Towards a Better Semantic Role Labeling of Complex Predicates

Glorianna Jagfeld
Institute for Natural Language Processing
University of Stuttgart
Pfaffenwaldring 5b, 70569 Stuttgart, Germany
jagfelga@ims.uni-stuttgart.de

Lonneke van der Plas
Institute of Linguistics
University of Malta
Tal-Qroqq, Msida, Malta
lonneke.vanderplas@um.edu.mt

Abstract

We propose a way to automatically improve the annotation of verbal complex predicates in PropBank which until now has been treating language mostly in a compositional manner. In order to minimize the manual re-annotation effort, we build on the recently introduced concept of *aliasing* complex predicates to existing PropBank rolesets which encompass the same meaning and argument structure. We suggest to find aliases automatically by applying a multilingual distributional model that uses the translations of simple and complex predicates as features. Furthermore, we set up an annotation effort to obtain a frequency balanced, realistic test set for this task. Our method reaches an accuracy of 44% on this test set and 72% for the more frequent test items in a lenient evaluation, which is not far from the upper bounds from human annotation.

1 Introduction

Semantic Role Labeling (SRL) aims at determining 'who' did 'what' to 'whom' in sentences by identifying and associating predicates with their semantic arguments. This information is useful for many downstream applications, for example for question answering (Shen, 2007). The PropBank corpus (PB) (Palmer et al., 2005) is one of the most widely used resources for training SRL systems. It provides senses of (mostly verbal) predicates with their typical semantic arguments annotated in a corpus and accompanied by a lexical resource. The sense of a predicate is referred to as a 'roleset' because it lists all required and possible semantic roles for the predicate used in a specific sense.

The 12K rolesets in PB describe mostly single word predicates, to a great part leaving aside multiword expressions (MWEs). Complex predicates (CPs), 'predicates which are multi-headed: they are composed of more than one grammatical element' (Ramisch, 2012), are most relevant in the context of SRL. Light verb constructions (LVCs), e.g. *take care*, and verb particle constructions (VPCs), e.g. *watch out*, are the most frequently occurring types of CPs. As Bonial et al. (2014) stated 'PB has previously treated language as if it were purely compositional, and has therefore lumped the majority of MWEs in with lexical verb usages'. For example the predicates in the CPs *take a hard line, take time* and many others are all annotated with a sense of *take*, meaning *acquire, come to have, chose, bring with you from somewhere*. This results in a loss of semantic information in the PB annotations.

This is especially critical because CPs are a frequent phenomenon. The Wiki50 corpus (Vincze et al., 2011), which provides a full coverage MWE annotation, counts 814 occurrences of LVCs and VPCs in 4350 sentences. This makes for one CP in every fifth sentence.

Recently, Bonial et al. (2014) have introduced an approach to improve the handling of MWEs in PB while keeping annotation costs low. The process is called *aliasing*. Instead of creating new frames for CPs, human annotators map them to existing PB rolesets which encompass the same semantic and argument structure. For example, the CP *give (a) talk* could be mapped to the alias *lecture.01*. While this

Proceedings of NAACL-HLT 2015 Student Research Workshop (SRW), pages 33–39,
Denver, Colorado, June 1, 2015.

method significantly reduces the effort to create new rolesets, the time consuming manual mapping is still required. To address this problem, our work extends this approach by proposing a method to find the aliases automatically.

One way to find the most suitable alias roleset for a given CP is to group predicates by their rolesets assigned by an automatic SRL system and compute the most similar roleset group by searching for (near-) synonymous predicates of the CP. The roleset of the most similar roleset group is selected as alias for the CP.

Finding synonyms, both single-word and multiword, from corpora has been done successfully with the multilingual variant of the distributional hypothesis (Van der Plas and Tiedemann, 2006; Van der Plas et al., 2011). The idea behind this approach is that words or MWEs that share many translations are probably synonymous. We use the word alignments in a parallel corpus to find the translations of CPs and single predicates. The predicates are automatically annotated with rolesets by an SRL system. This allows us to compute the most suitable roleset for a given CP fully automatically.

Our contributions are as follows: To the best of our knowledge, this work is the first to address the handling of CPs for SRL in an automatic way. We are thus able to scale up previous work that relies on manual intervention. In addition, we set up an annotation effort to gather a frequency-balanced, data-driven evaluation set that is larger and more diverse than the annotated set provided by Bonial et al. (2014).

