PBML

The Prague Bulletin of Mathematical Linguistics

NUMBER 105 APRIL 2016 17–50

# CzEngVallex: a Bilingual Czech-English Valency Lexicon

Zdeňka Urešová, Eva Fučíková, Jana Šindlerová

Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

## Abstract

This paper introduces a new bilingual Czech-English verbal valency lexicon (called CzEngVallex) representing a relatively large empirical database. It includes 20,835 aligned valency frame pairs (i.e., verb senses which are translations of each other) and their aligned arguments. This new lexicon uses data from the Prague Czech-English Dependency Treebank and also takes advantage of the existing valency lexicons for both languages: the PDT-Vallex for Czech and the EngVallex for English. The CzEngVallex is available for browsing as well as for download in the LINDAT/CLARIN repository.

The CzEngVallex is meant to be used not only by traditional linguists, lexicographers, translators but also by computational linguists both for the purposes of enriching theoretical linguistic accounts of verbal valency from a cross-linguistic perspective and for an innovative use in various NLP tasks.

## 1. Introduction

The CzEngVallex lexicon[1] is a result of the project called "A comparison of Czech and English verbal valency based on corpus material (theory and practice)".[2] In this project, two main goals were pursued: hands-on work with corpus data resulting in an explicit representation of cross-lingual meaning relations, and a theoretical comparative study particularly focused on differences between the Czech and English verbal valency structure. Theoretical aspects include both the description of verbal valency and the description of interlinking the translational verbal equivalents, focusing on comparison of the existing approaches in the two languages. This project is

[1] http://lindat.mff.cuni.cz/services/CzEngVallex

[2] A research grant supported by the Grant Agency of the Czech Republic under the id GP13-03351P

Corresponding author: uresova@ufal.mff.cuni.cz
Cite as: Zdeňka Urešová, Eva Fučíková, Jana Šindlerová. CzEngVallex: a Bilingual Czech-English Valency Lexicon. The Prague Bulletin of Mathematical Linguistics No. 105, 2016, pp. 17–50. doi: 10.1515/pralin-2016-0001.

based on the Functional Generative Description Valency Theory (FGDVT) and on its application to a corpus, namely to the Prague Czech-English Dependency Treebank (PCEDT)[3] (Hajič et al., 2011). This theoretical approach is highly suitable for the proposed specification of relations of verbal valency frames in both languages. The work with the data includes the creation of a parallel Czech-English valency lexicon which is interlinked with real examples of valency usage in the broad context of the PCEDT.

The underlying idea of the project builds on the assumption that verbal valency is the core structural property of the clause, therefore, capturing the alignment of the translationally equivalent verbs, as well as the mappings[4] of their valency positions, should provide a valuable model of basic patterns within cross-lingual semantic relations. Moreover, such a resource that stores interlingual valency relations for several thousands of verbs and verb pairs might enable us making predictions (on the basis of semantic relatedness, or verb classes) about the verbs unseen in the text.

This article is structured as follows: after a theoretical background (Sec. 2) we present the basic structure of the CzEngVallex lexicon (Sec. 3, published in part in Urešová et al. (2015)). The annotation environment and process description follows (Sec. 4, Sec. 5). Linguistic issues related to the annotated data using CzEngVallex are described in Sec. 6 and in Sec. 7 (of which Sec. 7.1 to 7.3 have been published in part in Šindlerová et al. (2015)). We conclude with suggestions concerning possible applications and future work.

## 2. Theoretical background

Our approach to the issues of valency of Czech and English verbs applied in this project is based on the following points of view and uses the following principles and features (Sec. 2.1–2.2).

### 2.1. Valency in the FGD

The project draws on the Functional Generative Description Valency Theory. In this dependency approach, valency is seen as the property of some lexical items, verbs above all, to select for certain complementations in order to form larger units of meaning. The governing lexical unit then governs both the morphological properties of the dependent elements and their semantic interpretation (roles). The number and realization of the dependent elements constituting the valency structure of the phrase (or sentence) can be represented by valency frames, which can be listed in valency lexicons.

[3] http://hdl.handle.net/11858/00-097C-0000-0015-8DAF-4

[4] Here, we often use the terms "mapping" and "alignment" interchangeably. Though by "mapping", we usually refer to the abstract notion of semantic equivalence of expressions between languages, and by "alignment", we refer to its practical implementation in the data.

The basics of the FGDVT can be found, e.g., in Panevová (1974). The FGD approaches valency as a special relation between a governing word and its dependents.[5] This relation belongs to the level of deep syntax (tectogrammatical layer of linguistic description). It combines a syntactic and a semantic approach for distinguishing valency elements. The verb is considered to be the core of the sentence (or clause, as the case may be). The relation between the dependent and its governor at the tectogrammatical layer is represented by a *functor*, which is a label representing the semantic value of a syntactic dependency relation and expresses the function of the complementation in the clause. For a full list of all dependency relations and their labels, see Mikulová et al. (2006a).

The FGDVT works with a systematic classification of verbal valency complementations (arguments)[6] along two axes. The first axis represents the opposition between inner complementations (actants) and free complementations (adjuncts) and it is determined independently of any lexical unit. The other axis relates to the distinction between obligatory and optional complementations, for each verb sense separately.

There are five "inner participants" (actants) in the FGDVT: Actor/Bearer (ACT), Patient (PAT), Addressee (ADDR), Origin (ORIG) and Effect (EFF). Which functors are considered actants has been determined according to two criteria. The first one says that actants can occur at most once as a dependent of a single occurrence of a particular verb (excluding apposition and coordination). According to the second criterion, an actant is restricted to only a relatively closed class of verbs.

Out of the five actant types, the FGDVT states that the first two are connected with no specific globally defined semantics, contrary to the remaining three ones. The first actant is always the Actor (ACT), the second one is always the Patient (PAT). The Addressee (ADDR) is the semantic counterpart of an indirect object that serves as a recipient or simply an "addressee" of the event described by the verb. The Effect (EFF) is the semantic counterpart of the second indirect object describing typically the result of the event (or the contents of an indirect speech, for example, or a state as described by a verbal attribute). The Origin (ORIG) also comes as the second (or third or fourth) indirect object, describing the origin of the event (in the "creation" sense, such as *to build from metal sheets*.ORIG, not in the directional sense).

The FGDVT has further adopted the concept of shifting of "cognitive roles". According to this special rule, semantic Effect, semantic Addressee and/or semantic Origin are shifted to the Patient position in case the verb has only two actants. Similarly, any of the actant roles are shifted to the Actor position in case the verb has only a single valency position.

[5] For the sake of brevity, we will further refer only to the valency of verbs, since the CzEngVallex contains so far only the alignment of verb pairs.

[6] In the following sections, we will use the term "argument" for any of the complementations of a particular verb (sense) entry in the lexicon, i.e., for actants and adjuncts included in such a valency frame.

The repertory of adjuncts (free modifications) is much larger (about 50) than that of actants (see again Mikulová et al. (2006a)). Adjuncts are always determined semantically; their set is divided into several subclasses, such as temporal (TWHEN, TSIN, TTILL, TFL, TFHL, THO, TPAR, TFRWH, TOWH), local (LOC, DIR1, DIR2, DIR3), causal (such as CAUS for cause, AIM for purpose, CRIT for 'according to', etc.) and other free complementations (MANN for general 'manner', ACMP for accompaniment, EXT for extent, MEANS, INTF for intensifier, BEN for benefactor, etc.). Adjuncts may be seen as deep-layer counterparts of surface adverbial complementations. More adjuncts of the same type can occur as dependents on a particular occurrence of the verb and adjuncts may modify in principle any verb – this is also where their name ('free complementations') comes from. Unlike actants, morphemic realization of adjuncts is rarely (if ever) restricted by a particular verb.

Due to this "free nature" of adjuncts, only the presence of actants (obligatory or optional) and obligatory adjuncts is considered necessary in any verbal valency frame (the FGDVT is thus said to use the notion of valency in its "narrow" sense): optional adjuncts are (as a general rule) not listed in the valency frame. As mentioned above, both actants and adjuncts can be in their relation to a particular word either obligatory (that means obligatorily present at the tectogrammatical level) or optional (that means not necessarily present in any sentence where the verb is used). It must be said that this definition of obligatoriness and optionality does not cover surface deletions but only semantically necessary elements.

Since the surface appearance of a complementation does not really help to distinguish between obligatory and optional elements, other criteria must be used. Specifically, the 'dialogue test' is used. It is a method based on asking a question about the element that is supposed to be known to the speaker because it follows from the meaning of the verb: if the speaker can answer the hearer's follow-up wh-question about the given complementation with *I don't know* (without confusing the hearer), it means that the given complementation is semantically optional. On the other hand, if the answer *I don't know* is disruptive in the (assumed) conversation, then the given complementation is considered to be semantically obligatory. For further details, see Urešová (2011a).

### 2.2. Comparative character and corpus approach to cross-language research

We are interested in differences in the expression of the same contents in two typologically different languages, namely Czech and English. The initial hypothesis is that even in relatively literal or exact translation, where the information and the meaning the sentences carry in both languages is essentially the same–as exemplified in economic, news, and similar non-artistic genres–the core sentence structure (i.e., the main verb of a clause and its arguments) often differs due to intrinsic language differences. Comparing Czech and English valency frames and their arguments, based on their usage in a parallel corpus, is expected to enable not only the detection of the types of divergences of expression in the core sentence structure but also a quantitative analysis of their similarities and differences, thanks to the substantial size of the corpora available.

