Syntactic-Semantic Classes of Context-Sensitive Synonyms Based on a Bilingual Corpus

Zdeňka Urešová, Eva Fučíková, Eva Hajičová, Jan Hajič

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranske nam. 25
11800 Prague, Czech Republic
{uresova,fucikova,hajicova,hajic}@ufal.mff.cuni.cz

Abstract

This paper summarizes first findings of a three-year study (an ongoing research project) on verb synonymy based on both syntactic and semantic criteria. Primary language resources used for the study are existing lexical and corpus resources, namely the Prague Dependency Treebank-style valency lexicons, FrameNet, VerbNet, PropBank and Czech and English WordNets and the parallel Prague Czech-English Dependency Treebank, which contains deep syntactic and partially semantic annotation of running texts. The resulting lexicon, called CzEngClass, and all associated resources linked to the existing lexicons and corpora resulting from this project will be made publicly and freely available. While the project proper assumes manual annotation work, we expect to use the resulting resource (together with the existing ones) as a necessary resources for developing automatic methods for extending such a lexicon, or creating similar lexicons for other languages.

1. Introduction

The goal of the project is to group verbs used as synonyms in Czech and English into (cross-lingual) synonym classes. For the purpose of this work, we use the term "synonym" in the "loose" interpretation (Lyons, 1968), i.e., the necessary semantic equivalence takes also wider context into account. The novel feature is the use of a richly annotated bilingual corpus to get more insight into the usage of verbs (together with their arguments) in translation. In the present paper, initial results are discussed based on a sample of 60 classes manually processed and linked to the existing resources, the relevant features of which are also described.

While not being the goal of this very project, the ultimate use of such resource is both for followup linguistic studies and for use in natural language applications. The resulting lexicon, together with the existing resources to which it will be linked, will be used as a "gold standard" for evaluating automatic methods that should mimic the laborious manual work performed in this project (and possibly also as training data for systems based on deep learning, depending on its final extent). That way, it will serve as a seed resource for future, automatically extracted, cross-lingual lexicons with the same properties.

2. Resources Used

2.1. Lexical Resources

- PDT-Vallex (Czech) is a Czech valency lexicon used for the annotation of the Prague Dependency Treebank (Hajič et al., 2006) family of treebanks (Hajič et al., 2003; Urešová, 2011)[1] based on the Functional Generative Description theory (Sgall et al., 1986). PDT-Vallex contains 7,121 verbs structured into 11,933 valency frames (verb senses), and it is available as part of the PDT 2.0 distribution.[2]
- EngVallex (English)[3] is an English valency lexicon with 7,148 valency frames for 4,337 verbs, using the same valency framework as PDT-Vallex. It was built by a (largely manual) adaptation of the PropBank Lexicon (Palmer et al., 2005) to the PDT labeling standards and principles (Cinková, 2006).
- CzEngVallex (Czech-English) (Urešová et al., 2016) is a Czech-English bilingual valency lexicon. It contains 20,835 explicitly linked verb senses (frame-to-frame pairs) and their aligned arguments (argument-to-argument pairs). It is linked, entry by entry and frame by frame, to the Prague Czech-English Dependency Treebank (Hajič et al., 2012)[4] and to the two monolingual valency lexicons mentioned above: PDT-Vallex and EngVallex.
- Berkeley FrameNet (English) (Baker et al., 1998; Ruppenhofer et al., 2006) is a lexical database of English[5], containing about 13,000 word senses from more than 200,000 manually annotated sentences linked to more than 1,200 Semantic Frames. FrameNet is based on the Frame Semantics theory (Fillmore, 1976; Fillmore, 1977); each lexical unit evokes a Semantic Frame (SF) which lists relevant Frame Elements (FEs), or Semantic Roles (SRs).
- VerbNet (English) (Schuler, 2006; Duffield et al., 2007; Kipper et al., 2006) is a class-based verb lex-

[1] https://lindat.mff.cuni.cz/services/PDT-Vallex/

[2] http://www.ldc.upenn.edu/LDC2006T01

[3] http://hdl.handle.net/11858/00-097C-0000-0023-4337-2

[4] PCEDT 2.0 is available from http://hdl.handle.net/11858/00-097C-0000-0015-8DAF-4)

[5] https://framenet.icsi.berkeley.edu

icon[6] with mappings to other lexical resources such as WordNet or FrameNet. VerbNet contains syntactic and semantic information on English verbs. It extended Levin (Levin, 1993) verb classes by refinement and addition of subclasses (Kipper et al., 2006). Each verb class is described by thematic roles, selectional restrictions on the arguments, and frames. Currently, VerbNet contains about 5,257 verb senses structured in 274 classes.

