Assessing SRL Frameworks with Automatic Training Data Expansion

Silvana Hartmann[†‡] **Éva Mújdricza-Maydt**[*‡] **Ilia Kuznetsov**[†]
Iryna Gurevych[†‡] **Anette Frank**[*‡]

[†]Ubiquitous Knowledge Processing Lab
Department of Computer Science
Technische Universität Darmstadt
[‡]Research Training School AIPHES
{hartmann,kuznetsov,gurevych}
@ukp.informatik.tu-darmstadt.de

[*]Department of
Computational Linguistics
Heidelberg University
[‡]Research Training School AIPHES
{mujdricza,frank}
@cl.uni-heidelberg.de

Abstract

We present the first experiment-based study that explicitly contrasts the three major semantic role labeling frameworks. As a prerequisite, we create a dataset labeled with parallel FrameNet-, PropBank-, and VerbNet-style labels for German. We train a state-of-the-art SRL tool for German for the different annotation styles and provide a comparative analysis across frameworks. We further explore the behavior of the frameworks with automatic training data generation. VerbNet provides larger semantic expressivity than PropBank, and we find that its generalization capacity approaches PropBank in SRL training, but it benefits less from training data expansion than the sparse-data affected FrameNet.

1 Introduction

We present the first study that explicitly contrasts the three popular theoretical frameworks for semantic role labeling (SRL) – FrameNet, PropBank, and VerbNet[1] in a comparative experimental setup, i.e., using the same training and test sets annotated with predicate and role labels from the different frameworks and applying the same conditions and criteria for training and testing.

Previous work comparing these frameworks either provides theoretical investigations, for instance for the pair PropBank–FrameNet (Ellsworth et al., 2004), or presents experimental investigations for the pair PropBank–VerbNet (Zapirain et al., 2008; Merlo and van der Plas, 2009). Theoretical analyses contrast the richness of the semantic model of FrameNet with efficient annotation of PropBank labels and their suitability for system training. VerbNet is considered to range between them on both scales: it fulfills the need for semantically meaningful role labels; also, since the role labels are shared across predicate senses, it is expected to generalize better to unseen predicates than FrameNet, which suffers from data sparsity due to a fine-grained sense-specific role inventory. Yet, unlike PropBank and FrameNet, VerbNet has been neglected in recent work on SRL, partially due to the lack of training and evaluation data, whereas PropBank and FrameNet were popularized in shared tasks. As a result, the three frameworks have not been compared under equal experimental conditions.

This motivates our contrastive analysis of *all three frameworks* for German. We harness existing datasets for German (Burchardt et al., 2006; Hajič et al., 2009; Mújdricza-Maydt et al., 2016) to create *SR3de (Semantic Role Triple Dataset for German)*, the first benchmark dataset labeled with FrameNet, VerbNet and PropBank roles in parallel.

Our motivation for working on German is that – as for many languages besides English – sufficient amounts of training data are not available. This clearly applies to our German dataset, which contains about 3,000 annotated predicates. In such a scenario, methods to extend training data automatically or making efficient use of generalization across predicates (i.e., being able to apply role labels to unseen predicates) are particularly desirable. We assume that SRL frameworks that generalize better across predicates gain more from automatic training data generation, and lend themselves better to cross-predicate SRL. System performance also needs to be correlated with the semantic expressiveness of frameworks: with the ever-growing expectations in semantic NLP applications, SRL frameworks also need to be judged with regard to their contribution to advanced applications where expressiveness may play a role, such as question answering or summarization.

[1] See Fillmore et al. (2003), Palmer et al. (2005) and Kipper-Schuler (2005), respectively.

Proceedings of the 11th Linguistic Annotation Workshop, pages 115–121,
Valencia, Spain, April 3, 2017. ©2017 Association for Computational Linguistics

Our work explores the generalization properties of three SRL frameworks in a contrastive setup, assessing SRL performance when training and evaluating on a dataset with parallel annotations for each framework in a uniform SRL system architecture. We also explore to what extent the frameworks benefit from training data generation via annotation projection (Fürstenau and Lapata, 2012).

Since all three frameworks have been applied to several languages,[2] we expect our findings for German to generalize to other languages as well.

