Bilingual English-Czech Valency Lexicon Linked to a Parallel Corpus

Zdeňka Urešová Ondřej Dušek Eva Fučíková Jan Hajič Jana Šindlerová
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské nám. 25
CZ-11800 Praha 1
Czech Republic
{uresova,odusek,fucikova,hajic,sindlerova}@ufal.mff.cuni.cz

Abstract

This paper presents a resource and the associated annotation process used in a project of interlinking Czech and English verbal translational equivalents based on a parallel, richly annotated dependency treebank containing also valency and semantic roles, namely the Prague Czech-English Dependency Treebank. One of the main aims of this project is to create a high-quality and relatively large empirical base which could be used both for linguistic comparative research as well as for natural language processing applications, such as machine translation or cross-language sense disambiguation. This paper describes the resulting lexicon, CzEngVallex, and the process of building it, as well some interesting observations and statistics already obtained.

1 Introduction

The present paper describes a cross-language verbal valency mapping between Czech and English and the process of capturing it in an annotated language resource. The result thereof is our Czech-English verbal valency lexicon called CzEngVallex, which explicitly links corresponding verbal senses and their valency arguments. As this mapping is based on the parallel Prague Czech-English Dependency Treebank (PCEDT), which also contains monolingual valency annotation on each side, we are getting a powerful, real-text-based complex of interlinked resources for a comparative description of verb senses and their argument structure in the context of translation equivalents.

While having the aforementioned relations captured in an explicit way will help cross-language linguistic comparison studies, it will also serve as training and testing material for multilingual natural language processing applications, most notably machine translation in systems using deep analysis with semantic elements (such as argument and semantic role labeling).

We are not aware of similar work which links aligned valency lexicons to a parallel dependency treebank, even though the resources as such do exist: a Japanese–English lexicon is described in (Fujita and Bond, 2004b). Similar lexicons have been suggested by Dorr (1997), Uszkoreit (2002) or Baldwin et al. (1999). Fujita and Bond (2004a) suggest an automatic extraction of valency from plain bilingual lexicons, but no subjective evaluation of the valency entries themselves is given.

The overview of the aim of the project described here is given in Sect. 2. In Sect. 3, we introduce the basis for building CzEngVallex—the underlying parallel Prague Czech-English Dependency Treebank and the corresponding monolingual valency lexicons. The CzEngVallex lexicon itself and the process of annotating it is described in Sect. 4, and we conclude with Sect. 5.

2 Comparing Czech and English Valency

This idea of a bilingual valency lexicon linked to a treebank comes from an exploratory and theoretically-oriented project for comparison of valency behavior of Czech and English verbs, which, of course, needs an annotated corpus material. Generalizing over the collected data—several thousand aligned verbs, linked to tens of thousand corpus occurrences—should give us more insight into the basic patterns of cross-language relations.

2.1 Valency in the FGD

This project is based on the valency theory of the Functional Generative Description (FGD) (Sgall et al., 1986) and on its application to the Prague Dependency Treebank (PDT) annotation style (Hajič et al., 2006). In this dependency approach, valency is seen as the ability of some lexical items (in general, not only verbs) to select for certain complementations in order to form larger units of meaning (Panevová, 1974). The governing lexical unit then governs both the morphosyntactic properties[1] of the dependent elements and their semantic interpretation (roles). The number and form of the dependent elements

[1] Morphological properties of verb arguments, or rather constraints on their use specific to every verb/argument combination, are very prominently present in inflectional languages such as Czech.

Proceedings of LAW IX - The 9th Linguistic Annotation Workshop, pages 124–128,
Denver, Colorado, June 5, 2015.

constituting the valency structure of a given verb sense is represented by a *valency frame*, which is listed in a valency lexicon.

According to FGD, the valency relation is a part of deep syntax (*tectogrammatic layer* of linguistic description). Every head-dependent relation is labeled by a *functor* denoting the role of the dependent relative to its head. While the FGD describes two dimensions of valency complementation, we can simplify to say that each verb frame (for a given verb sense) contains both verb arguments as well as adjuncts. The main functors used for verb arguments are Actor/Bearer (ACT), Patient (PAT), Addressee (ADDR), Origin (ORIG) and Effect (EFF).[2] The set of adjuncts (free modifications) is about 50 large (Mikulová et al., 2006; Urešová, 2011a).

