PVs from corpora (Baldwin and Villavicencio, 2002; Baldwin, 2005; Villavicencio, 2005) (Baldwin and Villavicencio, 2002; Baldwin, 2005; Villavicencio, 2005) and the determination of compositionality (McCarthy et al., 2003; Baldwin et al., 2003; Bannard, 2005) (McCarthy et al., 2003; Baldwin et al., 2003, Bannard, 2005).

For German, Schulte im Walde (2004, 2005) presented preliminary y distributional studies to explore salient features at the syntax-semantics interface that determine the semantic nearest neighbours of German PVs. Relying on the insights of those studies, Schulte im Walde (2006) and Hartmann (2008) presented preliminary experiments on modeling the subcategorisation transfer of German PVs with respect to their BVs, in order to strengthen PV-BV distributional similarity. They capture the idea that there are semantic classes of PV-BV pairs, which undergo similar syntactic transfers (e.g. from an intransitive BV to a transitive corresponding PV) and are also semantically similar. Kühner and Schulte im Walde (2010) used unsupervised clustering to determine the degree of compositionality of German PVs, via common PV-BV cluster membership. They part from the idea that PVs tend to occurr more often in the same clusters as their corresponding BVs the more compositional they are. This approach is lexically dis-

work. We are thankful to an anonymous reviewer for pointing to this question.

[2]The ambiguity of PV can be so strong that in extreme cases two readings of the same PV can even be antonyms (Fleischer and Barz, 2012) like in the case of *abdecken*, which - depending on context - can mean both *cover* and *uncover*.

tributional in the sense that the assessment of compositionality is done on the basis of lexical information. They use, however, syntactic information as a filter and only use lexical heads as cooccurrence features which also correspond to certain syntactic functions. They compare different feature configurations and conclude that the best results can be obtained with information stemming from direct objects and PP-objects. The incorporation of syntactic information in the form dependency arc labels (concatenated with the head nouns) yields less satisfactory results, putting the syntactic transfer problem in evidence, again. Nevertheless, they admit that an incorporation of syntactic transfer information between BVs and PVs could possibly improve the results.

3. Research Questions and Motivation

According to the distributional hypothesis (Firth, 1957; Harris, 1968) semantically similar words tend to occur in similar contexts. On this basis we can hypothesize that the more similar a PV and a BV are in their meanings, the more similar the contexts are in which both can be found. For the PV *umbringen* (*to kill*) in (2) we would expect it to occur frequently with words that express agents, patients and manners of killing events. For its BV *bringen* (*to bring*), however, we would expect strong cooccurrences of words which are bringers and things which can be brought. In other words, we can expect only a weak lexical overlap in the typical context in which the BV and the PV occur. This should indicate us a very low compositionality of the PV. For the PV *aufessen* in (1) (*to eat up*), on the other hand, we would expect a high lexical overlap of its typical contexts with the typical contexts of its corresponding BV *essen* (*to eat*): both will contain typical eaters and typical things which can be eaten.

A well known problem for lexical distributional models is that not all of the words which can be found in the contexts are equally predictive. Words which are very frequent by themselves tend to contribute little discriminative information. To remedy this, some form of term weighting, such as LMI (Evert, 2004), can be applied. Also purely grammatical, non-content, words may have little discriminative power and filtering them out may make a model more adequate.

These differences in contexts can be captured in terms of distributional semantic models, such as the Word Vector Space Model (Sahlgren, 2006; Turney and Pantel, 2010) which we use below. Distributional Semantics Models are unsupervised and can predict the measure of similarities between contexts (i.e. the distributional semantic similarity between words) but they cannot predict the exact degree of compositionality directly (as for example on a scale from 0 to 10). What we can expect, however, is that for a series of PV-BV pairs the ranking of these pairs in terms of semantic similarity will significantly correlate with the ranking of compositionality judgements from a gold-standard of human judgements. Once the two rankings are mapped onto each other, the distributional distance between a previously unseen PV-BV pair may then be mapped onto a scale of compositionality values.