- PropBank (English) (Palmer et al., 2005) is not only a lexicon but also a corpus[7] of one million words of English text, annotated with argument role labels for verbs (113,000 tokens, 3,324 frames files/types). Arguments are linked to their semantic roles (Palmer et al., 2005).
- SemLink (English) (Palmer, 2009)[8] links together different lexical resources (PropBank, VerbNet, FrameNet) through sets of mappings. The Semlink lexicon can be browsed online using the Unified Verb Index.[9]
- WordNet(s) (Miller, 1995; Fellbaum, 1998) is a semantic network[10] of English. Words (nouns, verbs, adjectives, adverbs) are hierarchically grouped into sets of synonyms (117,000 synsets). Each synset contains word forms (referring to a given concept), a definition gloss and an example sentence. Czech WordNet 1.9[11] (Pala and Smrž, 2004) will be used in future work when extending the classes on the Czech side.

2.2. Corpus Resources

The Prague Czech-English Dependency Treebank (PCEDT)[12] is a parallel treebank with over 1.2 million tokens in almost 50,000 sentences for each side. The PCEDT is based on the texts of the Wall Street Journal part of the Penn Treebank[13] and their manual translations. Each language part is annotated in the Prague Dependency Treebank style; the annotation is dependency-style with argument structure of verbs (syntactic and semantic labeling), which corresponds to the associated valency lexicons for both languages: the PDT-Vallex (for Czech) and the EngVallex (for English); see Sect. 2.1.

In addition, we also use various monolingual corpora, such as the COCA corpus[14], corpora avalilable in the SketchEngine[15] and corpora accessible and searchable through the Kontext tool in the LINDAT/CLARIN repository.[16]

[6] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

[7] http://propbank.github.io/

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

[9] http://verbs.colorado.edu/verb-index/

[10] https://wordnet.princeton.edu/

[11] http://hdl.handle.net/11858/00-097C-0000-0001-4880-3

[12] http://ufal.mff.cuni.cz/pcedt2.0/en/index.html

[13] https://catalog.ldc.upenn.edu/LDC99T42

[14] http://corpus.byu.edu/coca/

[15] https://www.sketchengine.co.uk/

[16] http://lindat.mff.cuni.cz/services/kontext/

3. Structure of CzEngClass Lexicon

As part of the preliminary study, a structure of the resulting lexicon (Fig. 1) has been designed.

The CzEngClass lexicon builds upon the existing resources, as described above: CzEngVallex, PDT-Vallex and EngVallex lexicons and the PCEDT parallel corpus. In addition, the other lexicons listed (FrameNet, VerbNet, PropBank and WordNet) are used as additional sources, and links will be kept between their entries and the CzEngClass entries.

At the core of the CzEngClass lexicon, there are Synonym Classes, which are, for the purpose of this project, defined as (multilingual, or rather cross-lingual)[17] groups of verb senses (of different lexemes/words) that have the same meaning *and* the arguments of which can be mapped to a common set of semantic roles.[18] The term "same meaning" is understood with regard to a context, the relevant information about which is expected (at least in some cases) to be part of the argument mapping in the form of certain restrictions (lexical, syntactic, semantic) put on the arguments or even to a wider context. This is the case of most light verb constructions (*hold a meeting - meet*), idiomatic verbal MWEs (*cut loose - sever*), in some cases clear cases of hyperonymy (*pay [back] - repay*), and more (e.g. *return [call] - call back*), with clear patterns emerging.

