German particle verbs: compositionality at the syntax-semantics interface

Stefan Bott and Sabine Schulte im Walde
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Germany

ABSTRACT

Particle verbs represent a type of multi-word expression composed of a base verb and a particle. The meaning of the particle verb is often, but not always, derived from the meaning of the base verb, sometimes in quite complex ways. In this work, we computationally assess the levels of German particle verb compositionality by applying distributional semantic models. Furthermore, we investigate properties of German particle verbs at the syntax-semantics interface that influence their degrees of compositionality: (i) regularity in semantic particle verb derivation and (ii) transfer of syntactic subcategorization from base verbs to particle verbs. Our distributional models show that both superficial window co-occurrence models as well as theoretically well-founded syntactic models are sensitive to subcategorization frame transfer and can be used to predict degrees of particle verb compositionality, with window models performing better even though they are conceptually and computationally simpler.

Keywords: particle verbs, multi-word expressions, compositionality, distributional semantics

1 INTRODUCTION

Particle verbs (PVs), such as the German *auf|essen* (*to eat up*) and the English *to blow up*, represent a type of multi-word expression (MWE) composed of a base verb (BV) and a particle. While particle verbs exist in many languages, German PVs are particularly frequent and form a highly productive paradigm which often produces neologisms and is subject to creative language use in puns and word plays.

German PVs, similarly to other MWEs, exhibit a varying degree of compositionality, as illustrated in examples (1) vs. (2). The meaning of the highly compositional PV *nach|drucken* (*to reprint*) is closely related to its BV *drucken* (*to print*), while the PV *nach|geben* (*to give in*) has little meaning in common with the BV *geben* (*to give*).

(1) *Der Verlag DRUCKTE das Buch NACH.*
the publisher PRINTED the book PRT_{nach}
'The publisher reprinted the book.'

(2) *Peter GAB ihrer Bitte NACH.*
Peter GAVE her request PRT_{nach}
'Peter gave in to her request.'

From a computational point of view, addressing the compositionality of PVs (and multi-word expressions in general) is a crucial ingredient for lexicography and Natural Language Processing (NLP) applications, in order to know whether the expression should be treated as a whole or as the sum of its constituents, and what the expression means. For example, studies such as Cholakov and Kordoni (2014), Weller *et al.* (2014) and Cap *et al.* (2015) have integrated the prediction of multi-word compositionality into statistical machine translation.

Assessing PV compositionality requires one to assess the semantic contributions of both the BV and the verb particle (Lechler and Roßdeutscher 2009; Haselbach 2011; Kliche 2011; Springorum 2011). This is obvious in highly compositional cases as in example (1): the meaning of *nach|drucken* (*to reprint*) is a straightforward composition of the meanings of *nach* (*again*) and *drucken* (*to print*).[1] Non-compositional cases such as *nach|geben* in example (2) behave differently: they are not semantically transparent with respect to the meaning of the BV, and the meaning contributed by the particle *nach* is not straightforward.

Compositionality is not a binary property of PVs, however. The levels of compositionality are distributed over a continuous scale,

[1] An evident problem is that the particle *nach* here means more than simply *again*: it implies that an additional copy is created. In addition, *nach*, like most particles, is semantically ambiguous. These issues will be addressed below.

where examples (1) and (2) refer to two extremes of the continuum, rather than prototypical cases. In contrast, *ab|segnen* in (3) represents an example which is judged as semi-compositional by human raters, meaning *to approve* rather than *to bless*.

(3) *Der Chef* SEGNETE *die Pläne* AB.
the boss BLESSED the plans PRT_{ab}
'The boss approved the plans.'

In this article, we investigate the factors that influence the prediction of PV compositionality from a corpus-linguistic perspective. We start with a series of hypotheses that are then investigated by a series of experiments. First, we argue that PVs can be grouped into semantically coherent classes that share the same semantic derivation when BVs from the same class are combined with a certain particle type. This combination typically selects a specific sense of the particle. Second, we address the prediction of compositionality by applying distributional semantic methods. After verifying a novel approach to model syntactic subcategorization changes, we compare window-based models with models that integrate syntactic transfer. Our main contributions are at the interface between a theoretical study of PV compositionality and the computational use of distributional semantic methods, to identify a theoretically reliable and computationally useful framework.

2 MOTIVATION AND HYPOTHESES

In this section we describe the theoretical foundations of our assumptions and analyses. We first discuss in more detail the notions of PV compositionality (Section 2.1), semantic derivation (Section 2.2), and syntactic transfer (Section 2.3). Section 2.4 then describes our distributional semantic approach, and Section 2.5 defines our hypotheses.

2.1 *Particle verb compositionality*

We illustrated above that compositionality is a scalar property: Apart from highly compositional PVs such as *nach|drucken*, PVs such as *ab|segnen* are not fully transparent with respect to their BVs, but still integrate meaning components attributed by the particle and the BV.

We refer to PVs that are semantically related to their BVs (in contrast to non-compositional PVs, which are semantically unrelated to their BVs) as *semantically derived PVs*.

Semantic derivation takes place not only for highly frequent PVs but also for infrequent or domain-specific PVs as well as neologisms. For example, while *nach|schneiden* in (4a) is a common verb in everyday language, *nach|sägen* in (4b) is more restricted to a specific domain and much less frequent; *nach|töten* in (4c) is a neologism.[2] The meanings of all three PVs in (4) are semantically derived from the meanings of the respective BVs, and the meaning contribution of the particle is productive and regular: All of the *nach*-PVs in (4) have a common semantic component which implies some kind of *correction to a previous action of BV by performing BV again*.

(4) a. *Der Friseur SCHNITT ihr die Haare NACH.*
the hairdresser CUT her the hair PRT_{nach}
'The hairdresser trimmed her hair.'

b. *Einfach mit der richtigen Größe NACH|SÄGEN ist nicht.*
simply with the right size PRT_{nach}|SAW is not
'You cannot simply resaw it with the right size.'

c. *Das Reh war noch nicht tot und wurde NACH|GETÖTET.*
the deer was yet not dead and was PRT_{nach}|KILLED
'The deer was not dead yet and had to be finished off.'

The same BVs from (4) can also combine with other particles, such as *an*, and undergo a different but also regular semantic derivation, as illustrated in (5). Here, all of the *an*-PVs have a common semantic component that refers to a partitive meaning, *to start a first bit of BV*.

(5) a. *Du musst das Messer abwaschen, bevor du das nächste*
you must the knife clean before you the next
Stück Torte AN|SCHNEIDEST.
piece cake PRT_{an}|CUT
'You have to clean the knife before you start cutting the next piece of the cake.'

[2] Examples with PV neologisms are taken from a sentence generation experiment by Springorum *et al.* (2013a), where the experiment participants generated sentences for existing and non-existing PVs.

b. *Max und Moritz* SÄGEN *die Brücke* AN.
Max and Moritz SAW the bridge PRT_{an}
'Max and Moritz start sawing the bridge.'

c. *Bring ihn nicht gleich um. Du solltest ihn erst* AN|TÖTEN.
bring him not already PRT_{um} you shall him first PRT_{an}|KILL
'Don't kill him right away. You should start killing him first.'

Often, similar semantic derivations apply to semantically similar BVs, such as *schneiden* and *sägen* in examples (4) and (5), which both refer to a cutting event. In these cases, we find regular semantic shifts, where combining semantically similar BVs with specific particle types results in semantically similar PVs (Springorum *et al.* 2013b; Köper and Schulte im Walde 2018). We refer to these regular semantic shifts as *semantic transfer patterns*.

(6) *Semantic Transfer Pattern*
Taking a BV from semantic group α and a particle β with meaning μ, we will derive a PV from semantic group δ.

Note that it is not the particle type that is responsible for the meaning shift, but a particular sense μ of the particle type. For example, the particle *nach* is ambiguous and does not only mean *again* (roughly corresponding to the English prefix *re*), cf. Haselbach 2011. Accordingly, the meaning of a PV may be ambiguous along the lines of the senses of the particle.

In contrast to semantically derived PVs, we refer to completely non-compositional PVs as fully lexicalized, such as *nach|geben* in (2) and *um|bringen* (*to kill*, while the BV *bringen* means *to bring*). Without diachronic considerations, the meanings of these PVs cannot directly be inferred from the meanings of their verbal bases *geben* and *bringen* and the meanings of the verb particle types *um* and *nach*.

Treating each PV as an independent lexical entry would require a large number of unrelated lexical entries and thus disregard generalizations about the semantic classes of PVs and the meaning contributions of the verb particles. Further on, a pure lexical listing approach does not explain the productivity of the PV paradigm regarding neologisms, whose meanings are derived from regular semantic transfer patterns. The semantic pattern approach is therefore appealing, since it reduces idiosyncracy in the lexicon, and accounts for the productivity of German PVs and the ease of native speakers to produce and interpret PV neologisms.

2.2 *Semantic derivation and the meanings of particles*

What is the meaning of verb particles? Some particle senses are parallel to homophonic prepositions or adverbs (Stiebels 1996). But it is not clear if such a treatment can be extended to all particles and particle meanings. It is thus difficult to assign particles a lexical entry rather than taking whole PVs into account (Lechler and Roßdeutscher 2009; Kliche 2011; Springorum 2011).

For a more comprehensive example, consider the particle *an*. PVs with *an* can express, among other things, a direction of an action, a fixation, a manner of communication, and a partitive event, as exemplified in (7a–d) (Springorum 2011; Bott and Schulte im Walde 2014a). The particle is highly ambiguous, and its meanings are sometimes difficult to capture, but assuming (6) *Semantic Transfer Patterns* ties them closely to common underlying semantic derivations.

(7) a. *A BLICKT/SCHAUT/STARRT/STIERT B AN.*
A LOOKS/STARES/GAZES B PRT$_{an}$
'A looks/stares/gazes at B.'

b. *A BRÜLLT/FAUCHT/BELLT/MECKERT B AN.*
A ROARS/HISSES/BARKS/BLEATS B PRT$_{an}$
'A brawls/hisses/scolds at B.'

c. *A KLEBT/HEFTET/SCHRAUBT B an C AN.*
A GLUES/AFFIXES/SCREWS B at/onto C PRT$_{an}$
'A glues/affixes/screws B onto C.'

d. *A SCHNEIDET/BRICHT/REIẞT B AN.*
A CUTS/BREAKS/TEARS B PRT$_{an}$
'A cuts/breaks/tears the first piece of B.'

The semantic class of the PV and individual particle meanings are also tied together by specific selectional restrictions. This is most apparent in cases like (7d): the particle *an* refers to *the first bit of BV*, which is only applicable if the BV belongs to a semantic class that allows for a partitive meaning, such as *consumption, cutting,* etc. Also, it is not trivial to decide if two PVs share the same sense of a particle or not, as in (7a) vs. (7b). Does *an* only express some kind of directionality or are the two semantic transfer patterns sufficiently different to assume two particle meanings? Note that our definition of semantic derivation does not make any claim about how to discriminate between particle senses and how to establish a number of senses.

