Automatic Semantic Classification of German Preposition Types: Comparing Hard and Soft Clustering Approaches across Features

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

Abstract

This paper addresses an automatic classification of preposition types in German, comparing hard and soft clustering approaches and various window- and syntax-based co-occurrence features. We show that (i) the semantically most salient preposition features (i.e., subcategorised nouns) are the most successful, and that (ii) soft clustering approaches are required for the task but reveal quite different attitudes towards predicting ambiguity.

1 Introduction

In the last decades, an impressive number of semantic classifications has been developed, both regarding manual lexicographic and/or cognitive classifications such as *WordNet* (Fellbaum, 1998), *FrameNet* (Fillmore et al., 2003), *VerbNet* (Kipper Schuler, 2006) and *PrepNet/The Preposition Project* (Litkowski and Hargraves, 2005; Saint-Dizier, 2005), as well as regarding computational classifications for *nouns* (Hindle, 1990; Pereira et al., 1993; Snow et al., 2006), *verbs* (Merlo and Stevenson, 2001; Korhonen et al., 2003; Schulte im Walde, 2006) and *adjectives* (Hatzivassiloglou and McKeown, 1993; Boleda et al., 2012).

Semantic classifications are of great interest to computational linguistics, specifically regarding the pervasive problem of data sparseness in the processing of natural language. Such classifications have been used in applications such as *word sense disambiguation* (Dorr and Jones, 1996; Kohomban and Lee, 2005; McCarthy et al., 2007), *parsing* (Carroll et al., 1998; Carroll and Fang, 2004), *machine translation* (Prescher et al., 2000; Koehn and Hoang, 2007; Weller et al., 2014), and *information extraction* (Surdeanu et al., 2003; Venturi et al., 2009).

Regarding prepositions, comparably little effort in computational semantics has gone beyond a specific choice of prepositions (such as spatial prepositions), towards a systematic classification of preposition senses, as in *The Preposition Project* (Litkowski and Hargraves, 2005). Distributional approaches towards preposition meaning and sense distinction have only recently started to explore salient preposition features, but with few exceptions (such as Baldwin (2006)) these approaches focused on token-based classification of preposition senses (Ye and Baldwin, 2006; O'Hara and Wiebe, 2009; Tratz and Hovy, 2009; Hovy et al., 2010; Hovy et al., 2011).

This paper addresses an automatic classification of preposition types in German, comparing various clustering approaches. We aim for an unsupervised setting that does not require predefined expensive resources, such as a token-based annotation of preposition senses. Our task is challenging, because (i) prepositions are notoriously ambiguous, (ii) the interpretation of out-of-context preposition type classification is more difficult than context-embedded token interpretation, (iii) there are no established lexical resources for type-based semantic classification other than for English, and (iv) there are no established evaluation measures for ambiguous linguistic classifications. We accept the challenges, identify salient preposition features, and demonstrate the inevitability to apply soft (rather than hard) clustering in order to explore linguistic ambiguity.

2 Experiments

2.1 Preposition Data

In the absence of any large-scale semantic hierarchical type classification, the German grammar book by Helbig and Buscha (1998) represents our gold standard. We selected those preposition

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 256–263,
Berlin, Germany, August 7-12, 2016.

Class		Size
lokal	'local'	27
modal	'modal'	24
temporal	'temporal'	21
kausal	'causal'	5
distributiv	'distributive'	6
final	'final'	4
urheber	'creator'	3
konditional	'conditional'	3
ersatz	'replacement'	2
restriktiv	'restrictive'	2
partitiv	'partitive'	2
kopulativ	'copulative'	2

Table 1: Preposition classes.

classes that contained more than one preposition, and deleted prepositions that appeared <10,000 times in our web corpus containing 880 million words (cf. Section 2.2). This selection process resulted in 12 semantic classes covering between 2 and 27 prepositions each (cf. Table 1), and a more fine-grained version that sub-divided the three largest classes 'local', 'modal' and 'temporal' into 6/10/7 sub-classes, respectively, and resulted in a total of 32 classes.[1][2] The prepositions in the fine-grained version exhibit ambiguity rates of 1 (monosemous) up to 10. Out of the 49 preposition types, 23 are polysemous (46.9%).