2 Representing CPs for SRL

Previous work on representing CPs for SRL has mostly focused on PB. The currently available version of the PB corpus represents most CPs as if they were lexical usages of the verb involved in the predicate. Figure 1 shows an example for the annotation of the LVC *take care* in PB.[1] The CP is split up into its two components that are each assigned their own roleset. This annotation ignores the semantic unity of the CP and is unable to capture its single meaning of *being concerned with* or *caring for* something.

[1]We show an excerpt of the original sentence found in the currently available version of PB (Proposition Bank I).

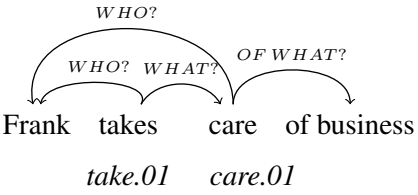

Figure 1: Current PB representation of the CP *take care*

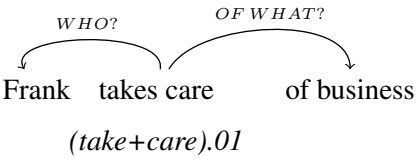

Figure 2: Improved representation of the CP *take care* adopted from (Hwang et al., 2010; Duran et al., 2011)

In contrast to this, Hwang et al. (2010) suggest a new annotation scheme for LVCs that assigns the argument structure of the LVC independently from the argument structure of its components. First, the arguments of the light verb and true predicate are annotated with roles regarding their relationship to the *combination* of the light verb and true predicate. Then, the light verb and predicate lemmas are joined into a single predicate. The result of this process is shown in Figure 2.

Duran et al. (2011) discuss the analysis of Brazilian Portuguese CPs. Similarly to Hwang et al. (2010) they argue that CPs should be treated as single predicates, not only for LVCs but for all CPs. They automatically extract CP candidates from a corpus and represent, if possible, the meaning of the CPs with one or more single-verb paraphrases.

Atkins et al. (2003) describe a way in which LVCs can be annotated in FrameNet (Baker et al., 1998), a framework that describes the semantic argument structure of predicates with semantic roles specific to the meaning of the predicate. In contrast to the proposals for PB by Hwang et al. (2010) and Duran et al. (2011), they suggest to annotate the light verb and its counterpart separately.

The *aliasing* process introduced by Bonial et al. (2014) tries to extend the coverage of PB for CPs while keeping the number of rolesets that should be newly created to a minimum. Bonial et al. (2014) conducted a pilot study re-annotating 138 CPs involving the verb *take*. As a first step, the annotators

determined the meaning(s) of the CP by looking at their usage in corpora. If they found that the CP is already adequately represented by the existing rolesets for *take*, no further action was needed (18/138). Otherwise, they were instructed to propose as alias an existing PB entry that encompasses the same semantics and argument structure as the CP (100/138). If unable to find an alias, they could suggest to create a new roleset for this CP (20/138). Expressions for which the annotators were unable to determine the meaning were marked as idiomatic expressions that need further treatment (4/138).[2]

According to this process, *take care* could be aliased to the existing PB roleset *care.01* whose entry is shown in Figure 3. This alias replaces *(take+care).01* shown in Figure 2 and thus avoids the creation of a new roleset.

Roleset id: care.01, *to be concerned*
Arg0: carer, agent
Arg1: thing cared for/about

Figure 3: alias PB roleset for the predicate *take care*

Encouraged by the high proportion of CPs that could successfully be aliased in the pilot study by Bonial et al. (2014), we created a method to automatically find aliases for CPs in order to decrease the amount of human intervention, thereby scaling up the coverage of CPs in PB.

3 Method

The task of finding aliases for CPs automatically is related to finding (near-) synonymous predicates and their accompanying roleset for the CPs. To find the near-synonyms, we apply the distributional hypothesis which states that we can assess the similarity of expressions by looking at their contexts (Firth, 1957). As previous work (Van der Plas and Tiedemann, 2006) has shown that multilingual contexts work better for synonym acquisition than monolingual syntactic contexts, we use the translations of the CPs and other predicates to all 20 languages available via the word alignments in a multilingual parallel corpus as context.

