Factoring Ambiguity out of the Prediction of Compositionality for German Multi-Word Expressions

Stefan Bott and **Sabine Schulte im Walde**
Institut für Maschinelle Sprachverabeitung
Universität Stuttgart
Pfaffenwaldring 5b, 70569 Stuttgart, Germany
{stefan.bott,schulte}@ims.uni-stuttgart.de

Abstract

Ambiguity represents an obstacle for distributional semantic models (DSMs), which typically subsume the contexts of all word senses within one vector. While individual vector space approaches have been concerned with sense discrimination (e.g., Schütze (1998), Erk (2009), Erk and Pado (2010)), such discrimination has rarely been integrated into DSMs across semantic tasks. This paper presents a soft-clustering approach to sense discrimination that filters sense-irrelevant features when predicting the degrees of compositionality for German noun-noun compounds and German particle verbs.

1 Introduction

Addressing the compositionality of complex words is a crucial ingredient for lexicography and NLP applications, to know whether the expression should be treated as a whole, or through its constituents, and what the expression means. For example, studies such as Cholakov and Kordoni (2014), Weller et al. (2014), Cap et al. (2015), and Salehi et al. (2015b) have integrated the prediction of multi-word compositionality into statistical machine translation.

We are interested in predicting the degrees of compositionality of two types of German multi-word expressions: (i) German noun-noun compounds (NCs) represent nominal multi-word expressions (MWEs), e.g., *Feuer|werk* 'fire works' is composed of the constituents *Feuer* 'fire' and *Werk* 'opus'. (ii) German particle verbs (PVs) are complex verbs such as *an|strahlen* ('beam/smile at') which are composed of a separable prefix particle (*an*) and a base verb (*strahlen* 'beam'/'smile'). Both types of German MWEs are highly frequent and highly productive in the lexicon. Table 1 presents some example MWEs and their constituents with human ratings on compositionality.[1]

Automatic approaches to predict compositionality degrees typically exploit distributional semantic models (DSMs), i.e. vector representations relying on the *distributional hypothesis* (Harris, 1954; Firth, 1957), that words with similar distributions have related meanings. Regarding the compositionality prediction, DSMs represent the meanings of the MWEs and their constituents by distributional vectors, and the similarity of a compound–constituent vector pair is taken as the predicted degree of compound-constituent compositionality. Existing approaches addressed the compositionality of NCs (Reddy et al., 2011; Salehi and Cook, 2013; Schulte im Walde et al., 2013; Salehi et al., 2014) and complex verbs (Baldwin, 2005; Bannard, 2005; Bott and Schulte im Walde, 2015), mainly dfor English and for German.

A major obstacle for DSMs is their conflation of contexts across individual word senses. DSMs typically subsume evidence of cooccurring items within one vector for the target word type, rather than discriminating contextual evidence for the specific target word senses. Taking the German noun-noun compound *Blatt|salat* 'leaf salad' as an example, its modifier constituent *Blatt* has at least four senses: 'leaf', 'sheet of paper', 'newspaper' and 'hand of cards'. If we had individual sense vectors for each sense of *Blatt*, a DSM might successfully predict a strong compositionality for the compound *Blatt|salat* regarding this constituent, when comparing the compound vector with the 'leaf' sense vector, because the vectors agree on

[1] The scales for mean ratings were 1–7 for noun-noun compounds, and 1–6 for particle verbs. Examples were taken from the two gold standards described in section 2.

Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), pages 66–72,
Valencia, Spain, April 4.