2.2 Preposition Features

The corpus-based features for the German prepositions were induced from the *SdeWaC* corpus (Faaß and Eckart, 2013), a cleaned version of the German web corpus *deWaC* (Baroni et al., 2009) containing approx. 880 million words. We compare three categories of distributional features:

(1) ***bag-of-words window co-occurrence features***: we apply a standard bag-of-words model (*BOW*) relying on a window of 2 words to the left and to the right, and a continuous bag-of-words model (*CBOW*) using negative sampling with K=15 (Mikolov et al., 2013);

(2) ***direct syntactic dependency***: we compare the most salient preposition-related dependencies: preposition-subcategorised nouns (*nouns-dep*, e.g., *in Buch* 'in book'), preposition-subcategorising nouns (*nouns-gov*, e.g., *Buch von* 'book by'), and preposition-subcategorising verbs (*verbs-gov*, e.g., *reisen nach* 'to travel to');

(3) ***2nd-order syntactic co-occurrence***: adjectives that modify nouns subcategorised by the prepositions, and adverbs that modify verbs subcategorising the prepositions.

The dependency information was extracted from a parsed version of the SdeWaC using Bohnet's MATE dependency parser (Bohnet, 2010; Scheible et al., 2013). All but the CBOW features were weighted according to positive pointwise mutual information.

2.3 Clustering Approaches

As we wanted to explore hard vs. soft clustering approaches on the same task, we chose *k-Means* as a standard hard clustering approach (relying on WEKA's spherical k-Means implementation), and compared it to various soft clustering approaches.

We transfered the hard k-Means cluster analyses to soft cluster analyses, using two alternative methods. (1) The ***prep-based soft k-Means*** method (Springorum et al., 2013) calculated the mean cosine distance $\bar{d}$ for each preposition p to the centroids z_c of the clusters c, and assigned a preposition to a specific cluster if its distance to the respective cluster centroid was below a threshold t multiplied with the mean distance, with $t = 0.05, 0.1, 0.15, \ldots, 0.95$. Additionally, (2) we propose a hard-to-soft clustering transfer ***prob-based soft k-Means*** that converts the cosine distances between the prepositions and the hard cluster centroids to membership probabilities.

Instead of transferring a hard clustering to a soft clustering we also directly applied soft clustering approaches: (1) The fuzzy ***c-Means*** algorithm extends k-means by a cluster membership function for each preposition, $f_m \in [0, 1]$. (2) We applied ***Latent Semantic Clustering (LSC)***, an instance of the Expectation-Maximisation (EM) algorithm (Baum, 1972) for unsupervised training on unannotated data (Rooth et al., 1999). The cluster analyses define two-dimensional soft clusters (in our case: preposition–feature clusters) with cluster membership probabilities, which are able to generalise over hidden data. (3) We used ***Non-negative matrix factorization (NMF)***, a factorisation approach with an inherent (soft) clustering property (Ding et al., 2005).

All variants of our hard-to-soft clustering approaches and the direct soft clustering approaches (except for k-Means/prep)[3] resulted in a

[1] While we also conducted experiments using the coarse-grained class distribution in Table 1, the experiments in this paper focus on the fine-grained inventory.

[2] The gold standard was previously used in Springorum et al. (2013) and in Köper and Schulte im Walde (2014).

[3] *k-Means/prep* directly provides binary membership.

preposition–cluster membership matrix with values $\in [0, 1]$. We transfered the real membership values to binary membership by applying a threshold t to decide about the cluster membership, again with $t = 0.05, 0.1, 0.15, \ldots, 0.95$. For each clustering approach and for each number of clusters k we then identified the best threshold.

2.4 Evaluation

We chose the fuzzy extension of *B-Cubed* (Bagga and Baldwin, 1998) as evaluation measure, because it is (a) a pair-wise evaluation, which is considered as most suitable for soft clustering evaluations, and (b) distinguishes between homogeneity and completeness of a clustering, and thus resembles an evaluation by precision and recall. Amigó et al. (2009) demonstrated the strengths of B-Cubed, and a similar version has been used in SemEval 2013 for Word Sense Induction (Jurgens and Klapaftis, 2013).

