Complex Verbs are Different: Exploring the Visual Modality in Multi-Modal Models to Predict Compositionality

Maximilian Köper and **Sabine Schulte im Walde**
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart, Germany
{maximilian.koeper,schulte}@ims.uni-stuttgart.de

Abstract

This paper compares a neural network DSM relying on textual co-occurrences with a multi-modal model integrating visual information. We focus on nominal vs. verbal compounds, and zoom into lexical, empirical and perceptual target properties to explore the contribution of the visual modality. Our experiments show that (i) visual features contribute differently for verbs than for nouns, and (ii) images complement textual information, if (a) the textual modality by itself is poor and appropriate image subsets are used, or (b) the textual modality by itself is rich and large (potentially noisy) images are added.

1 Introduction

Distributional semantic models (DSMs) rely on the *distributional hypothesis* (Harris, 1954), that words with similar distributions have related meanings. They represent a well-established tool for modelling semantic relatedness between words and phrases (Bullinaria and Levy, 2007; Turney and Pantel, 2010). In the last decade, standard DSMs using bag-of-words or syntactic co-occurrence counts have been enhanced by integration into neural networks (Baroni et al., 2014; Levy et al., 2015; Nguyen et al., 2016), or by integrating perceptual information (Silberer and Lapata, 2014; Bruni et al., 2014; Kiela et al., 2014; Lazaridou et al., 2015). While standard DSMs have been applied to a variety of semantic relatedness tasks such as word sense discrimination, selectional preferences, relation distinction (among others), multi-modal models have predominantly been evaluated on their general ability to model semantic similarity as captured by *SimLex* (Hill et al., 2015), *WordSim* (Finkelstein et al., 2002), etc.

In this paper, we compare a neural network DSM relying on textual co-occurrences with a multi-modal model extension integrating visual information. We focus on the prediction of compositionality for two types of German multi-word expressions: noun-noun compounds and particle verbs. Differently to most previous multi-modal approaches, we thus address a semantically specific task that was traditionally addressed by standard DSMs, mainly for English and German (Baldwin, 2005; Bannard, 2005; Reddy et al., 2011; Salehi and Cook, 2013; Schulte im Walde et al., 2013; Salehi et al., 2014; Bott and Schulte im Walde, 2014; Bott and Schulte im Walde, 2015; Schulte im Walde et al., 2016a). Furthermore, we zoom into factors that might influence the quality of predictions, such as lexical and empirical target properties (e.g., ambiguity, frequency, compositionality); and filters to optimise the visual space, such as dispersion and imageability filters (Kiela et al., 2014), and a novel clustering filter.

Our experiments demonstrate that the contributions of the textual and the visual models differ for predictions across the nominal vs. verbal compositions. The visual modality adds complementary features in cases where (a) the textual modality performs poorly, and images of the most imaginable targets are added, or (b) the textual modality performs well, and all available –potentially noisy– images are added. In addition, we demonstrate that perceptual features of verbs, such as abstractness and imageability, have a different influence on multi-modality than for nouns, presumably because they are more difficult to grasp.

2 Data

Target Multi-Word Expressions (MWEs) German noun-noun compounds represent two-part multi-word expressions where both con-

Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), pages 200–206,
Valencia, Spain, April 4.

(a) Complete set of images.

(b) Images in largest cluster.

Figure 1: Clustering filter for *abzupfen* 'to pick'.

stituents are nouns, e.g., *Feuerwerk* 'fire works' is composed of the nominal constituents *Feuer* 'fire' and *Werk* 'opus'. German particle verbs are complex verbs such as *anstrahlen* 'beam/smile at' which are composed of a separable prefix particle (such as *an*) and a base verb (such as *strahlen* 'beam/smile'). Both types of German MWEs are highly frequent and highly productive in the lexicon. In addition, the particles are notoriously ambiguous, e.g., *an* has a partitive meaning in *anbeißen* 'take a bite', a cumulative meaning in *anhäufen* 'pile up', and a topological meaning in *anbinden* 'tie to' (Springorum, 2011). We rely on two existing gold standards annotated with compositionality ratings: GS-NN, a set of 868 German noun-noun compounds (Schulte im Walde et al., 2016b), and GS-PV, a set of 400 particle verbs across 11 particle types (Bott et al., 2016).

