M A T E M A T I C K O - F Y Z I K Á L N Í   F A K U L T A

P R A H A

CzEngVallex: Mapping Valency between Languages

Zdeňka Urešová, Eva Fučíková, Jana Šindlerová

ÚFAL Technical Report
TR-2015-58

ISSN 1214-5521

U N I V E R S I T A S   C A R O L I N A   P R A G E N S I S

Copies of ÚFAL Technical Reports can be ordered from:

Institute of Formal and Applied Linguistics (ÚFAL MFF UK)
Faculty of Mathematics and Physics, Charles University
Malostranské nám. 25, CZ-11800 Prague 1
Czech Republic

or can be obtained via the Web: http://ufal.mff.cuni.cz/techrep

Abstract

This report presents a guideline for building a resource connected with the 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 parallel Prague Czech-English Dependency Treebank. One of the main aims of this project is to create a high-quality and relatively large empirical base, a bilingual valency lexicon, the `CzEngVallex`, which could be used both for linguistically oriented comparative research, as well as for natural language processing applications, such as machine translation or cross-language sense disambiguation.

Acknowledgements

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.

Contents

1 Introduction

The presented report describes the way verbal valency is mapped between languages, in particular between English and Czech. It is primarily meant to be used by `CzEngVallex` annotators; it serves as an annotation guideline to the project called "A comparison of Czech and English verbal valency based on corpus material (theory and practice)", a research grant by the Grant Agency of the Czech Republic under the id GP13-03351P.

The overview of the project and the summary of the system of valency representation in the Functional Generative Description approach (FGD) is given in Section 2. Section 3 describes the data being worked with, Section 4 briefly summarizes the goals and the structure of the resulting resource, and Section 5 is concerned with the annotation tool used and project relevant features of the annotation environment. Starting with Section 6, the proper annotation guidelines are presented. First, general issues concerning the recommended order and ways of execution of the individual annotation steps are described. Section 7.3 comments on possible problems and errors in the annotation, considering the alignment of verbs and their modifications. Section 8 elaborates on concrete difficult issues an annotator might encounter in the data and suggests consistent ways of their treatment. Finally, Section 9 addresses the usability of the `CzEngVallex` data in linguistic research and applications, namely the current use of the annotated data in running machine learning experiments.

2 Project of a comparison of Czech and English verbal valency

The `CzEngVallex` is an output of a project of building a bilingual valency lexicon that would possibly be helpful in machine translation tasks. The aim of the project is a cross-linguistic comparison of valency behavior of Czech and English verbs. In the project, two main goals are 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 in both languages and the description of interlinking the translational verbal equivalents with drawing a follow-up comparison of the achieved results.

The project is based on the valency theory of the Functional Generative Description (FGD) and on its application to a corpus, namely to the Prague Dependency Treebank (PDT) [8]. This theoretical approach is highly suitable for the proposed specification of relations of verbal valency frames in both languages, relating to the semantic and morphosyntactic level. The work with the data includes the creation of a parallel Czech-English valency lexicon which will be interlinked with real examples of valency usage in the broad context of the Prague Czech-English Dependency Treebank (PCEDT) [7].

The underlying idea of the project is the following: since verbal valency is the core structural property that builds the clause, capturing the mappings[1] of

[1]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 individual valency positions, as well as the alignment of the translational equivalents of the verbs, should model the basic patterns of cross-lingual semantic relations. Moreover, having a resource that stores such relations for several thousands of verbs and verb-pairs, we may be able to generalize (on the basis of semantic relatedness, or classes) about the unseen verbs in a text.

2.1 The contents of the project

The project covers two major areas of research:

(i) the theoretical part of the research includes the specification of verbal valency relations between Czech and English, in particular between Czech and English valency frames including their arguments (from the FGD point of view) and the contrastive description of the above mentioned relations based on a richly annotated parallel corpus, namely the PCEDT, and

(ii) the practical part of the research includes detailed work with electronically created and accessible data, namely with the `PDT-Vallex` and the `EngVallex` lexicons and the PCEDT, in order to interlink the entries in both valency lexicons based on the real usage in the texts of the PCEDT.

Our approach to the issues of valency of Czech and English verbs applied in the project is based on the following points of view and uses the following principles and features (2.1.1 - 2.1.3):

2.1.1 Valency in the FGD

The project draws on the valency theory developed within the Functional Generative Description approach – the Functional Generative Description Valency Theory (FGDVT). 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 (phrase, sentence etc.). 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 selected dependent elements constituting the valency structure of the phrase (or sentence) can be represented by valency frames, which can be listed in valency dictionaries.

The basics of the FGD approach to valency can be found, e.g., in [20]. The FGD approaches valency as a special relation between a governing word and its dependents.[2] 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 with 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.