Multi-Word Expressions				Mean Ratings	
				Modifier	Head
Ahorn|blatt	'maple leaf'	maple	leaf	5.64	5.71
Blatt|salat	'green salad'	leaf	salad	3.56	5.68
See|zunge	'sole'	sea	tongue	3.57	3.27
Löwen|zahn	'dandelion'	lion	tooth	2.10	2.23
Fliegen|pilz	'toadstool'	fly/bow tie	mushroom	1.93	6.55
Fleisch|wolf	'meat chopper'	meat	wolf	6.00	1.90
an|leuchten	'illuminate'	an$_{PRT}$	illuminate	–	5.95
auf|horchen	'listen attentively'	auf$_{PRT}$	listen	–	4.55
aus|reizen	'exhaust'	aus$_{PRT}$	provoke	–	3.62
ein|fallen	'remember/invade'	ein$_{PRT}$	fall	–	2.54
an|stiften	'instigate'	an$_{PRT}$	create	–	1.80

Table 1: Examples of German noun-noun compounds and German particle verbs, accompanied by translations and human mean ratings on the degrees of compound-constituent compositionality.

salient features such as *green* and *fresh*. But traditionally, the constituent vector contains distributional information across all *Blatt* senses, and the similarity between the conflated word type vector and the compound vector is most probably determined by the predominant sense of the word type (which does not necessarily coincide with the relevant sense).

While individual vector space approaches have been concerned with sense discrimination (e.g., Schütze (1998), Erk (2009), Erk and Pado (2010)), the approaches have rarely been integrated into DSMs across semantic tasks. Alternatively, sense disambiguation/discrimination approaches have been developed for *SemEval* tasks on Word Sense Disambiguation/Discrimination and (Cross-lingual) Lexical Substitution (McCarthy and Navigli, 2007; Mihalcea et al., 2010; Jurgens and Klapaftis, 2013). As to our knowledge, few systems have attempted to distinguish between word senses and then address various semantic relatedness tasks, such as Li and Jurafsky (2015) and Iacobacci et al. (2015). Computational compositionality assessment has been studied for NCs (Reddy et al., 2011; Schulte im Walde et al., 2013; Salehi and Cook, 2013; Schulte im Walde et al., 2016a) and PVs (McCarthy et al., 2003; Baldwin et al., 2003; Bannard, 2005; Kühner and Schulte im Walde, 2010). Most similar to our current work is Salehi et al. (2015a), who addressed the problem of semantic ambiguity in MWEs by using a multi-sense skip gram model with two to five embeddings per word. They expected multiple embeddings to capture different word senses. They could, however, not find an improvement over the use of single-word embeddings.

In this paper, we suggest soft clustering as an approximation to separate the different senses of a word type. We expect that the assignments of compound and constituent words to clusters reflect the differences between word senses, and that the underlying features refer to the features of the respective word sense. We assume further that if we find a pair $<\mu, \kappa>$ of an MWE μ and one of its constituents κ with high distributional similarity in the same cluster, this indicates closeness in meaning and therefore strong compositionality. We exploit the soft clusters by (a) identifying the relevant senses of the MWE and constituents based on overlap in cluster assignment, and by (b) comparing reduced vectors of MWEs and constituents when taking into account only a subset of cluster-based salient sense features, in order to optimize the prediction of compositionality.

2 Experiment Setup

Distributional Semantics Models Our DSM is a word space model that uses lemmatized words as dimensions in the high-dimensional vectors space (Sahlgren, 2006; Turney and Pantel, 2010). The associative strength between target and context words is measured as Local Mutual Information (LMI) (Evert, 2004), based on context word frequency. The context of the targets is defined as a window of n words to the left and the right of the target. We use the cosine value of the angle between two vectors as a measure for semantic similarity and compositionality. For technical reasons we ignore context words with a count of 5 or less or an LMI value below 0.

We use the word vectors in three ways here: (a) we use them directly as *window models* in order to measure the distance between vector pairs for an MWE and each of its components (e.g. *Blatt|salat*

vs. *Blatt*). We also use them (b) as an input matrix for soft clustering and (c) we build word vector models for each cluster.

LSC for Soft Clustering We use Latent Semantic Classes (LSC) as a soft clustering algorithm (Rooth, 1998; Rooth et al., 1999). LSC is a two-dimensional soft-clustering algorithm which learns three probability distributions: (a) across the clusters, (b) for the output probabilities of each element within a cluster and (c) for each feature type with regard to a cluster. The access to all three probability distributions is crucial for our approach, since it allows to determine which features are salient for individual clusters.