The ambiguity of particles often leads to different senses of PVs, even if the PVs are compositional with respect to the same meaning of the BV. For example, the PV *an|fahren* can have at least three meanings. It is ambiguous between *to drive into* as in (8a), *to start driving* as in (8b), and *to approach by driving* as in (8c). These particle meanings of *an* are shared among semantically similar PVs, respectively, e.g., *an|rempeln* (*to bump into*), *an|laufen* (*to start running*) and *an|steuern* (*to approach by steering*, e.g. a ship).

(8) a. *Das Auto FUHR den Fußgänger AN.*
the car DROVE the pedestrian PRT_{an}
'The car ran into the pedestrian.'

b. *Das Auto FUHR AN, als die Ampel grün wurde.*
the car DROVE PRT_{an}, when the light green turned
'The car went when the light turned green.'

c. *Der Bus FUHR die Haltestelle AN.*
the bus DROVE the stop PRT_{an}
'The bus approached the bus stop.'

We also find cases where a new non-standard meaning is enforced by the semantic interpretation of a PV. (9) is an example from an advertisement campaign for a soft drink which carries the word *Sonne* (*sun*) in its name. Here the PV *zu|gehen* (*to close*) is used, along with the PV *auf|gehen* (*to rise* and *to open*). The sun cannot *close*, but the new type of package – which is advertised here – can.

(9) *Die Sonne GEHT AUF. Und ZU.*
the sun GOES PRT_{auf} and PRT_{zu}
'The sun rises/opens. And closes.'

A definition of particle meaning in terms of semantic transfer patterns as expressed by (6) is compatible with all of the findings listed above, while it does not define precise lexical entries for particles and does not make claims about the number of senses per particle.

2.3 *Syntactic transfer*

So far, we have only discussed the semantic aspects of PVs, but the shifts from BVs to PVs also influence the syntactic behavior of the PVs, which in turn may provide a helpful approximation to the semantics of PVs (Levin 1993). To illustrate the syntactic aspect, consider the examples in (10). Although the PV *an|leuchten (to shine at)* is rather compositional, the means for the illumination *Lampe (lamp)* is represented by the subject of the BV in (10a) vs. a PP complement headed by *mit*$_{dat}$ of the PV in (10b). PV and BV thus behave syntactically differently with respect to their argument structures and the syntactic functions of identical semantic roles.

(10) a. *Die Lampe* LEUCHTET.
the lamp shines
'The lamp shines.'

b. *Peter* LEUCHTET *das Bild mit der Lampe* AN.
Peter SHINES the picture with the lamp PRT$_{an}$
'Peter illuminates the picture with the lamp.'

In addition to changes in the predominant syntactic functions for semantic arguments when comparing PVs to their BVs, we also find *extension* and *incorporation* of syntactic complements, as illustrated by (11) and (12), respectively. The BV *bellen* (*to bark*) in (11) is intransitive, while the corresponding PV *an|bellen* (*to bark at*) is transitive and takes an additional accusative object to express the entity being barked at. This is a case of argument extension within PV subcategorization with respect to its BV. The PV *an|schrauben* (*to screw on*) in (12) shows argument incorporation: it rarely selects an argument to express the location onto which something is screwed, while its BV *schrauben* (*to screw*) adds a complement (here: a PP) to express the direction.

(11) a. *Der Hund* BELLT.
the dog$_{nom}$ BARKS
'The dog barks.'

b. *Der Hund* BELLT *den Postboten* AN.
the dog$_{nom}$ BARKS the postman$_{acc}$ PRT$_{an}$
'The dog barks at the postman.'

(12) a. *Der Mechaniker* SCHRAUBT *die Abdeckung auf die Öffnung.*
the mechanic$_{nom}$ screws the cover on the opening$_{acc}$
'The mechanic screws the cover on the opening.'

b. *Der Mechaniker* SCHRAUBT *die Abdeckung* AN.
the mechanic$_{nom}$ SCREWS the cover PRT$_{an}$
'The mechanic fixes the cover.'

Usually, groups of verbs which are similar in meaning also have similar subcategorization frames and selectional preferences (Schulte im Walde 2000; Merlo and Stevenson 2001; Korhonen *et al.* 2003; Schulte im Walde 2006; Joanis *et al.* 2008). But in (10)–(12) we can observe that this is not necessarily the case for pairs of PVs and their BVs, even if the meaning of the PV is highly transparent.

The problem illustrated here is what we call the *syntactic transfer problem*: the subcategorization frame of the BV must be mapped onto the subcategorization frame of the PV, and the semantic arguments are not necessarily realized as the same syntactic complements by the two verbs. Note that such syntactic transfer patterns tend to be quite stable within groups of PVs with the same semantic shift (Aldinger 2004; Bott and Schulte im Walde 2014c).

One way to computationally address the syntactic transfer problem is by measuring the overlap between all complement slot combinations of any given PV–BV pair and to identify the best correspondences between the slots. We suggest distributional semantic models to support us in the assessment of PV compositionality, while paying attention to syntactic PV–BV transfer: if the PV is non-compositional, we expect a large distributional distance between the correspondences of PV–BV subcategorization slots. For example, in (13b) the PV *an|drehen* (*to palm off sth. on so.*) is opaque with respect to the

BV *drehen* (*to turn*). The typical patients of *turning* (*drehen*) events may be *knobs, wheels* and *heads*, cf. (13a), which are different from the typical patients of a *selling* event as in *an|drehen*. We thus expect to find very different words as typical fillers of the direct object slot of the two verbs, signalling that the two slots do not express the same type of semantic argument, and that the PV is thus non-compositional.

(13) a. *Eulen können ihren Kopf nach hinten* DREHEN.
owls can their head$_{acc}$ to the back TURN
'Owls can turn their heads around backward.'

b. *Der Verkäufer hat ihm das Auto* AN|GEDREHT.
the seller has him the car$_{acc}$ PRT$_{an}$|TURNED
'The salesman has palmed the car off on him.'

The strength of the syntactic transfer will be taken as a proxy for semantic classes and compositionality. We hypothesize that the higher the distributional associative strength between the slots within a syntactic transfer pattern, the stronger the PV compositionality. We further hypothesize that the semantic transfer patterns expressed by (6) are paralleled by regular syntactic transfer patterns.

2.4 *Distributional information*

In order to test our assumptions against empirical data we use distributional semantic models. According to the distributional hypothesis, the meaning of a word is characterized by the distribution of its contexts (Harris 1954; Firth 1957). Intuitively, this corresponds to the idea that we expect to find a word such as *driver* in the context of the word *car*, and the word *captain* in the context of the word *ship*.

One way of defining the concept of *context* is a vector in a high-dimensional space, where each dimension represents an aspect of contextual distribution, such as context words (Sahlgren 2006; Turney and Pantel 2010). Each target word is represented by a vector, and each vector dimension is determined by the co-occurrence strength with context words. For example, if *bone* occurs c times in the local context of *dog*, the dimension *bone* in the vector of *dog* will be c. If each vector dimension refers to a context word, the unreduced vector space has as many dimensions as there are word types in the corpus.

It is possible to reduce the dimensionality and thus abstract over individual lexical items by applying dimensionality reduction techniques, such as Random Indexing (Sahlgren 2005), Singular Value Decomposition (Landauer and Dumais 1997) and Latent Dirichlet Allocation (Blei *et al.* 2003). It is also possible to use more complex units of context than simple words as vector dimensions, e.g., by relying on subcategorization functions (Padó and Lapata 2007), where verbs can, for example, be characterized by the kinds of subjects or objects they typically take. An obvious example is that we expect to find *dog* as a typical subject of the verb *to bark* and *cat* as a typical subject of *to meow*. The distributional similarity/distance between two lexical items can be measured as the geometrical distance between their vectors, e.g. by computing the cosine of the angles of said vectors.

While distributional methods cannot provide clear-cut lexical definitions, they are convenient and successful proxies for comparing words semantically: words which are similar in meaning have a strong tendency to appear in similar contexts. Applied to the problem of PV compositionality, we can expect that distributional closeness of PVs and BVs signals high compositionality. For our experiments, we use the following configurations of context representations:

- *Windows* of surrounding lemmatized words: we use *n* words to the left and to the right of each target word, where *n* is a variable. Vector components represent words from the context, and the extension in each dimension represents frequency or *local mutual information* (LMI) as association strength (Evert 2004).

- *Complement slot fillers* for syntactic subcategorization models: vectors represent subcategorization slots for each verb (either BV or PV); vector components correspond to slot filler words or abstractions of slot fillers (such as latent dimensions).

- *Subcategorization frames*: dimensions represent subcategorization frames for each PV–BV pair. Each vector component corresponds to the observed frequency of a subcategorization frame. The distance between different PV–BV pairs can be used as a criterion for grouping together verb pairs with similar patterns.

From a practical point of view, the window approach has an advantage over the syntactic approach because it can use much more evidence mass: it is not restricted to verb arguments and can thus use all words in local contexts. From a theoretical point of view, however, the window approach does not integrate the linguistic generalizations we discussed above: regularity of semantic shifts and instances of syntactic transfer.

2.5 *Hypotheses*

The goal of this article is to empirically test hypotheses H1–H3 which we have derived on a theoretical basis:

H1 *Semantic Transfer:* For PVs that are not fully lexicalized there are groups of BVs which undergo the same semantic derivation when they combine with the same particle type, cf. Sections 2.1 and 2.2.

H2 *Syntactic Transfer:* The semantic transfer patterns are paralleled by syntactic transfer patterns, cf. Section 2.3.

H3 *Distributional Transfer:* The degree of PV compositionality can be assessed by comparing distributional PV and BV contexts at the syntax-semantics interface, cf. Section 2.4.

Following an overview of related previous work on particle verbs in Section 3, Section 4 will define and conduct three experiments according to our three hypotheses.

3 PREVIOUS APPROACHES TO PARTICLE VERBS

German PVs have been studied extensively from a theoretical point of view (Stiebels and Wunderlich 1994; Stiebels 1996; Lüdeling 2001; Dehé *et al.* 2002; Müller 2002, 2003; McIntyre 2007).[3] Lüdeling (2001) investigated whether PVs are morphological objects or phrasal constructions and how they can be distinguished from secondary predicate constructions or adverbial constructions. She revealed a series of theoretical problems and analyzed PVs as lexicalized phrasal constructions, considering separability the strongest argument for this analysis. Olsen (1997) studied German PVs at the morpho-syntactic interface

[3] Also see a bibliography on verb particle constructions, as maintained by Nicole Dehé until 2015: `http://ling.uni-konstanz.de/pages/home/dehe/bibl/PV.html`.

and analyzed cases in which an explicit argument of a BV becomes implicit in the formation of a PV. Müller (2002, 2003), in turn, argued for an analysis of PVs as verbal complexes at the morpho-syntactic interface, and provided lexical interpretations. Under his view, PVs are seen as both morphological and syntactic objects. For the present work, the status of PVs on the morphological vs. the syntactic level is not relevant, so we will not commit ourselves to a specific perspective in this respect.

Research addressing the semantics of verb particles has mostly focused on specific particle types, such as *auf* (Lechler and Roßdeutscher 2009), *nach* (Haselbach 2011), *ab* (Kliche 2011), and *an* (Springorum 2011). Springorum *et al.* (2012) and Rüd (2012) presented automatic classification methods for PVs with the particles *an* and *auf*, respectively. Springorum *et al.* (2013b) provided a case study of regular meaning shifts in PVs where they argue that particles have a meaning which is implicit in the semantic transfer pattern, in a similar way as we argue here.