In the present work we attempt to discover whether the compositionality of the meaning of German PVs is predictable from purely distributional information and in which ways a distributional model may be optimized. We started out from a standard general purpose Word Vector Model and compared this to an improved version of the same type of model which optimizes some technical parameters, which are, however, linguistically motivated. We investigate the influence of term weighting and filtering out non-content words. We also use a model which restores PV lemmata and thus better separates the occurrences of BVs from the occurrences of the homophonous verbs which are actually part of a PV, in those cases where its particle has been separated and usually appears in clause final position. We were interested in the following questions:

- With which degree of success can the compositionality of PVs be assessed with the use of purely lexical distributional information and without the use of any syntactic features?
- To which degree can the predictive power of a model improve if we use syntactic information in order to better separate BVs from homographic PVs which occur in syntactic separation from their particle?[3] Even if this may seem a technical question, it is linguistically motivated and may potentially affect other computational approaches to German PV compositionality.
- Can the distributional semantic model be improved by taking only content words in the context into account?
- What role does the ambiguity of the PV play in the predictability of compositionality? One could expect that ambiguous verbs are harder to rank than unambiguous ones, because of the varying degree of compositionality among different senses of the same PV.
- What role does the frequency of PVs (and also the frequency of BVs) play? We expect that compositionality of high frequency verbs is easier to predict than that of low frequency verbs, simply because of data sparsity issues.
- Which is the ideal size of context to be taken into consideration for building predictive models?

In contrast to previous approaches (Hartmann, 2008; Kühner and Schulte im Walde, 2010), we use no information about syntactic subcategorization, nor do we filter cooccurrents by their syntactic relation to the target verb.

4. Data and Models

We used a lemmatized and tagged version of the SDeWaC corpus (Faaß and Eckart, 2013), a corpus of nearly 885 million words. SDeWaC is a large resource, but has the disadvantage that it is distributed in a form where sentences are ordered alphabetically and the origianl contexts of sentences are not preserved. As a consequence our windows

[3] Note that we only use syntactic information for the reconstruction and repair of the lemma information of PVs, not in form of syntactic features.

for word vector extraction are truncated must be truncated the beginning and the end of the sentence. For linguistic preprocessing we used the Mate parser (Bohnet, 2010), which also allowed us to use syntactic information for the correction of PV lemmata in the separated paradigm.

For the evaluation of our models, we used the gold standard created by Hartmann (2008). This gold standard consists of 99 particle verbs, corresponding to 11 different particles, with 9 PVs representing each particle. All particles used in this set stem from closed class word, namely prepositions, adverbials and adjectives. The set of PVs is balanced for frequency and verbs are taken from 8 different frequency bands ((2,5], (5,10], (10,18], (18,30], (30,55], (55,110], (100,300] and (300,6000] occurrences in the HGC corpus of 180 million words). This allows to study compositionality for different frequency bands. The PVs were evaluated by 4 human judges for the degree of compositionality. The judges were also asked to annotate any lexical ambiguity they could detect, which makes PV ambiguity information for the gold standard available. Hartman used this gold standard for experiments of the prediction of compositionality with the use of subcategorization frames. The same gold standard was used for evaluation purposes in Kühner and Schulte im Walde (2010), who already applied a word vector space model to these verbs, but applying syntactic filters as described in section 2.

For the modeling of lexical semantics, we use Word Vector Space models (Sahlgren, 2006), which project a vector space where every possible context word is represented by a vector dimension. Each target word (either a PV or a BV) is represented as a vector in this space. The extension of the vector in each of the dimensions represents the number of times the cooccurrent has been seen in the local contexts (of a term weighted value based on that count), where local context is defined as a window of *n* words to the left and the right of the target word. In order to predict the compositionality of PVs we take the cosine of the angle between the vectors for each PV and its corresponding BV. The cosine value cannot directly be mapped onto a compositionality value, but we can take the rank order of cosine values to predict an order of the PVs, from the lowest to the highest compositionality rank. The correlation of the predicted compositionality with human judgements from the gold standard is measured as Spearman's rank correlation coefficient (with tie corrections) (Siegel and Castellan, 1988).

5. Experiments and Evaluation

We build an array of baseline models using only raw counts for words from the context, without applying neither POS filtering (of non-content words), lemma corrections nor term weighting.