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

In principle, a valency complementation can bear any of the functors listed in Table 1. For a full list of all dependency relations and their labels, i.e., the functors, as they are used in the PDT (based on those described and used in the FGDVT), see also [12].

The FGDVT works with a systematic classification of verbal valency modifications along two axes. The first axis represents the opposition between inner modifications (arguments) and free modifications (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" (arguments) in the FGDVT: Actor/Bearer (`ACT`), Patient (`PAT`), Addressee (`ADDR`), Origin (`ORIG`) and Effect (`EFF`). Which functors are considered arguments has been determined according to two criteria. The first one says that arguments 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 argument is restricted to modify only a relatively closed class of verbs.

Out of the five argument types, the FGDVT states that the first two are connected with no specific globally defined semantics, contrary to the remaining three ones. The first argument 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 complement). 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 being shifted to the Patient position in case the verb has only two arguments. Similarly, any of the argument roles are shifted to the Actor position in case the verb has only a single valency position. I.e., the position of the first and the second argument (if there is any) in the structure must always bear the `ACT` and `PAT` labels respectively, disregarding the actual semantic role of the argument. In the sentence *Peter has dug a hole*, the semantic Effect (*a hole*) is labeled a Patient; similarly, in the sentence *The teacher asked the pupil* the semantic Addressee (*the pupil*) is shifted to the Patient position. In *The book came out* the deep object (semantic Patient, *the book*) is shifted to the Actor position due to the fact that the Actor position is not taken by any other lexical candidate and would otherwise remain unoccupied. This rule, when viewed from the annotation point of view, helps to keep consistency at the expense of lower "semantic precision".

The repertory of adjuncts (free modifications) is much larger than that of arguments (see again Table 1). The FGD distinguishes about 50 types of adjuncts (for the full list of adjuncts see [12]). Adjuncts are always determined semantically; their set might be divided into several subclasses, such a temporal (`TWHEN`, TSIN, TTILL, TFL, TFHL, THO, TPAR, TFRWH, TOWH), local (LOC, DIR1, DIR2, DIR3), causal (such as CAUS for cause, AIM, CRIT for 'according to', etc.) and other free modifications (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 modifications') comes from. Unlike arguments, morphemic realization of adjuncts is rarely (if ever) restricted by the particular verb.

Due to this "free nature" of adjuncts, only the presence of arguments (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 not listed in the valency frame.[3] As mentioned above, both arguments 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 modification is semantically optional. On the other hand, if the answer *I don't know* is disruptive in the (assumed) conversation, then the given modification is considered to be semantically obligatory. For further details, see [29].

2.1.2 Terminology note

In this technical report, from now on, we use the term "argument" in a simplifying manner, deviating a bit from the FGDVT approach as described in the previous section. We use the term "argument" for any element which is included in a valency frame in either `PDT-Vallex` or `EngVallex`. We will continue to use the "obligatory" term for any element in a valency frame which is unmarked, and the term "optional" for those marked in valency frames as optional (see 3.2.1 and 3.2.2). Only when necessary, we will distinguish between arguments (inner participants) and adjuncts (free modifications) as defined in the FGDVT.

2.1.3 Comparative character of the project

In the project, we search for differences in the expression of the same contents in two typologically different languages, which Czech and English undoubtedly are. The initial hypothesis is that even in relatively literal or exact translation, where the information and the meaning the sentences carry in both languages

[3] Note that the `EngVallex` sometimes includes optional adjuncts in the frame specification. This is a leftover from the automatic conversion procedure (starting from the PropBank frame files) which has been used to pre-process it for the manual re-annotation to fit the FGDVT scheme (see Sec. 3.2.2).