Predicting degrees of PV compositionality from a computational perspective has been addressed previously, mainly for English. Most prominently, Baldwin *et al.* (2003) defined a word-based model of Latent Semantic Analysis for English particle verbs and their constituents, and measured the distributional similarity of the models to evaluate the resulting degrees of compositionality against various WordNet-based gold standards. McCarthy *et al.* (2003) exploited measures on syntax-based distributional descriptions as well as selectional preferences, to predict the compositionality of English particle verbs. Bannard (2005) describes a distributional approach that compared word-based co-occurrences within the British National Corpus for English particle verbs with those of the respective base verbs and particles. Cook and Stevenson (2006) addressed the compositionality and the meaning of English particle verbs by a distributional model encoding standard verb semantic features (especially subcategorization-based information) and PV-specific heuristics. A larger multifactorial study of idiomacity within a construction grammar framework (Wulff 2010) introduced a measure to compute compositionality with respect to both PV constituents.

Regarding computational approaches to German PVs, Aldinger (2004) and Schulte im Walde (2004, 2005) were the first to study them from a corpus-based perspective, with an emphasis on the subcategorization behavior and syntactic change. Aldinger (2004) investigated the regularity in syntactic subcategorization transfer. Schulte im Walde (2005) explored salient features at the syntax-semantics interface that determined the nearest semantic neighbors of German PVs. Relying on the insights of this study, Hartmann (2008) presented preliminary experiments on modeling the subcategorization transfer of German PVs by measuring the overlap of argument heads, in order to strengthen PV–BV distributional similarity. The results of that study were not conclusive due to data sparseness. Kühner and Schulte im Walde (2010) used unsupervised clustering to determine the degree of compositionality of German PVs. They hypothesized that compositional PVs tend to occur more often in the same clusters with their corresponding BVs than opaque PVs. Their approach relied on nominal complement heads in two modes, (i) with and (ii) without explicit reference to the syntactic functions. The explicit incorporation of syntactic information (i) yielded less satisfactory results, since a given subcategorization slot for a PV complement does not necessarily correspond to the same semantic type of complement slot for the BV, thus putting the syntactic transfer problem in evidence, again.

Bott and Schulte im Walde (2014b) showed that a window-based model can predict degrees of compositionality and establish a ranking of PVs accordingly, to significantly correlate with human ratings. Within this study, we focused on the influence of various linguistic factors, such as the ambiguity and the overall frequency of the verbs and syntactically separate occurrences of verbs and particles that typically cause difficulties for the correct lemmatization of PVs.

Köper and Schulte im Walde (2017) combined similar textual distributional information with images, to improve the prediction of compositionality for German noun compounds and particle verbs. Bott and Schulte im Walde (2014c) argued that the semantic classes of PVs can be predicted by purely syntactic features. They showed that automatically derived semantic classes overlap significantly with class distinctions based on human ratings. In Bott and Schulte im Walde (2014a), we showed that a computational assessment of syntactic transfer patterns is feasible and that a computational model can predict slot correspondences. Finally, in Bott and Schulte im Walde (2015) we presented preliminary work on predicting PV compositionaly on the basis of the modeling of syntactic transfer patterns.

4 EXPERIMENTS

Up to now, we motivated our research hypotheses from a theoretical perspective. In this section, we assess our hypotheses within three computational experiments. In Section 4.1, we approximate semantic transfer and the meaning of particles by semantically clustering PVs that share semantic transfer patterns, while using syntactic features in the form of subcategorization frames. In Section 4.2, we verify that syntactic transfer can be predicted in isolation, and in Section 4.3, we compare window-based models and models integrating syntactic transfer information to determine the compositionality of PVs. The experiments presented here are based on preliminary investigations in Bott and Schulte im Walde (2014b,c,a, 2015), which we now extend and discuss in more detail and depth.

4.1 *Experiment 1: Modeling semantic transfer*

The first experiment explores semantic derivation and the meanings of particles. Based on our theoretical considerations, we expect PV–BV pairs to group such that both BVs and PVs are semantically similar, and that the relation between them (i.e. a particle meaning) is captured as a consistent semantic transfer pattern. Since we also assume that semantic derivation is reflected by syntactic transfer patterns, we aim to automatically derive semantic groups on the basis of the syntactic behavior of PV–BV pairs.

As argued above, it is difficult both to define the meanings of particles and to clearly distinguish between them. For this reason, supervised classification techniques are reasonable, as they require training and test sets which reliably reflect distinctions between particle senses. Such data sets are expensive to create, however, and it is difficult to agree on exact numbers and definitions of particle senses on theoretical grounds. For these reasons, we believe that the derivation of groups of PV–BV pairs (and different particle senses) can be addressed more efficiently by means of clustering techniques.

4.1.1 Gold standard classification

We created a gold standard of 32 PVs listed in Fleischer and Barz (2012), including 14 PVs with the particle *an* and 18 PVs with the particle *auf*. We focused on two particle types in order to have a small and controlled test bed which allows us to study the syntactic transfer in detail. The selected verbs were considered highly compositional, in order to investigate the correspondences between subcategorization properties. The PV set contains PVs with argument slots that are typically realized through different syntactic subcategorizations, as in example (10) with *an|leuchten*. In addition, the PV set contains PVs exhibiting argument incorporation or extension. We excluded PVs which are clearly polysemous.

The full gold standard is presented in Table 1. The first part of the *semantic class* labels was taken from Fleischer and Barz (2012); we further distinguished between the classes based on the meanings of the BVs (second part of the labels), by breaking down the general classes into more detailed classes, such as verbs of *tying*, *gaze* and *sound*. The selected verbs have a clear subcategorization pattern for BVs and PVs.

In order to validate the gold standard, we assessed it with the help of six human expert raters,[4] all German native speakers with a linguistic background. The raters were not directly asked to group PVs into categories. Instead, the PVs were presented in pairs,[5] and the raters decided whether or not the pairs belonged to the same semantic category, taking semantic similarity of the PVs as the basis for their decision. For example, the PVs *an|schneiden* (*to start cutting*) and *an|ketten* (*to chain at*) were presented as a pair to be rated. In this case, the decision that they *do not belong to the same semantic class* was expected. No pre-defined categories were provided, and the raters were not asked to provide a name or description of the categories. We did *not* ask participants to take any syntactic criteria into consideration, which were the criteria we actually used for the compilation of the gold standard.

The inter-annotator agreement was substantial (Landis and Koch

[4] All human ratings in this article exclude the authors as raters.

[5] All possible PV combinations were generated, while keeping PVs with *an* separate from those with *auf*.

Table 1: The gold standard PV–BV classes, with sub-categorization patterns

Particle	Typical frames for the BV	Typical frames for the PV	Semantic class	Verbs in class	
an	NPnom +NPacc +PP-an	NPnom +NPacc +PP-an	locative/ relational tying	an|binden an|ketten	to tie at to chain at
	NPnom +PP-zu/ in/nach/ auf	NPnom +NPacc	locative/ relational gaze	an|blicken an|gucken an|starren	to glance at to look at to stare at
	NPnom +NPacc +PP-mit	NPnom +NPacc +PP-mit	ingressive consump-tion	an|brechen an|reißen an|schneiden	start to break start to tear start to cut
	NPnom	NPnom +NPacc	locative/ relational sound	an|brüllen an|fauchen an|meckern	to roar at to hiss at to bleat at
	NPnom +NPacc +PP-an	NPnom +NPacc	locative/ relational fixation	an|heften an|kleben an|schrauben	to stick at to glue at to screw at
auf	NPnom	NPnom	locative/ blaze-bubble	auf|brodeln auf|flammen auf|lodern auf|sprudeln	to bubble up to light up to blaze up to bubble up
	NPnom +PP-zu/ in/nach/ auf	NPnom	locative/ gaze	auf|blicken auf|schauen auf|sehen	to glance up to look up to look up
	NPnom +NPacc	NPnom +NPacc	locative/ dimensional instigate	auf|hetzen auf|hetzen	to instigate to rouse
	NPnom +NPacc +PP-auf	NPnom +NPacc	locative/ relational fixation	auf|heften auf|kleben auf|pressen	to staple on to glue on to press on
	NPnom	NPnom	ingressive sound	auf|brüllen auf|heulen auf|klingen auf|kreischen auf|schluchzen auf|stöhnen	suddenly roar suddenly howl suddenly sound suddenly scream suddenly sob suddenly moan

Table 2:
Inter-annotator agreement and comparison of the gold standard and the human ratings (Fleiss' κ)

	an	auf	an + auf
Inter-annotator agreement	0.79	0.64	0.70
Average agreement between annotators and gold standard	0.73	0.74	0.73

1977) with Fleiss' $\kappa = 0.68$ (Fleiss 1971).[6] As a measure of agreement between raters and the previously created gold standard, we performed pair-wise calculations. For this assessment, the gold standard was transformed into PV pairs, and the value *true* was assigned if the two verbs of a pair belonged to the same category, and *false* otherwise. κ scores were calculated for each annotator, and the average of the agreement scores was taken.

Table 2 presents the human–gold comparison, separately for *an* and *auf* and also for the gold standard as a whole. While for the particle *an* the inter-annotator agreement is higher than the agreement between raters and gold standard, the reverse is true for the particle *auf*, and on average the human agreement with the gold standard is similar to the agreement among the annotators. We conclude that our gold standard provides a valid representation of human language intuition. Most importantly, the annotators did not use syntactic criteria and still validated a gold standard whose creation was explicitly based on syntactic subcategorization frames. In other words: there is an apparent syntax-semantics relation for our selected PVs.

4.1.2 Feature selection

As basis for corpus-based features, we used a lemmatized and tagged version of the SdeWaC corpus (Faaß and Eckart 2013), a web corpus of ≈880 million words. For linguistic pre-processing, we used the MATE parser (Bohnet 2010) to extract syntactic subcategorization frames.

For each PV–BV pair, we extracted two parallel sets of features, one for the BV and one for the PV. This allowed us to model the syntactic transfer. For example, we expected that an ideal transfer from a group of transitive BVs to a group of intransitive PVs should be re-

[6] One of the six raters showed low agreement with the other raters. Eliminating this rater from the calculation of agreement, we achieved an even higher inter-annotator agreement score of $\kappa = 0.76$.

flected in high values for the features *BV:transitive* and *PV:intransitive*[7] and, in turn, low values for *BV:intransitive* and *PV:transitive*.

We distinguished between two ways of selecting the feature types from the corpus: manually and automatically. For the manual feature selection, we extracted only those features from the parsed frames which we already used in the creation of the gold standard and which are listed in Table 1. This resulted in a small feature set of 30 features (15 features for PVs and BVs, respectively). For the automatic feature selection, we used the *n* most frequent frames in the corpus, as determined across the set of verbs in the gold standard. In order to create an artificial upper bound, we used the typical frames as defined in Table 1 as a set of idealized "lexicographic" descriptions.

Regarding the syntactic dependency representation provided by the parser, we excluded subjects and modifiers from the representation of subcategorization frames. We, however, included PP modifiers because quantitative information on PP adjuncts has proven successful next to that of PP arguments (Schulte im Walde 2006; Joanis *et al.* 2008).