Both lexicons, which we used as a starting point, are based on the same theoretical foundations (cf. Sec. 2.1). Our task was thus slightly simplified in that we were not comparing two different valency theories, but rather an application of a single theoretical (and formal) framework to two particular languages (and to a translated, i.e., parallel corpus material). Such approach has, we believe, a major advantage: we are able to pinpoint the differences much more clearly against a unified theoretical background, as opposed to a possibly fuzzy picture which widely differing valency theories might give.

Our approach to the comparative study of valency builds on the growing role of computer corpora in linguistic research. Our study is based on corpus examples with natural contexts, which gives well-founded research results backed also by quantitative findings. Therefore, a detailed and thorough work with electronically created and accessible data, namely, with the PDT-Vallex and the EngVallex lexicons and the PCEDT, are the foundations we build our research on.

## 3. CzEngVallex reference data

For the CzEngVallex project, two treebanks are most relevant: the PDT[7] and the PCEDT[8] which contain manual annotation of morphology, syntax and tectogrammatics (semantics).

Next, we work with the PDT-Vallex verbal valency lexicon for Czech (Urešová, 2011b) and with a similar resource for English called EngVallex (Cinková, 2006).

These data resources are the "input" material for the creation of the CzEngVallex. Also, they are heavily referred to from the resulting CzEngVallex and can thus be considered an integral part of it.

### 3.1. Czech-English parallel corpus

The CzEngVallex primary data source is the parallel Prague Czech-English Dependency Treebank (PCEDT). The PCEDT is a sentence-parallel treebank based on the texts of the Wall Street Journal part of the Penn Treebank[9] and their manual (human) translations.

It is annotated on several layers, of which the tectogrammatical layer (layer of deep syntactic dependency relations) includes also the annotation of verbal valency relations. The tectogrammatical annotation of this corpus includes also links to two va-

[7] http://ufal.mff.cuni.cz/pdt/

[8] https://catalog.ldc.upenn.edu/LDC2004T25

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

lency lexicons, the PDT-Vallex (for Czech) and the EngVallex (for English), see their detailed description below.

## 3.2. Czech and English valency lexicons

### 3.2.1. PDT-Vallex - Czech valency lexicon

The Czech valency lexicon, called PDT-Vallex,[10] is publicly available as a part of the one-million-word Prague Dependency Treebank (PDT) version 2 published by the Linguistic Data Consortium.[11] It has been developed as a resource for valency annotation in the PDT; for details, see Urešová (2011b). As such, it has been designed in close connection to the specification of the treebank annotation. The "bottom up", data-driven practical approach to the forming of the valency lexicon had made it possible for the first time to confront the already existing FGDVT and the real usage of language. Precise linking of each verb occurrence to the valency lexicon has made it possible to verify the information contained in the valency lexicon entry against the corpus by automatic means, making it a reliable resource for further research.

Each valency entry in the lexicon contains a headword, according to which the valency frames are grouped, indexed, and sorted. The valency frame contains the following specifications: the number of valency frame members, their labels, the obligatoriness feature and the surface form of valency frame members. Any concrete lexical realization of the particular valency frame is exemplified by an appropriate example, i.e., an understandable fragment of a Czech sentence, taken almost exclusively from the PDT. Notes help to delimit the meaning of the individual valency frames inside the valency entry. Typically, synonyms, antonyms and aspectual counterparts serve as notes. For a detailed information about the actual structure of the PDT-Vallex entry, see Urešová (2011a).

The version of the PDT-Vallex used for the CzEngVallex contains 11,933 valency frames for 7,121 verbs. The verbs and frames come mostly from the data appearing in the PDT, version 2.0, and the PCEDT, version 2.0. The lexicon is being constantly enlarged with data provided by further annotations.

### 3.2.2. EngVallex - English valency Lexicon

The EngVallex[12] is a lexicon of English verbs, also built on the grounds of the FGDVT. It was created by a (largely manual) adaptation of an already existing resource for English with similar purpose, namely the PropBank Lexicon (Palmer et al., 2005; Kingsbury and Palmer, 2002), to the PDT labeling standards (see also Cinková (2006)). During the adaptation process, arguments were re-labeled, obligatoriness was marked

[10] http://hdl.handle.net/11858/00-097C-0000-0023-4338-F

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

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

for each valency slot, frames with identical meaning were unified and sometimes, frames with a too general meaning were split. Links to PropBank frames have been preserved wherever possible. The EngVallex was used for the valency annotation of the Wall Street Journal part of the Penn Treebank during its manual annotation on the tectogrammatical layer; the result is the English side of the PCEDT.

The EngVallex currently contains 7,148 valency frames for 4,337 verbs.

## 4. Building CzEngVallex

### 4.1. The annotation goal

To meet the goals stated in Sec. 1, an explicit linking between valency frames of Czech and English verbs based on a parallel corpus is needed. This has been accomplished by creating the bilingual Czech-English Valency Lexicon (CzEngVallex).[13]

The CzEngVallex stores alignments between Czech and English valency frames and their arguments. The resulting alignments are captured in a stand-off mode (in a file called `frames_pairs.xml`). This file is the "entry point" to the CzEngVallex; it cannot be used independently, since it refers to the valency frame descriptions contained in both the PDT-Vallex and the EngVallex, and it also relies on the PCEDT as the underlying corpus.

The idea of CzEngVallex builds on Šindlerová and Bojar (2009) and Bojar and Šindlerová (2010). However, only a pilot experiment has been described in these two papers; the actual process of creating CzEngVallex differed from suggestions in these papers in several substantial aspects.

### 4.2. CzEngVallex structure

The CzEngVallex builds on all the resources mentioned in Sec. 3. It is technically a single XML file `frames_pairs.xml` (shown in Fig. 1) which lists for each included English verb (identified by a verb `id`) a list of its valency frames (identified by a valency frame `id`), and for each English valency frame all the collected frames-pairs, and for each of the collected frames-pairs (identified by a pair `id`) the pairings of their valency slots (identified by functors).

Aligned pairs of individual verb frames are grouped by the English verb frame (`<en_frame>`) (cf. Fig. 1), and for each English verb sense, 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 (the PDT-Vallex for Czech, the EngVallex for English).

[13] Available for browsing and searching at `http://lindat.mff.cuni.cz/services/CzEngVallex`, download from `https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1512`

```
<frames_pairs owner="...">
 <head>...</head>
 <body>
  <valency_word id="vw1484" vw_id="ev-w1869">
   <en_frame id="vw1484f1" en_id="ev-w1869f1">
    ...
    <frame_pair id="vw1484f1p8" cs_id="v-w8735f1">
     <slots>
      <slot en_functor="ACT" cs_functor="ACT"/>
      <slot en_functor="PAT" cs_functor="PAT"/>
      <slot en_functor="EFF" cs_functor="---"/>
     </slots>
    </frame_pair>
    ...
   </en_frame>
  </valency_word>
 </body>
</frames_pairs>
```

*Figure 1. Structure of the CzEngVallex (part of limit pairing)*

In the example in Fig. 1, for the pair *limit*[14] - *zabránit* (lit. *limit/prevent*) we can observe a match of the first two actants (ACT:ACT, PAT:PAT) and a zero alignment (cf. Sec. 6.2.2) of the third frame element: EFF,[15] which does not match any verb argument for this particular Czech counterpart.

It is crucial to mention here that while all verb–verb pairs have been aligned, annotated and then collected in this pairing lexicon, there are also many verb–non-verb or non-verb–verb pairs, which have been left aside for the first version of the CzEngVallex, since none of the underlying lexicons has enough entries covering nominal valency included.

## 5. Annotation environment

### 5.1. Prerequisites

The annotation was done over the bilingual data from the parallel PCEDT 2.0.[16] The annotation interface for building the CzEngVallex was constructed as an extension of the tree editor TrEd (Pajas and Fabian, 2011)[17] environment.

[14] Frame ID ev-w1869f1, which has been created from limit.01 in the PropBank, as in *... which.*ACT *limits any individual holding.*PAT *to 15%.*EFF

[15] Marked as optional in EngVallex but optional actants must still be aligned.

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

[17] http://ufal.mff.cuni.cz/tred

TrEd is a fully customizable and programmable graphical editor and viewer for tree-like structures. Among other projects, it was used as the main annotation tool for the tectogrammatical annotation of both source treebanks (PDT and PCEDT). It allows displaying and annotating sentential tree structures on multiple linguistic layers with a variety of tags using either the Prague Markup Language (PML) format[18] or the Treex format.[19]

Treex (formerly TectoMT) (Žabokrtský, 2011; Popel and Žabokrtský, 2010) is a development framework for general as well as specialized NLP tasks (such as machine translation) working with many representations of text and sentence structure, including tectogrammatically annotated structures. It offers its own file format, which is capable of storing and displaying (using TrEd) multiple tree structures at once, hence it is a fitting environment when cross-lingual relations are involved.

We have tried to keep the annotation environment as simple and transparent as possible, though still leaving all its important features available (see Fig. 2). It provides an annotation mode for valency frames alignment between the PDT-Vallex and the EngVallex. This extension builds on previously used TrEd extensions: the pdt2.0 extension (for the annotation of the PDT 2.0), the PDT-Vallex extension, and the pedt extension (for annotating the English side of the PCEDT); all these extensions offer functions necessary for browsing Czech and English treebanks and their valency lexicons, while the CzEngVallex extension itself provides the cross-lingual interlinking function.