The Pipeline We create two types of models: The *window models* are simple word space models which use LMI values based on counts of context words. The *clustering models* apply soft clustering as a previous step to the determination of distributional similarity. For their construction, we use the window-based models as an input to the LSC algorithm. The clusters produced by LSC are used to create individual models for each cluster C in a way that each of these cluster-specific models only contain vectors for the target words which are contained in C and represent only those features as dimensions which are predicted to be salient features for C. The models vary with respect to the number of clusters created.

With this, we expect that in our example of *Blatt|salat* some clusters will capture the *leaf*-sense and others the *sheet*- or other senses. The comparison between the vectors for *Blatt* and *Blatt|salat* is then done separately for each cluster, where the context dimensions of the vectors are reduced to only those context words which are also salient features of each cluster. We expect that the pair of our example only occur in clusters which can be attributed to the *leaf*-sense.

Comparison across Clusters In cases like the NC *Blatt|salat* it appears that the word sense which should be considered for compositionality assessment is the one which is distributionally closest to the target MWE. But this is not necessarily the case for all MWEs. The PV *zu|schlagen* is one example: it can mean both *to hit hard and quickly* or *to take advantage of a good offer/bargain*; in this case the MWE itself is ambiguous. The base verb *schlagen* means *to hit*, so one sense of the PV is highly compositional and the other sense is less so; nevertheless none of the senses is predominant. We use three methods to compare the distributional similarity across clusters: *highest*, *lowest* and *average*. In the first two methods (*highest/lowest*) we select the cluster with the highest/lowest distributional similarity between μ and κ and use its similarity value. In the last method (*average*) the average similarity is computed among those clusters which contain both the MWE μ and the target component κ, while clusters which do not contain the pair <μ, κ> are ignored.

Thresholds The fact that LSC outputs probabilities for both targets and features allows to set two different thresholds on these probabilities. The threshold on the target output probability (t-threshold) controls the number of clusters to which a target element will be assigned. The lower the threshold is set, the more elements each cluster will contain. Lower threshold values also lead to higher average numbers of clusters to which each element is assigned. The t-threshold influences the predictions of our models in that low values also increase the likelihood for each Cluster C and for each pair <α, β> of a MWE and a constituent word that both α and β are included in C. The threshold on the feature output probability (f-threshold) allows to filter the vectors for both elements of <α, β> according to each cluster C so that only the dimensions representing the salient features for C are included in the vectors.

Corpus For the extraction of features we use the SdeWaC (v.3, 880 million words) corpus (Faaß and Eckart, 2013), in a tokenized (Schmid, 2000), POS-tagged and lemmatized (Schmid, 1994) version.

Gold Standards For NCs and PVs we use the following gold standards:

- GS-NN: 868 German NCs (Schulte im Walde et al., 2016b) randomly selected from different frequency ranges, different ambiguity levels of the heads and different levels of modifier and head productivity. NCs were annotated by eight native speakers on a scale from 1 to 6 for compositionality with respect to both head and modifier constituents.
- GS-PV: 354 PVs, for 11 verb particles. PVs were randomly selected, balanced over 3 frequency bands. The PVs were automatically

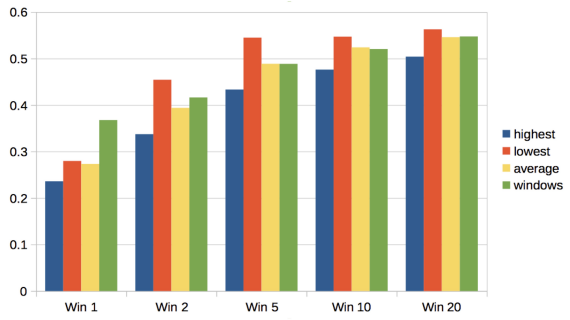

Figure 1: Results (in ρ values) for different window sizes for the NC-head gold standard