Figure 4 shows an overview of the architecture of

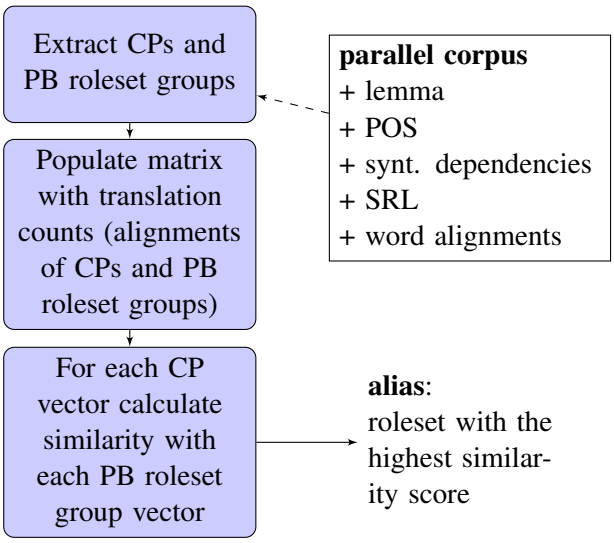

Figure 4: Overview of the alias finder

our system. First, we extract the CPs and all predicates that share a PB roleset (PB roleset groups) from the parallel corpus. For example, all verbs that were assigned to the roleset *care.01* by the SRL system belong to the PB roleset group of *care.01*. The CPs stem from the gold standard MWE annotation in the Wiki50 corpus (Vincze et al., 2011). We parsed this corpus to obtain lemmas, POS and syntactic dependencies and extracted this information for all VPCs and LVCs annotated in the corpus.[3] Figure 5 shows the two patterns we identified that the majority of the CPs followed.[4] We used these two patterns to search for occurrences of the CPs in Europarl.

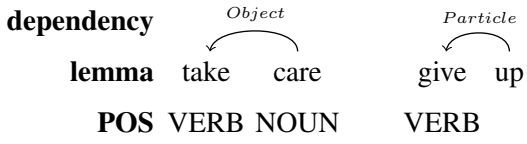

Figure 5: Patterns used for finding occurrences of CPs

Next, we build a co-occurrence matrix containing as head terms the CP and all PB roleset groups found in the parallel corpus. Figure 6 shows a toy example of such a matrix for the CP *take care*. The

[2]Note that the numbers do not add up to 138 because four MWEs obtained two different strategies.

[3]We concentrate on VPCs and LVCs because they are the most frequent types of CP in English.

[4]Here we use the example CPs *take care* and *give up*, but the lemmas were of course introduced as variables.

head words are listed in the rows, the translations (i.e. features) in the columns. Note that in contrast to previous work on distributional semantics we include PB roleset groups as head words. These contain several distinct verbal predicates but they share the same sense. Consequently, polysemous verbs are found in several distinct PB roleset groups.

	ter cuidado (es)	achten (de)	prendre soin (fr)	penser a (fr)
take care	3	3	5	0
care.01	4	3	7	1
think.01	0	2	1	6

Figure 6: Toy example co-occurrence matrix

Finally, we measure the similarity between CPs and roleset groups using the cosine similarity because it worked best in previous experiments for finding synonyms (Van der Plas, 2008). This results in a similarity ranking of PB roleset groups for each CP, from which we select the roleset with the highest cosine value as alias.

4 Experiments

4.1 Tools and Data

We processed the English section of the Europarl corpus (Koehn, 2005) (about 2 million sentences) with the MATE tools (Björkelund et al., 2010) to obtain lemmas, part-of-speech (POS) tags, dependency structures and semantic role labels. These annotations are used to find occurrences of the CPs and words assigned with PB rolesets in the English part. The word alignments produced with the grow-diag-final-and-heuristics (Koehn et al., 2003) provided by the OPUS project (Tiedemann, 2012) are then used to find their alignments to all other 20 languages in the corpus and exploited as features in the distributional model.

4.2 Evaluation Framework

Human Annotation. In order to evaluate our system, we set up an annotation effort loosely following the guidelines provided by Bonial et al. (2014). We selected 50 LVCs and 50 VPCs from the Wiki50 corpus (Vincze et al., 2011) divided equally over two frequency groups: Half of the expressions occur only once in the Wiki50 corpus (low-frequency subgroup) and the other half occur at least twice (high-frequency subgroup). All occurrences of these 100 CP types in the corpus were selected to account for the polysemy of CPs. Different instances of the same CP could get assigned to different aliases. This resulted in a total of 197 annotated instances.