### 5.2. Preprocessing and data preparation

The following steps were taken before the start of the annotation proper:

- automatic alignment on the word level of the PCEDT 2.0;
- preliminary collection of all verb-verb alignments and alignments of their complementations based on the referred-to valency lexicon entries, as they had been included in the PCEDT;
- preparation of lists grouping together all verb-sense pairs for every English verb as collected within the previous step.[20]

For the word alignment of the PCEDT data, the GIZA++[21] algorithm was used, and subsequently, this alignment was mapped to the nodes of the corresponding (deep/tectogrammatical) dependency trees representing the original and the translated sentence.

[18] http://ufal.mff.cuni.cz/jazz/PML

[19] http://ufal.mff.cuni.cz/treex

[20] These lists of verb occurrences in the parallel treebank are technically called ‘filelists’.

[21] https://code.google.com/p/giza-pp

The resulting pairs were grouped by these references, one group for each English verb, and stored as *filelists*, which can be fed directly into the annotation tool TrEd (described in Sec. 5.4). Thus, the annotator was able to inspect the same verb occurrences together in a single data block. Similarly, the individual pairs for the same source verb sense were sorted in succession within the groups. The process of correcting, re-aligning (when necessary) and finally collecting the verb–verb alignments followed, based on the EngVallex and the PDT-Vallex references contained already in the treebank data for both translation sides.

### 5.3. The filelists

The corresponding pairs of Czech and English verbs were looked up in the PCEDT, using a btred[22] script. The script searches through the alignment attribute of the English verb nodes, where the information about the connection to the Czech counterpart is usually stored. All instances of individual verb pairs in the PCEDT were then listed in the form of filelists containing treebank position identifiers of the corresponding nodes. As such, they can be browsed alphabetically, or on the basis of pair frequency in a treebank, or employing other useful criteria.

Filelists were sorted by the English verb lemma and organized alphabetically into folders according to the first letter of the source verb. If a single English verb corresponded to more than one Czech verb, those verbs were placed in the same folder - the name of the folder then consists of the name of the English verb, the number of corresponding Czech verbs and the number of occurrences in the parallel corpus (e.g., *abate.3v.4p*). The filelists' names were designed according to the following rules:

(i) if there exist more Czech verbs to a given English verb in the parallel corpus, the filelist corresponding to one of the pairs will be placed in a directory named after the English verb, and will bear a name containing the Czech verb and the number of occurrences of this pair in the parallel corpus (e.g., for the pair *abate-polevit*, a filelist named *polevit.2.fl* is in a directory *abate.3v.4p*);

(ii) if there exists only a single Czech verb to a given English verb in the parallel corpus, the name of the filelist for this pair will contain both the English and Czech verb and the number of occurrences of this pair in the parallel corpus (e.g., *abide_by.1v.2p.dodržovat.2.fl*).

The annotator received a set of all available sentences for each verb pair at once. In total, there were 92,889 sentences, which were split into 15,931 filelists with an average number of sentences in one filelist 5,83 (median 1). The most frequent pair is *be→být*, which has 10,287 instances in its filelist.

Single-instance filelists[23] have been, for the sake of annotation efficiency, unified into a single filelist within the corresponding folder, e.g., for the verb *abate* the filelists

[22] http://ufal.mff.cuni.cz/pdt2.0/doc/tools/tred/bn-tutorial.html

[23] By single-instance filelists we mean verb pairs with only a single occurrence in the parallel corpus.

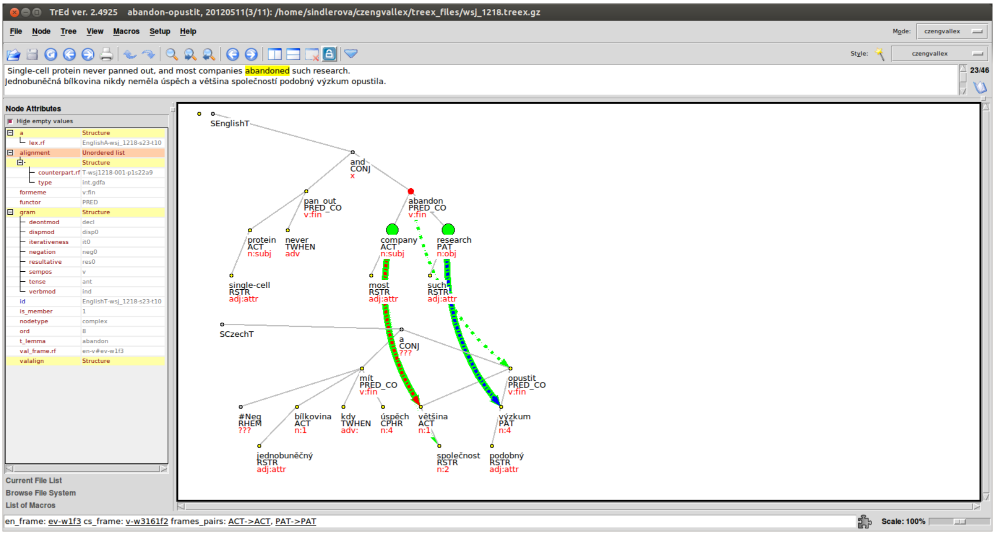
*zmírnit.1.fl* and *zmírnit_se.1.fl* merge into one filelist *abate.1_1.2.fl*; similarly, the filelists *abdicate.1v.1p.zbavovat_se.1.fl*, *abet.1v.1p.podporovat.1.fl*, *abort.1v.1p.potratit.1.fl* etc. are absorbed in a single filelist *a.1_1.30.fl*).

The annotators thus eventually processed 7,891 filelist in total, with the average number of sentences in the filelist 11,77 (median 3).[24]

## 5.4. The annotation process

During the actual annotation process, English and Czech verbs and their arguments were manually aligned or re-aligned, and after checking carefully all the occurrences of any given pair in the PCEDT data, the corresponding arguments were captured in the CzEngVallex lexicon, using the structure described in Sec. 4.2.

Even though all PCEDT occurrences of all verb–verb pairs were inspected manually, the process was helped substantially by several automatic preprocessing steps, as described in Sec. 5.2.


TrEd ver. 2.4925 abandon-opustit, 20120511(3/11): /home/sindlerova/czengvallex/treex_files/wsj_1218.treex.gz

Mode: czengvallex

Style: czengvallex

Single-cell protein never panned out, and most companies abandoned such research.
Jednobuněčná bílkovina nikdy neměla úspěch a většina společností podobný výzkum opustila.

23/46

Node Attributes

☑ Hide empty values

| | |
|---|---|
| ⊟ a | Structure |
| └ lex.rf | EnglishA-wsj_1218-s23-t10 |
| ⊟ alignment | Unordered list |
| ⊟- | Structure |
| ├ counterpart.rf | T-wsj1218-001-p1s22a9 |
| └ type | int.gdfa |
| formeme | v:fin |
| functor | PRED |
| ⊟ gram | Structure |
| ├ deontmod | decl |
| ├ dispmod | disp0 |
| ├ iterativeness | it0 |
| ├ negation | neg0 |
| ├ resultative | res0 |
| ├ sempos | v |
| ├ tense | ant |
| └ verbmod | ind |
| id | EnglishT-wsj_1218-s23-t10 |
| is_member | 1 |
| nodetype | complex |
| ord | 8 |
| t_lemma | abandon |
| val_frame.rf | en-v#ev-w1f3 |
| valalign | Structure |

en_frame: ev-w1f3 cs_frame: v-w3161f2 frames_pairs: ACT->ACT, PAT->PAT

Scale: 100%

*Figure 2. Annotation environment at work*

[24] For detailed work with filelists see Urešová et al. (2015).

### 5.5. Manual alignment - the starting point

The environment described in Sec. 5 was used to display, edit, collect, and store the alignments between Czech and English valency frames.

Each annotator had her/his own copy of the PDT-Vallex, the EngVallex and the PCEDT and the filelists to work on (Sec. 5.2).[25]

S/he was expected to go through all verb occurrences in the filelist and build a typical valency frame alignment for each verb sense. S/he was also expected to deal with the potential conflicting cases (choose the most probable alignment option, mark complicated issues, such as missing or inappropriate frames or wrong tree structure in a note, etc.). Once collected, the frame alignment was automatically extended to all occurrences of the pair of the valency frames; it was the annotator's responsibility to check all the occurrences of such a pair if they correspond to the collected alignment, as recorded in the CzEngVallex.

Direct changes (changing the tree structure or frame adjustments) in the treebank were disallowed, though the extension allowed storing some minor type of changes (change of functor label) in specific CzEngVallex-related attributes. Also, the annotator reported problems through a note system for later corrections,[26] and s/he was allowed to change the valency frame link if considered inappropriate.

## 6. Understanding CzEngVallex

While this paper is not a substitute for the annotation guidelines, the basic rules for aligning verbs and their arguments will be described here so that the reader can understand the CzEngVallex data - what was annotated, what was not, in which cases examples were not included, treatment of convention differences in both valency lexicons, and more.

All details regarding annotation guidelines, annotation workflow and functionality of the annotation extension of TrEd are given in the CzEngVallex Technical Report (Urešová et al., 2015).

### 6.1. Verb pairs to include (or exclude)

As explained previously, CzEngVallex contains only those verb pairs for which a reasonable alignment was found in the treebank; sometimes, all occurrences (one or more) of the same frame pair align such diverging structures that they could not be aligned.

These cases include:

[25] A subversion system has been used for easy synchronization between annotators' laptops and the main data store.

[26] The CzEngVallex extension offers specific pre-defined "note" attributes to the annotator, which can be extended by free text, cf. Urešová et al. (2015).

1. good translation but with too different syntax which can be the result of
   - (a) the use of a language-specific syntactic structure,
   - (b) translation of a single verb by multiple verbs and consequent untypical argument distribution between these verbs;
2. semantically incorrect or too loose translation resulting in a syntactic difference.