The Mate parser does not give the correct lemmas for PVs in those cases where the particle is separated from the verb. Instead the lemma of the BV is given. We can expect that if only the lemma information from the parser is used for the training of the distributional semantic model, the vector entries will poorly separate the true BVs from the their derivate PVs. In previous experiments we found out that around 20% of the PVs occur in this separated setting. In order to address the problem of inadequate lemmatization we also built a new set of models for which we extracted the window information from the SDeWaC in the same way as with the baseline models, except for an additional treatment of separately occurring PVs. If we could find a verb particle which the parser resolved as directly depending on a verb, we concatenated the particle with the verb lemma in order to derive the lemma of the PV. Table 1 shows the relevant part of a dependency parse tree, containing the PV abschauen, where the lemma is *schauen* and the particle *ab* has a syntactic dependency to the head verb in position 20. From this information we can repair the wrong lemma and form the correct lemma *abschauen*, which will then be used for the model extraction.

ID	Word	Lemma	POS tag	Morph	Syn Dep
17	denn	denn	KON		4
18	der	der	ART	nom|sg|masc	19
19	Zweite	zweiter	NN	nom|sg|masc	20
20	schaut	schauen	VV-FIN	sg|3|pres|ind	17
21	sich	sich	PRF	acc|sg|3	20
22	sehr	sehr	ADV		23
23	viel	viel	PIS		20
24	vom	von	APP-RART	dat|sg|masc	20
25	Ersten	ersten	NN	dat|sg|masc	24
25					
26	ab	ab	PTK VZ		20

Table 1: Example of a partial dependency parse tree containing a PV (abschauen) with a syntactically separated particle (ab)

Each semantic model depends on the size of the context which was taken into consideration. We defined the context as a window of *n* words to the left and right of the target word and built models for {1,2,5,10,20}. The variable size of the context window also allowed us to study the influence of context size on the degree of success in assessing compositionality. Term weighting was applied and carried out with local mutual information (LMI, Evert, 2005). As a test we also built the same models without the application of term weighting. The models were applied in two different settings: in one we use all words from the context and in the other we use only content words (verbs, adjectives and common nouns).

6. Result

The first two important results concern the technical parameters we study. First of all, restricting the context to content words (adjectives, common nouns and verbs) has a very high impact. The models trained on all context words nearly consistently fail to predict compositionality to a degree which reaches statistical significance and perform much worse than the ones trained on the filtered POS

Window size	1	2	5	10	20
Original	0.2102*	0.2507*	0.2308*	0.2416*	0.2668**
Restored	0.3058**	0.2910**	0.3696***	0.3008**	0.1859

Table 2: Comparison between baseline models vs. models with lemma correction (values are given in Spearman rank order correlation coefficients)

set. The same is true for term weighting with LMI. If no term weighting was applied to the models the prediction of compositionality in most cases failed to shows statistically significant correlations to the gold standard. For this reason we only present comparisons with restricted context and term weighting in the rest of this section.

Frequency	Spearman' rho 1
(2,5]	0.16
(5,10]	0.27
(10,18]	0.26
(18,55]	0.59
(55,110]	0.25
(110,300]	0.06
(300,6000]	0.13

Table 3: Spearman rho values for different frequency ranges (models with restored lemma information, window size 5)

Table 2 and its graphical representation in figure 1 show the direct comparison between the two sets of models: with (restored) and without (original) lemma correction for separated PVs. The x-axis represents the use of context windows of different size. It is clearly visible that our expectation was met: applying lemma correction improves the prediction of compositionality, except for very large window sizes. The values are given in Spearman's rho values, the critical values for $p<0.025$, $p<0.005$ and $p<0.001$ are 0.199, 0.260 and 0.310, respectively (n=98, values that exceed the critical values are marked with *, ** and ***).

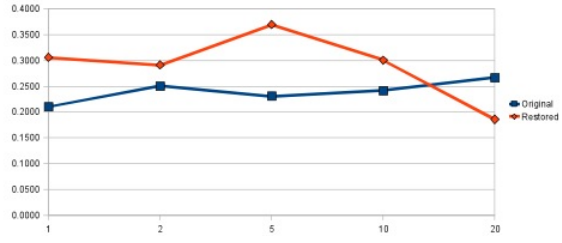

Figure 1: Comparison between baseline models vs. models with lemma correction (x-axis represents window size in number of words, y-axis represents values in Spearman's rho)

It is interesting to distinguish between the performance of the models for ambiguous and unambiguous verbs. Figure 2 show these for the original model without lemma restoration and Figure 3 for the improved model with such restoration applied. From theoretic considerations we expected that the compositionality of unambiguous PVs should be easier to predict than that of ambiguous PVs. This prediction is borne out and it can be seen, again, in a much clearer way if models with restored lemma information is used. Once more, the best results are obtained with window size 5.