Our contributions are (i) novel resources: parallel German datasets for the three frameworks, including automatically acquired training data; and (ii) empirical comparison of the labeling performance and generalization capabilities of the three frameworks, which we discuss in view of their respective semantic expressiveness.

2 Related Work

Overview on SRL frameworks *FrameNet* defines frame-specific roles that are shared among predicates evoking the same frame. Thus, generalization across predicates is possible for predicates that belong to the same *existing* frame, but labeling predicates for unseen frames is not possible. Given the high number of frames – FrameNet covers about 1,200 frames and 10K role labels – large training datasets are required for system training.

PropBank offers a small role inventory of five core roles (A0 to A4) for obligatory arguments and around 18 roles for optional ones. The core roles closely follow syntactic structures and receive a predicate-specific interpretation, except for the Agent-like A0 and Patient-like A1 that implement Dowty's proto-roles theory (Dowty, 1991).

VerbNet defines about 35 semantically defined *thematic* roles that are not specific to predicates or predicate senses. Predicates are labeled with Levin-type semantic classes (Levin, 1993). VerbNet is typically assumed to range between FrameNet with respect to its rich semantic representation and PropBank with its small, coarse-grained role inventory.

Comparison of SRL frameworks Previous experimental work compares VerbNet and PropBank: Zapirain et al. (2008) find that PropBank SRL is more robust than VerbNet SRL, generalizing better to unseen or rare predicates, and relying less on predicate sense. Still, they aspire to use more meaningful VerbNet roles in NLP tasks and thus propose using automatic PropBank SRL for core role identification and then converting the PropBank roles into VerbNet roles heuristically to VerbNet, which appears more robust in cross-domain experiments compared to training on VerbNet data. Merlo and van der Plas (2009) also confirm that PropBank roles are easier to assign than VerbNet roles, while the latter provide better semantic generalization. To our knowledge, there is no experimental work that compares all three major SRL frameworks.

German SRL frameworks and data sets The SALSA project (Burchardt et al., 2006; Rehbein et al., 2012) created a corpus annotated with over 24,200 predicate argument structures, using English *FrameNet* frames as a basis, but creating new frames for German predicates where required.

About 18,500 of the manual SALSA annotations were converted semi-automatically to *PropBank*-style annotations for the CoNLL 2009 shared task on syntactic and semantic dependency labeling (Hajič et al., 2009). Thus, the CoNLL dataset shares a subset of the SALSA annotations. To create PropBank-style annotations, the predicate senses were numbered such that different frame annotations for a predicate lemma indicate different senses. The SALSA role labels were converted to PropBank-style roles using labels A0 and A1 for Agent- and Patient-like roles, and continuing up to A9 for other arguments. Instead of spans, arguments were defined by their dependency heads for CoNLL. The resulting dataset was used as a benchmark dataset in the CoNLL 2009 shared task.

For VerbNet, Mújdricza-Maydt et al. (2016) recently published a small subset of the CoNLL shared task corpus with *VerbNet*-style roles. It contains 3,500 predicate instances for 275 predicate lemma types. Since there is no taxonomy of verb classes for German corresponding to original VerbNet classes, they used GermaNet (Hamp and Feldweg, 1997) to label predicate senses. GermaNet provides a fine-grained sense inventory similar to the English WordNet (Fellbaum, 1998).

Automatic SRL systems for German State-of-the-art SRL systems for German are only available for PropBank labels: Björkelund et al. (2009) developed mate-tools; Roth and Woodsend (2014) and Roth and Lapata (2015) improved on mate-tools SRL with their mateplus system. We base our experiments on the *mateplus* system.

[2] Cf. Hajič et al. (2009), Sun et al. (2010), Boas (2009).

	Der Umsatz ’Sales	***stieg*** rose	*um 14 %* by 14%	*auf 1,9 Milliarden .* to 1.9 billion’
PB	A1	**steigen.1**	A2	A3
VN	Patient	**steigen-3**	Extent	Goal
FN	Item	**Change_position _on_a_scale**	Difference	Final_value

Figure 1: Parallel annotation example from SR3de for predicate ***steigen*** *(’rise, increase’)*.