Multi-Modal Vector Space Models For the textual representation we used two sets of embeddings. Based on *word2vec* (Mikolov et al., 2013), we obtained both representations using the skip-gram architecture with negative sampling. The sets differ with respect to window size (5 vs. 10) and dimensionality (400 vs. 500). As corpus resource we relied on the lemmatized version of the *DECOW14AX*, a German web corpus containing 12 billion tokens (Schäfer and Bildhauer, 2012).

The visual features rely on images downloaded from the *bing* search engine, following Kiela et al. (2016). We queried 25 images per word, and converted all images into high-dimensional numerical representations by using the caffe toolkit (Jia et al., 2014) and pre-trained models. In the default setting, a word is represented in the visual space by the mean vector of its 25 image representations. As image-recognition neural network models, we used: (i) GoogLeNet (Szegedy et al., 2015), a 22-layer deep network; we obtained vectors by using the value of the last layer before the final softmax, containing 1024 elements (= dimensionality). (ii) AlexNet (Krizhevsky et al., 2012), a neural network with five convolutional layers (4,096-dim).

The multi-modal representations were combined by applying mid-fusion between textual and visual representation, i.e., concatenation of the L2-normalized representations (Bruni et al., 2014)[1]

3 Experiments

Predicting Compositionality For the prediction of compositionality, we represented the meanings of the multi-word expressions and their constituent words by textual, visual and textual+visual (i.e., multi-modal) vectors. The similarity of a compound–constituent vector pair as measured by the *cosine* was taken as the predicted degree of compound–constituent compositionality, and the overall ranking of pair similarities was compared to the gold standard compositionality ratings using Spearman's Rank-Order Correlation Coefficient ρ (Siegel and Castellan, 1988).

[1]Experiments with other fusion techniques showed that mid-fusion performs best.

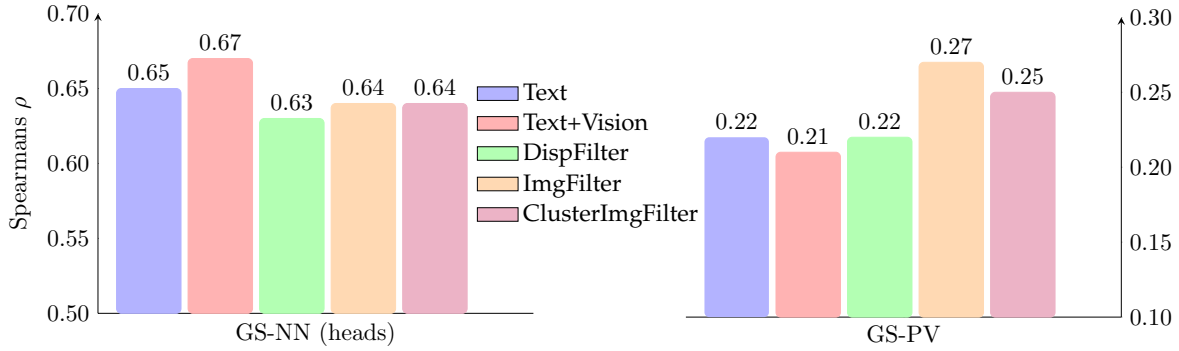

Figure 2: Overall prediction of compositionality for GS-NN (heads) and GS-PV.

Lexical, Empirical and Visual Filters The experiments compare the predictions of compositionality across all targets in the gold standards.[2] Furthermore, we zoom into factors that might influence the quality of predictions: (A) the ***impact of lexical and empirical target properties***, i.e., ambiguity (relying on the DUDEN dictionary[3], frequency (as provided by the gold standards), abstractness and imageability (as taken from Köper and Schulte im Walde (2016)); (B) ***optimisation of the visual space***: (i) In accordance with human concept processing (Paivio, 1990), including image representations should be more useful for words which are visual. We therefore apply the *dispersion-based filter* suggested by Kiela et al. (2014). The filter decides whether to include perceptual information for a specific word or not, relying on a pairwise similarity between all images of a concept. The underlying idea is that highly visual concepts are visualised by similar pictures and thus trigger a high average similarity between the word's images. Abstract concepts, on the other hand, are expected to provide a lower dispersion. For a given word, the filter decides about using only the textual representation, or both the textual and visual representations, depending on the dispersion value and a predefined threshold (set to the median of all the dispersion values). (ii) We apply an *imageability filter* based on external imageability norms (Köper and Schulte im Walde, 2016), to successively include only images for the most imaginable target words. This filter is applied in the same way as dispersion. (iii) We suggest a novel *clustering filter*, that performs a clustering of the 25 images for a given concept, using the algorithm from Apidianaki (2010), and includes only images from the largest image cluster, cf. Figure 1.