3 CzEngVallex Source Data

3.1 The Czech-English Parallel Corpus

The Prague Czech-English Dependency Treebank (PCEDT) (Hajič et al., 2011; Hajič et al., 2012)[3] is a sentence-aligned parallel treebank with automatic word alignments based on the Wall Street Journal (WSJ) section of Penn treebank[4] and its manual translation to Czech. It contains manual annotation of morphology and syntax for Czech and for English on approx. 50,000 sentences (about a million words), i.e., all the usual "merged" 2,312 files of the Penn Treebank WSJ corpus.

It is annotated on several layers, of which the tectogrammatical layer (cf. Sect. 2.1) includes also the annotation of verbal valency relations by referring, for each verb occurrence in the corpus, to the PDT-Vallex and EngVallex valency lexicons (see Sect. 3.2 and 3.3).[5]

3.2 PDT-Vallex – The Czech Valency Lexicon

The Czech valency lexicon PDT-Vallex (Hajič et al., 2003; Urešová, 2011b) has been developed as part of the PDT annotation effort. Valency frames representing verb senses in this lexicon are grouped by headwords (lemmas). Each frame contains the following fields: a unique ID, labeled valency frame members ("slots"), their obligatoriness and required surface forms. The frames are accompanied by example fragments of Czech sentences, taken almost exclusively from the PDT. Additional notes help to distinguish the meaning of the individual valency frames for the same headword.

The version of PDT-Vallex used to build CzEngVallex contains 11,933 valency frames for 7,121 verbs. The verbs and frames come mostly from the data appearing in the latest versions of the PDT and PCEDT.

```
<frames_pairs owner="...">
 <head>...</head>
 </head>
 <body>
  <valency_word id=... vw_id="ev-w1">
   <en_frame id=... en_id="ev-w1f2">
    <frame_pair id=... cs_id="v-w3161f1">
     <slots>
      <slot en_functor="ACT" cs_functor="ACT"/>
      <slot en_functor="PAT" cs_functor="PAT"/>
     </slots>
    </frame_pair>
    <frame_pair id=... cs_id="v-w9887f1">
     <slots>
      <slot en_functor="ACT" cs_functor="ACT"/>
      <slot en_functor="PAT" cs_functor="PAT"/>
      <slot en_functor="EFF" cs_functor="SUBS"/>
     </slots>
    </frame_pair>
   </en_frame>
  </valency_word>
 </body>
</frames_pairs>
```

Figure 1: Structure of CzEngVallex (part of *abandon* pairing)

3.3 EngVallex – The English Valency Lexicon

EngVallex has been created by a (largely manual) adaptation of an already existing similar resource for English, the PropBank (Kingsbury and Palmer, 2002), to the FGD valency format and to PDT labeling standards (Cinková, 2006). During the adaptation process, arguments were re-labeled, obligatoriness was marked for each valency slot and frames with identical meaning were merged (and some split as well). Links to the original PropBank frame file and roleset have been kept wherever possible.

EngVallex was used for the annotation of the English part of the PCEDT. It contains 7,148 valency frames for 4,337 verbs.

4 Building CzEngVallex

4.1 Structure of CzEngVallex

CzEngVallex builds on all the resources mentioned in Sect. 3. It connects pairs of valency frames in the PCEDT (verb senses) which are translations of each other, aligning their arguments as well. This resource cannot be used independently, since it refers to the valency frame descriptions contained in both PDT-Vallex and EngVallex, and it also relies on the PCEDT.

The structure of this new resource, which is technically a single XML file, is shown in Fig. 1.[6] Aligned pairs of verb frames are grouped by the English verb frame (`<en_frame>`), and for each English verb sense,

[2] These would roughly correspond to Arg0, Arg1, etc. in the PropBank style of argument labeling.

[3] `https://catalog.ldc.upenn.edu/LDC2012T08`

[4] `https://catalog.ldc.upenn.edu/LDC99T42`

[5] Both lexicons can be found at `http://ufal.mff.cuni.cz/pcedt2.0` and also online at `http://lindat.mff.cuni.cz/services/PDT-Vallex` and `.../EngVallex`.