The feature vectors were normalized to their unit vectors of length 1, because the frequency ratio between BVs and PVs potentially varied strongly. The vector combination for each PV–BV pair was done by simply concatenating the dimensions of the two BV and PV vectors. In this way, each subcategorization frame was represented for both the BV and the PV. For example, the vectors for the intransitive frame were represented as *BV:intransitive* and *PV:intransitive*.

4.1.3 Clustering methods

We wanted to assess and compare hard and soft clustering for our problem, so we applied the two clustering algorithms *K-means* and *Latent Semantic Classes (LSC)*. K-means is a widely used flat, hard-clustering algorithm; we used the Weka implementation (Witten and Frank 2005). LSC (Rooth 1998; Rooth *et al.* 1999) is a two-dimensional soft-clustering algorithm which learns three probability distributions: one for the clusters, and one for the output probabilities of each element and for each feature type with regard to a cluster. The latter two (elements and features) correspond to the two dimensions of the clus-

[7] Note that *transitive* and *intransitive* are only convenient abbreviations for the labels *NPnom* and *NPnom+NPacc*, which are used in Table 1.

tering. In our case the elements are the PV–BV pairs, and the features are normalized counts of the subcategorization frames.

4.1.4 Evaluation

We evaluated the clusterings in terms of *Purity* (Manning *et al.* 2008), *Rand Index* (Rand 1971) and *Adjusted Rand Index* (Hubert and Arabie 1985). Purity assesses individual clusters in terms of the ratio between the number of elements of the majority class and the total number of elements in the data set. A perfect clustering has a Purity of 1 while the lower bound is 0. Since Purity does not capture the amount of clusters over which each target class is distributed, also non-perfect clusterings may have a Purity of 1. However, as long as the number of clusters is constant, Purity provides an intuitive means to evaluate our cluster analyses.

The Rand Index (RI) looks at pairs of elements and assesses whether they have been correctly placed in the same cluster. RI is sensitive to the number of non-empty clusters and can capture both the quality of individual clusters and the amount to which elements of target categories have been grouped together. Since RI looks at pairwise decisions, it is also applicable to the human ratings. The Adjusted Rand Index (ARI) is a variant of RI which is corrected for chance. RI has values between 0 and 1; ARI can have negative values.

We evaluated the cluster analyses of the verbs with the particles *an* and *auf* separately and for the gold standard as a whole (*an* + *auf*). We set the number of clusters equal to the number of target gold categories: 5 clusters for both the *an*-set and the *auf*-set and 10 clusters for the whole gold standard.

For the evaluation of LSC clusters with respect to Purity, RI and ARI, we transferred each soft clustering to a hard clustering by applying a cutoff value to the output probabilities for cluster membership. We tried various cutoff levels and found that for the sets of *an* and *auf* PVs 0.1 provided a reasonable trade-off between coverage (the total number of elements retained in all clusters) and ARI. This is also the value used in Kühner and Schulte im Walde (2010) in a similar setup.

4.1.5 Results and discussion

The clustering results are presented in Table 3, with the best automatically obtained results in gray cells. The human rating scores are given

Table 3: Results across clustering methods and feature sets

		an			auf			an + auf		
		Purity	RI	ARI	Purity	RI	ARI	Purity	RI	ARI
Human ratings			0.93			0.92			0.92	
K-means	upper: bound: idealized features	0.83	0.91	0.70	0.88	0.92	0.72	0.93	0.97	0.82
	selected features	0.67	0.82	0.29	0.75	0.87	0.52	0.46	0.88	0.32
	20 feat	0.58	0.74	0.18	0.69	0.69	0.40	0.43	0.88	0.14
	50 feat	**0.67**	**0.80**	0.20	0.75	0.83	0.38	0.43	0.90	0.19
	100 feat	**0.67**	0.79	0.18	0.75	0.83	0.40	0.49	0.90	0.21
	200 feat	0.58	0.74	0.13	**0.81**	**0.86**	0.52	0.43	0.88	0.18
LSC	selected features; cutoff: 0.1	0.63	0.78	**0.22**	0.80	0.85	**0.55**	**0.85**	**0.92**	**0.59**

in the first row and allow for a direct comparison between automatic clustering and human decisions.[8] The second row shows the upper bound represented by the manually defined feature vectors. Note that this is an *artificial* upper bound and not an experimental result, even if obtained by clustering.

The third row corresponds to the evaluation results for the manually selected corpus-based features used within K-means, in comparison to the following rows concerning the results based on the automatically selected n most frequent features, with $n = \{20, 50, 100, 200\}$. The last part of the table shows the results obtained with the LSC soft clustering algorithm, when applying the cutoff of 0.1 to the cluster membership probability. Note that the Purity values are comparable to each other because the number of clusters was held constant.

The results relying on our manual features as provided by Table 1 do not get perfect scores of 1 because of lexicographic differences concerning individual entries. They are, however, highly similar to the results obtained by the human validation of the gold standard, and thus demonstrate the feasibility of our approaches. The automatic cluster-

[8] Differently to RI, Purity and ARI are not based on pair-wise decisions and thus not applicable to the human ratings.

ing results relying on corpus-based features result in lower scores, of course, but they still represent a very strong tendency to group together PV–BV pairs into semantic classes. We can achieve relatively high Purity and RI scores, thus demonstrating that our approach is generally valid.

Concerning the corpus-based features, the manually selected set seems to perform only slightly better than the automatic feature selection settings. This is surprising, since the manually selected set was "tuned" to use the most salient features for our task. So while the noise adds potentially unrelated features, it does not considerably harm the cluster analyses. There appears to be no optimal setting for n to provide the best results across all settings. It is clear from the table, however, that the lowest number of features ($n = 20$) tends to be outperformed by a larger number of features.

As a general tendency, the soft clusterings by LSC perform on a comparable level with the hard clusterings by K-means. For the joint gold standard set *an* + *auf* and a cutoff point of 0.1, LSC performs even much better than K-means. But this comes at the cost of a very low coverage: Only 20 verbs are retained in the converted clusters, while the target size is 32.

Given that (i) the automatic clustering was performed on the basis of syntactic features while the annotators in the human classification task focused on purely semantic criteria, and that (ii) the cluster analyses were rather successful, we conclude that the semantic and the syntactic perspectives led to the creation of similar classes. We therefore provided empirical evidence for both hypotheses H1 and H2.

4.2 *Experiment 2: Modeling syntactic transfer*

In Section 2, we hypothesized that syntactic transfer patterns can be detected with distributional methods. If subcategorization slots from a PV–BV pair correspond to each other and realize the same semantic argument, we expect them to be distributionally similar. This hypothesis was tested with the following experiment.

4.2.1 Automatic prediction of slot correspondences

We rely on the same gold standard as in the previous experiment (cf. Table 1). Most importantly, the dataset contains PV–BV verb pairs whose argument slots are typically realized by different syntactic subcategorizations, as described by the expected "typical frames". The differences in the typical frames for PV vs. BV groups represent the expected transfer patterns.

The aim of this experiment was to predict transfer patterns by correspondences between syntactic slots in PV and BV subcategorization frames. Firstly, we extracted all subcategorization frames for both BVs and PVs from the parsed version of the SdeWaC corpus. We then selected the n most frequent subcategorization frames, where n was limited to 5. Each of these frames is a set of subcategorization slots of the form $\{\sigma_1, \ldots, \sigma_m\}$. If $\mathrm{frame}_{v,i}$ refers to the set of subcat slots of the i^{th} most frequent subcategorization frame for a verb v, we then define the set $\mathrm{slots}_{v,n}$ as follows:

(14) $\mathrm{slots}_{v,n} := \{\sigma_j | \sigma_j \in \mathrm{frame}_{v,i}, 0 < i \leq n\}$

Informally, $\mathrm{slots}_{v,n}$ is the set of subcat slots which appear in any of the n most frequent frames of v. The simple transitive frame, for example, contains a subject slot and an accusative object slot.

We built a vector space model for all possible combinations of BV slots and PV slots for each PV–BV pair $<pv, bv>$. The dimensions of the vector were instantiated by the head nouns of the respective syntactic function. The best matching slot $\hat{\sigma}'$ of a PV for a given slot σ_i (with slot vector $\vec{\sigma}_i$) of the corresponding BV is then defined as the maximum slot cosine score:

(15) $\hat{\sigma}' := \underset{\sigma_j | \sigma_j \in \mathrm{slots}_{pv,n}}{\arg\max} \cos(\vec{\sigma}_i, \vec{\sigma}_j)$

Table 4 shows the most frequent dimensions in the vectors corresponding to PP arguments headed by *an* for the verbs *heften* (*to attach*) and *an|heften* (*to attach to*). The two verbs can be used in similar contexts with similar arguments. For example, both vectors include head nouns expressing typical places to attach things to, such as a *pin board* (*Pinnwand*), a *wall* (*Wand*), and a *board* (*Brett*). Accordingly, the two vectors are similar to each other. Note that although both vectors correspond to PP slots headed by the preposition *an*, a syntactic transfer from the accusative to the dative case takes place. In addition, the example vectors demonstrate that the features are often sparse.

A variable threshold was applied to the cosine similarity, to sep-

Table 4: Most frequent dimensions for two sample vectors representing subcategorization slots of the verbs *heften* (*to attach*) and *an|heften* (*to attach to*)

anheften-an$_{dat}$	count	**heften-an**$_{acc}$	count
Oberfläche (surface)	3	Ferse (heel)	154
Gerichtstafel (court notice board)	3	Brust (breast)	48
Stelle (spot)	2	Revers (lapel)	43
Schluss (end)	2	Kreuz (cross)	32
Unterlage (document)	1	Wand (wall)	30
Kirchentür (church door)	1	Spur (trace)	12
Brett (board/plank/shelf)	1	Tafel (board)	11
Pinnwand (pin board)	1	Fahne (flag)	11
Körper (body)	1	Tür (door)	11
Punkt (point)	1	Pinnwand (pin board)	9
Bauchdecke (abdominal wall)	1	Kleid (dress)	6
Baum (tree)	1	Brett (board/plank)	6
Schleimhautzelle (epithelial cell)	1	Mastbaum (mast tree)	6
Himmel (heaven/sky)	1	Körper (body)	5
Spur (trace)	1	ihn (him)	5
Sphäre (sphere)	1	Kleidung (clothing)	5
Wand (wall)	1	Oberfläche (surface)	5
Hauptreaktor (main reactor)	1	Stelle (spot)	4
Engstelle (constriction)	1	Baum (tree)	4
Pflanze (plang)	1	Jacke (jacket)	4
Protein (protein)	1	Mantel (coat)	4
Unterseite (down side)	1	Teil (part)	3
Zweig (twig)	1	Krebszelle (cancer cell)	3
Geist (spirit)	1	mich (me)	3
Pin-Wand (pin board)	1	schwarz (black)	3

arate corresponding from non-corresponding subcategorization slots. This is important for the detection of argument incorporation and extension. If, for example, for a given BV slot no PV slot can be found with a cosine value above the threshold, we interpret this as a case of argument incorporation. In contrast, a slot from a PV which cannot be matched to a slot of its BV is taken to signal argument extension.