Pair-wise precision P determines the homogeneity of a cluster analysis, by calculating for each individual preposition p the amount of prepositions p' in the same cluster c that also belong to the same gold-standard class g, cf. Equation (1). Pair-wise recall R determines the completeness of a cluster analysis, by calculating for each individual preposition p the amount of prepositions p' in the same gold-standard class g that also belong to the same cluster c, cf. Equation (2). The overall B-Cubed precision and recall scores are the averages over all preposition-wise scores. We combined precision and recall by their harmonic mean, the f-score.

$$P(p, p') = \frac{min(|c(p) \cap c(p')|, |g(p) \cap g(p')|}{|c(p) \cap c(p')|} \quad (1)$$

$$R(p, p') = \frac{min(|c(p) \cap c(p')|, |g(p) \cap g(p')|}{|g(p) \cap g(p')|} \quad (2)$$

2.5 Baselines

We created two baselines for our preposition clusterings: The ***hard baseline*** was computed for every number of clusters k=[5, 40]. For each k, each preposition was randomly assigned to one of the k clusters, and the resulting hard cluster analysis was evaluated. The hard cluster assignments were repeated 1,000 times for each k, and the overall evaluation score for k clusters is the average score of the 1,000 runs. The ***soft baseline*** was also created by random assignment across 1,000 runs for each k, but –integrating the fuzzy component– each preposition was assigned to n clusters, with n a random number between 1 and the number of gold-standard classes for that specific preposition. Note that this baseline is more informed than an entirely random baseline, because the information about the number of gold-standard classes for each preposition is very helpful. For example, the baseline assigns monosemous prepositions to only one cluster, and prepositions with three senses to a random integer in $[1, 3]$.

3 Results

Figure 1 compares the fuzzy B-Cubed f-score values across the hard and soft clustering approaches, relying on the preposition-subcategorised nouns as one of the best features (cf. Figure 2 below). The plot demonstrates that (i) the hard k-Means clustering approach is the only one resulting in f-scores below the soft baseline, while (ii) the vast majority of soft clustering results lies above the soft baseline. Furthermore, (iii) there is a clear tendency for all soft clustering approaches to provide the best f-scores for similar values of k clusters: $15 \leq k \leq 19$. The overall best result is reached by NMF for a clustering with 17 clusters.

Figure 2 compares the f-scores across feature types, relying on NMF as the best clustering approach. The plot confirms that (i) –across features–, the vast majority of soft clustering results lies above the soft baseline. In addition, (ii) in the previously most successful range for $15 \leq k \leq 19$ clusters, the preposition-subcategorised nouns represent the best features. (iii) The best cluster analyses relying on window vs. syntax features are similarly successful, and outperform 2nd-order co-occurrence features.

We checked the overall best cluster analysis (NMF, $k = 17$, nouns-dep) on the predicted degree of ambiguity (cf. Figure 3): for 23 out of the 26 monosemous prepositions, we *correctly* predicted one preposition sense; for 7 out of the 23 polysemous prepositions, we predicted the *correct* number of senses; for 9 out of the 23 polysemous prepositions, we predicted *less* senses than the gold standard defines; and for 7 out of the 23 polysemous prepositions, we predicted *more* senses than the gold standard defines.

Our best soft-clustering approach to the preposition classification task thus demonstrates its usefulness through quantitative B-Cubed evaluation and through reliable predictions of ambiguity.