[6] Similar scheme is used in (Hansen-Schirra et al., 2006).

their Czech counterparts are listed (<frame_pair>). For each of such pairs, all the aligned valency slots are listed and referred to by the functor assigned to the slot in the respective valency lexicon. In this example, for the pair *abandon*[7] – *opustit* (lit. *leave [alone]*) the first two arguments match perfectly (ACT:ACT, PAT:PAT) and the third argument in English (EFF) does not match any argument for this particular Czech counterpart, while for the pair *abandon* – *zříci se* (lit. *get rid of [for sth]*), the third English argument maps to a Czech adjunct (SUBS, substitution).

It must be noted here that while all verb–verb pairs have been aligned, annotated, and included in this pairing, there are also many verb–non-verb or non-verb–verb pairs, which have been left aside for this first version of CzEngVallex as none of the underlying lexicons include a complete description of other parts-of-speech.

4.2 The Annotation Process

During the actual annotation process, we have manually aligned English and Czech verbs and their arguments (and in some clear cases also adjuncts). After carefully checking all occurrences of any given valency frame pair in the PCEDT, we included it in CzEngVallex using the structure described in Sect. 4.1, which is based on (Šindlerová and Bojar, 2009; Bojar and Šindlerová, 2010).[8] The process is helped by automatic preprocessing steps.

4.2.1 Preprocessing and Data Preparation

The following steps had been taken before the manual annotation proper started:

- automatic pre-alignment using GIZA++ word alignment (Och and Ney, 2003) and a projection to deep dependency trees (taken from the original PCEDT);
- grouping the occurrences of the same verb sense pairs together to simplify annotation.

4.2.2 Annotation Environment

The annotation interface for manual valency frame alignment[9] has been built as an extension of the TrEd annotation environment (Pajas and Fabian, 2011). TrEd is a fully customizable and programmable graphical editor and viewer for any tree-like structures. It allows displaying and editing sentential tree structures annotated on multiple linguistic layers. The new CzEngVallex TrEd extension uses the data format of the Treex NLP framework (Žabokrtský, 2011; Popel and Žabokrtský, 2010) and pre-existing TrEd extensions for PCEDT, PDT-Vallex, and EngVallex.

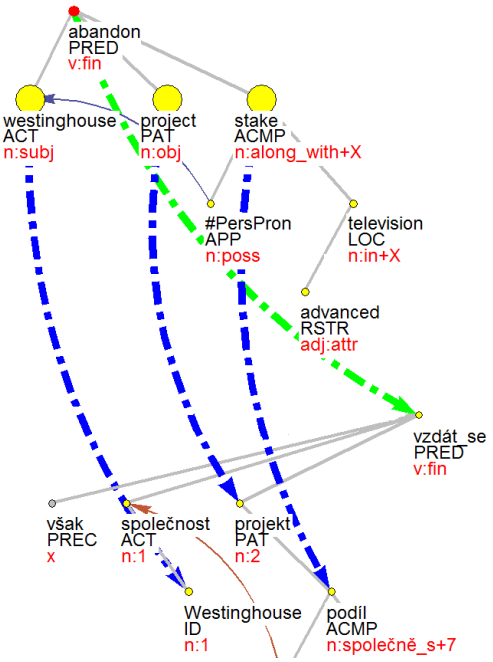

Figure 2: Highlighted alignment in the annotation tool TrEd; color-coding: green for verbs, blue for arguments/adjuncts

The annotation interface includes keyboard macros to change values of individual attributes or to add or delete whole nodes from the structure. Links between English and Czech nodes are added or changed in a drag-and-drop fashion.

4.2.3 Manual Annotation Workflow

The environment described in Sect. 4.2.2 is used to display, edit, collect, and store the alignments between Czech and English valency frames.

Each annotator has their own copy of the treebank, the lexicons, and the valency frame pairing to work on. The changes done by the annotators are merged in the last stage of the process. Any problems encountered, such as wrong annotation in the treebank, or wrong translation, are reported by the annotators through a note system for later corrections.