For initializing the BV and PV vector dimensions, we relied on the subcategorization database compiled by Scheible *et al.* (2013), which provides a convenient access to subcategorisation information in the same dependency-parsed version of the SdeWaC corpus as used in the previous experiment. Once the verb vectors were built, we used them to predict subcategorization transfer. The baseline for the predictions was obtained by a random PV–BV slot correspondence. The results will be presented in Section 4.2.3, after introducing the gold ratings.

4.2.2 Human ratings on slot correspondences

Each pair of subcategorization slots described in Section 4.2.1 was rated by human judges. The pairs were presented as

<BV-subcategorization-slot, PV-subcategorization-slot>

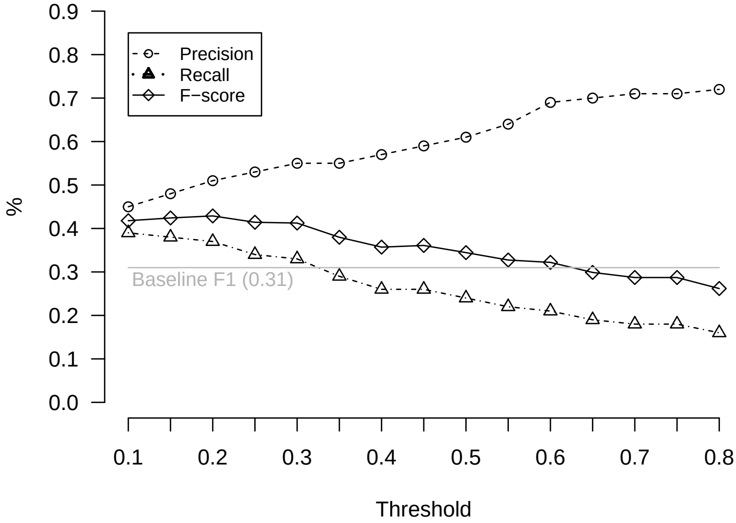
and in blocks corresponding to identical BV subcategorization slots, such that the raters could directly compare all PV subcategorization slots for a given BV slot. The order of the blocks was randomized.

The raters were asked to rate the pairs on their semantic correspondence. Three annotation examples were provided to guide the ratings, cf. (16). (16a) presents a negative example, as no grammatically correct sentence is possible for *durch|schwimmen* with a PP complement headed by *durch*. Accordingly, the sentence in (16a-iii) is ungrammatical. (16b) presents a positive example. In unclear cases, the raters were invited to produce example sentences.

(16) a. (i) <schwimmen-durch$_{acc}$, durchschwimmen-durch$_{acc}$>
(ii) *Der Hund* SCHWIMMT *durch den Fluss.*
the dog SWIMS through the river$_{acc}$
'The dog swims through the river.'
(iii) **Der Hund* DURCH|SCHWIMMT *durch den Fluss.*
the dog PRT$_{durch}$|SWIMS through the river$_{acc}$

b. (i) <schwimmen-durch$_{acc}$, durchschwimmen$_{acc}$>
(ii) identical to (16a-ii)

Figure 1: Trade-off between precision and recall across thresholds



1

(iii) *Der Hund* DURCH|SCHWIMMT *den Fluss.*
the dog PRT$_{durch}$|SWIMS the river$_{acc}$
'The dog swims through the river.'

The dataset was distributed over two annotation forms, and each annotation form was annotated by two native speakers. The annotators had a background in linguistics or computational linguistics. They described the annotation as difficult to perform. This was also reflected by inter-annotator agreement; we observed fair agreement, Fleiss' $\kappa = 0.31$ (Landis and Koch 1977).

4.2.3 Results and discussion

Figure 1 presents the results when predicting slot correspondences, as measured by precision, recall and the harmonic F-score when comparing the system output to the human ratings. True positives were obtained if the system selected the same slot correspondence for a given slot that the human raters had selected. Since a variable threshold was applied, we find a trade-off between precision and recall. As expected, precision improves with higher thresholds, but this comes at the cost of lower recall. The F-score decreases with an increasing threshold, with a local maximum around a threshold of 0.2. With threshold values >0.6 the F-score drops below the baseline.

Overall, the system manages to predict correspondences between syntactic subcategorization slots to a fair degree of success. Our hypothesis that correspondence between subcategorization slots can be predicted by distributional semantic similarity has thus been confirmed. Then again, the success was not as high as we initially expected. We assume that this is due to the difficulty of the task, as indicated by the low inter-annotator agreement.

Since the annotators gave detailed comments after the annotation was completed, we detected theoretical problems which also apply to the automatic matching process. For example, the pair (17a)/(17b) for the verb *kleben* (*to stick/glue*) exemplifies a syntactic transfer of the theme argument *Zettel* (*note*), which is realized as the accusative object of the PV in (17a) and as the subject of the BV in (17b). The system failed to predict this transfer. This can be attributed to the fact that *kleben* can undergo a causative/inchoative alternation (Levin 1993), as exemplied by (17b)/(17c). We can observe a one-to-many match here. This is a problem which is hard to solve with our approach because the correspondence of PV–BV slots interferes with a slot correspondence among different uses of the BV.

(17) a. *Gerda KLEBT den Zettel an die Tür AN.*
Gerda STICKS the note on the door PRT_{an}
'Gerda sticks the note on the door.'

b. *Der Zettel KLEBT an der Tür.*
the note STICKS at the door
'The note sticks to the door.'

c. *Gerda KLEBT den Zettel an die Tür.*
Gerda STICKS the note at the door
'Gerda sticks the note on the door.'

Finally, we found that many of the feature vectors were extremely sparse, such as the vector of the PP headed by an_{dat} for the verb *an|heften* in Table 4. The sparsity problem could be remedied by reducing the number of dimensions, e.g, by applying some kind of abstraction over the head nouns. For example, the concepts of *Tür* (*door*) and *Kirchentür* (*church door*) are strongly related and could be merged into one dimension of the feature vector. The same holds for the concepts of *Pinnwand* (*pin board*), *Wand* (*wall*) and *Tafel* (*blackboard*). We suspect that with a certain level of abstraction over such concepts, the vectors would be more reliable. For this reason, we used generalization techniques in the following experiment.

4.3 *Experiment 3: Modeling distributional transfer*

In Section 2, we argued for a distributional assessment for predicting the degrees of compositionality for German PVs. We hypothesized that the more compositional the PVs are, the more similar a PV and a BV are in their meanings and the more similar are their distributional properties. In the following, we suggest two types of distributional models in order to assess PV compositionality in a distributional manner:

1. *Window models*: If PVs occur in similar lexical contexts as their BVs, they are distributionally similar, which is taken as an indicator that the PVs are semantically similar to their BVs, hence highly compositional. In contrast, distributional distance should indicate lexical dissimilarity and thus low compositionality.
2. *Syntactic subcategorization models*: This approach models syntactic transfer: If PV subcategorization slots can be strongly mapped to subcategorization slots of their BVs, this indicates strong compositionality. The model thus integrates the prediction of slot correspondences between PVs and their BVs that was verified in the previous section.

The first option, *window models*, is conceptually very simple, since it compares unsorted local contexts. It does however not exploit the fact that local co-occurring words can be distinguished by their syntactic functions. Then again, window-based models accumulate an evidence mass which is proportionate to window size. One might suspect that this advantage in evidence mass comes at the cost of degraded quality, since windows represent bags of words.

The second option models the syntactic transfer and is thus theoretically more appealing because it distinguishes between context words according to their syntactic functions. Our hypothesis is that the degree of predicted associative strength of syntactic transfer represents an indicator of semantic transparency. If the complements of a PV strongly correspond to any complement of its BV, the PV is regarded as highly compositional, even if the PV complements are *not* realized as the same syntactic argument types, as long as a relation between these two subcategorization slots can be established. Conversely, if only a weak correspondence between the PV complements and the BV complements can be established, this is an indicator of low compositionality.

Our second approach is novel and exploits fine-grained syntactic transfer information, which is not accessible within a window-based approach. At the same time, it preserves an essential part of the information contained in context windows, since the head nouns within subcategorization frames typically appear in the local context.

The syntactic approach may however suffer from a practical problem, i.e., data sparseness. While in the case of window information every instance of a verb has *2*n* words in the local context, in the transfer approach each verb instance has just as many co-occurring words as it has subcategorization slots. To compensate for this inevitable data sparseness, we employed the lexical taxonomy *GermaNet* (Hamp and Feldweg 1997) and *Singular Value Decomposition (SVD)* to generalize over individual complement heads. Dimensionality reduction techniques have proven effective in previous distributional semantics tasks (e.g., Joanis *et al.* 2008, Brody and Elhada 2010, Ó Séaghdha 2010, Guo and Diab 2011, Bullinaria and Levy 2012, Turney 2012).

1. *GermaNet (GN)* (Hamp and Feldweg 1997) is the German version of WordNet (Fellbaum 1998). We used the n^{th} topmost taxonomy levels in the GermaNet hierarchy as generalizations of head nouns. In the case of multiple inheritance, the counts of a subordinate node were distributed over the superordinated nodes.
2. *Singular Value Decomposition (SVD)*: We used the DISSECT tool (Dinu *et al.* 2013) to apply singular value decomposition to the vectors of complement head nouns in order to reduce the dimensionality of the vector space.

GermaNet is a knowledge-driven way of mapping concepts to more general concepts; SVD learns abstract latent dimensions automatically.

4.3.1 Experimental setup

Window Model: For the assessment of PV compositionality based on windows we used a word vector space model (Sahlgren 2006; Turney and Pantel 2010). The experiment replicates and extends an approach presented in Bott and Schulte im Walde (2014b), where we demonstrated the reliability of window-based models to predict PV compositionality and assessed the effect of target frequency, ambiguity, and lemma restoration. For each target PV, we constructed a vector space with s_l dimensions, where s_l was the size of the vocabulary as extracted from a lemmatized corpus. The vector components represened co-occurrence counts in local context, which was defined as a window of n words to the left and to the right of the target PV. In our experiment with window-based models, words were lemmatized, but no dimensionality reduction was applied. Since PVs may occur in syntactically separated paradigms (i.e., the particle separated from the verb), but lemmatizers are blind to syntactic dependencies, we applied lemma correction: If we found 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. Our models vary (a) in the size of the context window, (b) by (not) applying term-weighting, and (c) by using all context words or only content words as vector dimensions. Windows did not go beyond sentence boundaries, because our corpora were sentence-shuffled for copyright reasons. The semantic similarity, which is taken as the associative strength of a PV–BV pair $<pv, bv>$ was calculated as the cosine between the vectors for pv and bv.

Syntactic Subcategorization Model: The rationale behind the use of syntactic slot correspondence to predict the degree of PV–BV compositionality is that we only try to match those semantic arguments which correspond to each other. This requires two steps: first, detecting the best matching slots in PV–BV pairs; second, determining their average distributional similarity. Relying on the five most frequent subcategorization frames, we first selected the best matching BV slot for each PV complement slot, as described in 4.2, and then calculated the associative strength as_{pv}^{bv} between a PV–BV pair $<pv, bv>$ as the average cosine score over the best matches for all PV slots and the best matches for all BV slots. The associative strength as_{pv}^{bv} is taken as a measure of the correspondence of PV–BV complement slots and their realization of the same semantic arguments. We thus take the strength to predict the degree of PV compositionality. To account for possible null correspondences in argument incorporation and argument extension cases, we applied a variable threshold on the cosine distance ($t = 0.1/0.2/0.3$). If the best matching BV complement slot of a PV complement slot had a cosine score below this threshold, it was not taken into account. $t = 0$ refers to setting no threshold.