4. Data Preparation

The work so far has been done in several steps. First, we have randomly selected a set of 200 Czech verbs (verb senses) from three categories based on their frequency in the PCEDT corpus (high, medium, low). We have used the bilingual valency lexicon CzEngVallex (Urešová et al., 2016) to determine a set of candidate English verbs for one synonym class, based simply on their parings with the original Czech verb. Since CzEngVallex is linked to the PCEDT corpus, this gave us also a set of usage examples of these verb pairs, i.e., the context in which they have been used in the the original English and in its Czech translation.

5. Annotation

5.1. Sense Determination

We have linked each English verb in the initial sample to the VerbNet sense as available from VerbNet "Groupings", in order to get more precise (even if not unique) links to the corresponding VerbNet entry(ies), PropBank id(s), FrameNet frame(s) and WordNet synset(s).[19] For example, for the English verb *set up* in a group extracted from the translation pairs linked to the Czech verb *budovat* (lit. *build*), VerbNet sense No. 4 of `set-v` has been assigned ("prepare (something) for a particular purpose"),

[17] For the time being, bilingual: in Czech and English.

[18] The term "sense" is used here for the differentiation of a single verb lexeme ("word") into one or more senses, represented technically by its valency frame ID, as it si done in the original valency lexicons (PDT-Vallex and EngVallex).