Results and Discussion Figure 2 present the prediction results for the two gold standards, GS-NN and GS-PV. For GS-NN, we focus on predicting the compositionality for compound–head pairs (ignoring compound–modifier pairs), in order to have a more parallel setup to GS-PV, where the particle verb compositionality focuses on the contribution of the base verb. The figures show the results across all targets. Note that the vertical axis, showing the range of Spearman's ρ are different for both results.

Figures 3 and 4 zoom into target subsets regarding target ambiguity (one sense vs. multiple senses), frequency, abstractness vs. concreteness, imageability, and compositionality. The bars refer to the textual model, the multi-modal model (including all images for all targets), and the best results obtained when using the dispersion/imageability/clustering[4] filters.

The plots demonstrate that overall the multi-modal model provides only a tiny gain for GS-NN in comparison to the text-only model, which is however significant using *Steiger*'s test ($p < 0.001$) (Steiger, 1980). All filters worsen the results. For GS-PV, we also obtain a significant improvement by the multi-modal model, but only when applying the imageability or the clustering filter to the visual information. The main differences in the overall noun and verb results are emphasised in Figure 5, comparing the successive increase of images to the multi-modal model in comparison to the textual model, based on the dispersion and imageability filters. Note that the textual

[2]We focus on the model with window 5 and 500 dimensions, and GoogLeNet as the overall best approach.

[3]`www.duden.de`

[4]For the clustering filter, we focus on a combination with the imageability filter, which provided the best results.

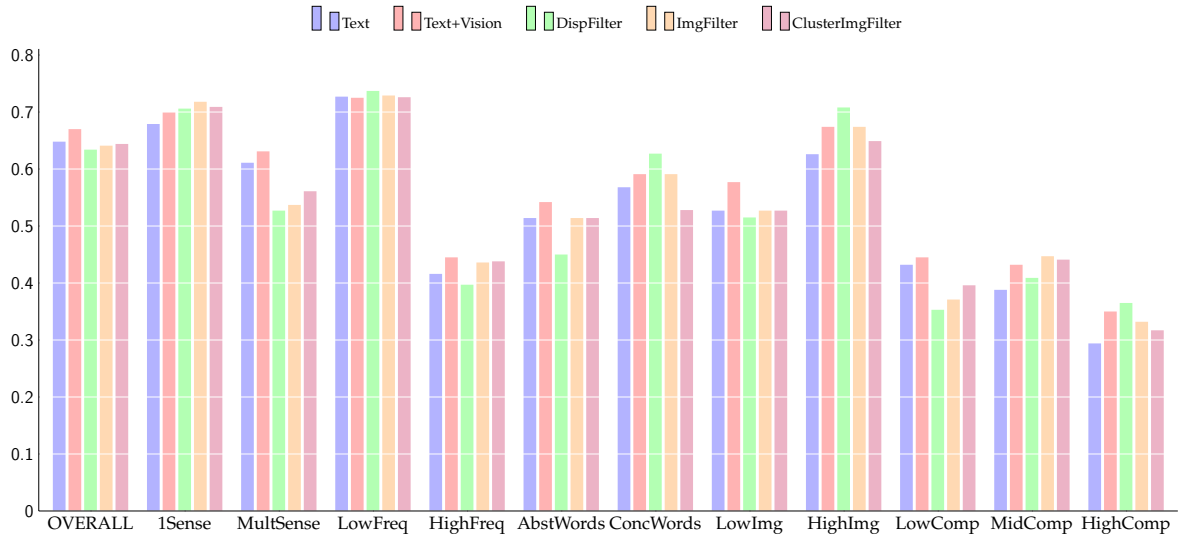

Figure 3: Prediction of compositionality for GS-NN heads.