4.3.2 Vector weighting and Generalization

Not all context words are equally predictive for lexical distributional models: Some words tend to occur frequently across many contexts, which makes them bad predictors. We thus leveraged information which stems from words that occur in specific contexts and were expected to represent salient predictors. To this end, we used *local mutual information* (LMI, Evert 2004) as a vector weighting method and test if term weighting has an effect on the prediction quality. To filter out the distortion introduced by non-content words, we used window models which only contain context information corresponding to nouns, verbs and adjectives. To address the second representation issue, data sparseness in syntactic subcategorization models, we applied GermaNet and SVD as generalizations.

4.3.3 Corpora

In order to estimate the effect that the amount of data has on the prediction quality, we compare vector spaces from two differently sized corpora. As in the previous two experiments, we used the dependency-parsed SdeWaC corpus with ≈880 million words. In comparison, we used the DECOW14[9] corpus (Schäfer and Bildhauer 2012) with ≈20 billion words. The DECOW14 data was pre-processed and dependency-parsed with a toolchain presented in Björkelund *et al.* (2013): Their pipeline used the graph-based MATE dependency parser (Bohnet 2010), which was also used for the preprocessing of the SdeWaC corpus. For morphological analysis MarMoT (Müller *et al.* 2013) and SMOR (Schmid *et al.* 2004) were applied.

4.3.4 Gold standards

We evaluated our models against three gold standards (GSs). Each of them contains PVs across different particles and was annotated by humans for the degree of compositionality:

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

1. ***GS1***: A gold standard collected by Hartmann (2008), consisting of 99 randomly selected PVs across 11 particles, balanced over 8 frequency ranges and judged by 4 experts on a scale from 0 to 10.
2. ***GS2***: A gold standard of 354 randomly selected PVs across the same 11 particles, balanced over 3 frequency ranges while taking the frequencies from 3 corpora into account. Ratings were collected with Amazon Mechanical Turk on a scale from 1 to 7.
3. ***GS3***: A cleaned subset of 150 PVs from GS2, after removing the most frequent and infrequent PVs as well as prefix verbs.[10]

We compared the rankings of the system-derived PV–BV cosine scores against the human ratings, using Spearman's rank-order correlation coefficient ρ (Siegel and Castellan 1988).

4.3.5 Results and discussion

Window Model: Figure 2 presents the general results for different window sizes and across the three gold standards. All of the ρ scores correspond to very high levels of statistical significance ($p<0.005$). The results tend to improve slightly with increasing window sizes. For very large windows, especially for sizes 15 and above, the results remain at the same level, except for GS1 which slightly drops. This is not surprising since windows were cut at sentence boundaries which in practice makes the sentence length the upper bound for the window size.

Results for GS1 based on the SdeWaC vs. the DECOW14 corpus are shown in Figure 3. The performance of the two groups of models is largely comparable, and no clear advantage of one over the other is observable. Given that DECOW is considerably larger than SdeWaC, we take this as evidence that window models are relatively robust against data sparseness.

Figure 4 compares models that use raw frequency counts for all context words with using only content words, combined with LMI weighting. Clearly, the latter type of model leads to far better results.

Syntactic Subcategorization Model: As for models that take syntactic transfer strength into account, Figure 5 shows the overall results for subcategorization models with a threshold of $t = 0.3$. The first set of bars represents the best window model as a point of comparison, i.e.,

[10] Some verbs such as *um|fahren* do exists as both PVs and prefix verbs.

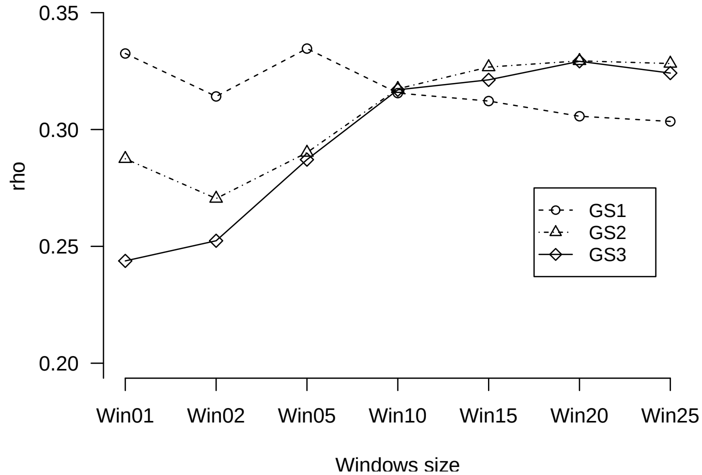


Figure 2:
Results for
erently sized
dow models
ss the three
d standards.
models rely
ntent words
and use LMI
weighting

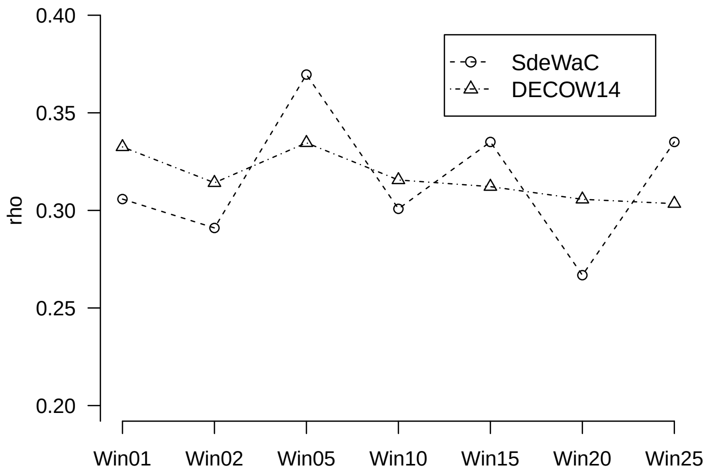
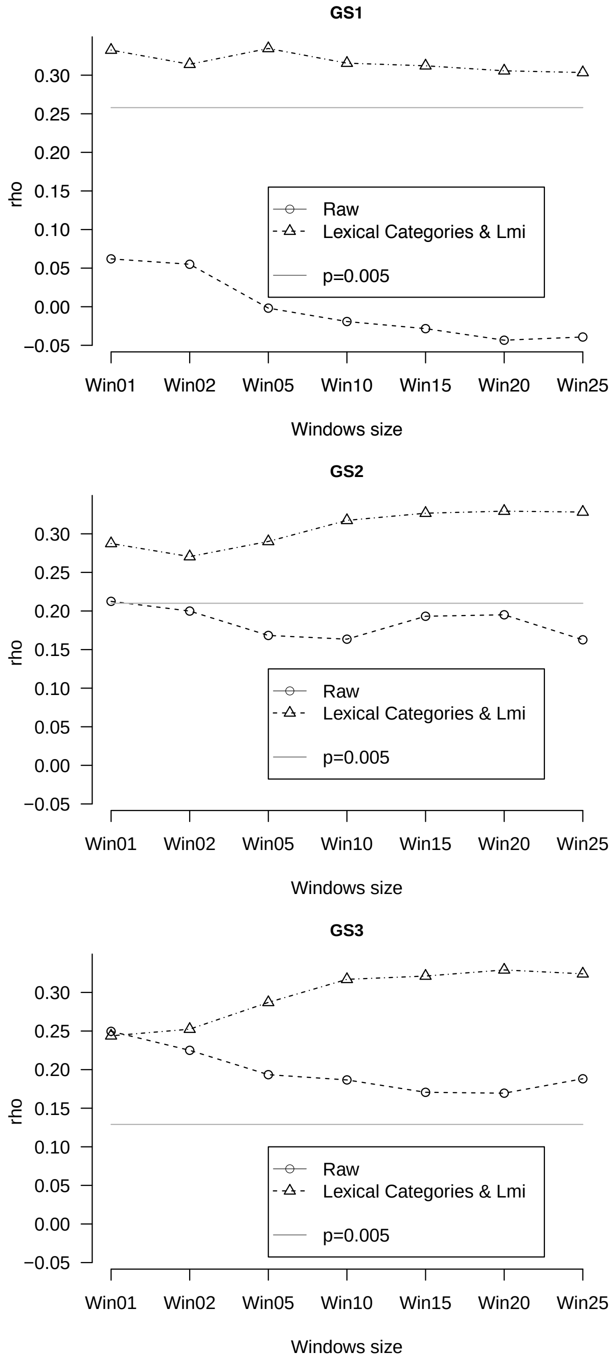


Figure 3:
Results for GS1
with window
models extracted
from two
different
corpora: SdeWaC
and DECOW14.
The models rely
on content words
and use LMI
weighting

using a window of 20 words, reduced to content words, and with LMI weighting. The following groups of bars represent syntactic transfer models with raw frequency counts, LMI weighting, GermaNet generalizations (gn.lv*x*) and SVD (svd_*dim*) dimensionality reductions.

Two observations can be made: firstly, none of the syntactic models reaches the level of performance of the window-based models. Second, the high-dimensional models based on raw frequency counts and LMI perform much worse than the models which apply generalization techniques. So, contrary to the window-based models, applying LMI

Figure 4:
Results for raw frequency models vs. models with content words and LMI weighting



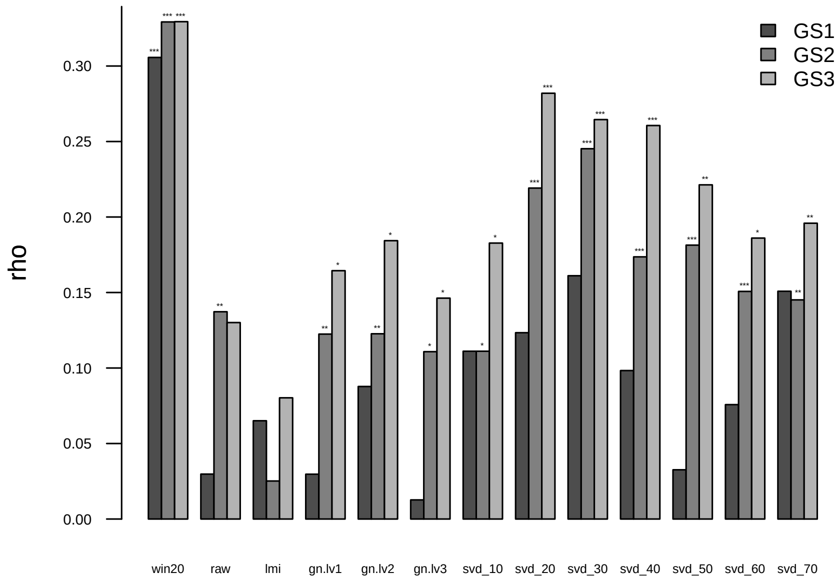


Figure 5: Results across gold standards, for t = 0.3 (*** p<0.001, ** p<0.01, * p<0.05)