Judging the degree of syntactic diversity has been fully up to the annotator. In case of complex and rare syntactic differences, the annotator was required not to include the sentence (or more sentences for a given frame pair) in the annotation. The reason for omission is usually described in the note attribute. For example, if the translation was substantially inaccurate or if the translation was too loose, the sentences remained manually "unannotated," i.e., there was no attempt to correct alignments in the data or to make other data adjustments. The annotator was required to leave a note saying, e.g., "too loose translation".

In case all occurrences of a verb pair were deemed unalignable, such a verb pair is not included in the `frames_pairs.xml` file.

### 6.2. Discrepancies and conflicts in annotation

Ideally, each pair of frames is supposed to have only a single way of argument alignments. This follows from the semantic character of the tectogrammatical structure. Due to the deep character of the description, it is also supposed that the alignment should be to a great extent "parallel," i.e., that the nodes of the two trees ideally correspond 1:1 and that their functors match.

Nevertheless, this is often not the case. There are discrepancies and conflicts of different kinds in the data, as the CzEngVallex annotation reflects.

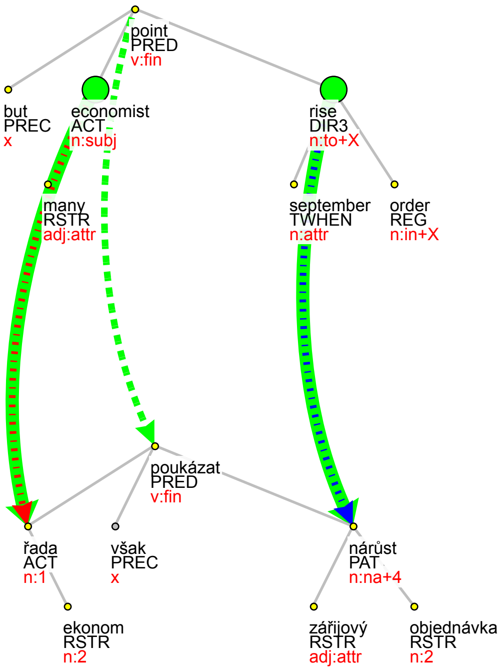
By discrepancies, we refer either to the so-called zero alignment (see Sec. 6.2.2), i.e., places where an argument node in one of the languages is translated in such a way that it is not a direct dependent (i.e., not an argument) of the aligned verb in the other language, or to the functor mismatch (6.2.1), i.e., when two aligned nodes have different tectogrammatical functor labels.

By conflicts in annotation (Sec. 6.2.3), we refer to cases where the alignment of the verb or its arguments looks differently in different sentences in the corpus. In other words, for that frame pair, one such alignment would be in conflict with another alignment observed elsewhere in the data.[27]

[27] The design of CzEngVallex (Sec. 4.2), as mirrored in the structure of the `frames_pairs.xml` file, does not allow for alternative argument alignments for the same verb frame pair. Please recall that verb frames already represent a single verb sense, thus this type of conflict should not be blamed on potentially mixed senses of the verb involved.

### 6.2.1. Functor mismatch

By functor mismatch, we mean alignment of nodes with different functor labels (see example in Fig. 3).[28] These alignments can involve either (proper) actant-actant mapping, or even an actant-adjunct mapping. The causes for functor mismatch often involve different morphosyntactic realization which was treated differently in the two languages, rather than a clear semantic difference.



En: But many economists pointed to a ... September rise in orders ...
Cz: Řada ekonomů však poukázala na ... zářijový nárůst objednávek, ...

*Figure 3. Functor mismatch* DIR3→PAT *in the data*

Though this is in most cases technically unproblematic, we provide some notes of the common causes of functor mismatch in the following paragraphs.

[28] In the examples displayed, the green lines connect either the annotated verb pair or the already collected argument pairs, the automatic node alignment suggestion is displayed as a blue arrow, the manually corrected alignment is marked as a red arrow. The images have been cropped or otherwise adjusted for the sake of clarity.

The data show that it is quite often the case that the alignment connects an actant (usually on the English side) to an adjunct (usually on the Czech side), for example ADDR to DIR3 or LOC, also EFF to COMPL, ACT to LOC, ACT to CAUS etc. These differences often have grounds in different morphosyntactic forms of the given modifications, which was taken as decisive for using an adjunct instead of an actant (mostly on the Czech side due to its richer morphology). This is a feature of the underlying linguistic theory that was perhaps a bit overstressed in the original treebank (PDT) annotation when assigning the functor(s) to slots in the valency frames.

Since the morphosyntactic forms of the valency complementations are to a great extent fixed with the given verb, the alignment for individual functor pairs seems to be quite consistent throughout certain verb pairs or even verb classes.[29] For example, (English) ADDR to (Czech) DIR3 appears with, e.g., the verbs *commit/svěřit* (En: *...committing more than half their funds to either.*ADDR *of those alternatives* / Cz: *...svěřilo více než polovinu svých prostředků do jediné.*DIR3 *z těchto alternativ*). Similarly, the link (English) EFF to (Czech) COMPL appears with the verb pair *consider/posoudit* (En: *...will be considered timely.*EFF *if postmarked no later than Sunday* / Cz: *...budou posouzeny jako včas podané nabídky.*COMPL).

This kind of functor mismatch can occur with any actant label, even with the ACT. For example, the case of ACT aligning to MEANS appears due to a known problem of the so-called instrument-subject alternation, here illustrated with the verb pair *please/potěšit*: En: *Pemex's customers are pleased with the company's new spirit.*MEANS / Cz: *Zákazníky společnosti Pemex rovněž potěšil nový elán.*ACT *společnosti*.

In case there is a "third" actant in the structure, this third (or higher-numbered) actant may also differ in labeling in English and Czech, even in cases where the semantic correspondence is clear. For example, see the following occurrence of the verb pair *insulate/chránit*: En: *...will further insulate them.*PAT *from the destructive effects.*ORIG / Cz: *...je.*PAT *bude dále chránit před destruktivními vlivy.*EFF. Here, the English ORIG corresponds to the Czech EFF. While this is not a technical problem, it signals unclear definitions of those actant labels in the Czech and English guidelines for valency entries. This deficiency was found both for actants, semantically close adjuncts and for actant/adjunct pairs, e.g., EFF/MEANS mapping: for the verb pair *outfit/vybavovat*: En: *…will outfit every computer with a hard drive.*EFF / Cz: *...bude vybavovat všechny počítače pevným diskem.*MEANS. The question of labeling the actants (PAT ORIG x ADDR PAT) arose also in the following example for the verb pair *rid/zbavit*: En: *...to clean up Boston Harbor or rid their beaches.*PAT *of medical waste.*ORIG / Cz: *...zbavit pláže.*ADDR *nemocničního odpadu.*PAT.

An example of semantically close functors mismatch is the problem of a "dynamic versus static expression of location", i.e., DIR3/LOC mismatch: for the verb pair *include/zahrnout*, the data offer the following example: En: *...real-estate assets are in-*

[29] At this time, we have not fully investigated this interesting issue in a quantitative way, leaving it for future research.

*cluded in the capital-gains provision*.DIR3 / Cz.: *…nemovitý majetek je v ustanovení*.LOC *o kapitálových ziscích zahrnut*; or: En: *...prime minister ordered to deposit 57 million in bank*.LOC / Cz: *…ministerský předseda nařídil uložit asi 57 milionů dolarů do banky*.DIR3. Note that the theory based on deep syntactic frames does not allow to reinterpret labels in semantic changes caused by syntactic shifts such as passivization.

The fact that the functor mismatch often occurs when semantically parallel structures differ in morphological realization only, and in some cases even allow alternative interpretation, leads us to the need to reconsider the valency slot labeling schemes for both English and Czech, and more precisely define the "semantics" of these labeling schemes, since often the differences in argument and/or adjunct labels do not seem warranted.

### 6.2.2. Zero alignment

By zero alignment we mean such structural configurations that involve different number of arguments in the corresponding syntactic structures, i.e., an alignment of "something" on one side of the translation to "nothing" on the other side. There are various reasons for zero alignment, e.g., a simple absence of a lexical or structural counterpart in the translation, or deeper embedding of an argument counterpart in a subtree.

In Fig. 4, the reason is that in English the word *earnings* is treated as an argument of the light verb *have*, whereas in Czech its counterpart (*výdělky*) depends on the nominal part of the light verb constructions (the word *dopad* - lit. *impact*).

A slightly different case appears for the verb pair *call/volat*, En: *...this calls into question the validity of the R... theory* / Cz: *...to volá po otázce po správnosti R... teorie*: the Czech equivalent *správnost* to the English *validity*.PATient is embedded, since the English construction is considered an idiom (*calls into question*), marking *into question* as DPHR. In Czech, *správnost* carries the RSTR label and depends not on the verb, but on the noun *otázka* (lit. *question*).

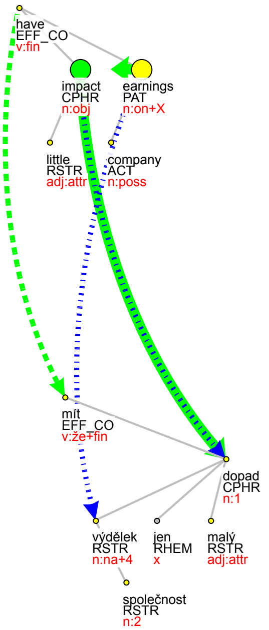
The usual way of treating zero alignment is keeping the alignment of the appropriate "superfluous" node to "no specific node".