Corpus	train			dev			test		
	type		token	type		token	type		token
predicate	198		2196	121		250	152		520
	sense	role	role	sense	role	role	sense	role	role
SR3de-PB	506	10	4,293	162	6	444	221	8	1,022
SR3de-VN	448	30	4,307	133	23	466	216	25	1,025
SR3de-FN	346	278	4,283	133	145	456	176	165	1,017

Table 1: Data statistics for SR3de (PB, VN, FN).

Training data generation In this work, we use a corpus-based, monolingual approach to training data expansion. Fürstenau and Lapata (2012) propose monolingual annotation projection for lower-resourced languages: they create data labeled with FrameNet frames and roles based on a small set of labeled seed sentences in the target language. We apply their approach to the different SRL frameworks, and for the first time to VerbNet-style labels.

Other approaches apply cross-lingual projection (Akbik and Li, 2016) or paraphrasing, replacing FrameNet predicates (Pavlick et al., 2015) or PropBank arguments (Woodsend and Lapata, 2014) in labeled texts. We do not employ these approaches, because they assume large role-labeled corpora.

3 Datasets and Data Expansion Method

SR3de: a German parallel SRL dataset The VerbNet-style dataset by Mújdricza-Maydt et al. (2016) covers a subset of the PropBank-style CoNLL 2009 annotations, which are based on the German FrameNet-style SALSA corpus. This allowed us to create SR3de, the first corpus with parallel sense and role labels from SALSA, PropBank, and GermaNet/VerbNet, which we henceforth abbreviate as FN, PB, and VN respectively. Figure 1 displays an example with parallel annotations.

Data statistics in Table 1 shows that with almost 3,000 predicate instances, the corpus is fairly small. The distribution of role types across frameworks highlights their respective role granularity, ranging from 10 for PB to 30 for VN and 278 for FN. The corpus offers 2,196 training predicates and covers the CoNLL 2009 development and test sets; thus it is a suitable base for comparing the three frameworks. We use SR3de for the contrastive analysis of the different SRL frameworks below.

Training data expansion To overcome the data scarcity of our corpus, we use *monolingual annotation projection* (Fürstenau and Lapata, 2012) to generate additional training data. Given a set of labeled *seed sentences* and a set of unlabeled *expansion sentences*, we select suitable expansions based on the predicate lemma and align dependency graphs of seeds and expansions based on lexical similarity of the graph nodes and syntactic similarity of the edges. The alignment is then used to map predicate and role labels from the seed sentences to the expansion sentences. For each seed instance, the *k* best-scoring expansions are selected. Given a seed set of size n and the maximal number of expansions per seed k, we get up to $n \cdot k$ additional training instances. Lexical and syntactic similarity are balanced using the weight parameter α.

Our adjusted re-implementation uses the mate-tools dependency parser (Bohnet, 2010) and word2vec embeddings (Mikolov et al., 2013) trained on deWAC (Baroni et al., 2009) for word similarity calculation. We tune the parameter α via intrinsic evaluation on the SR3de dev set. We project the seed set SR3de-train directly to SR3de-dev and compare the labels from the *k*=1 best seeds for a dev sentence to the gold label, measuring F1 for all projections. Then we use the best-scoring α value for each framework to project annotations from the SR3de training set to deWAC for predicate lemmas occurring at least 10 times. We vary the number of expansions *k*, selecting k from $\{1, 3, 5, 10, 20\}$. Using larger k values is justified because a) projecting to a huge corpus is likely to generate many high-quality expansions, and b) we expect a higher variance in the generated data when also selecting lower-scoring expansions.

Intrinsic evaluation on the dev set provides an estimate of the projection quality: we observe F1 score of 0.73 for PB and VN, and of 0.53 for FN. The lower scores for FN are due to data sparsity in the intrinsic setting and are expected to improve when projecting on a large corpus.

4 Experiments

Experiment setup We perform extrinsic evaluation on SR3de with parallel annotations for the three frameworks, using the same SRL system for each framework, to a) compare the labeling performance of the learned models, and b) explore their behavior in response to expanded training data.

We employ the following settings (cf. Table 2):

#BL: ***Baseline*** We train on SR3de train, which is small, but comparable across frameworks.

#FB: ***Full baseline*** We train on the full CoNLL-training sections for PropBank and SALSA, to compare to state-of-the-art results and contrast the low-resource #BL to full resources.[3]

#EX: ***Expanded*** We train on data expanded via annotation projection.