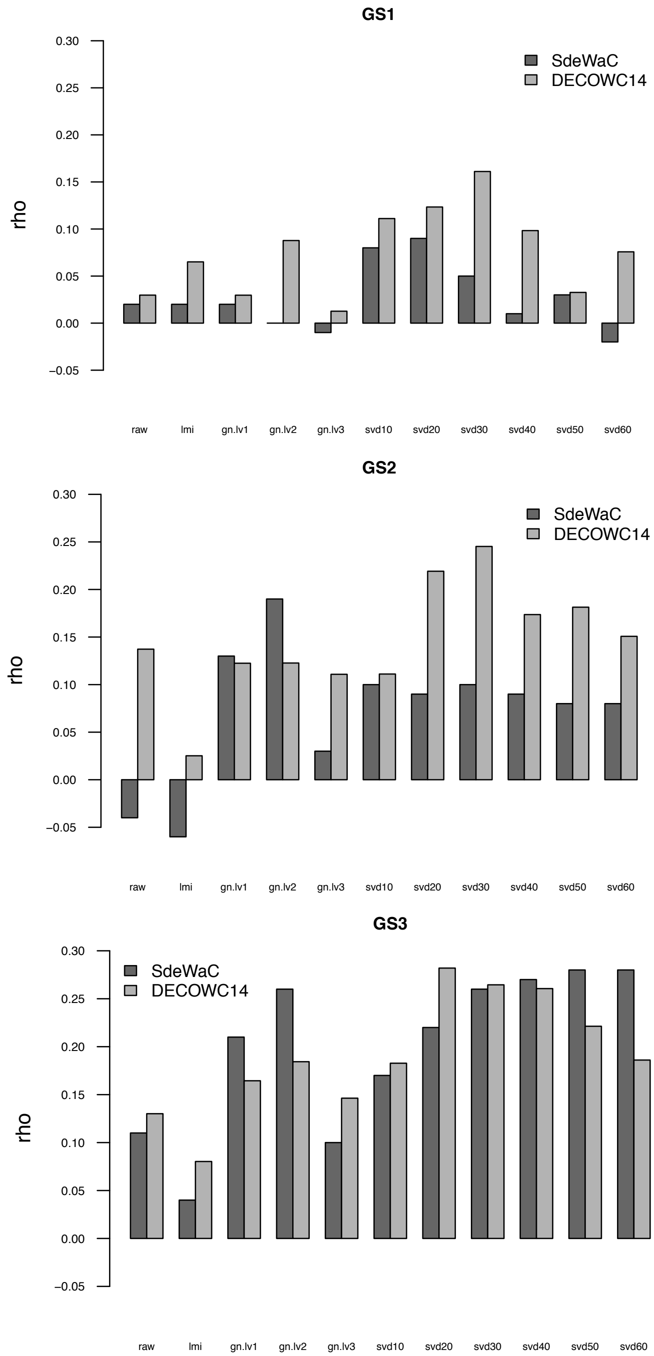
weighting does not improve the predictions. But generalizations boost the quality of the predictions in many conditions.

The fact that the concentration of evidence mass through generalization by GermaNet and SVD greatly benefits the results suggests that the major problem of the syntactic subcategorization approach is data sparseness. The use of GermaNet generalizations already tends to improve the performance, although not consistently. But the use of such taxonomy-based generalizations is clearly limited by the fact that taxonomies notoriously lack coverage and, in the frequent case of semantic ambiguity, are not able to provide reliable estimates on the probabilities of different word readings. A major boost in performance can be observed with the use of SVD, which does not run into coverage problems. The best SVD results are obtained in the range of twenty dimensions (svd_20), which seems to be the best equilibrium between the concentration of evidence mass and over-generalization.

A similar effect can also be observed for GermaNet generalizations: the highest level of distinction in the taxonomy (gn.lv1) is too general to be useful while the third (gn.lv3) is too specific; the second level of the taxonomy (gn.lv2) appears to be the best compromise.

The assumption that data sparseness plays a major role in the performance of the syntactic subcategorization models is also backed

Figure 6:
Results for syntactic models extracted from two different corpora: SdeWaC vs. DECOW14



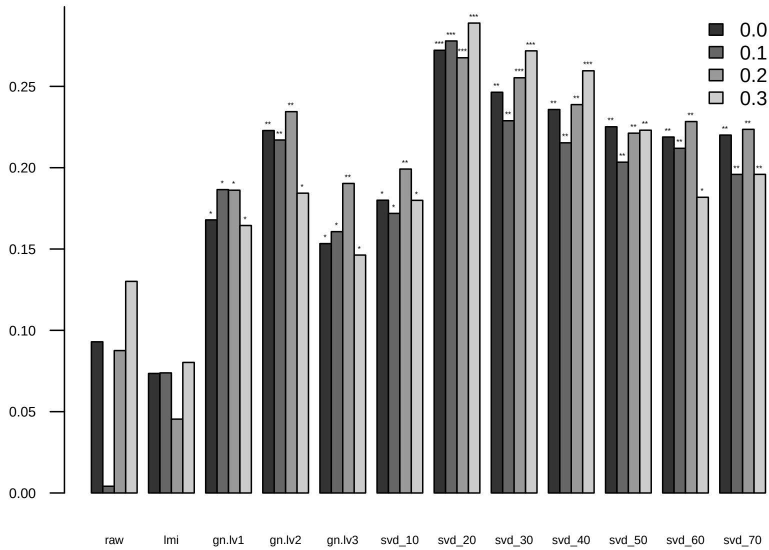
up by a comparison between models extracted from our differently sized corpora, as presented in Figure 6. It is important to keep in mind that the SdeWaC corpus itself is not a small corpus, but the use of the much larger DECOW14 leads to better results in most cases. This stands in sharp contrast to the window-based models which, as we have seen above, apparently do not improve with the larger corpus and do not run into data sparseness problems.

As discussed earlier, we suspected that information stemming from window models provides semantic evidence of a somewhat degraded quality. For this reason, the evidence extracted from syntactic slot fillers should in theory be qualitatively better. But if we assume that information stemming from the argument grid and the heads of syntactic relations is qualitatively more valuable information for our task, we should expect that larger window sizes do not predict compositionality as well as small or medium-sized windows, since small windows tend to contain more concentrated material from arguments than very large windows. What we found in Figure 2, however, is that in general large windows lead to a better performance than small windows. This strongly suggests that words from the general context, which are not necessarily syntactically linked to our target verbs in a direct way, are also very valuable predictors for the semantic similarity between PV–BV pairs and, thus, their level of compositionality. This also means that building our theoretical considerations about the matching of argument slots between PV–BV pairs does not outweight the larger mass of unsorted evidence contained in the window models.

A further problem of our syntax-aware approach is revealed if we look at Figure 7, which compares the prediction results across thresholds t. We can see that a threshold of 0.2 or 0.3 often leads to a slightly better performance than 0.1 or no threshold, but no globally optimal value for t can be established. If the threshold is set too low, many non-correspondences are interpreted as semantic links (false positives). If the threshold is set too high, many semantic links are discarded (false negatives). There seems to be no optimal point of equilibrium between the filtering of false positives and false negatives. A dynamic threshold for individual PV–BV pairs and the average cosine distances of a target slot to all given complementary candidate slots would be beneficial, but at present we see no way to compute this reliably.

Finally, and with respect to the last problem, our syntax-based

Figure 7: Results across thresholds, GS3 (*** p<0.001, ** p<0.01, * p<0.05)



approach somewhat naively neglects the possibility of one-to-many and many-to-one correspondences between subcategorization slots, and always tries to establish a one-to-one link. In reality, however, many subcategorization slots with more than one correct correspondence can be found. For example, the PV–BV pair *leuchten/an|leuchten* as in example (10) happens to be a classification outlier in many of the syntax-based prediction models. The subject slot (SB) of the BV *leuchten* (e.g., *Lampe (lamp)*) is usually matched to a PP subcategorization slot of the PV *an|leuchten* headed by the preposition *mit*, which requires the dative case (e.g., *mit der Lampe (with the lamp)*). Our system computed the following two slots for *leuchten* which receive high cosine values in correspondence to the PP `mit-dat` slot of *an|leuchten*.

```
anleuchten-mit-dat vs leuchten-SB:     0.8931
anleuchten-mit-dat vs leuchten-in-dat: 0.6386
```

One slot is the subject (SB), as expected, and the second is a PP headed by the preposition *in* and the dative case. The latter option represents a linguistically plausible complement of *leuchten* indicating the location where the illumination takes place (e.g., *leuchtet in dem Raum (shines in the room)*), but without semantic correspondence to the target PV slot. A possible remedy for our prediction model could be to include an estimation about how many links have to be established, but this is not a trivial problem in itself and will not be pursued here.

In sum, we provided empirical evidence for hypothesis H3: we found that both window models and syntactic models that are sensitive to subcategorization frame transfers can be used to predict degrees of PV compositionality. Window-based models perform better, even though they are conceptually and computationally simpler. The worse performance of the syntactic models is presumably due to data sparseness and underlying linguistic problems which are difficult to solve computationally.

5 CONCLUSION

At the beginning of this article, we hypothesized that for PVs that are not fully lexicalized there are groups of BVs which undergo the same semantic derivation when they combine with the same particle type, and that the semantic transfer patterns are paralleled by syntactic transfer patterns. We further hypothesized that syntactic transfer between pairs of PVs and BVs, as well as the degree of PV compositionality, can be predicted with distributional methods.

Our first experiment in Section 4.1 addressed the hypothesis that particle meaning and semantic derivation are closely related. We found evidence that there are groups of PVs which share the same semantic transfer patterns and also the same syntactic transfer patterns. This shows that the PVs in the same semantic classes (i) are semantically coherent, (ii) share semantically coherent BVs and the same particle senses, and (iii) undergo parallel shifts regarding syntactic and semantic properties. We thus contributed both to the theoretical understanding and to an empirical verification of German PV composition at the syntax-semantics interface.

Our second experiment in Section 4.2 addressed the empirical prediction of PV–BV syntactic subcategorization transfer, which we argued is necessary to integrate into a prediction of PV compositionality from a theoretical point of view. While modeling slot correspondences in the syntactic transfer was challenging for humans and suffered from severe data sparseness, we verified our distributional approach using hard and soft cluster analyses.

Finally, our third experiment in Section 4.3 integrated the idea of slot correspondence into a syntactic transfer model of PV compositionality, and compared the syntactic model against window models. Although the syntactic transfer approach is much more elaborate and theoretically well-founded, it could not outperform the conceptually simpler window-based approach. We argued that local windows contain information which is useful in the prediction of semantic similarity between PV–BV pairs, and which apparently captures aspects of the verb meanings that the syntactic complements are missing. The window-based approach also proved more robust to data sparseness. Overall, we found that both models can be used to predict degrees of PV compositionality, and the comparison between the two approaches allowed important theoretical insights: many of the misclassifications produced by the syntax-based models could be traced to underlying linguistic problems, the complexity of which makes computational analysis infeasable given the available resources.

ACKNOWLEDGMENTS

The research was supported by the DFG Research Grant SCHU 2580/2 (Stefan Bott) and the DFG Heisenberg Fellowship SCHU-2580/1 (Sabine Schulte im Walde). We thank the anonymous reviewers and our colleagues Jeremy Barnes and Diego Frassinelli for their helpful feedback, and our annotators for the tedious manual annotations of particle verb syntax and semantics.