Zero alignment is caused, i.a., systematically by certain linguistic phenomena, such as different complexity of verbal meaning expression or loose or specific translation. Some of the cases are treated in Sec. 7.1 to 7.3.

### 6.2.3. Conflicts

Conflicts, as defined above, arise if the verb argument annotation at one place in the data is inconsistent with another occurrence in the data.

First, there may be problems with the granularity of verb senses as represented by the verb frames in the PDT-Vallex and EngVallex lexicons, which is then displayed in the aligned PCEDT data (as opposed to the Czech and English sides when taken separately, where it cannot be seen easily). With some verbs, the alignment as displayed



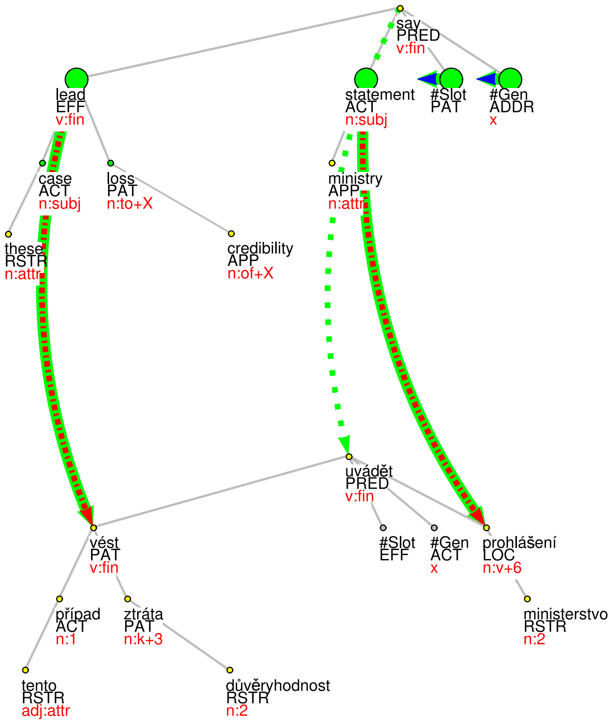
En: ... have little impact on the company's earnings.
Cz: ... bude mít na výdělky společnosti jen malý dopad.

*Figure 4. Zero alignment (embedded argument)* PAT⟶- - -

in the parallel data might show that two separate frames for two separate verb senses are needed, instead of the currently used one frame for both (or more), often due to certain overgeneralization in either of the lexicons. That is, the parallel data give a reason for more fine-grained distinctions in verb senses (i.e., more verb frames) for that particular verb in that valency lexicon.

For example, the English verb *bite* when translated as *kousnout* generates a conflict in the data. In one, rather idiomatic, occurrence, *bite one's lip*.PAT is translated with *kousnout se*.PAT *do rtu*.DIR3, thus aligning the English PAT with a Czech DIR3 functor. In another occurrence, arguably the more general one, the PAT actants of the verbs on both sides are aligned. Thus the data give evidence of a possible need of establishing a new frame for certain (for example, idiomatic) uses of the verb.

Second, conflicts arise in rather specific syntactic constructions, i.e., for two syntactic constructions, a default one and a specific one, which are otherwise considered to represent the same valency frame, though having a different placement of semantic modifications in the syntactic structure.



En: ”These cases lead to the loss of ... credibility,” a ministry statement said.
Cz: ”Tyto případy vedou ke ztrátě důvěryhodnosti ...,” uvádělo se v prohlášení ministerstva.

*Figure 5. Conflicting occurrence of an* ACT→LOC *alignment (vs.* ACT→ACT*)*

An example documenting this case is shown in Fig. 5, where we see a conflicting alignment for the pair *say–uvádět* (in the appropriate senses). In many (other) instances, the standard alignment of ACT (ACT→ACT) applies (*The president*.ACT *said that ...–Prezident*.ACT *uváděl, že ...*). However, in the parallel sentences depicted in Fig. 5: the same frame pair would lead to a different, non-identical mapping (ACT→LOC). This locative representation of the *medium of information transfer* modification (Cz: *prohlášení*), combined with a reflexive passive of the verb, is a syntactically typical alternation for Czech (but *only* for such a “medium” class of words, as opposed to persons etc.), whereas in English, the *medium* (En: *statement*) usually takes the subject (ACT in a canonical active sentence form) position in the sentence.

Third, conflicts can be lexically motivated, depending on the translation variant chosen by the translator. This differs from the first case above in that it is not possible to classify this as a difference in granularity of the valency frame(s), since the expression(s) used may not be considered clear idioms.

Conflicts have not been resolved on solid theoretical grounds in the current version of CzEngVallex, but notes from the annotation process have been preserved internally to reflect in future releases of the underlying treebanks, valency lexicons, or both (and, consequently, in CzEngVallex itself).

## 7. Specific linguistic issues

In the following sections, we describe some specific linguistic issues found in the data, we comment on their linguistic background and on the way they are annotated.

### 7.1. Catenative and modal verbs

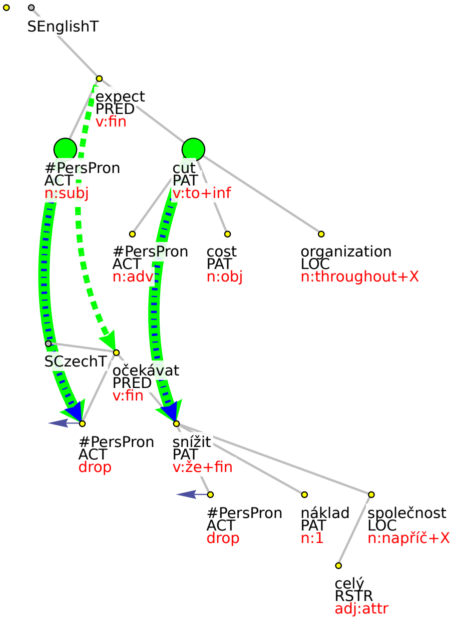
Special attention in the annotation was paid to verbs that form, together with another verb, a single homogeneous verb phrase, i.e., they precede another verb and function either as a chain element (catenative) or as an auxiliary (modal) verb. Catenative verbs are usually defined as those combining with non-finite verbal forms, with or without an intervening NP that might be interpreted as the subject of the dependent verbal form. Most of the classes described in Palmer (1974); Mindt (1999) can premodify main verbs and occupy the same syntactic position as auxiliaries or modals. They often cause some kind of structural discrepancy in the data.[30]

#### 7.1.1. ECM Constructions, Raising to Object

Most Czech linguistic approaches do not recognize the term Exceptional Case Marking (ECM) in the sense of "raising to object", instead they generally address similar constructions under the label "accusative with infinitive". The difference between ECM and control verbs is not being taken into account in most of Czech grammars. In short, raising and ECM are generally considered a marginal phenomenon in Czech and are not being treated conceptually (Panevová, 1996), except for several attempts to describe agreement issues, e.g., the morphological behaviour of predicative complements described in a phrase structure grammar formalism (Przepiórkowski and Rosen, 2005).

The reason for this particular approach to ECM is probably rooted in the low frequency of ECM constructions in Czech. Czech sentences corresponding to English sentences with ECM mostly do not allow catenative constructions. They usually involve a standard dependent clause with a finite verb, see Fig.6, or they include a nominalization, thus keeping the structures strictly parallel.

[30] By a structural discrepancy in dependencies, we mean such structural configurations that involve different number of dependencies in the corresponding syntactic structures, i.e., an alignment of "something" on one side of the translation to "nothing" on the other side, see also Sec. 6.2.



En: They expect him to cut costs...
Cz: Očekávají, že sníží náklady...

*Figure 6. Alignment of the ECM construction*

The only exception are verbs of perception (*see*, *hear*), which usually allow both ways of Czech translation – with an accusative NP followed by a non-finite verb form (1a), or with a dependent clause (1b), not speaking about the third possibility involving an accusative NP followed by a dependent clause (1c).

(1) He saw Peter coming.

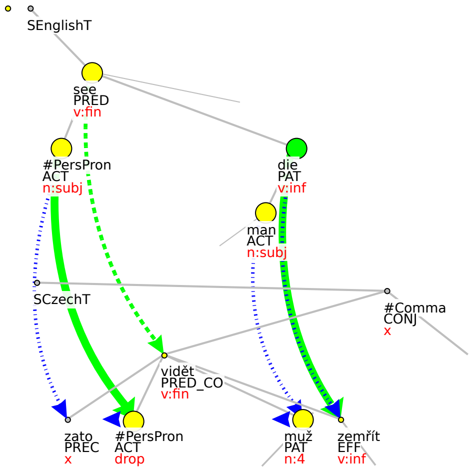
a. Viděl Petra přicházet.
He saw Peter.ACC to come.

b. Viděl, že Petr přichází.
He saw that Peter.NOM is coming.

c. Viděl Petra, jak přichází.
He saw Peter.ACC, how is coming.

In this type of accusative-infinitive sequence, the accusative element is in FGDVT analysed consistently as the direct object of the matrix verb (PAT) and the non-finite verb form then as the predicative complement of the verb (EFF).

The PCEDT annotation of verbs of perception is shown in Fig. 7, with frame arguments mapped in the following way: ACT→ACT; PAT→EFF; - - -→PAT. The corresponding arguments man-muž are interpreted as belonging to verbs in different levels of the structure.



En: I have seen [one or two] men die...
Cz: Zato jsem viděla [jednoho nebo dva] muže zemřít...

*Figure 7. Alignment of the perception verbs' arguments.*

The literature mentions two ways of ECM structural analysis, a flat one, representing the NP as dependent on the matrix verb, and a layered one, representing the intervening NP as the subject of the dependent verb. This mirrors the opinion that verbs allowing ECM usually have three syntactic, but only two semantic arguments. The practical solution is then a matter of decision between a syntactic and semantic approach to tree construction.