During the annotation process, the annotator is handed a set of all available sentences for a given verb sense pair. Since verb nodes and their complementations in the PCEDT are automatically pre-aligned (see Sect. 4.2.1), a verb sense pairing suggestion is displayed for each sentence by visually highlighting the pre-alignments (Fig. 2).

The annotator then manually corrects the automatic pre-alignments in the sentence. Then, if the pair is seen for the first time, it is inserted into CzEngVallex by the

[7]Frame ID ev-w1f2, which has been created from abandon.02 in the PropBank, as in *Noriega abandoned command ... for an exile*.

[8]These papers describe only a pilot experiment; the current process differs from their suggestions in several substantial respects.

[9]There are other environments for manual alignment, such as (Melamed, 1998; Samuelsson and Volk, 2007; Ahrenberg et al., 2002), but they work on plain text or phrases, not dependency trees.

annotator (a new CzEngVallex entry is created). For subsequent occurrences, the annotation environment is used to check the pair against the already existing CzEngVallex entry. If any conflict arises, annotators can mark material for further analysis. Typically, errors in either the PCEDT annotation or in the valency lexicons are implied in such cases.

4.3 Lexicon and Corpus: Statistics

	Verb	Frame	PCEDT Tokens	
Language	types	types	verbs	aligned
English	3,288	4,967	130,514	86,573
Czech	4,192	6,776	118,189	85,606

Table 1: Alignment coverage statistics - CzEngVallex/PCEDT

Table 1 contains some statistics about the new resource. It shows that the financial domain of the WSJ (866,246 English tokens/953,187 Czech tokens) is not very rich in terms of different verbs used: only 4,967 different verb frames (which correspond to a medium-grained sense inventory) on the English side and 6,776 different verb frames on the Czech side have been aligned. However, 19,916 different alignment pairs have been collected: this shows that in translation, even if in a restricted domain, translators use a very rich set of synonyms. The verbs with the highest number of different alignments are *be* (353 different verbs aligned to it in Czech), *make* (203) and *take* (171); conversely, it is *být* (184), *mít* (104) and *získat* (70) (lit. *be, have* and *gain*, respectively).

Comparing the aligned pairs with the complete monolingual valency lexicons (see Sect. 3), about 57% of PDT-Vallex (Czech) verb frames are covered, compared to about 69% of covered EngVallex frames. Token-wise, over 66% of English verb nodes (over 72% Czech ones) have been successfully aligned and match CzEngVallex pairings; the rest are aligned to nouns or other parts-of-speech, or impossible to align at all. These numbers jump to 75/86% (English/Czech) if we discount verbs not aligned to any node.

Statistics for the number of differing members are shown in Table 2. We can see that only about 45% frames match fully, i.e., have the same number of arguments and the same labels. Many frames differ in one or two members (47%) while more divergent pairings are a relatively rare occurrence. The differences can be in part explained by the different behavior of the verbs (i.e., not a full semantic match), but a large number of them can be attributed to a certain degree of ambiguity in label assignments, which could be harmonized in future versions of the valency dictionaries (Šindlerová et al., 2014).

		# Pairs
Full match		9,033
# Differences:	1	6,288
	2	3,135
	3	1,138
	4	261
	5	50
	6	10
	7	1

Table 2: Pairing statistics

5 Conclusions

While the statistics themselves provoke an inquiry into translation practice, the goal is to investigate primarily the cases where the straightforward alignment did **not** happen, i.e., those 25/14% verbs not aligned to a verb, or not matching CzEngVallex pairings. Some of these cases can be extracted by inspecting the data where comments have been added by the annotators, and others by simple technical means (finding verbs with no matching alignment, finding verbs aligned to nouns, adjectives, or other structurally divergent structures).

In addition, we plan to use the newly created resource for NLP tasks, such as MT, or to provide features for cross-language machine learning tasks, such as verb sense disambiguation.

The new resource itself, as described here, after necessary quality check and corrections of the underlying data for consistency reasons, will be published under a Creative Commons license and included with the next edition of the PCEDT.

Acknowledgements

The work described herein has been supported by the Grant No. GP13-03351P of the Grant Agency of the Czech Republic, the Grant No. DF12P01OVV022 of Ministry of Culture of the Czech Republic, and SVV project No. 260 224. It is using language resources hosted by the LINDAT/CLARIN Research Infrastructure, project No. LM2010013 of the Ministry of Education, Youth and Sports of the Czech Republic.