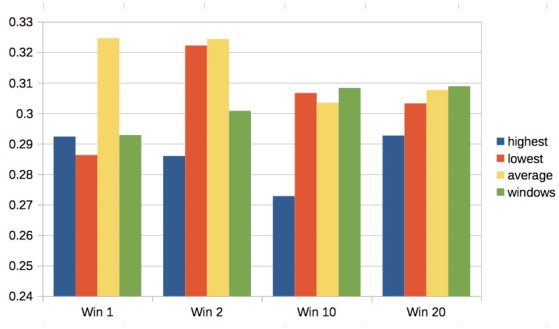

Figure 2: Results (in ρ values) for different window sizes for the PV gold standard

harvested from various corpora, assigned to 3 different frequency ranges per particle and then automatically selected. Some manual revision was done to filter out non-existing PVs resulting from lemmatization errors. Ratings were obtained with Amazon Mechanical Turk.[2]

Feature Sets We were interested in which parts of speech provide the best predictive features for compositionality. We use only content-word categories: adjectives, nouns and verbs. We use four different combinations: all content words and categories in isolation.

Measures Distributional similarity is measured with the cosine between vectors. The cosine similarity values are used to rank the compared pairs from lowest to highest. For the evaluation, system rankings and human judgment rankings of MWEs are compared to each other with Spearman's rank order correlation ρ (Siegel and Castellan, 1988). Spearman's ρ is a non-parametric measure which assesses monotonic relationships of ranks that range between -1 (inverse correlation) and 1 (perfect correlation); a ρ value of 0 indicates a lack of correlation. Significance is determined with the use of the Fisher transformation.

Soft clustering does not guarantee that each of the pairs of NCs and a constituent word is placed together in at least one of the clusters. This may potentially lead to problems of coverage. In practice, however, we experience coverage problems only for very restrictive threshold settings.

[2]This gold standard is a preliminary, but not identical, version of the one presented in Bott et al. (2016). It was also used in Bott and Schulte im Walde (2014).

3 Results and Discussion

Figure 1 and 2 show the results for different window sizes for NCs and PVs. The two figures have different scales and higher ρ scores are obtained for NCs. The values are compared to the results of the window-based models. The predictions of compositionality levels become more accurate with increasing window sizes. For NC compositionality apparently more general information about the larger context plays an important role. Interestingly, no negative effect from larger contexts can be observed, even if smaller contexts tend to concentrate on closely related words such as complements, modifiers and the complementary parts of collocations in which the target word takes part. All ρ values above 0.108 are statistically highly significant ($p<0.001$ for n=868), which applies to nearly all of the observed values.

Regarding PV compositionality, window models increase their performance with larger context sizes, but this is not true for clustering models. The latter tend to perform better with small to medium window sizes and in this range clustering models clearly outperform window models. Also NC compositionality tends to be better predicted with the clustering-based models, but to a degree. It is also interesting to note that the successful combination cluster methods are different for NC (where *highest* performs best) and PVs (for which the *average* method yields the best results). This suggests a more fundamental difference in the two types of MWEs. One of the possible differences lies in the average degree of ambiguity of the MWEs and their constituents. NCs have a strong tendency to be less ambiguous than their constituent nouns. PVs, on the other hand, are often highly ambiguous themselves.

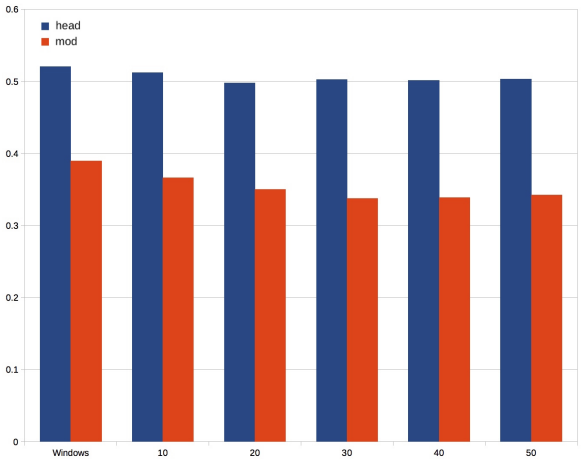

Figure 3: Results for different numbers of clusters for the NC gold standard (heads vs. modifiers)