[19] Using the Unified Verb Index, http://verbs.colorado.edu/verb-index/

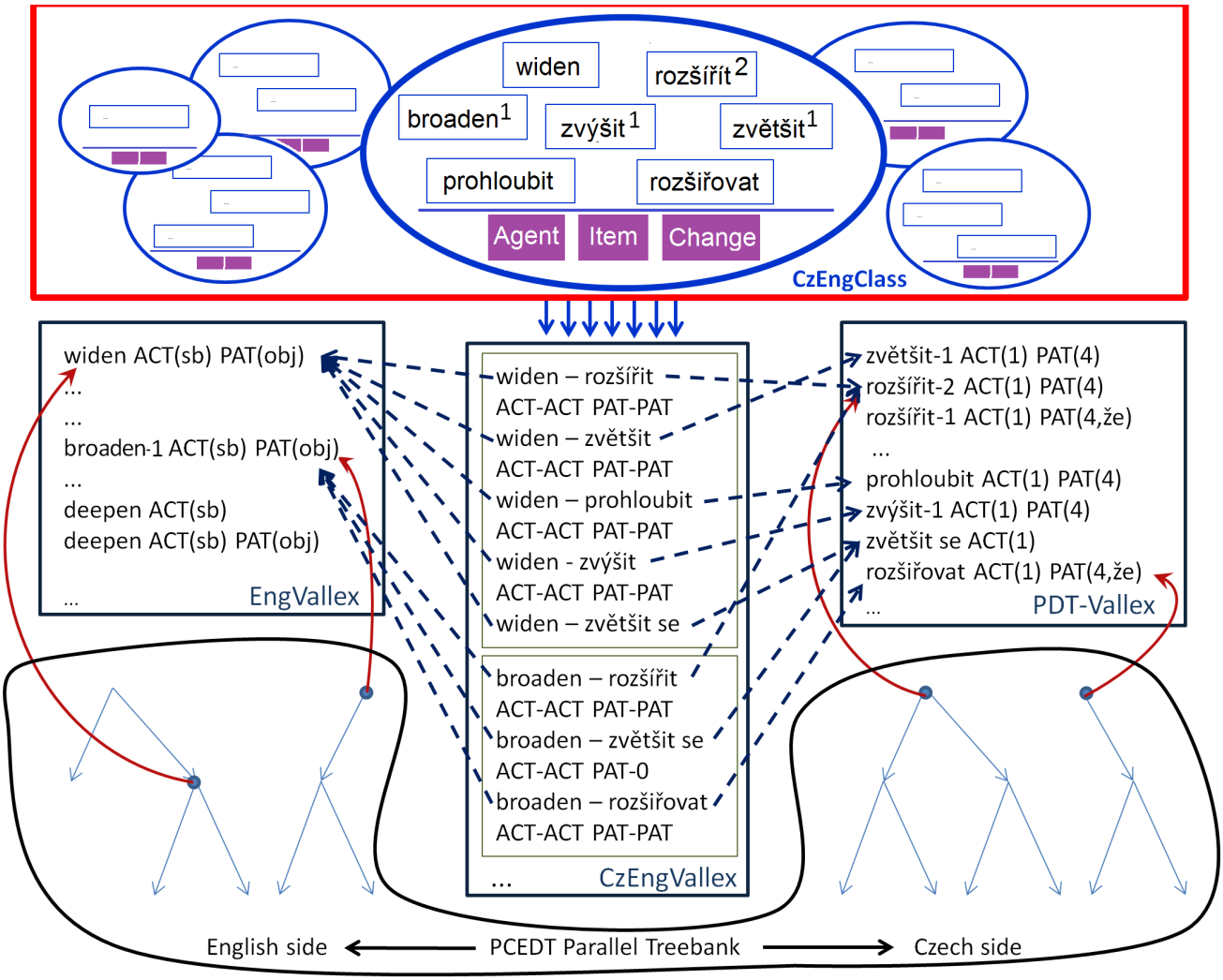

Figure 1: CzEngClass lexicon & relation to core resources

attaching it to FrameNet frames ARRANGING and INTENTIONALLY_CREATE, VerbNet classes `braid-41.1.2` and `preparing-26.3-2`, PropBank ids `set.03` and `set.08` and WordNet senses `6, 7, 21, 22, 25` (of *set*).

5.2. Common Semantic Roles (SRs)

In the next phase, we have started devising a common set of SRs for each group (candidate synonym class) and mapping them to the valency frame slots from the PDT-Vallex and EngVallex lexicons for the Czech source verb and all the English verbs. In devising these roles, we have used FrameNet's FE labels and descriptions as the initial pool of roles.[20] Many of the verbs (verb senses) in our candidate synonym sets have been found in FrameNet, so that we have started with the core FEs in the frame(s) associated with these verb senses (to be pruned later during a reconciliation phase also with VerbNet thematic roles). For example, for the *set up* example, we have listed `Agent`, `Configuration` and `Theme`, `Creator` and `Created_Entity` from the ARRANGING and INTENTIONALLY_CREATE frames.

[20]Using FrameNet v1.7, there are 1,168 different FE labels available across all frames. Later, VerbNet's thematic roles will be compared with the selected FEs and a common set used, provided a suitable common theoretical framework can be found.

5.3. Argument - Role Mapping

For each verb in the candidate class, we have then paired each of the the FEs (roles) to a valency slot as found in PDT-Vallex (for the Czech verb) and to EngVallex (for the English verbs). This has not, as expected, been straightforward, as will be described in more details in the next section; we have also used the other English resources to help clarify the relations if necessary.

6. Initial Findings

This section is based on a (sub)sample of 60 candidate synonym groups that have been created and mapped so far during the annotation process.

6.1. Synonym Classes Composition

While the translation pairs extracted from the parallel corpus should have been clear synonyms, in some cases, even if the particular context has been taken into account, verbs had to be deleted from the group. For example,

sometimes the parallel corpus correctly identified hyperonyms as translation equivalents, but there was no specific context that would restrict the hyperonym to the particular sense on the other side of the translation (this has happened in both directions):

- *That would hold*.PRED *spending*.PAT *on the program at about the previous year's level.*
- *To by znamenalo*.PRED *investice*.PAT *do programu přibližně ve stejné výši jako loni.*

In the above example, the Czech verb *znamenat* (form *znamenalo*) (in the sense of lit. *mean*, *imply*, *indicate*) is aligned with English *hold* (*... hold spending ... at about the same level*, Czech lit. *... mean spending ... etc.*). This is (almost) an equivalent translation in this context,[21] but since the context cannot be described just in terms of verb arguments, it has been decided to delete *hold* from this synonym class.