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, Colin BANNARD, Takaaki TANAKA, and Dominic WIDDOWS (2003), An Empirical Model of Multiword Expression Decomposability, in *Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp. 89–96, Sapporo, Japan.

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

Anders BJÖRKELUND, Özlem ÇETINOĞLU, Richárd FARKAS, Thomas MÜLLER, and Wolfgang SEEKER (2013), (Re)ranking Meets Morphosyntax: State-of-the-art Results from the SPMRL 2013 Shared Task, in *Proceedings of the*

4th Workshop on Statistical Parsing of Morphologically-Rich Languages, pp. 135–145, Seattle, WA, USA.

David BLEI, Andrew NG, and Michael JORDAN (2003), Latent Dirichlet Allocation, *Journal of Machine Learning Research*, 3:993–1022.

Bernd BOHNET (2010), Top Accuracy and Fast Dependency Parsing is not a Contradiction, in *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 89–97, Beijing, China.

Stefan BOTT and Sabine SCHULTE IM WALDE (2014a), Modelling Regular Subcategorization Changes in German Particle Verbs, in *Proceedings of the 1st Workshop on Computational Approaches to Compound Analysis*, pp. 1–10, Dublin, Ireland.

Stefan BOTT and Sabine SCHULTE IM WALDE (2014b), Optimizing a Distributional Semantic Model for the Prediction of German Particle Verb Compositionality, in *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pp. 509–516, Reykjavik, Iceland.

Stefan BOTT and Sabine SCHULTE IM WALDE (2014c), Syntactic Transfer Patterns of German Particle Verbs and their Impact on Lexical Semantics, in *Proceedings of the 3rd Joint Conference on Lexical and Computational Semantics*, pp. 182–192, Dublin, Ireland.

Stefan BOTT and Sabine SCHULTE IM WALDE (2015), Exploiting Fine-grained Syntactic Transfer Features to Predict the Compositionality of German Particle Verbs, in *Proceedings of the 11th Conference on Computational Semantics*, pp. 34–39, London, UK.

Samuel BRODY and Noemie ELHADA (2010), An Unsupervised Aspect-Sentiment Model for Online Reviews, in *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 804–812, Los Angeles, CA, USA.

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.

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*, pp. 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*, pp. 196–201, Doha, Qatar.

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, pp. 45–53, Sydney, Australia.

Nicole DEHÉ, Ray JACKENDOFF, Andrew MCINTYRE, and Silke URBAN (2002), Introduction, in Nicole DEHÉ, Ray JACKENDOFF, Andrew MCINTYRE, and Silke URBAN, editors, *Verb-Particle Explorations*, Interface Explorations, pp. 1–20, Mouton de Gruyter, Berlin, New York.

Georgiana DINU, Nghia THE PHAM, and Marco BARONI (2013), DISSECT – DIStributional SEmantics Composition Toolkit, in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 31–36, Sofia, Bulgaria.

Stefan EVERT (2004), The Statistical Analysis of Morphosyntactic Distributions, in *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pp. 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*, pp. 61–68, Darmstadt, Germany.

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.

Wolfgang FLEISCHER and Irmhild BARZ (2012), *Wortbildung der deutschen Gegenwartssprache*, Walter de Gruyter, 4th edition.

Joseph L. FLEISS (1971), Measuring Nominal Scale Agreement among Many Raters, *Psychological Bulletin*, 76(5):378–382.

Weiwei GUO and Mona DIAB (2011), Semantic Topic Models: Combining Word Distributional Statistics and Dictionary Definitions, in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 552–561, Edinburgh, UK.

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*, pp. 9–15, Madrid, Spain.

Zellig HARRIS (1954), Distributional Structure, *Word*, 10(23):146–162.

Silvana HARTMANN (2008), Einfluss syntaktischer und semantischer Subkategorisierung auf die Kompositionalität von Partikelverben, Studienarbeit, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Boris HASELBACH (2011), Deconstructing the Meaning of the German Temporal Verb Particle *nach* at the Syntax-Semantics Interface, in *Proceedings of Generative Grammar in Geneva*, pp. 71–92, Geneva, Switzerland.

Lawrence HUBERT and Phipps ARABIE (1985), Comparing Partitions, *Journal of Classification*, 2:193–218.

Eric JOANIS, Suzanne STEVENSON, and David JAMES (2008), A General Feature Space for Automatic Verb Classification, *Natural Language Engineering*, 14(3):337–367.

Fritz KLICHE (2011), Semantic Variants of German Particle Verbs with *ab-*, *Leuvense Bijdragen*, 97:3–27.

Maximilian KÖPER and Sabine SCHULTE IM WALDE (2017), Complex Verbs are Different: Exploring the Visual Modality in Multi-Modal Models to Predict Compositionality, in *Proceedings of the 13th Workshop on Multiword Expressions*, pp. 200–206, Valencia, Spain.

Maximilian KÖPER and Sabine SCHULTE IM WALDE (2018), Analogies in Complex Verb Meaning Shifts: The Effect of Affect in Semantic Similarity Models, in *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, LA, USA, to appear.

Anna KORHONEN, Yuval KRYMOLOWSKI, and Zvika MARX (2003), Clustering Polysemic Subcategorization Frame Distributions Semantically, in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 64–71, Sapporo, Japan.

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*, pp. 47–56, Saarbrücken, Germany.

Thomas K. LANDAUER and Susan T. DUMAIS (1997), A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge, *Psychological Review*, 104(2):211–240.

J. Richard LANDIS and Gary G. KOCH (1977), The Measurement of Observer Agreement for Categorical Data, *Biometrics*, 33:159–174.

Andrea LECHLER and Antje ROßDEUTSCHER (2009), German Particle Verbs with *auf*. Reconstructing their Composition in a DRT-based Framework, *Linguistische Berichte*, 220:439–478.

Beth LEVIN (1993), *English Verb Classes and Alternations*, The University of Chicago Press.

Anke LÜDELING (2001), *On German Particle Verbs and Similar Constructions in German*, Dissertations in Linguistics, CSLI Publications, Stanford, CA.

Christopher D. MANNING, Prabhakar RAGHAVAN, and Hinrich SCHÜTZE (2008), *Introduction to Information Retrieval*, Cambridge University Press.

Diana McCarthy, Bill Keller, and John Carroll (2003), Detecting a Continuum of Compositionality in Phrasal Verbs, in *Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp. 73–80, Sapporo, Japan.

Andrew McIntyre (2007), Particle Verbs and Argument Structure, *Language and Linguistics Compass*, 1(4):350–397.

Paola Merlo and Suzanne Stevenson (2001), Automatic Verb Classification Based on Statistical Distributions of Argument Structure, *Computational Linguistics*, 27(3):373–408.

Stefan Müller (2002), Syntax or Morphology: German Particle Verbs Revisited, in Nicole Dehé, Ray Jackendoff, Andrew McIntyre, and Silke Urban, editors, *Verb-Particle Explorations*, Interface Explorations, pp. 119–140, Mouton de Gruyter, Berlin, New York.

Stefan Müller (2003), Solving the Bracketing Paradox: An Analysis of the Morphology of German Particle Verbs, *Journal of Linguistics*, 39(2):275–325.

Thomas Müller, Helmut Schmid, and Hinrich Schütze (2013), Efficient Higher-Order CRFs for Morphological Tagging, in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 322–332, Seattle, WA, USA.

Diarmuid Ó Séaghdha (2010), Latent Variable Models of Selectional Preference, in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 435–444, Uppsala, Sweden.

Susan Olsen (1997), Prädikative Argumente syntaktischer und lexikalischer Köpfe—Zum Status der Partikelverben im Deutschen und Englischen, *Folia Linguistica*, 31(3–4):301–330.

Sebastian Padó and Mirella Lapata (2007), Dependency-based Construction of Semantic Space Models, *Computational Linguistics*, 33(2):161–199.

William M. Rand (1971), Objective Criteria for the Evaluation of Clustering Methods, *Journal of the American Statistical Association*, 66(336):846–850.

Mats Rooth (1998), Two-Dimensional Clusters in Grammatical Relations, in *Inducing Lexicons with the EM Algorithm*, AIMS Report 4(3), Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil (1999), Inducing a Semantically Annotated Lexicon via EM-Based Clustering, in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 104–111, Maryland, MD.

Stefan Rüd (2012), *Untersuchung der distributionellen Eigenschaften der Lesarten der Partikel 'auf' mittels Clustering-Methoden*, Master's thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Magnus SAHLGREN (2005), An Introduction to Random Indexing, in *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering*, volume 5.

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.

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*, pp. 486–493, ELRA, Istanbul, Turkey.

Silke SCHEIBLE, Sabine SCHULTE IM WALDE, Marion WELLER, and Max KISSELEW (2013), A Compact but Linguistically Detailed Database for German Verb Subcategorisation Relying on Dependency Parses from a Web Corpus: Tool, Guidelines and Resource, in *Proceedings of the 8th Web as Corpus Workshop*, pp. 63–72, Lancaster, UK.

Helmut SCHMID, Arne FITSCHEN, and Ulrich HEID (2004), SMOR: A German Computational Morphology Covering Derivation, Composition and Inflection, in *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pp. 1263–1266.

Sabine SCHULTE IM WALDE (2000), Clustering Verbs Semantically According to their Alternation Behaviour, in *Proceedings of the 18th International Conference on Computational Linguistics*, pp. 747–753, Saarbrücken, Germany.

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*, pp. 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*, pp. 608–614, Borovets, Bulgaria.

Sabine SCHULTE IM WALDE (2006), Experiments on the Automatic Induction of German Semantic Verb Classes, *Computational Linguistics*, 32(2):159–194.

Sidney SIEGEL and N. John CASTELLAN (1988), *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill, Boston, MA.

Sylvia SPRINGORUM (2011), DRT-based Analysis of the German Verb Particle *an*, *Leuvense Bijdragen*, 97:80–105.

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*, pp. 73–80, Istanbul, Turkey.

Sylvia SPRINGORUM, Sabine SCHULTE IM WALDE, and Antje ROßDEUTSCHER (2013a), Sentence Generation and Compositionality of Systematic Neologisms of German Particle Verbs, Talk at the 5th Conference on Quantitative Investigations in Theoretical Linguistics.

Sylvia SPRINGORUM, Jason UTT, and Sabine SCHULTE IM WALDE (2013b), Regular Meaning Shifts in German Particle Verbs: A Case Study, in *Proceedings of the 10th International Conference on Computational Semantics*, pp. 228–239, Potsdam, Germany.

Barbara STIEBELS (1996), *Lexikalische Argumente und Adjunkte. Zum semantischen Beitrag von verbalen Präfixen und Partikeln*, Akademie Verlag, Berlin.

Barbara STIEBELS and Dieter WUNDERLICH (1994), Morphology Feeds Syntax: The Case of Particle Verbs, *Linguistics*, 32:913–968.

Peter D. TURNEY (2012), Domain and Functions: A Dual-Space Model of Semantic Relations and Compositions, *Journal of Artificial Intelligence Research*, 44:533–585.

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

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*, pp. 81–90, Dublin, Ireland.

Ian H. WITTEN and Eibe FRANK (2005), *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann.

Stefanie WULFF (2010), *Rethinking Idiomaticity: A Usage-Based Approach*, A&C Black.