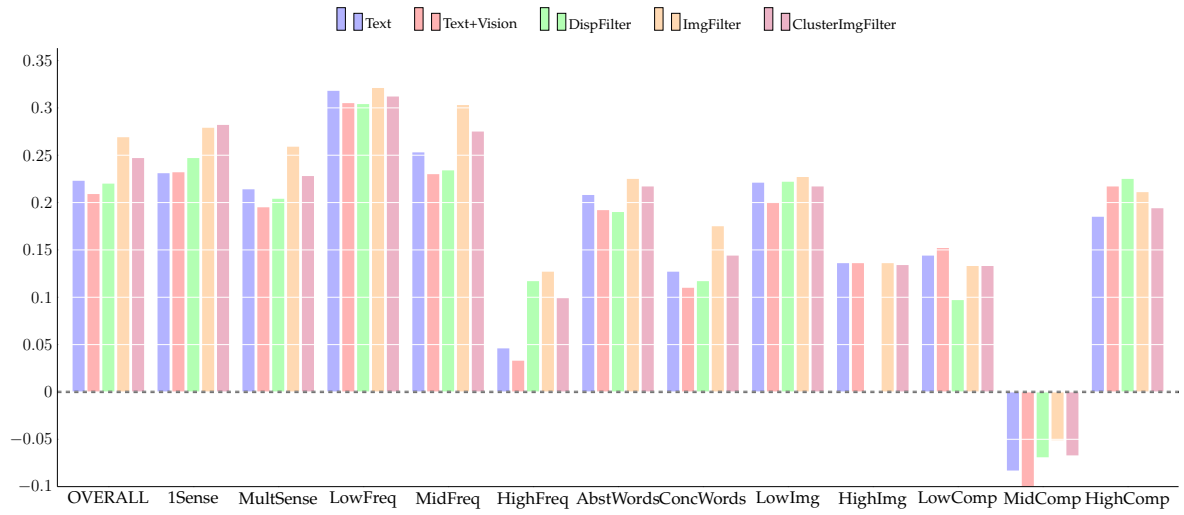

Figure 4: Prediction of compositionality for GS-PV.

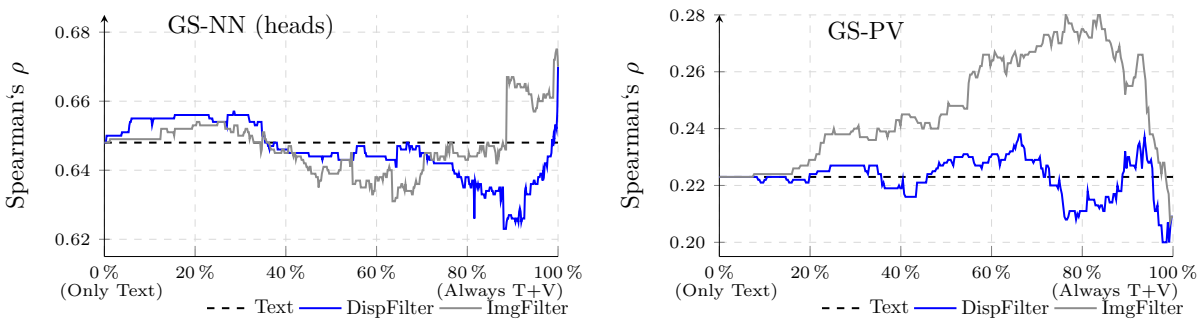

Figure 5: Prediction of compositionality: effect of dispersion and imageability filters.

model baselines are very different for the two gold standards, $\rho = .65$ for GS-NN and $\rho = .22$ for GS-PV. Regarding the nouns, the multi-modality improves the textual modality when adding the images for the ≈35% most imaginable words, and when adding all images. Regarding the verbs, the multi-modality improves the textual modality in most proportions, reaching its maximum when adding images for ≈80% of the most imaginable verbs; when adding the ≈10% of the least imaginable verbs, the model strongly drops in its performance. For the dispersion filter, the tendencies are less clear. We conclude that the visual information adds to the textual information either by adding all (potentially noisy) images because the textual information is rich by itself; or by adding a selection of images (unless they are overly dissimilar to each other, or for non-imaginable targets), because the textual information by itself is poor.

Zooming into target subsets, the predictions for monosemous targets are better than those for ambiguous targets (significant for GS-NN), see Figure 3; ditto for low-frequency vs. high-frequency targets. Taking frequency as an indicator of ambiguity, these differences are presumably due to the difficulty of distinguishing between multiple senses in vector spaces that subsume the features of all word senses within one vector, which applies to our textual and multi-modal models.