6.2. Roles and Argument Mapping

The initial set of roles for each class has been a union of FrameNet's Frame Elements (FEs) of all frames in which the appropriate English verbs have been found (see also Sect. 5.3.). The goal was to establish a common set of roles for a given class, carefully considering both the FrameNet's FEs and the corresponding valency slots in the valency lexicons associated with the parallel corpus, including their use in the bilingual texts themselves. Merging FrameNet-provided FEs has been the most frequent operation, even within the same frame. For example, verbs inheriting from the STATEMENT frame might in some cases have merged the `Topic` FE (the subject matter to which the `Message` pertains) with the `Message` FE (what the `Speaker` is communicating to the `Addressee`)[22], since in our view, these typically occupy just one "slot" (Kettnerová, 2009), with `Topic` being often part of the `Message`. For example: *She said about her past* (`Topic`) *that it was wild* (`Message`) - *She said that her past was wild* (`Topic+Message`). Similarly, FEs differing only in animateness have often been merged, such as in the `Agent/Cause` case in the class represented by "widen" in Fig. 1.

The mapping of SRs to valency slots is mostly 1:1, as in the example of the synonym class corresponding most closely to the FrameNet's COMMERCE_PAY frame (Tab. 1): the arguments of *cover, pay, reimburse* are mapped 1:1 (if we tentatively add some of the missing ones into EngVallex, e.g. EFF/Effect to *reimburse* and EFF/Effect and BEN/Benefactor to *settle*). `Buyer` typically maps to the valency frame argument Actor, `Goods` to Effect, `Seller` to Addressee and `Money` to Patient.

However, the correspondence of SRs and valency arguments is not necessarily always 1:1 - SRs have been occasionally merged (cf. `Topic` and `Message`) or split. For example, in BECOMING_AWARE, `Phenomenon` is mapped either to valency frame argument Effect or to valency frame argument Patient as shown in the following example *...to know details*.Effect *of one side only*, where `Phenomenon` is mapped to Effect and for another class member of the same class, the verb *hear*, `Phenomenon` is mapped to Patient: *she heard about the artery-clogging hazards*.Patient. Similarly, the mapping of `Goods` in the *pay* class is either to Patient (for *cover*) or Effect for PAY and other verbs; see also Tab. 1).

There are also examples with some specific context restrictions when a mapping can be applied. E.g., for the idiomatic verb *foot [the bill]* of the PAY synonym class (Tab. 1), the restriction to this idiomatic meaning (using *bill*) must be recorded. As another example, the Patient mapping of the verb *drill* in the BUILDING-related synonym class must be restricted to *drill a well (or other [large] hole-like thing)*. For light verb constructions, the nominal argument often maps to the same role as the Patient argument does for non-light verbs.

	Roles			
	`Buyer`	`Goods`	`Seller`	`Money`
hradit	ACT	EFF	ADDR	PAT
cover	ACT	PAT	BEN?	MANN
foot	ACT	DPHR(*bill*)	BEN?	MANN
pay	ACT	EFF	ADDR	PAT
reimburse	ACT	EFF?	ADDR	PAT
settle	ACT	EFF?	BEN?	PAT

Table 1: Argument mapping for PAY class

7. Conclusions and Next Steps

We have described preliminary findings on synonymy of verb senses of generally different verbal lexemes in a bilingual setting, and specifically, we focused on their valency behavior and possibly common semantic roles.

Our future research will be aimed at extending it to more verbs, at further refinement of our semantic roles and their explicit mappings from valency arguments to the semantic roles, and at formalizing the additional restrictions. We will analyze in more detail the relation of valency and semantic roles also from their morphosyntactic realization point of view. We will also confront the findings as supported by the corpus material to the underlying theoretical framework(s), in order to possibly refine them in their approach to verb sense distinctions, valency and argument description. We will also compare our results with automatic approaches to cross-lingual semantic similarity detection, such as in (Wu et al., 2010), which is very much related to our work.

Finally, we plan to publish the resulting lexicon as an open source dataset.

Acknowledgments

This work has been supported by the grant No. GA17-07313S of the Grant Agency of the Czech Republic, and

[21] We can only speculate why the translator has used *znamenat* here; possibly because literal translations of *hold* are awkward in Czech (in this context), and the translator also determined that in fact the semantics of *hold* is already contained in the phrase *previous year's level*, and thus a translation of a hyperonym of *hold* can be used instead.