We train mateplus using the reranker option and the default featureset for German[4] excluding word embedding features.[5] We explore the following role labeling tasks: predicate sense prediction (pd in mateplus), argument identification (ai) and role labeling (ac) for predicted predicate sense (pd+ai+ac) and oracle predicate sense (ai+ac). We report F1 scores for all three role labeling tasks.

We assure equivalent treatment of all three SRL frameworks in mateplus and train the systems only on the given training data without any framework-specific information. Specifically, we do not exploit constraints on predicate senses for PB in mateplus (i.e., selecting sense.1 as default sense), nor constraints for licensed roles (or role sets) for a given sense (i.e., encoding the FN lexicon). Thus, mateplus learns predicate senses and role sets only from training instances.

Experiment results for the different SRL frameworks are summarized in Table 2.[6] Below, we discuss the results for the different settings.

#BL: for role labeling with oracle senses (ai+ac), PB performs best, VN is around 5 percentage points (pp.) lower, and FN again 5 pp. lower. With predicate sense prediction (pd+ai+ac), performance only slightly decreases for VN and PB, while FN suffers strongly: F1 is 17 pp. lower than for VN, despite the fact that its predicate labeling F1 is similar to PB and higher than VN. This indicates that generalization across senses works much better for VN and PB roles. By contrast, FN, with its sense-dependent role labels, is lacking generalization capacity, and thus suffers from data sparsity.

no	train	sense (pd)	sense+role (pd+ai+ac)	role only (ai+ac)
	#BL: SR3de training corpora			
(1)	#BL-PB	58.84	73.70	74.76
(2)	#BL-VN	55.19	69.66	69.86
(3)	#BL-FN	58.26	52.76	64.72
	#FB: CoNLL training sections			
(4)	#FB-CoNLL	82.88	84.01	86.26
(5)	#FB-SALSA	84.03	78.03	84.34
	#EX: SR3de train with data expansion			
(1)	**#BL-PB**	58.84	73.70	74.76
(6)	#EX-k=1	58.65	75.09*	76.65**
(7)	#EX-k=3	58.65	75.43	77.71**
(8)	#EX-k=5	59.03	**76.30***	**78.27****
(9)	#EX-k=10	59.03	74.65	77.95**
(10)	#EX-k=20	**59.42**	74.36	78.15**
(2)	**#BL-VN**	55.19	69.66	69.86
(11)	#EX-k=1	55.00	68.75	68.86
(12)	#EX-k=3	55.19	**69.14**	**69.02**
(13)	#EX-k=5	55.19	68.49	68.57
(14)	#EX-k=10	55.19	66.34**	66.84**
(15)	#EX-k=20	**55.38**	65.70**	66.91**
(3)	**#BL-FN**	58.26	52.76	64.72
(16)	#EX-k=1	57.88	**55.47****	69.18**
(17)	#EX-k=3	58.65	54.13	69.37**
(18)	#EX-k=5	57.88	54.54**	**70.41****
(19)	#EX-k=10	58.26	53.97	69.15**
(20)	#EX-k=20	**58.84**	54.43	70.19**

Table 2: F1 scores for predicate sense and role labeling on the SR3de test set; *pd*: predicate sense labeling; *pd+ai+ac*: sense and role labeling (cf. official CoNLL scores); *ai+ac*: role labeling with oracle predicate sense. We report statistical significance of role labeling F1 with expanded data *#EX* to the respective *#BL* (*: $p < 0.05$; **: $p < 0.01$).

#FB: The full baselines #FB show that a larger training data set widely improves SRL performance compared to the small #BL training sets. One reason is the extended sense coverage in the #FB datasets, indicating the need for a larger training set. Still, FN scores are 6 pp. lower than PB (pd+ai+ac).

#EX: Automatically expanding the training set for PB leads to performance improvements of around 3 pp. to #BL for k=5 (pd+ai+ac and ai+ac), but the scores do not reach those of #FB. A similar gain is achieved for FN with k=1. Contrary to initial expectations, annotation projection tends to create similar instances to the seen ones, but at the same time, it also introduces noise. Thus, larger k (k>5) results in decreased role labeling performance compared to smaller k for all frameworks.