The English part of the PCEDT data was annotated in the layered manner,[31] thus most of the pairs in the treebank appear as strictly parallel. The consistency of structures is one of the most important advantages of the layered approach; there is no need of having two distinct valency frames for the two syntactic constructions of the verb, therefore, the semantic relatedness of the verb forms is kept.

[31] The annotation followed the original phrasal annotation of the data in the Penn Treebank.

On the other hand, the Czech part of the PCEDT data uses flat annotation, partly because the catenative construction with raising structure is fairly uncommon in Czech (cf. Sect. 7.1.1). The flat structure is easier to interpret, or translate in a morphologically correct way to the surface realization, but it requires multiple frames for semantically similar verb forms (the instances of the verb *to see* in *see the house fall* and *see the house* are in the FGD valency approach considered two distinct lexical units) and it also leaves alignment mismatches in the parallel data.

The treatment of ECM constructions in English and in Czech is different. It reflects both the differences internal to the languages and their consequences in theoretical thinking. Contrary to English, Czech nouns carry strong indicators of morphology – case, number and gender. The rules for the subject-verb agreement block overt realization of subjects of the infinitives. The accusative ending naturally leads to the interpretation of the presumed subject of the infinitive as the object of the matrix verb. The morphosyntactic representation is taken as a strong argument for using a flat structure in the semantic representation, and a covert co-referential element for filling the "empty" ACTor position of the infinitive. In English, in general, there is no such strong indication and therefore the layered structure is preferred in the semantic representation.

### 7.1.2. Object control verbs, equi verbs, causatives

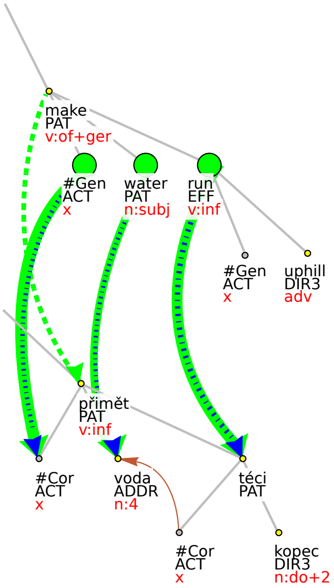
Contrary to the ECM constructions, object control verbs constructions (OCV), involving verbs such as *make, cause, or get*, are analyzed strictly as double-object in both languages. OCV constructions are similarly frequent in Czech and English and their alignment in the PCEDT data is balanced, see Fig. 8.[32]

Interestingly, it is sometimes the case that English control verbs in the treebank are translated with non-control, non-catenative verbs on the Czech side, and the intervening noun phrase is transformed to a dependent of the lower verb of the dependent clause (see Fig. 9).

The verb involved in this kind of translation shift may be either a more remote synonym, or a conversive verb.[33] Such a translation shift brings about (at least a slight) semantic shift in the interpretation, usually in the sense of de-causativisation of the meaning (*prompt*→*lead to*). of (any) language to suppress certain aspects of meaning without losing the general sense of synonymity.

[32] In Fig. 8, English ACT of *run* does not show the coreference link to *water* since the annotation of coreferential relations has not yet been completed on the English side of the PCEDT, as opposed to the Czech side (cf. the coreference link from ACT of *téci* to *voda*).

[33] Semantic conversion in our understanding relates different lexical units, or different meanings of the same lexical unit, which share the same situational meaning. The valency frames of conversive verbs can differ in the number and type of valency complementations, their obligatoriness or morphemic forms. Prototypically, semantic conversion involves permutation of situational modifications.



En: ...making water run uphill...
Cz: ...přimět vodu téct do kopce...

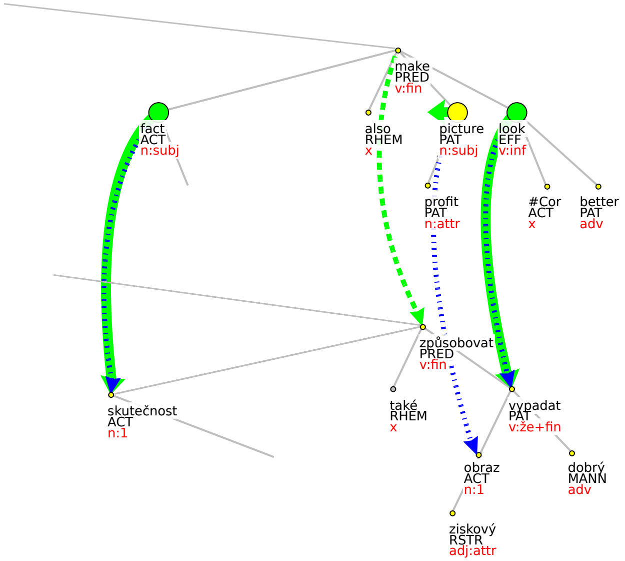
*Figure 8. Alignment of the control verbs' arguments*

Such occurrences have been treated as typical examples of zero alignment (see Sec. 6.2.2).

## 7.2. Complex Predication

By "complex predication" we mean a combination of two lexical units, usually a (semantically empty, or "light") verb and a noun (carrying main lexical meaning and marked with CPHR functor in the data), forming a predicate with a single semantic reference, e.g., *to make an announcement*, *to undertake preparations*, *to get an order*. There are some direct consequences for the syntactically annotated parallel data where we encounter two types of zero alignment.

First type of zero alignment is connected to the fact that a complex predication in one language can be easily translated with a one-word reference, and consequently aligned to a one-word predication, in the other language. This is quite a trivial case. In the data, then, one component of the complex predication remains unaligned. There



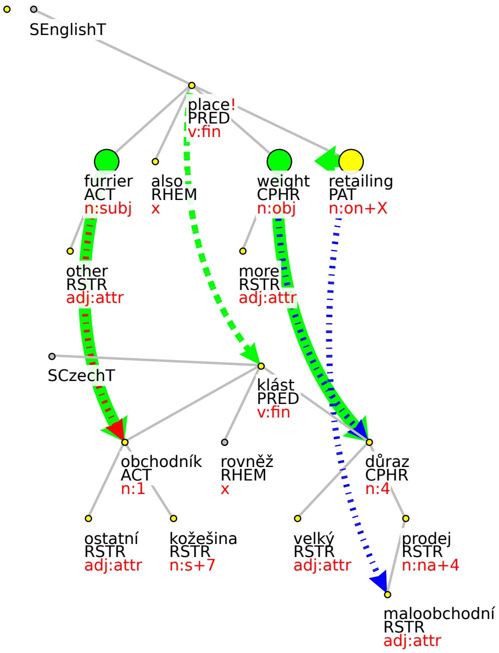
En: The fact ... will also make the profit picture look better.
Cz: Skutečnost ... způsobuje, že ziskový obraz vypadá lépe.

*Figure 9. Alignment of English OCV with Czech non-OCV construction*

are basically two ways of resolving such cases: either one can align the light verb with the full verb in the other language, or one can align the full verb with the dependent noun in the complex predication, based on the similarity of semantic content. In the CzEngVallex, the decision was to align the verbs, reflecting the fact that the verb and the noun phrase form a single unit from the semantic point of view.

The second type of zero alignment is connected to the presence of a "third" element within the complex predication structure, structured as dependent on the verb on one side, and on the predicative noun on the other side of the translation, e.g., En: *placed weight on retailing* - Cz: *klást důraz na prodej*, see Fig. 10.

Complex predicates have been annotated according to quite a complicated set of rules on the Czech side of the PCEDT data (Mikulová et al., 2006b). Those rules include also the so-called dual function of a valency complementation. There are two possible dependency positions for the "third" argument of the complex predicate: either it is modelled as the dependent of the semantically empty verb, or as a dependent of the nominal component. The decision between the two positions relies on multiple factors, such as valency structure of the semantically full use of the verb, valency



En: Other furriers have also placed more weight on retailing.
Cz: Ostatní obchodníci s kožešinami rovněž kladou větší důraz na maloobchodní prodej.

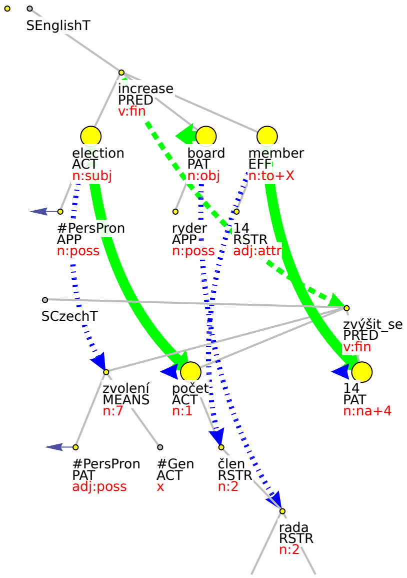
*Figure 10. Mismatch due to complex predication solution*

structure of the noun in other contexts, behaviour of synonymous verbs etc. On the Czech side, the "third" argument was strongly preferred to be a dependent of the nominal component. On the English side of the PCEDT, the preferred decision was different. The "third" argument was annotated as a direct dependent of the light verb (probably due to lower confidence of non-native speaker annotators in judging verb valency issues).

There is probably no chance of dealing with the dependencies in one of the two above stated ways only. The class of complex predicates in the data is wide and heterogeneous with respect to semantic and morphosyntactic qualities. Nevertheless, though resigning on the absolute consistency of the class, we may reach at least the consistency within the treatment of the individual light verbs throughout the corpus.