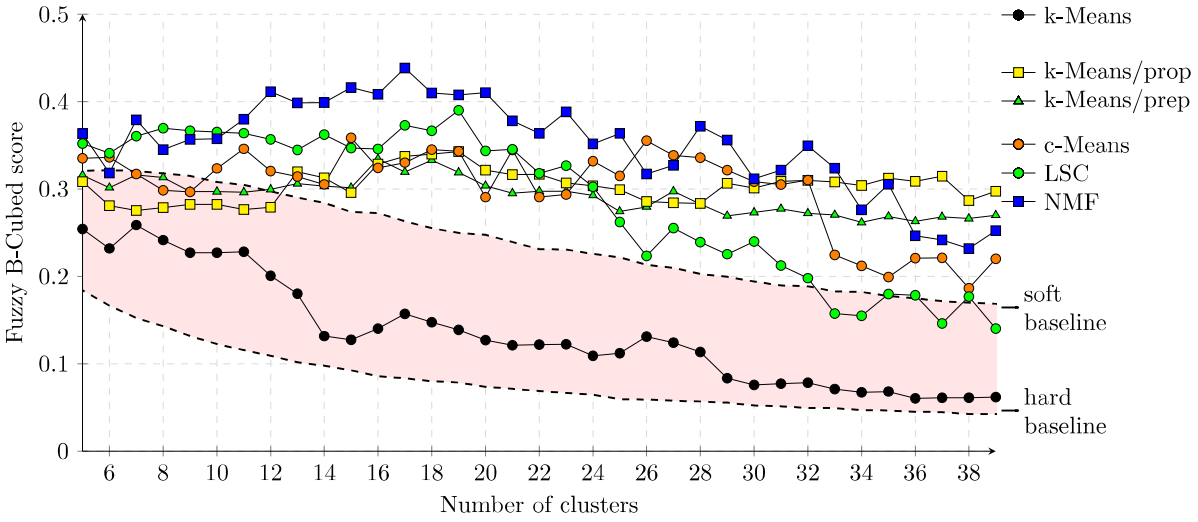

Figure 1: Fuzzy B-Cubed f-score using the subcategorised noun feature set (nouns-dep), across soft clustering approaches.

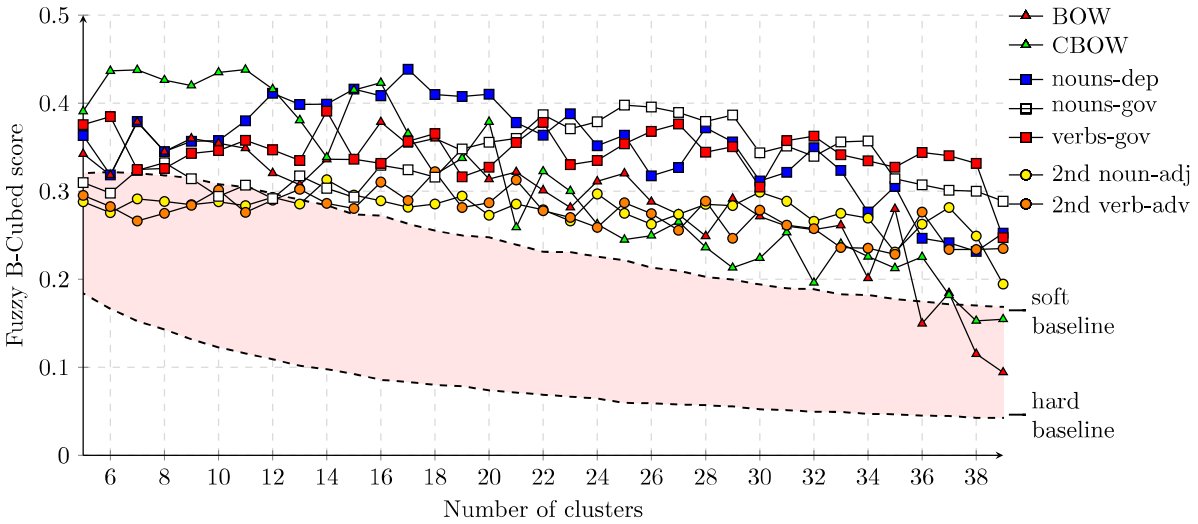

Figure 2: Fuzzy B-Cubed f-score using NMF soft clustering, across feature sets.