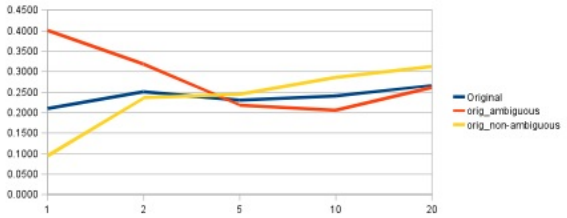

Figure 2: The performance of the baseline models for ambiguous and unambiguous PVs (x-axis represents window size in number of words, y-axis represents values in Spearman's rho)

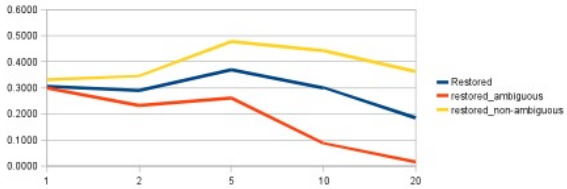

Figure 3: The performance of the models with lemma restoration for ambiguous and unambiguous PVs (x-axis represents window size in number of words, y-axis represents values in Spearman's rho)

Finally, we were interested in how far the frequency of the PVs influences the performance of the models. We expected that the compositionality of more frequent PVs would be easier to predict. But the results, given in Table 3 speak a different language: As expected, the values for low frequency PVs are hard to predict, but we obtained also very low scores for high frequency verbs ($\geq$110).

7. Discussion

The experiments show that to a large extend the relative degree of PV compositionality can be predicted from the lexical information of the contexts of PVs and their corresponding BVs without the use of syntactic information. By the application of such a conceptionally simple and computationally efficient method we can obtain results comparable to those presented in Kühner and Schulte im Walde

Rank	*Frequency*	*LMI + non-content-word filtering*					
	bringen	**bringen**	**umbringen**	**betonieren**	**betonieren**	**essen**	**aufessen**
1	Ausdruck	Ausdruck	Mensch	asphaltieren	Landschaft	trinken	Kind
2	neu	Verbindung	versuchen	Fläche	Fläche	Fleisch	Teller
3	Weg	nah	Frau	Schalung	ganz	Brot	Brot
4	Verbindung	Einklang	Jude	Boden	Mühlenberger	gehen	Lamm
5	Mensch	Weg	lassen	neu	asphaltieren	geben	umbringen
6	gut	Markt	gegenseitig	Frost	riesig	Gemüse	Portion
7	Jahr	Vorteil	drohen	mauern	Küstenstreifen	Obst	töten
8	nah	nahe	Kind	gelastern	Fluß	essen	bestreich
9	groß	Krankenhaus	Mann	fest	komplett	Mittag	besäen
10	Welt	Erfahrung	Million	Piste	flächendeckend	satt	verschimmeln
11	Kind	Punkt	Vater	gepflastert	Arbeitsmarkt	schlafen	Spinat
12	Markt	Geltung	Leute	durchgehend	sauer	bekommen	braten
13	Einklang	Ordnung	Weise	Fundament	bullern	Fisch	Schokolade
14	anderer	neu	bestialisch	Baugrube	Raketensilos	Tag	ausbrüten
15	Punkt	Sicherheit	Zivilist	hydraulisch	Wald	gut	knabbern

Table 4: The most dominant dimensions of some sample vectors, ranked by raw frequency and the values obtained by LMI term weighting and reduced to content words (windows size: 5

(2010), evaluating on the same gold standard. Our results are not fully comparable to theirs because of the difference in the methods applied, which also impose a different way of evaluating. Kühner and Schulte im Walde obtain Spearman's rho values, but they have to weight them against a coverage value, since they cannot predict the compositionality of all PVs. Their rho-scores tend to be slightly higher than ours, but this comes on the cost of a reduced coverage. Our results with respect to the optimal size of the context window are, however, along the lines of this study: we found that the optimal size of the window is 5 words to the right and the left, a range which usually includes the arguments of the verb and tends to ignore material which is less direcly related to the target verb.