The gold standard predictions strongly differ regarding the influence of target abstractness, imageability and compositionality. For GS-NN, the compositionality of concrete and imaginable targets is predicted better than for abstract and less imaginable targets, as one would expect and has been shown by Kiela et al. (2014); for GS-PV, the opposite is the case. Similarly, while for GS-NN highly compositional targets are predicted worse than low- and mid-compositional targets, for GS-PV mid-compositional targets are predicted much worse than low- and high-compositional targets. These differences in results point to questions that have still been unsolved across research fields: while humans can easily grasp intuitions about the abstractness, imageability and compositionality of nouns, the categorisations are difficult to define for verbs (Glenberg and Kaschak, 2002; Brysbaert et al., 2014). Particle verbs add to this complexity, especially since compositionality (rating) is typically reduced to the semantic relatedness between the complex verb and the base verb, ignoring the particle that however contributes a considerable portion of meaning to the complex verb.

4 Conclusion

The paper demonstrated strong differences in the effect of adding visual information to a textual neural network model, when predicting the compositionality for nominal vs. verbal MWE targets. The visual modality adds complementary features in cases where (a) the textual modality performs poorly, and images of the most imaginable targets are added, or (b) the textual modality performs well, and all available –potentially noisy– images are added. Image filters relying on imageability and a novel clustering filter positively affect the verbal but not the nominal perceptual feature spaces.

Acknowledgments

The research was supported by the DFG Collaborative Research Centre SFB 732 (Maximilian Köper) and the DFG Heisenberg Fellowship SCHU-2580/1 (Sabine Schulte im Walde).

References

Marianna Apidianaki. 2010. An Algorithm for Cross-lingual Sense Clustering tested in a MT Evaluation Setting. In *Proceedings of the 7th International Workshop on Spoken Language Translation*, pages 219–226, Paris, France.

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.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A Systematic Comparison of Context-counting and Context-predicting Semantic Vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 238–247, Baltimore, MD.

Stefan Bott and Sabine Schulte im Walde. 2014. 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*, pages 509–516, Reykjavik, Iceland.

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*, pages 34–39, London, UK.

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*, pages 125–133, Osaka, Japan.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal Distributional Semantics. *Journal of Artificial Intelligence Research*, 49(1):1–47.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness Ratings for 40 Thousand generally known English Word Lemmas. *Behavior Research Methods*, 64:904–911.

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

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing Search in Context: The Concept Revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Arthur M. Glenberg and Michael P. Kaschak. 2002. Grounding Language in Action. *Psychonomic Bulletin & Review*, 9(3):558–565.

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

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation. *Computational Linguistics*, 41(4):665–695.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the International Conference on Multimedia*, pages 675–678, New York, NY, USA.

Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark. 2014. Improving Multi-Modal Representations Using Image Dispersion: Why Less is Sometimes More. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 835–841, Baltimore, MA.

Douwe Kiela, Anita L. Verő, and Stephen Clark. 2016. Comparing Data Sources and Architectures for Deep Visual Representation Learning in Semantics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Austin, TX.

Maximilian Köper and Sabine Schulte im Walde. 2016. Automatically Generated Affective Norms of Abstractness, Arousal, Imageability and Valence for 350 000 German Lemmas. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 2595–2598, Portoroz, Slovenia.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining Language and Vision with a Multimodal Skip-gram Model. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163, Denver, Colorado, USA.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of Computational Linguistics*, 3:211–225.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Kim-Anh Nguyen, Sabine Schulte im Walde, and Thang Vu. 2016. Integrating Distributional Lexical Contrast into Word Embeddings for Antonym-Synonym Distinction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 454–459, Berlin, Germany.

A. Paivio. 1990. *Mental Representations: A Dual Coding Approach*. Oxford Psychology Series. Oxford University Press.

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

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

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

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

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

Sabine Schulte im Walde, Anna Hätty, 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 5th Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 148–158, Berlin, Germany.

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 10th International Conference on Language Resources and Evaluation*, pages 2285–2292, Portoroz, Slovenia.

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

Carina Silberer and Mirella Lapata. 2014. Learning Grounded Meaning Representations with Autoencoders. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 721–732, Baltimore, Maryland.

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

James H Steiger. 1980. Tests for Comparing Elements of a Correlation Matrix. *Psychological Bulletin*.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going Deeper with Convolutions. In *Computer Vision and Pattern Recognition*.

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