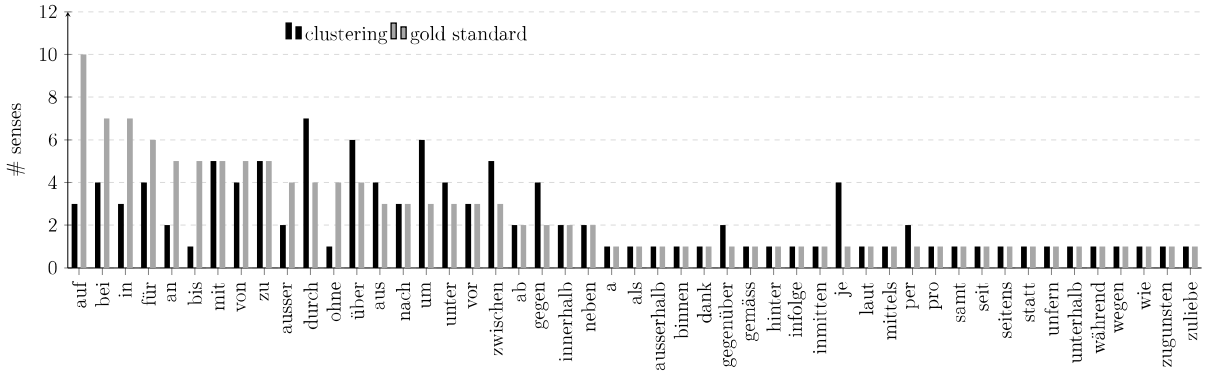

Figure 3: Predicting polysemy across prepositions (NMF, $k = 17$, *nouns-dep*).

4 Discussion

While the results in the previous section demonstrate the success of the type-based clustering, we were interested in two specific questions: (i) Where do the differences in the quality of the cluster analyses come from? (ii) Do the best cluster analyses present linguistically reliable and useful semantic classes?

From a quantitative point of view, both questions have been addressed by the evaluation measure, fuzzy B-Cubed, which we chose for reasons outlined in Section 2.4. One should keep in mind, however, that there is an ongoing discussion about cluster comparison and cluster evaluation (Meila, 2007; Rosenberg and Hirschberg, 2007; Vinh and Bailey, 2010; Utt et al., 2014), which demonstrates uncertainty about an optimal measure, and which concerns us, expecially regarding the linguistic aspects of soft clustering. In the following, we therefore provide qualitative analyses and discussions of the cluster approaches and analyses.

Ambiguity rate of soft-clustering approaches: We looked into the best cluster analysis for each soft-clustering approach, and checked the ambiguities. While the number of preposition types in the cluster analyses is similar across approaches (between 44 and 48), the ambiguity rate (i.e., the number of cluster assignments per preposition type) and the number of ambiguous preposition types (i.e., the number of prepositions assigned to more than one cluster) differ strongly. For example, k-Means/prob and NMF perform an average of 3.1/3.7 assignments for each preposition, in comparison to 2.2–2.4 assignments by the other approaches. On the other hand, while k-Means/prob defines almost all preposition types (43 out of 48) as ambiguous, NMF only defines 28 out of 46 prepositions as ambiguous. NMF (best approach) thus shows a high ambiguity rate, but only 60% of the prepositions are ambiguous.

Cluster sizes: Looking into the actual cluster analyses reveals that the sizes and the structures within the individual clusters differ strongly. The best k-Means/prep and k-Means/prob analyses ($k = 16$, $F = 0.33$, and $k = 19$, $F = 0.34$), for example, each contain 7 large clusters with 10–25 prepositions. All other clusters contain only 1–3 prepositions. In comparison, the best NMF analysis ($k = 17$, $F = 0.43$) contains only one cluster with three prepositions, and all other clusters but one contain ≥ 5 and ≤ 14 prepositions. The cluster sizes of the best NMF analysis are therefore more homogeneous than for other clustering approaches.

Optimal k: While fuzzy B-Cubed determined the numbers of clusters $[15, 19]$ as optimal for the soft-clustering approaches, we also looked into the NMF cluster analysis with $k = 32$, with NMF as the best approach and 32 as the number of gold standard classes. The clusters are, again, very similar in size, including only one singleton and only one cluster with 9 prepositions. All other clusters contain $2 - 6$ prepositions. The smaller cluster sizes allow manual evaluations. We can indeed find reliable semantic clusters, such as {*an, auf, hinter, in, mit, nach, neben, um, vor*}, where 7 out of 9 prepositions belong to the gold-standard class *local: not target-oriented* containing a total of 12 prepositions.