In order to understand why a Vector Space model works well for our problem setting it is interesting to see some sample vectors. Table 4 shows the 15 strongest dimensions for the verbs corresponding to three PV-BV pairs: the ones corresponding to the initial examples (1) and (2) and the pair *betonieren*/*zubetonieren* (*to apply concrete* / *to cover a surface with concrete*), which is one of the 3 examples with the highest human compositionality ratings from the gold standard. The vectors for the highly non-compositional pair *bringen*/*umbringen* (cf. (2) above) show very little overlap in their most dominant dimension. The strongest dimensions for the vector for *bringen* (*to bring*) includes typical places where one could bring something or someone (*Markt*/*market* and *Krankenhaus*/*hospital*) and some nouns which typically occur in idioms together with this verb (*etwas zum Ausdruck bringen* / *to express s.th.*). The vector for *umbringen* (*to kill*) has strong dimension for typical killers and killees (*humans*, *men*, *women*, *millionaires*, etc.) and manners of killing (e.g. *bestialisch*/*crudely*). If we compare this to the other two BV-PV pairs, which are both highly compositional, we can observe a much higher lexical overlap there (e.g. *asphaltieren*/*to asphalt* and *Fläche*/*surface* in the case of *betonieren*/*zubetonieren*).

The lexical distributional approach has also limits, however. If we analyse outliers of the predicted ranking we can find different possible error sources. The pair *rüsten*/*austrüsten* (*to stock up on arms*/*to equip*), for example, is predicted by our model to be highly compositional, while human raters tend to judge it as being compositional on a medium level (4.5 on a scale from 1 to 9). The strongest dimensions for *ausrüsten* include military terms like *Waffe* (*arm*) or *Flotte* (*armada*), just because these are also typical things which might be equiped in some way. This leads to a very low vector distance which wrongly predicts a high compositionality. Also the reverse case can be found in which a PV is highly compositional, but occurs typically in very different contexts. The verb *durchwinken* (*to wave through*, i.e. to signal someone or a vehicle to pass on by waving) is hightly compositional, but the BV *winken* (*to wave*) occurs in many figurative or idiomatic contexts, such as "*ein Preis winkt*" ("*a price is promised*"), ultimately leading to a prediction of low compositionality.

The experiments also show that the compositionality of unambiguous PVs is easier to assess than the compositionality of the full set, including ambiguous ones. This is not a surprise and corresponds to our initial hypothesis. It is more surprising that the assessment of the highest frequency band of PVs is harder than for medium frequency bands. We attribute this finding to the tendency of high frequency verbs to be more ambiguous and also to the fact that, they being more frequent, more of the word senses are represented in the corpus.

As for the more technical parameters, the influence of term weighting and the filtering of non-content words can be easily explained, although the magnitude of the effect is surprising. LMI reduces the influence of those words which are very frequent by themselves and, thus, not very predictive for the target verbs. The first two columns of table 4 show the 15 strongest vector dimensions for the verb *bringen* (*to bring*), the first one based on pure frequency information (i.e. without LMI application) and the second

one with LMI application.[4] It can be seen that there is a tendency for the frequency based vectors to over-represent word with weak semantic selectional preferences, such as *neu* (*new*), *gut* (*good*) and *anderer* (*another*). The influence of LMI reduces the strength of these dimensions. Interestingly, in rank 5 of the frequency based vector dimensions the noun *Mensch*, appears, which is also a highly predictive context word for the PV *umbringen* (*to kill*). *Mensch* is very frequent and relatively unspecific word and LMI application also reduces its influence for the vector of *bringen*, thus making its vector more distinct from the vector for the PV *umbringen*. This is desirable, since in this case the PV is highly opaque with respect to the BV. This explains the importance of LMI application to a certain extend, but still the strength of it's influence is surprising.

8. Conclusions

In this work we have shown that a purely windows-based lexical distributional model can capture the prediction of the semantic compositionality of German PVs. The ranking of PVs according to the distributional distance they show to their corresponding BVs correlate to the human ranking of perceived semantic compositionality with high levels of statistical significance. We could also show that the word vector models we used for this purpose if some technical, but linguistically motivated, issues are resolved. The performance of the models can be greatly improved if syntactic information is used in order to restore lemma information in cases where the parser outputs a lemma which corresponds not to the PV, but to the underlying BV. We also investigated to which extend the use of a context limited to a reduced POS set and the use of LMI term weighting benefit the prediction of compositionality and we show that both factors improve the performance substantially.