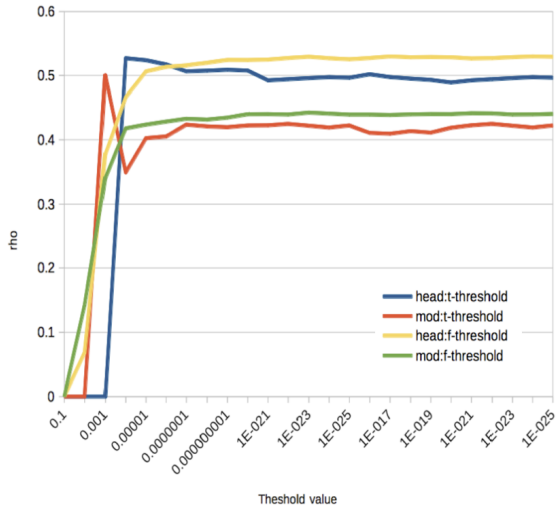

Figure 4: ρ values for variations over thresholds (NC gold standard)

Figure 3 shows the effect of the number of clusters which are used in the clustering stage. The graphic shows that the number of clusters has not a strong influence on performance, but slightly better results can be observed with smaller numbers of clusters. This might be due to the fact that larger numbers of clusters split up the feature space into smaller segments and the feature vectors tend to suffer from sparseness. Figure 3 also shows that the predictions for the noun compound compositionality with respect to the heads are generally better than with respect to the modifiers. This is probably a consequence of the fact that meaning of NCs is in most cases more strongly determined by the meanings of their heads than their modifiers. This might explain the observed asymmetry. This finding is in line with earlier studies (Hätty, 2016; Schulte im Walde et al., 2016a) which investigated the asymmetry between the properties of heads and modifiers in noun-noun compounds. They showed that head constituent properties, such as their ambiguity or frequency, influence the predictability of NC compositionality to a much larger degree than modifier constituent properties.

As for feature selection, we found that adjectives represented the least reliable predictive features for compositionality assessment, while nouns were the most reliable ones. The use of the latter even leads to a slightly better performance than the use of the full feature set that contains all content word categories.

Figure 4 shows the influence of the target and the feature thresholds on compositionality prediction. As expected, very high threshold values lead to poor performance since they cause very sparse vector representations. Lowering the threshold the performance curve raises steeply and reaches a stable plateau which is observable in this figure.

4 Conclusions

We started this paper with a theoretical justification to factor out the influence of ambiguity from the prediction of compositionality across multi-word expressions. We applied soft clustering to extract word-sense vectors from word-type vectors, in order to strengthen salient sense features and improve the prediction of compound–constituent compositionality. Both NCs and PVs benefit from the use of clustering in distributional modeling, but in different ways. First, PVs benefit much more than NCs. Second, the optimal type of the combination method which calculates a global similarity score per compound–constituent pair from the cluster-specific DSMs differs between the two types of MWEs. This suggests an underlying difference between them.

In future work we will explore alternative ways to treat the ambiguity of constituent words more adequately. We further plan to examine why different types of MWEs tend to benefit from the clustering approach but with different cluster combination methods. We will also extend our investigation to other semantic relatedness tasks, such as the distinction between semantic relations, which potentially suffer from the same ambiguity issue.

References

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions*, pages 89–96, Sapporo, Japan. Association for Computational Linguistics.

Timothy Baldwin. 2005. Deep lexical acquisition of verb–particle constructions. *Computer Speech and Language*, 19:398–414.

Collin Bannard. 2005. Learning about the meaning of verb–particle constructions from corpora. *Computer Speech and Language*, 19:467–478.

Stefan Bott and Sabine Schulte im Walde. 2014. Syntactic transfer patterns of German particle verbs and their impact on lexical semantics. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 182–192, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

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 International Conference on Computational Semantics*, pages 34–39, London, UK. Association for Computational Linguistics.

Stefan Bott, Nana Khvtisavrishvili, Max Kisselew, and Sabine Schulte im Walde. 2016. $G_h$ost-pv: A representative gold standard of German particle verbs. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)*, pages 125–133, Osaka, Japan. The COLING 2016 Organizing Committee.