Four annotators were presented with the CP in their original sentence context and were asked to propose one or several PB aliases which encompass the same meaning and argument structure. One annotator (A, one of the authors of this article) labeled the whole set of 100 expressions. The three other annotators (B,C,D) each labeled one third of the expressions assigned randomly, so that every expression was annotated by two annotators.

First, they were asked to decide if there is already an appropriate PB roleset for the CP and then provide it. The annotators were requested to divide these cases into semantically compositional CPs (e.g. obtain permission with the roleset *obtain.01*) and uncompositional CPs for which PB already provides a multi-word predicate (e.g. *open.03* for *open_up*). For the remaining CPs, they were asked to suggest PB rolesets (aliases) that share the same semantics and argument structure as the CP.

The simple inter-annotator agreement[5] was 67% for annotator A%B, 51% for A&C and 44% for A&D. These agreement figures are higher than the figures in Bonial et al. (2014), and actual agreement is probably even higher, because synonymous rolesets are regarded as disagreements. Annotator A discussed the annotations with the other annotators and they were able to reach a consensus that resulted in a final agreed-upon test set.

Table 1 shows the final decisions with respect to the complete set of 197 expressions. In line with the results from Bonial et al. (2014) who aliased 100 out of 138 uncompositional *take* MWEs, we were also able to alias most of the CPs in our annotation set.

The final Wiki50 set consists of 154[7] instances of

[5]Kappa scores (Cohen, 1960) are not suited to the present multi-label and multi-class setting: Annotators could choose from roughly 6K classes and were encouraged to provide multiple synonymous rolesets.

[6]Discarded CPs contained spelling or annotation errors in the Wiki50 corpus.

[7]We removed two CPs from the 'aliased' group because our extraction patterns do not cover LVCs formed with an adjective.

Decision	Count	MWE example
aliased	96	take part
multi-word PB pred.	60	open up
compositional	18	obtain permission
no alias found	16	go into politics
discarded[6]	7	take conrol

Table 1: Final decisions on the 197 annotated expressions

CPs from the categories 'aliased' and 'multi-word PB predicate' (low-frequency: 34, high-frequency: 120). The latter were included because the predicted roleset of the SRL only coincides with the gold standard for 23 out of 60 instances. This means that for the majority of the CPs, even if an adequate PB roleset exists, this roleset was not selected by the SRL system. We hope to also improve these cases with our method. All CPs were labeled with one to four appropriate PB alias rolesets.

In addition, we evaluated our system on the dataset from Bonial et al. (2014), restricted to the type of CP our system handles (LVCs and VPCs) and verb aliases (as opposed to aliases being a noun or adjective roleset). We used 70 of the 100 MWEs from their annotations.

Evaluation Measures and Baseline. We report the accuracy of our system's predictions as compared to the gold standard. For the STRICT ACCURACY, an alias is counted as correct if it corresponds exactly to one of the gold aliases. This evaluation is very rigid and regards synonymous rolesets as incorrect. Thus, we also compute a more LENIENT ACCURACY, which counts an alias as correct if it belongs to the same VerbNet (Kipper-Schuler, 2006) verb class as the gold alias. VerbNet (VN) is a hierarchically organized lexicon of English verbs. It consists of syntactically and semantically coherent verb classes, which are extensions of the classes proposed by Levin (1993). For the PB-VN mappings, we rely on the resource provided by the SemLink project[8] (Loper et al., 2007) and use the most-specific (deepest) layer of the verb classes. Since the mapping provided in SemLink is not complete (only 58% of the rolesets found in PB have a mapping to a corresponding VN class), we discard rolesets that are not found in SemLink, unless they are correct according to the gold standard in the first place.

We compared our system with a baseline system that distinguishes between VPCs and LVCs. For VPCs, it checks whether there exists a PB multi-word predicate for the expression and selects the first roleset of that predicate (e.g. there exists a predicate called *open_up (open.03)* for the VPC 'open up'). For LVCs, it checks whether the noun has a corresponding verb predicate in PB and selects the first roleset of this predicate (e.g. *walk.01* for *take a walk*). Note that this is an informed baseline that is very hard to beat and only fails in case of lack in coverage.