FN benefits most from training data expansion, with a performance increase of 5 pp. to #BL, reach-

[3] Both #FB training sets contain ≈ 17,000 predicate instances. There is no additional labeled training data for VN.

[4] `https://github.com/microth/mateplus/tree/master/featuresets/ger`

[5] Given only small differences in mateplus performance when using word embeddings, we report results without them.

[6] Significance is computed using approximation randomization, i.e., SIGF (Padó, 2006) two-tailed, 10k iterations.

ing similar role labeling scores as VN for the oracle sense setting. For predicted senses, performance increase is distinctly smaller, highlighting that the sparse data problems for FN senses do not get solved by training data expansion. Performance improvements are significant for FN and PB for both role labeling settings. Against expectation, we do not observe improved role labeling performance for VN. We believe this is due to the more complex label set compared to PB and perform a analyses supporting this hypothesis below.

Analysis: complexity of the frameworks We estimate the role labeling complexity of the frameworks by computing $C(d)$, the average ambiguity of the role instances in the dataset d, d $\in \{$PB, VN, FN$\}$. $C(d)$ consists of the normalized sum s over the number n of role candidates licensed for each role instance in d by the predicate sense label; for role instances with unseen senses, n is the number of distinct roles in the framework. The sum s is then divided by all role instances in dataset d.

Results are C(PB)=4.3, C(VN)=9.7, C(FN)=60. $C(d)$ is inversely correlated to the expected performance of each framework, and thus predicts the role labeling performance for #BL (pd+ai+ac).

When considering only seen training instances, complexity is 1.67 for both PB and VN, and 1.79 for FN. This indicates a larger difficulty for FN, but does not explain the difference between VN and PB. Yet, next to role ambiguity, the number of instances seen in training for individual role types is a decisive factor for role labeling performance, and thus, the coarser-grained PB inventory has a clear advantage over VN and FN.

The sense labeling performance is lower for VN systems compared to FN and PB. This correlates with the fact that GermaNet senses used with VN are more fine-grained than those in FN, but more abstract than the numbered PB senses. Still, we observe high role labeling performance independently of the predicate sense label for both VN and PB. This indicates high generalization capabilities of their respective role sets.[7]

The 5 pp. gap between the VN and PB systems is small, but not negligible. We expect that a suitable sense inventory for German VN, analogous to VerbNet's classes, will further enhance VN role labeling performance. Overall, we conclude that the higher complexity of the FrameNet role inventory causes data sparsity, thus FN benefits most from the training data expansion for seen predicates. For the other two frameworks, cross-predicate projection could be a promising way to increase the training data coverage to previously unseen predicates.

5 Discussion and Conclusion

We perform the first experimental comparison of all three major SRL frameworks on a small German dataset with parallel annotations. The experiment settings ensure comparability across frameworks.

Our baseline experiments prove that the generalization capabilities of the frameworks follow the hypothesized order of FrameNet $<$ VerbNet $<$ PropBank. Comparative analysis shows that PropBank and VerbNet roles generalize well, also beyond predicates. Taking into account the semantic expressiveness of VerbNet, these results showcase the potential of VerbNet as an alternative to PropBank. By contrast, FrameNet's role labeling performance suffers from data sparsity in the small-data setting, given that its role inventory does not easily generalize across predicates.

While VerbNet generalizes better than FrameNet, it does not benefit from our automatic training data generation setup. Currently, annotation projection only applies to lemmas seen in training. Thus, the generalization capacities of VerbNet – and PropBank – are not fully exploited. Relaxing constraints in annotation projection, e.g., projecting across predicates, could benefit both frameworks.

FrameNet suffers most from sparse-data problems and thus benefits most from automatic training data expansion for seen predicates, yet sense labeling persists as its performance bottleneck.

In future work we plan to a) further evaluate cross-predicate generalization capabilities of VerbNet and PropBank in cross-predicate annotation projection and role labeling, b) explore semi-supervised methods and constrained learning (Akbik and Li, 2016), and c) explore alternative sense inventories for the German VerbNet-style dataset.

We publish our benchmark dataset with strictly parallel annotations for the three frameworks to facilitate further research.[8]

[7] This is confirmed when replacing the predicate sense label with the lemma for training: the role labeling results are fairly close for PB (74.34%) and VN (68.90%), but much lower for FN (54.26%).