Fabienne Cap, Manju Nirmal, Marion Weller, and Sabine Schulte im Walde. 2015. How to account for idiomatic German support verb constructions in statistical machine translation. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 19–28, Denver, Colorado. Association for Computational Linguistics.

Kostadin Cholakov and Valia Kordoni. 2014. Better statistical machine translation through linguistic treatment of phrasal verbs. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 196–201, Doha, Qatar. Association for Computational Linguistics.

Katrin Erk and Sebastian Pado. 2010. Exemplar-based models for word meaning in context. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 92–97, Uppsala, Sweden. Association for Computational Linguistics.

Katrin Erk. 2009. Representing words in regions in vector space. In *Proceedings of the 13th Conference on Computational Natural Language Learning*, pages 57–65, Boulder, CO.

Stefan Evert. 2004. The statistical analysis of morphosyntactic distributions. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004)*, pages 1539–1542, Lisbon, Portugal. European Language Resources Association (ELRA). ACL Anthology Identifier: L04-1509.

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*, pages 61–68, Darmstadt, Germany.

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

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

Anna Hätty. 2016. Vector space models of compositionality for German and English noun-noun compounds. Master Thesis. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. SensEmbed: Learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 95–105, Beijing, China. Association for Computational Linguistics.

David Jurgens and Ioannis Klapaftis. 2013. SemEval-2013 Task 13: Word sense induction for graded and non-graded senses. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 290–299, Atlanta, Georgia, USA. Association for Computational Linguistics.

Natalie Kühner and Sabine Schulte im Walde. 2010. Determining the degree of compositionality of German particle verbs by clustering approaches. In *Proceedings of the 10th Conference on Natural Language Processing*, pages 47–56, Saarbrücken, Germany.

Jiwei Li and Dan Jurafsky. 2015. Do multi-sense embeddings improve natural language understanding? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1722–1732, Lisbon, Portugal. Association for Computational Linguistics.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 Task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic. Association for Computational Linguistics.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan.

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. 2010. SemEval-2010 Task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 9–14, Uppsala, Sweden. Association for Computational Linguistics.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

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*, pages 104–111, College Park, Maryland, USA. Association for Computational Linguistics.

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.

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.

Bahar Salehi and Paul Cook. 2013. Predicting the compositionality of multiword expressions using translations in multiple languages. In *Second Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 266–275, Atlanta, Georgia, USA. Association for Computational Linguistics.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014. Using distributional similarity of multi-way translations to predict multiword expression compositionality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 472–481, Gothenburg, Sweden. Association for Computational Linguistics.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015a. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 977–983, Denver, Colorado. Association for Computational Linguistics.

Bahar Salehi, Nitika Mathur, Paul Cook, and Timothy Baldwin. 2015b. The impact of multiword expression compositionality on machine translation evaluation. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 54–59, Denver, Colorado. Association for Computational Linguistics.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49. Manchester, UK.

Helmut Schmid. 2000. Unsupervised learning of period disambiguation for tokenisation. Technical report, Universität Stuttgart.

Sabine Schulte im Walde, Stefan Müller, and Stefan Roller. 2013. Exploring vector space models to predict the compositionality of German noun-noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 255–265, Atlanta, Georgia, USA. Association for Computational Linguistics.

Sabine Schulte im Walde, Anna Hätty, and Stefan Bott. 2016a. The role of modifier and head properties in predicting the compositionality of English and German noun-noun compounds: A vector-space perspective. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 148–158, Berlin, Germany. Association for Computational Linguistics.

Sabine Schulte im Walde, Anna Hätty, Stefan Bott, and Nana Khvtisavrishvili. 2016b. G$_h$ost-NN: A representative gold standard of German noun-noun compounds. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2285–2292, Portoroz, Slovenia. European Language Resources Association (ELRA).

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123. Special Issue on Word Sense Disambiguation.

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

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

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 First Workshop on Computational Approaches to Compound Analysis (ComAComA 2014)*, pages 81–90, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.