[22] `https://framenet2.icsi.berkeley.edu/fnReports/data/frameIndex.xml?frame=Statement`

it uses resources hosted by the LINDAT/CLARIN Research Infrastructure, project No. LM2015071, supported by the Ministry of Education of the Czech Republic. The last co-author has been supported by the project VIADAT, No. DG16P02R019, by the Ministry of Culture of the Czech Republic.

8. References

Baker, Collin F., 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: Association for Computational Linguistics.

Cinková, Silvie, 2006. From PropBank to EngValLex: Adapting the PropBank-Lexicon to the Valency Theory of the Functional Generative Description. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*. Genova, Italy: ELRA.

Duffield, Cecily Jill, Jena D. Hwang, Susan Windisch Brown, Dmitriy Dligach, Sarah E. Vieweg, Jenny Davis, and Martha Palmer, 2007. Criteria for the manual grouping of verb senses. In *Proceedings of the Linguistic Annotation Workshop*, LAW '07. Stroudsburg, PA, USA: Association for Computational Linguistics.

Fellbaum, Christiane (ed.), 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. Cambridge, MA: MIT Press.

Fillmore, Charles J., 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, 280(1):20–32.

Fillmore, Charles J., 1977. *Scenes-and-frames semantics*. Number 59 in Fundamental Studies in Computer Science. North Holland Publishing.

Hajič, Jan, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský, 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the 8th LREC 2012)*. Istanbul, Turkey: ELRA.

Hajič, Jan, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová, and Petr Pajas, 2003. PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In Erhard Nivre, Joakim//Hinrichs (ed.), *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9 of *Mathematical Modeling in Physics, Engineering and Cognitive Sciences*. Vaxjo, Sweden: Vaxjo University Press.

Hajič, Jan, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, Magda Ševčíková Razímová, and Zdeňka Urešová, 2006. *Prague Dependency Treebank 2.0*. Number LDC2006T01. Philadelphia, PA, USA: LDC.

Kettnerová, Václava, 2009. Konstrukce s rozpadem tématu a dikta v češtině (Constructions with Topic and Message Separation in Czech). *Slovo a slovesnost*, 70(3):163–174.

Kipper, Karin, Anna Korhonen, Neville Ryant, and Martha Palmer, 2006. Extending VerbNet with novel verb classes. In *Proceedings of LREC*, volume 2006.

Levin, B., 1993. *English Verb Classes and Alternations*. Chicago and London: The University of Chicago Press.

Lyons, John, 1968. *Introduction to Theoretical Linguistics*. Cambridge University Press.

Miller, George A., 1995. WordNet: A Lexical Database for English. *Commun. ACM*, 38(11):39–41.

Pala, Karel and Pavel Smrž, 2004. Building Czech Wordnet. 2004(7):79–88.

Palmer, Martha, 2009. Semlink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference*.

Palmer, Martha, Daniel Gildea, and Paul Kingsbury, 2005. The proposition bank: An annotated corpus of semantic roles. *Comput. Linguist.*, 31(1):71–106.

Ruppenhofer, Josef, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk, 2006. FrameNet II: Extended theory and practice. *Unpublished Manuscript*.

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

Sgall, Petr, Eva Hajičová, and Jarmila Panevová, 1986. *The meaning of the sentence in its semantic and pragmatic aspects*. Dordrecht: D. Reidel.

Urešová, Zdeňka, 2011. *Valence sloves v Pražském závislostním korpusu*. Studies in Computational and Theoretical Linguistics. Praha, Czechia: Ústav formální a aplikované lingvistiky.

Urešová, Zdeňka, Eva Fučíková, and Jana Šindlerová, 2016. CzEngVallex: a bilingual Czech-English valency lexicon. *The Prague Bulletin of Mathematical Linguistics*, 105:17–50.

Wu, Shumin, Jinho D. Choi, and Martha Palmer, 2010. Detecting Cross-lingual Semantic Similarity Using Parallel PropBanks. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas*, AMTA'10. Denver, CO.