5 Results and Discussion

We evaluated our approach on the 160 CPs annotated in the course of this work (Wiki50 set), as well as on the 70 *take* CPs from Bonial et al. (2014) (*take* set) and compare our results to the baseline. Table 2 shows percentage coverage, accuracy and the harmonic mean of coverage and accuracy for our system and the baseline. We report results on the two evaluation sets in the strict and lenient evaluation.

The first five rows of Table 2 show the results for the Wiki50 set and its subsets. We see that our system scores 44.1 accuracy on the whole test set in the strict evaluation and 69.0 in the lenient evaluation. These numbers seem quite low, but they are not that far apart from the micro averaged IAA from our annotation effort (53%). Our system outperforms the baseline with very high coverage numbers. It beats the baseline in terms of the harmonic mean for all subsets except the multiword PB predicate subset. This is not surprising as the test items in this subset have a corresponding multiword PB predicate and all the baseline has to do is select the right sense. The high performance of the baseline on the multiword PB predicates leads to the high accuracy numbers for the baseline in all (sub-)sets except from the alias subset, which contains the expressions for which a true alias was provided. Our system beats the baseline in terms of strict accuracy for the alias subset. This is good news because the actual task is to find new aliases for CPs that are not covered in PB. The performance on the low-frequency subset is lower than on the high-frequency subset, as expected for a distributional method.

[8] http://verbs.colorado.edu/semlink/

Set	Strict Cov	Strict Acc	Strict Hm	Lenient Cov	Lenient Acc	Lenient Hm
Wiki50 all	**98.7** (65.6)	44.1 (**54.5**)	**60.9** (59.5)	**98.0** (59.5)	69.0 (**85.9**)	**81.0** (70.3)
alias	**98.9** (50.0)	**36.6** (34.0)	**53.4** (40.5)	**98.4** (40.5)	60.0 (**68.8**)	**74.5** (51.0)
mw. PB pred.	**98.3** (86.7)	55.9 (**71.2**)	71.3 (**78.1**)	**97.6** (84.6)	82.5 (**97.7**)	89.4 (**90.7**)
high-freq.	**100.0** (68.3)	45.0 (**52.4**)	**62.1** (59.3)	**100.0** (62.7)	72.0 (**84.4**)	**83.7** (72.0)
low-freq.	**94.1** (50.0)	40.6 (**58.5**)	**56.8** (54.1)	**92.6** (41.4)	60.0 (**91.7**)	**72.8** (57.0)
take	67.1 (**71.4**)	25.5 (**32.0**)	37.0 (**44.2**)	56.6 (**64.9**)	**60.0** (45.0)	**58.3** (53.8)

Table 2: Percentage coverage (Cov), accuracy (Acc) and the harmonic mean (Hm) of coverage and accuracy of the predicted aliases in the Wiki50 set (+ four of its subsets) and the *take* set; The results of the baseline are in brackets

The results on the *take* set are shown in the last row of Table 2. Compared to the Wiki50 set, they are substantially lower. We would like to stress that the *take* set is far from what we expect to find in an actual corpus. This set comprises only CPs that contain the word *take*. Many test items have been extracted from WordNet and possibly have a very low frequency in a general corpus. This is reflected in the coverage number, which shows the proportion of CPs for which our system was able to suggest at least one alias: It is above 94% for all Wiki50 (sub)sets, but only 67% for the take set. We constructed the Wiki50 set to allow us to get a better estimate of how our method would fare in a natural setting.

5.1 Error analysis

We examined all expressions from the full Wiki50 set for which the top ranked predicted alias was incorrect. Due to space limitations we only mention the main reasons for errors we identified. First of all, the limited language domain of the Europarl corpus caused a low frequency of some rolesets selected as gold alias, like *fuse.01* ('melt into lump') for the VPC *melt down*. This problem could be solved by adding more parallel data from different domains.

Another source of errors is the fact that our approach requires the output of an SRL system which, in turn, we want to improve. For 45 out of 160 CPs our system suggested the roleset as alias that was assigned to the verb by the SRL system, e.g. *leave.02* for *leave for*. But the automatically attributed roleset is only correct in 21 cases, which means that we reproduced the errors of the SRL in 24 cases.