5 Conclusion

We presented variants of hard and soft clustering approaches across several sets of preposition features, to automatically classify preposition types into semantic classes. While type-based classifications for highly ambiguous word classes are a computational challenge, our best approach (NMF-based classification with 17 clusters) reached an f-score of 0.43. The clustering experiments showed that (i) the semantically most salient preposition features are indeed the most successful, and that (ii) the clustering of highly ambiguous words requires soft rather than hard clustering approaches.

Most interestingly, a qualitative analysis zoomed into the assignment behaviour of the soft clustering approaches, and revealed different attitudes towards predicting ambiguity. NMF as the best approach predicted a high ambiguity rate but only for a restricted proportion of 60% of the preposition types. Furthermore, the distribution of cluster sizes was less skewed than for other approaches.

Acknowledgments

The research was supported by the DFG Research Project "Distributional Approaches to Semantic Relatedness" (Maximilian Köper) and the DFG Heisenberg Fellowship SCHU-2580/1 (Sabine Schulte im Walde).

References

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A Comparison of Extrinsic Clustering Evaluation Metrics Based on Formal Constraints. *Information Retrieval*, 12(4):461–486.

Amit Bagga and Breck Baldwin. 1998. Entity-based Cross-document Coreferencing Using the Vector Space Model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pages 79–85, Montréal, Canada.

Timothy Baldwin. 2006. Distributional Similarity and Preposition Semantics. In Patrick Saint-Dizier, editor, *Syntax and Semantics of Prepositions*, chapter 1, pages 197–210. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.

Leonard E. Baum. 1972. An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes. *Inequalities*, III:1–8.

Bernd Bohnet. 2010. Top Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 89–97, Beijing, China.

Gemma Boleda, Sabine Schulte im Walde, and Toni Badia. 2012. Modelling Regular Polysemy: A Study on the Semantic Classification of Catalan Adjectives. *Computational Linguistics*, 38(3):575–616.

John Carroll and Alex C. Fang. 2004. The Automatic Acquisition of Verb Subcategorisations and their Impact on the Performance of an HPSG Parser. In *Proceedings of the 1st International Joint Conference on Natural Language Processing*, pages 107–114, Sanya City, China.

John Carroll, Guido Minnen, and Ted Briscoe. 1998. Can Subcategorisation Probabilities Help a Statistical Parser? In *Proceedings of the 6th ACL/SIGDAT Workshop on Very Large Corpora*, pages 118–126, Montréal, Canada.

Chris Ding, Xiaofeng He, and Horst D. Simon. 2005. On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering. In *Proceedings of the SIAM International Conference on Data Mining*, pages 606–610, Newport Beach, CA, USA.

Bonnie J. Dorr and Doug Jones. 1996. Role of Word Sense Disambiguation in Lexical Acquisition: Predicting Semantics from Syntactic Cues. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 322–327, Copenhagen, Denmark.

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.

Christiane Fellbaum, editor. 1998. *WordNet – An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16:235–250.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1993. Towards the Automatic Identification of Adjectival Scales: Clustering Adjectives According to Meaning. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 172–182, Columbus, OH.

Gerhard Helbig and Joachim Buscha. 1998. *Deutsche Grammatik*. Langenscheidt – Verlag Enzyklopädie, 18th edition.

Donald Hindle. 1990. Noun Classification from Predicate-Argument Structures. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 268–275, Pittsburgh, PA.

Dirk Hovy, Stephen Tratz, and Eduard Hovy. 2010. What's in a Preposition? Dimensions of Sense Disambiguation for an Interesting Word Class. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 454–462, Beijing, China.

Dirk Hovy, Ashish Vaswani, Stephen Tratz, David Chiang, and Eduard Hovy. 2011. Methods and Training for Unsupervised Preposition Sense Disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 323–328, Portland, OR.

David Jurgens and Ioannis Klapaftis. 2013. SemEval-2013 Task 13: Word Sense Induction for Graded and Non-Graded Senses. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics*, pages 290–299, Atlanta, Georgia, USA.