References

Lars Ahrenberg, Mikael Andersson, and Magnus Merkel. 2002. A system for incremental and interactive word linking. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pages 485–490.

Timothy Baldwin, Francis Bond, and Ben Hutchinson. 1999. A Valency Dictionary Architecture for Machine Translation.

In *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, pages 207–217, Chester, UK.

Ondřej Bojar and Jana Šindlerová. 2010. Building a bilingual vallex using treebank token alignment: First observations. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 304–309, Valletta, Malta. ELRA, European Language Resources Association.

Silvie Cinková. 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)*, pages 2170–2175, Genova, Italy. ELRA.

Bonnie J. Dorr. 1997. Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. *Machine Translation*, 12(4):271–322.

Sanae Fujita and Francis Bond. 2004a. An automatic method of creating valency entries using plain bilingual dictionaries. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-2004)*, pages 55–64, Baltimore, MD, USA.

Sanae Fujita and Francis Bond. 2004b. A method of creating new bilingual valency entries using alternations. In Gilles Sérasset, editor, *COLING 2004 Multilingual Linguistic Resources*, pages 41–48, Geneva, Switzerland, August 28. COLING.

Jan Hajič, 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, LDC2006T01.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2011. Prague Czech-English Dependency Treebank 2.0, LDC2012T08.

Jan Hajič, 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 International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3153–3160, İstanbul, Turkey. ELRA, European Language Resources Association.

Jan Hajič, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová, and Petr Pajas. 2003. PDT-Vallex: creating a large-coverage valency lexicon for treebank annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9, page 57–68.

Silvia Hansen-Schirra, Stella Neumann, and Mihaela Vela. 2006. Multi-dimensional annotation and alignment in an english-german translation corpus. In *Proceedings of the 5th Workshop on NLP and XML (NLPXML-2006): Multi-Dimensional Markup in Natural Language Processing, at EACL 2006*, pages 35–42, Trento, Italy.

Paul Kingsbury and Martha Palmer. 2002. From Treebank to Propbank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pages 1989–1993.

I. Dan Melamed. 1998. Manual annotation of translational equivalence: The blinker project. *CoRR*, cmp-lg/9805005.

Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, and Zdeněk Žabokrtský. 2006. Annotation on the tectogrammatical level in the prague dependency treebank. annotation manual. Technical Report 30, Prague, Czech Rep.

F. J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51, March.

Petr Pajas and Peter Fabian. 2011. Tred 2.0 – newly refactored tree editor. `http://ufal.mff.cuni.cz/tred`.

Jarmila Panevová. 1974. On verbal frames in Functional generative description I. *Prague Bulletin of Mathematical Linguistics*, (22):3–40.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: modular NLP framework. *Advances in Natural Language Processing*, pages 293–304.

Yvonne Samuelsson and Martin Volk. 2007. Alignment tools for parallel treebanks. In *GLDV Frühjahrstagung*.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht, Reidel, and Prague, Academia, Prague.

Jana Šindlerová and Ondřej Bojar. 2009. Towards English-Czech parallel valency lexicon via treebank examples. In *Proceedings of 8th Treebanks and Linguistic Theories Workshop (TLT)*, pages 185–195, Milano, Italy. Università Cattolica del Sacro Cuore, Università Cattolica del Sacro Cuore.

Jana Šindlerová, Zdeňka Urešová, and Eva Fučíková. 2014. Resources in conflict: A bilingual valency lexicon vs. a bilingual treebank vs. a linguistic theory. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

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

Zdeňka Urešová. 2011b. *Valenční slovník Pražského závislostního korpusu (PDT-Vallex)*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czechia, ISBN 978-80-904571-1-9, 375 pp.

Hans Uszkoreit. 2002. New chances for deep linguistic processing. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02), August 24 - September 1*. Morgan Kauffmann Press.

Zdeněk Žabokrtský. 2011. Treex – an open-source framework for natural language processing. In Markéta Lopatková, editor, *Information Technologies – Applications and Theory*, volume 788, pages 7–14, Košice, Slovakia. Univerzita Pavla Jozefa Šafárika v Košiciach.