### 7.3. Conversive Verbs

A considerable number of unaligned arguments in the data is caused by the translator's choice of a verb in a conversive relation to the verb used in the original language. For some reason (e.g., frequency of the verbal lexical unit, topic-focus articulation etc.), the translator decides not to use the syntactically most similar lexical unit, but uses a conversive one (cf. also Sect. 7.1.2), thus causing the arguments to relocate in the deep syntactic structure, see Fig. 11.

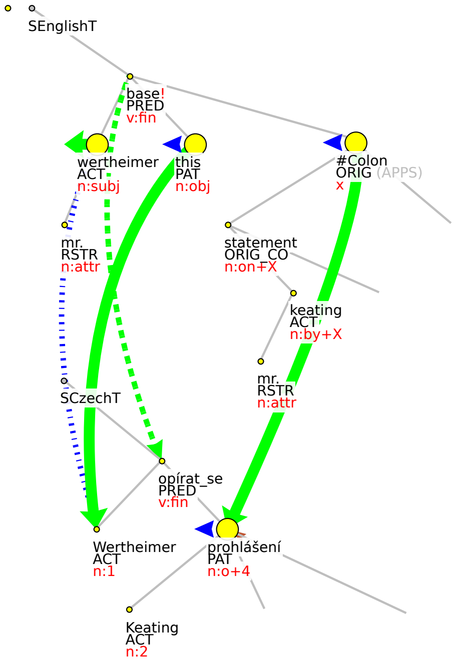


En: His election increases Ryder's board to 14 members.
Cz: Jeho zvolením se počet členů správní rady společnosti Ryder zvýšil na 14.

*Figure 11. Mismatch due to the use of conversive verbs*

The relocation of arguments frequently goes together with backgrounding of one of the arguments, which then either disappears from the translation, or is transformed into an adjunct, or into a dependent argument embedded even lower in the structure.

The first actant (ACT) in the FGD approach is strongly underspecified. It is mostly delimited by its position in the tectogrammatic annotation. Its prevalent morphosyntactic realization is nominative case, but certain exceptions are recognized (verbs of feeling etc.). Also, the ACT position is subject to the process called "shifting of cognitive roles” (Panevová, 1974), cf. Sec. 2.1, i.e., other semantic roles can take the nominative case and the corresponding place in the structure in case there is no semantic agent in the structure. Thus we get semantically quite different elements (e.g., +anim vs. -anim) in the ACT position, even with formally identical verb instances (Fig. 12 and 13).

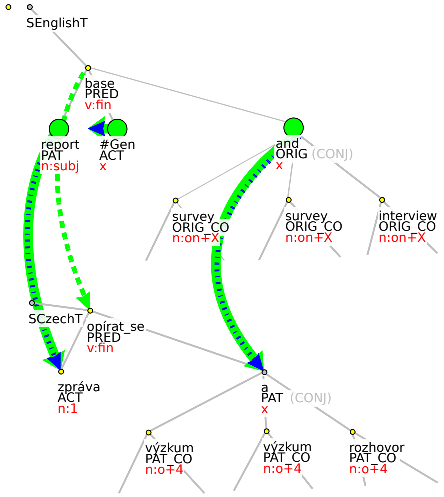


En: Mr. Wertheimer based this on a statement by Mr. Keating...
Cz: Wertheimer se opírá o prohlášení Keatinga...

*Figure 12. Conflict due to the underspecification of the ACT position*

This formal feature of the FGDVT gives rise to a number of conflicts in the parallel structures considering structures that undergo semantic de-agentization or (milder) de-concretization of the agent.

Here the question arises, whether such verb instances correspond to different meanings of the verb, or whether they correspond to a single meaning (represented by a single valency frame). It is often the case, that the Czech data tend to overgeneralize the valency frames through considering the different instances as realizations of a single deep syntactic valency frame, when there is no other modification intervening in the frame. Therefore, this approach chosen for the Czech annotation sometimes shows a conflict, as in Fig. 12 and 13.



En: The report was based on a telephone survey...
Cz: Zpráva se opírá o telefonický výzkum...

*Figure 13. Original collect for the verbs* base *and* opírat se

The valency structure for both instances of *base* (in Fig. 12 and 13) is identical, only in the first case, the verb is used in active voice, whereas in the second case, it is in passive voice. There are three semantic arguments in the structure. We will call them the Person that expresses an opinion, the Expressed Opinion and the Resource for the opinion. The Person bases the Expressed Opinion on the Resource. With the English verb, the Expressed Opinion always takes the PAT position and the Resource the ORIGin position in the valency structure. On the other hand, on the Czech side of the data, there is a conflict. In both Czech cases, there are seemingly only two arguments. In the first case, the Expressed Opinion is sort of backgrounded from the semantic structure. In the second case, on the other hand, the structure follows the passivized English structure in backgrounding the Person, the Expressed Opinion does not take the PAT position, but the ACT position in the structure, which is the cause of the conflict (for more details, see Šindlerová et al. (2015)).

The conflicts in annotation have a substantial reason – the ways in which English and Czech express backgrounding of the agent are multiple and they differ across the languages. Czech uses the *se*-morphemization often, in order to preserve the topic focus articulation (information) structure, whereas English does not have such a morpheme to work with, so it often uses simple passivization, or middle construction.

Moreover, the first valency position in Czech is often overgeneralized, allowing a multitude of semantically different arguments, which is, due to "economy of description", sometimes not reflected in the linguistic theory.

### 7.4. Head-dependent switch

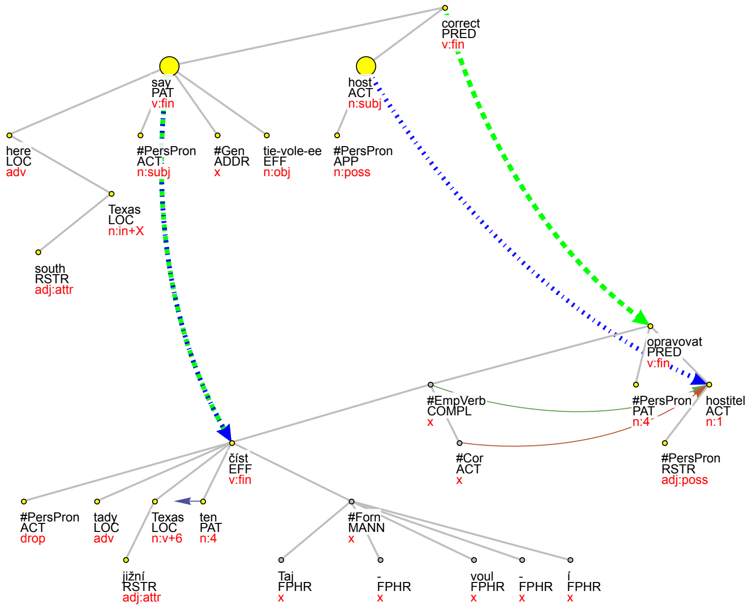
Due to some differences in annotation guidelines for the two languages, or due to translation issues, some slight semantic "switches" in alignments are allowed in order to map the arguments properly.

A frequent case of a head-dependent switch involves numerical expressions. For example, the English phrase *many economists* is annotated with *economist* as a head (labeled as argument) but in its Czech translation *řada ekonomů*, the word *řada* is, on the basis of its morphosyntactic behaviour, considered the head (labeled as valency argument), with *economist* in a dependent position. Numerical expressions overtaking the head position (with certain morphosyntactic consequences for the sentence) are called "container" expressions. With container expression of one side of translation, and modifying numeral on the other side, the alignment should be considered as encompassing a small subtree as opposed to a single node. Nevertheless, the annotators were asked to align head to head (i.e., align both direct daughters of the verb and arguments). In the above example, the word *economist* and *řada* are aligned instead of aligning the English head (*economist*) with the Czech dependent (*ekonom*) according to the very meaning of the lexical items, see Fig. 3 on page 30.

Another manifestation of the problem comes with the names of companies (e.g., *IBM*). Due to preservation of an appropriate inflection marking in the Czech translation, they are usually preceded with a generic name like *společnost* (*company*) in the Czech sentence, whereas they are used on their own in the English version of the sentence. In such cases, the alignment again is to be viewed as covering the whole subtree in Czech, and thus the nodes *IBM* and *společnost* are aligned.

### 7.5. Direct speech

According to the annotation guidelines, the annotation rules for direct speech in English (Cinková et al., 2006) and Czech (Mikulová et al., 2006a) on the tectogrammatical level are similar. Both languages add a new node representing the gerund (transgressive) of a verb of saying to the tectogrammatical annotation in cases where



En: "Here in south Texas we say Tie-vole-ee," my host ... corrects .
Cz: "Tady v jižním Texasu to čteme Taj-voul-í," ... mě opravuje můj hostitel.

*Figure 14. Direct speech alignment*

the direct speech is adjacent to a verb which cannot be considered a verb reporting the direct speech (none of the arguments of the valency frame of the verb can be expressed by the direct speech). This newly added node is assigned a `t_lemma` substitute `#EmpVerb` and the functor `COMPL`. An example of a direct speech paraphrasable with a verb of saying: *Vtrhl do dveří* #EmpVerb.`COMPL`: *„Kdy bude.*`EFF` *večeře?“* (*He burst in at the door: "When will the dinner be ready?"*)

Due to the same instructions, mismatches were not expected in collecting direct speech utterances. Nevertheless, the annotation process reveals some discrepancies, as shown in Fig. 14, where the collected frame pair is as follows: `ACT`→`ACT PAT`→`---`, `---`→`PAT`.