Karin Kipper Schuler. 2006. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania, Computer and Information Science.

Philipp Koehn and Hieu Hoang. 2007. Factored Translation Models. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 868–876, Prague, Czech Republic.

Upali S. Kohomban and Wee Sun Lee. 2005. Learning Semantic Classes for Word Sense Disambiguation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 34–41, Ann Arbor, MI.

Maximilian Köper and Sabine Schulte im Walde. 2014. A Rank-based Distance Measure to Detect Polysemy and to Determine Salient Vector-Space Features for German Prepositions. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 4459–4466, Reykjavik, Iceland.

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*, pages 64–71, Sapporo, Japan.

Kenneth C. Litkowski and Orin Hargraves. 2005. The Preposition Project. In *Proceedings of the 2nd ACL-SIGSEM Workshop on The Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, pages 171–179, Colchester, England.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2007. Unsupervised Acquisition of Predominant Word Senses. *Computational Linguistics*, 33(4):553–590.

Marina Meila. 2007. Comparing Clusterings – An Information-based Distance. *Journal of Multivariate Analysis*, 98(5):873–895.

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

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.

Tom O'Hara and Janyce Wiebe. 2009. Exploiting Semantic Role Resources for Preposition Disambiguation. *Computational Linguistics*, 35(2):151–184. Special Issue on Prepositions in Applications.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional Clustering of English Words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 183–190, Columbus, OH.

Detlef Prescher, Stefan Riezler, and Mats Rooth. 2000. Using a Probabilistic Class-Based Lexicon for Lexical Ambiguity Resolution. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 649–655, Saarbrücken, Germany.

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, Maryland, MD.

Andrew Rosenberg and Julia Hirschberg. 2007. V-Measure: A Conditional Entropy-based External Cluster Evaluation Measure. In *Proceedings of the joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 410–420, Prague, Czech Republic.

Patrick Saint-Dizier. 2005. An Overview of PrepNet: Abstract Notions, Frame, and Inferential Patterns. In *Proceedings of the 2nd ACL-SIGSEM Workshop on The Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, Colchester, England.

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*, pages 63–72, Lancaster, UK.

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

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2006. Semantic Taxonomy Induction from Heterogenous Evidence. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 801–808, Sydney, Australia.

Sylvia Springorum, Sabine Schulte im Walde, and Jason Utt. 2013. Detecting Polysemy in Hard and Soft Cluster Analyses of German Preposition Vector Spaces. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 632–640, Nagoya, Japan.

Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using Predicate-Argument Structures for Information Extraction. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 8–15, Sapporo, Japan.

Stephen Tratz and Dirk Hovy. 2009. Disambiguation of Preposition Sense using Linguistically Motivated Features. In *Proceedings of the NAACL-HLT Student Research Workshop and Doctoral Consortium*, pages 96–100, Boulder, CO.

Jason Utt, Sylvia Springorum, Maximilian Köper, and Sabine Schulte im Walde. 2014. Fuzzy V-Measure – An Evaluation Method for Cluster Analyses of

Ambiguous Data. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 581–587, Reykjavik, Iceland.

Giulia Venturi, Simonetta Montemagni, Simone Marchi, Yutaka Sasaki, Paul Thompson, John McNaught, and Sophia Ananiadou. 2009. Bootstrapping a Verb Lexicon for Biomedical Information Extraction. In Alexander Gelbukh, editor, *Linguistics and Intelligent Text Processing*, pages 137–148. Springer, Heidelberg.

Nguyen Xuan Vinh and James Bailey. 2010. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *Journal of Machine Learning Research*, 11:2837–2854.

Marion Weller, Sabine Schulte im Walde, and Alexander Fraser. 2014. Using Noun Class Information to model Selectional Preferences for Translating Prepositions in SMT. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, pages 275–287, Vancouver, Canada.

Patrick Ye and Timothy Baldwin. 2006. Semantic Role Labelling of Prepositional Phrases. *ACM Transactions on Asian Language Information Processing*, 5(3):228–244.