As for purely linguistic parameters, we could show that, in compliance to our initial hypothesis, unambiguous Particle Verbs are easier to assess in terms of compositionality than ambiguous ones. This finding can be easily explained, because ambiguous verbs coresond to vectors that do not represent specific word meanings, but a mixture of all meanings. Finally, we investigated the predictability of compositionality in different frequency range of PVs and found that, even if it is true that medium frequent verbs are easier to assess than low frequent verbs, contrary to what was expected, it is not true that the compositionality of highly frequent verbs is easier to predict. This suggests that high frequency verbs present additional problems that are difficult to capture within our type of model, most notably a higer tendency towards ambiguity.

The models we use here are only based on lexical distributional information. Many researchers (Aldinger, 2004; Schulte im Walde, 2006; Hartmann, 2008; Kühner and Schulte im Walde, 2010), however, have stressed the influence of syntactic factors on the compositionality of German Particle Verbs. In future work we plan to investigate this additional aspect and integrate it within our current approach.

[4] We use ranks here for expository purposes, rather than frequency or LMI values. Using ranks makes it easier to compare frequencies to LMI values.

Acknowledgements

This work was funded by the DFG Research Project "Distributional Approaches to Semantic Relatedness" (Stefan Bott, Sabine Schulte im Walde), and the DFG Heisenberg Fellowship SCHU-2580/1-1 (Sabine Schulte im Walde).

9. References

Nadine Aldinger. 2004. Towards a Dynamic Lexicon: Predicting the Syntactic Argument Structure of Complex Verbs. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

Timothy Baldwin and Aline Villavicencio. 2002. Extracting the Unextractable: A Case Study on Verb Particles. In *Proceedings of the Sixth Conference on Computational Natural Language Learning*, pages 98–104, Taipei, Taiwan.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An Empirical Model of Multiword Expression Decomposability. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96, Sapporo, Japan.

Timothy Baldwin. 2005. Deep Lexical Acquisition of Verb–Particle Constructions. *Computer Speech and Language*, 19:398–414.

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

Stefan Evert. 2004. The Statistical Analysis of Morphosyntactic Distributions. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1539–1542, Lisbon, Portugal.

Gertrud Faaß and Kerstin Eckart. 2013. SdeWaC – a Corpus of Parsable Sentences from the Web. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology*, Darmstadt, Germany. To appear.

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

Wolfgang Fleischer and Irmhild Barz. 2012. *Wortbildung der deutschen Gegenwartssprache*. Walter de Gruyter, 4th edition.

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

Silvana Hartmann. 2008. Einfluss syntaktischer und semantischer Subkategorisierung auf die Kompositionalität von Partikelverben. Studienarbeit. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart. Supervision: Sabine Schulte im Walde and Hans Kamp.

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.

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.

Magnus Sahlgren. 2006. *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces*. Ph.D. thesis, Stockholm University.

Sabine Schulte im Walde. 2004. Identification, Quantitative Description, and Preliminary Distributional Analysis of German Particle Verbs. In *Proceedings of the COLING Workshop on Enhancing and Using Electronic Dictionaries*, pages 85–88, Geneva, Switzerland.

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.

Sabine Schulte im Walde. 2006. The Syntax-Semantics Interface of German Particle Verbs. Panel discussion at the 3rd ACL-SIGSEM Workshop on Prepositions at the 11th Conference of the European Chapter of the Association for Computational Linguistics.

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

Sylvia Springorum, Sabine Schulte im Walde, and Antje Roßdeutscher. 2012. Automatic Classification of German *an* Particle Verbs. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 73–80, Istanbul, Turkey.

Sylvia Springorum, Sabine Schulte im Walde, and Antje Roßdeutscher. 2013. Sentence Generation and Compositionality of Systematic Neologisms of German Particle Verbs. Talk at the 5th Conference on Quantitative Investigations in Theoretical Linguistics.

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

Aline Villavicencio. 2005. The Availability of Verb-Particle Constructions in Lexical Resources: How much is enough? *Computer Speech & Language*, 19(4):415–432.