[8] `http://projects.cl.uni-heidelberg.de/SR3de`

Acknowledgements The present work was partially funded by a German BMBF grant to CLARIN-D and the DFG-funded Research Training School ”Adaptive Preparation of Information from Heterogeneous Sources” (GRK 1994/1). We thank the anonymous reviewers for their valuable comments and suggestions.

References

Alan Akbik and Yunyao Li. 2016. K-SRL: Instance-based Learning for Semantic Role Labeling. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 599–608, Osaka, Japan, December.

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.

Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 43–48, Boulder, CO, USA.

Hans C. Boas. 2009. *Multilingual FrameNets in computational lexicography: methods and applications*, volume 200. Walter de Gruyter.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In Chu-Ren Huang and Dan Jurafsky, editors, *Proceedings of Coling 2010*, pages 89–97.

Aljoscha Burchardt, Kathrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2006. The SALSA Corpus: a German Corpus Resource for Lexical Semantics. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 969–974, Genoa, Italy.

David Dowty. 1991. Thematic Proto-Roles and Argument Selection. *Language*, 67(3):547–619.

Michael Ellsworth, Katrin Erk, Paul Kingsbury, and Sebastian Padó. 2004. PropBank, SALSA, and FrameNet: How design determines product. In *Proceedings of the Workshop on Building Lexical Resources From Semantically Annotated Corpora, LREC-2004*, Lisbon, Portugal.

Christiane Fellbaum. 1998. *WordNet: an eletronic lexical database*. The Mit Press, Cambridge, Massachusetts.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250.

Hagen Fürstenau and Mirella Lapata. 2012. Semi-Supervised Semantic Role Labeling via Structural Alignment. *Computational Linguistics*, 38(1):135–171.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Nivre Joakim, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdenau, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL): Shared Task*, pages 1–18, Boulder, CO, USA.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a Lexical-Semantic Net for German. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15, Madrid, Spain.

Karin Kipper-Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania, Pennsylvania, Philadelphia, PA.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Paola Merlo and Lonneke van der Plas. 2009. Abstraction and Generalisation in Semantic Role Labels: PropBank, VerbNet or both? In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pages 288–296, Suntec, Singapore.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Éva Mújdricza-Maydt, Silvana Hartmann, Iryna Gurevych, and Anette Frank. 2016. Combining Semantic Annotation of Word Sense & Semantic Roles: A Novel Annotation Scheme for VerbNet Roles on German Language Data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3031–3038, Portorož, Slovenia.

Sebastian Padó, 2006. *User's guide to* `sigf`*: Significance testing by approximate randomisation*.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Ellie Pavlick, Travis Wolfe, Pushpendre Rastogi, Chris Callison-Burch, Mark Dredze, and Benjamin

Van Durme. 2015. FrameNet+: Fast Paraphrastic Tripling of FrameNet. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 408–413, Beijing, China.

Ines Rehbein, Joseph Ruppenhofer, Caroline Sporleder, and Manfred Pinkal. 2012. Adding nominal spice to SALSA - Frame-semantic annotation of German nouns and verbs. In *Proceedings of the 11th Conference on Natural Language Processing (KONVENS'12)*, pages 89–97, Vienna, Austria.

Michael Roth and Mirella Lapata. 2015. Context-aware frame-semantic role labeling. *Transactions of the Association for Computational Linguistics (TACL)*, 3:449–460.

Michael Roth and Kristian Woodsend. 2014. Composition of word representations improves semantic role labelling. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 407–413, Doha, Qatar.

Lin Sun, Thierry Poibeau, Anna Korhonen, and Cedric Messiant. 2010. Investigating the cross-linguistic potential of VerbNet-style classification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1056–1064.

Kristian Woodsend and Mirella Lapata. 2014. Text rewriting improves semantic role labeling. *Journal of Artificial Intelligence Research*, 51:133–164.

Beñat Zapirain, Eneko Agirre, and Lluís Màrquez. 2008. Robustness and generalization of role sets: PropBank vs. VerbNet. In *Proceedings of ACL-08: HLT*, pages 550–558, Columbus, OH, USA.