The mismatch occurs due to a different practical annotation approach to direct speech in the individual languages, most notably, the English annotation often deviates from the common guidelines. While in Czech the use of `#EmpVerb` and the functor `COMPL` is common, in English the addition of the `#EmpVerb` node is rarely done.

In case of such a discrepancy in the data, based on the presence of a `COMPL` node on just one side of the translation, the annotator is asked neither to align the direct argument of the other side to the COMPL node, nor to its lexical counterpart, but rather to collect the zero alignment (alignment to no specific node in the structure, see Sec. 6.2.2). Such structures are left for future treatment within possible tectogrammatical annotation revisions.

## 8. Use and future work

The CzEngVallex has been planned as a resource to be used both for the purposes of possibly revising theoretical linguistic accounts of verbal valency from a crosslinguistic perspective, and for an innovative use in various NLP tasks.

In both of these areas, the CzEngVallex has proved to be a valid resource. Our publications Šindlerová et al. (2013); Urešová et al. (2013); Šindlerová et al. (2014); Urešová et al. (2014a, 2015); Šindlerová et al. (2015); Urešová et al. (2015) show some interesting and important results concerning verbal valency from the Czech-English comparison perspective, while Dušek et al. (2014, 2015) shows that the inclusion of the CzEngVallex bilingual mapping feature into a word sense disambiguation task significantly improves the performance of the system. Our findings are also very useful when comparing different formal representations of meaning, see Xue et al. (2014); Urešová et al. (2014b); Oepen et al. (2015).

As for future work, a more detailed comparative description of the argument structure of translation equivalents found in the data would be needed. The attention should be paid especially to verb–non-verb or non-verb–verb pairs which were not included in the first version of CzEngVallex. And, of course, there exist many other manifestations of the above mentioned phenomena: functor mismatches, conflicts in data, zero alignments, which deserve our future attention and which might - on top of their better understanding from the linguistic point of view - lead to changes in the structure and content of the underlying valency lexicons towards a more universal valency description with less differences across languages. The results could also influence translation studies and the practice of translation, as well as deep methods in the area of natural language processing.

We also plan to create (manually but with substantial computational support) a class-based "superlexicon" over the CzEngVallex, grouping together synonyms or at least related sense pairs.

## Acknowledgements

We would like to thank the reviewers for their insightful comments, that help improve the paper. This work described herein has been supported by the grant GP13-03351P of the Grant Agency of the Czech Republic and it is using language resources hosted by the LINDAT/CLARIN Research Infrastructure, project LM2010013 funded by the MEYS of the Czech Republic.

## Bibliography

Bojar, Ondřej and Jana Šindlerová. 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, 2010. ELRA.

Cinková, Silvie. 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, 2006. ELRA, ELRA. ISBN 2-9517408-2-4.

Cinková, Silvie, Jan Hajič, Marie Mikulová, Lucie Mladová, Anja Nedolužko, Petr Pajas, Jarmila Panevová, Jiří Semecký, Jana Šindlerová, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. Annotation of English on the tectogrammatical level. Technical Report 35, UFAL MFF UK, 2006.

Dušek, Ondřej, Jan Hajič, and Zdeňka Urešová. Verbal Valency Frame Detection and Selection in Czech and English. In *The 2nd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 6–11, Stroudsburg, PA, USA, 2014. Association for Computational Linguistics. ISBN 978-1-941643-14-3.

Dušek, Ondřej, Eva Fučíková, Jan Hajič, Martin Popel, Jana Šindlerová, and Zdeňka Urešová. Using Parallel Texts and Lexicons for Verbal Word Sense Disambiguation. In Hajičová, Eva and Joakim Nivre, editors, *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 82–90, Uppsala, Sweden, 2015. Uppsala University, Uppsala University. ISBN 978-91-637-8965-6.

Hajič, Jan, 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. Prague Czech-English Dependency Treebank 2.0, 2011.

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

Mikulová, Marie, 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. Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual. Technical Report 30, ÚFAL MFF UK, Prague, Czech Rep., 2006a.

Mikulová, Marie, 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. Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual. Technical Report 30, ÚFAL MFF UK, Prague, Czech Rep., 2006b.

Mindt, Dieter. Finite vs. Non-Finite Verb Phrases in English. In *Form, Function and Variation in English*, pages 343–352, Frankfurt am Main, 1999. Peter Lang GmbH. ISBN 978-3-631-33081-4.

Oepen, Stephan, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Silvie Cinková, Dan Flickinger, Jan Hajič, and Zdeňka Urešová. SemEval 2015 Task 18: Broad-Coverage Semantic Dependency Parsing. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 915–926, Denver, Colorado, June 2015. Association for Computational Linguistics. ISBN 978-1-941643-40-2. URL `http://aclweb.org/anthology/S15-2153`.

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

Palmer, F. R. *The English verb / F. R. Palmer*. Longman London, 2d ed. edition, 1974. ISBN 058252458.

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

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

Panevová, J. On verbal Frames in Functional Generative Description. *Prague Bulletin of Mathematical Linguistics*, 22:3–40, 1974.

Panevová, Jarmila. More Remarks on Control. *Prague Linguistic Circle Papers*, 2(1):101–120, 1996.

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

Przepiórkowski, Adam and Alexandr Rosen. Czech and Polish Raising/Control with or without Structure Sharing. *Research in Language*, 3:33–66, 2005.

Šindlerová, Jana and Ondřej Bojar. Towards English-Czech Parallel Valency Lexicon via Treebank Examples. In *Proceedings of 8th Treebanks and Linguistic Theories Workshop (TLT)*, pages 185–195, Milano, Italy, 2009.

Šindlerová, Jana, Zdeňka Urešová, and Eva Fučíková. Verb Valency and Argument Non-correspondence in a Bilingual Treebank. In Gajdošová, Katarína and Adriána Žáková, editors, *Proceedings of the Seventh International Conference Slovko 2013; Natural Language Processing, Corpus Linguistics, E-learning*, pages 100–108, Lüdenscheid, Germany, 2013. Slovak National Corpus, L'. Štúr Institute of Linguistics, Slovak Academy of Sciences, RAM-Verlag. ISBN 978-3-942303-18-7.

Šindlerová, Jana, Zdeňka Urešová, and Eva Fučíková. Resources in Conflict: A Bilingual Valency Lexicon vs. a Bilingual Treebank vs. a Linguistic Theory. In Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, and Joseph Mariani, editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2490–2494, Reykjavík, Iceland, 2014. European Language Resources Association. ISBN 978-2-9517408-8-4.

Šindlerová, Jana, Eva Fučíková, and Zdeňka Urešová. Zero Alignment of Verb Arguments in a Parallel Treebank. In Hajičová, Eva and Joakim Nivre, editors, *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 330–339, Uppsala, Sweden, 2015. Uppsala University.

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

Urešová, Zdeňka. *Valenční slovník Pražského závislostního korpusu (PDT-Vallex)*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czechia, 2011b.

Urešová, Zdeňka, Eva Fučíková, Jan Hajič, and Jana Šindlerová. An Analysis of Annotation of Verb-Noun Idiomatic Combinations in a Parallel Dependency Corpus. In *The 9th Workshop on Multiword Expressions (MWE 2013)*, pages 58–63, Atlanta, Georgia, USA, 2013. Association for Computational Linguistics. ISBN 978-1-937284-47-3.

Urešová, Zdeňka, Eva Fučíková, Jan Hajič, and Jana Šindlerová. Verb-Noun Idiomatic Combinations in a Czech-English Dependency Corpus. In *PARSEME 2nd general meeting*, Athens, Greece, 2014a. Institute for Language and Speech Processing of the Athena Research Center.

Urešová, Zdeňka, Jan Hajič, and Ondřej Bojar. Comparing Czech and English AMRs. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014, at Coling 2014)*, pages 55–64, Dublin, Ireland, 2014b. Dublin City University, Association for Computational Linguistics and Dublin City University. ISBN 978-1-873769-44-7.

Urešová, Zdeňka, Ondřej Dušek, Eva Fučíková, Jan Hajič, and Jana Šindlerová. Bilingual English-Czech Valency Lexicon Linked to a Parallel Corpus. In *Proceedings of the The 9th Linguistic Annotation Workshop (LAW IX 2015)*, Stroudsburg, PA, USA, 2015. Association for Computational Linguistics.

Urešová, Zdeňka, Eva Fučíková, Jan Hajič, and Jana Šindlerová. CzEngVallex, 2015. URL `http://hdl.handle.net/11234/1-1512`. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Urešová, Zdeňka, Eva Fučíková, and Jana Šindlerová. CzEngVallex: Mapping Valency between Languages. Technical Report TR-2015-58, ÚFAL MFF UK, 2015.

Xue, Nianwen, Ondřej Bojar, Jan Hajič, Martha Palmer, Zdeňka Urešová, and Xiuhong Zhang. Not an Interlingua, But Close: Comparison of English AMRs to Chinese and Czech. In Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, and Joseph Mariani, editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1765–1772, Reykjavík, Iceland, 2014. European Language Resources Association. ISBN 978-2-9517408-8-4.

Žabokrtský, Zdeněk. Treex – an open-source framework for natural language processing. In Lopatková, Markéta, editor, *Information Technologies – Applications and Theory*, volume 788, pages 7–14, Košice, Slovakia, 2011. Univerzita Pavla Jozefa Šafárika v Košiciach. ISBN 978-80-89557-02-8.

**Address for correspondence:**
Zdeňka Urešová
`uresova@ufal.mff.cuni.cz`
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics, Charles University in Prague
Malostranské náměstí 25, 118 00 Praha 1, Czech Republic