Some LVCs keep their light verb structure in other languages, i.e. they receive multi-word translations. This diminishes the overlap of translations between the LVC and the PB roleset groups. PB rolesets are assigned to simplex verbs and therefore predominantly receive simplex translations. As more frequent rolesets have more diverse translations that contain more MWEs, these are promoted as aliases. Applying frequency weights to the roleset matrix could remedy this problem.

Lastly, our system adheres to the most frequent sense baseline due to lack of word sense disambiguation of the CPs and assigns the alias that fits the most dominant sense of the CP in the corpus.

6 Conclusions

We have presented an approach to handle CPs in SRL that extends on work from Bonial et al. (2014). We automatically link VPCs and LVCs to the PB roleset that best describes their meaning, by relying on word alignments in parallel corpora and distributional methods. We set up an annotation effort to gather a frequency-balanced, contextualized evaluation set that is more natural, varied and larger than the pilot annotations provided by Bonial et al. (2014). Our method can be used to alleviate the manual annotation effort by providing a correct alias in 44% of the cases (up to 72% for the more frequent test items when taking synonymous rolesets into account). These results are not too far from the upper bounds we calculate from human annotations.

In future work, we would like to improve our method by incorporating the methods discussed in the error analysis section. Additionally, we plan to evaluate the impact of the new CP representation on downstream applications by retraining an SRL system on the new annotations.

Acknowledgments

We thank Anna Konobelkina and two anonymous annotators for their efforts as well as the anonymous reviewers.

References

Sue Atkins, Charles J. Fillmore, and Christopher R. Johnson. 2003. Lexicographic relevance: selecting information from corpus evidence. *International Journal of Lexicography*, 16.3.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, Stroudsburg, PA, USA.

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Coling 2010: Demonstrations*, Beijing, China.

Claire Bonial, Meredith Green, Jenette Preciado, and Martha Palmer. 2014. An approach to take multi-word expressions. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, Gothenburg, Sweden.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1).

Magali Sanches Duran, Carlos Ramisch, Sandra Maria Aluísio, and Aline Villavicencio. 2011. Identifying and analyzing brazilian portuguese complex predicates. In *Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World*, MWE '11, Stroudsburg, PA, USA.

John Rupert Firth. 1957. A synopsis of linguistic theory 1930-55. 1952-59.

Jena D. Hwang, Archna Bhatia, Clare Bonial, Aous Mansouri, Ashwini Vaidya, Nianwen Xue, and Martha Palmer. 2010. Propbank annotation of multilingual light verb constructions. In *Proceedings of the Fourth Linguistic Annotation Workshop*, LAW IV '10, Stroudsburg, PA, USA.

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

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, Stroudsburg, PA, USA.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the Tenth Machine Translation Summit*, Phuket, Thailand.

Beth Levin. 1993. *English verb classes and alternations: a preliminary investigation*. University of Chicago Press, Chicago and London.

Edward Loper, Szu-Ting Yi, and Martha Palmer. 2007. Combining lexical resources: Mapping between propbank and verbnet. In *Proceedings of the 7th International Workshop on Computational Linguistics*, Tilburg, the Netherlands.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics Journal*, 31(1).

Carlos Ramisch. 2012. *A generic and open framework for multiword expressions treatment: from acquisition to applications*. Ph.D. thesis, University of Grenoble (France) and Federal University of Rio Grande do Sul (Brazil).

Dan Shen. 2007. Using semantic role to improve question answering. In *Proceedings of EMNLP 2007*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Lonneke van der Plas and Jörg Tiedemann. 2006. Finding synonyms using automatic word alignment and measures of distributional similarity. In *Proceedings of ACL-COLING 2006*, Sydney, Australia.

Lonneke van der Plas, Jörg Tiedemann, and Ismail Fahmi. 2011. Automatic extraction of medical term variants from multilingual parallel translations. In *Interactive Multi-modal Question Answering*, Theory and Applications of Natural Language Processing. Springer-Verlag, Berlin.

Lonneke van der Plas. 2008. *Automatic lexico-semantic acquisition for question answering*. Ph.D. thesis, University of Groningen.

Veronika Vincze, István Nagy T., and Gábor Berend. 2011. Multiword expressions and named entities in the wiki50 corpus. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, Hissar, Bulgaria.