Label	Function	Type	Class
ACT	Actor	Actant	Actant
PAT	Patient	Actant	Actant
EFF	Effect	Actant	Actant
ADDR	Addressee	Actant	Actant
ORIG	Origin	Actant	Actant
TWHEN	Temporal - when	Free modif.	Temporal
TFHL	Temporal - for how long	Free modif.	Temporal
TFRWH	Temporal - from when	Free modif.	Temporal
THL	Temporal - how long	Free modif.	Temporal
THO	Temporal - how often	Free modif.	Temporal
TOWH	Temporal - to when	Free modif.	Temporal
TPAR	Temporal - parallel	Free modif.	Temporal
TSIN	Temporal - since when	Free modif.	Temporal
TTILL	Temporal - till	Free modif.	Temporal
DIR1	Directional - from	Free modif.	Locative/Directional
DIR2	Directional - which way	Free modif.	Locative/Directional
DIR3	Directional - to	Free modif.	Locative/Directional
LOC	Locative	Free modif.	Locative/Directional
AIM	Aim	Free modif.	Implicational
CAUS	Cause	Free modif.	Implicational
CNCS	Concession	Free modif.	Implicational
COND	Condition	Free modif.	Implicational
INTT	Intent	Free modif.	Implicational
ACMP	Accompaniment	Free modif.	Manner
CPR	Comparison	Free modif.	Manner
CRIT	Criterion	Free modif.	Manner
DIFF	Difference	Free modif.	Manner
EXT	Extent	Free modif.	Manner
MANN	Manner	Free modif.	Manner
MEANS	Means	Free modif.	Manner
REG	Regard	Free modif.	Manner
RESL	Result	Free modif.	Manner
RESTR	Restriction	Free modif.	Manner
CPHR	Compound predicate	Free modif.	Multi-word
DPHR	Phraseme	Free modif.	Multi-word
BEN	Benefactor	Free modif.	Specific
HER	Heritage	Free modif.	Specific
SUBS	Substitution	Free modif.	Specific
CONTRD	Contradiction	Free modif.	Specific
COMPL	Complement	Free modif.	Predicative complement

Table 1: Functors allowed for valency complementations

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 produce not only the detection of the types of divergences of expression in the core sentence structure but also the quantitative analysis of their similarities and differences, thanks to the substantial size of the corpora available.

Both lexicons which we use as a starting point are based on the same theoretical approach (cf. Sec. 2.1.1). Our task is thus slightly simplified in that we are not comparing two different valency theories, but rather an application of a single theoretical (and formal) framework to the two languages in question (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.

2.1.4 Corpus approach to cross-language research

Our approach to the comparative study of valency in this project 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 Prague Czech-English Dependency Treebank (PCEDT) and the very large CzEng parallel corpus, is the cornerstone of our research. Specifically, the PCEDT, which is a 1-million word Czech-English parallel corpus, is systematically explored to get a complex picture of the relations between valency frames and their arguments. This corpus contains original English texts which are aligned with their corresponding translations, and manually analyzed on the traditional three layers of the Prague Dependency Treebanks: morphology, syntax, and tectogrammatics (semantics).

Moreover, we are able to take advantage of the much larger (yet only automatically analyzed) Czech-English parallel CzEng 1.0 corpus, which contains over 200 million words on both sides, while making smaller scale comparisons on the carefully selected texts and genres as contained in the InterCorp corpus. We are convinced that these corpora have the potential to help us to get enough material for the comparative description of verbal valency in Czech and English, and thus to reliably interpret our findings and individual hypotheses. However, we are also aware of the danger of misinterpretation due to the possible shortage of characteristic samples or examples, as follows from the infamous Zipf law; therefore, we might be forced to limit ourselves to study only a subset of verbs (or their senses) with enough corpus evidence.

3 CzEngVallex reference data

For the `CzEngVallex` project, two treebanks are most relevant: the PDT[4] and the PCEDT [7]),[5] which contain manual annotation of morphology, syntax and tectogrammatics (semantics).

The project assumes the use of (monolingual) valency lexicons as the starting point. In our project, we work with the verbal valency lexicon called `PDT-Vallex` [30] and with a similar resource for English called `EngVallex` [3].[6]

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

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

[6]The contents of both lexicons can be found at `http://ufal.mff.cuni.cz/pcedt2.`

All these data resources (the PDT corpus, the parallel PCEDT corpus and the two valency lexicons) are the “input” material for the creation of the new resource, `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 Wall Street Journal part of Penn treebank[7] 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 valency 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`, has been developed as a resource for valency annotation in a large-scale syntactically annotated corpus, the Prague Dependency Treebank. This lexicon is publicly available as a part of the PDT version 2 published by the Linguistic Data Consortium.[8]

The `PDT-Vallex` [30] is a valency lexicon containing Czech verbs, some Czech nouns and adjectives. It has been designed in close connection to the annotation of the PDT. The “bottom up” practical approach to the forming of the valency lexicon made it possible for the first time to confront the already existing valency theory 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 which comprises an understandable fragment of a Czech sentence, taken almost exclusively from the PDT corpus. The notes help to distinguish the meaning of the individual valency frames inside the valency entry. Typically, synonyms, antonyms and aspectual counterparts serve as notes.

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 coming from further annotations.

`0/en/documentation.html`, their structure in detail is described at `http://ufal.mff.cuni.cz/pcedt2.0/publications/t-man-en.pdf` and `http://ufal.mff.cuni.cz/pcedt2.0/en/valency.html`.

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

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

For a detailed information about the actual structure of the PDT-Vallex entry, see [29].

3.2.2 EngVallex - English valency Lexicon

The EngVallex is a lexicon of English verbs, also built according to the FGD theoretical framework. It was created by a (largely manual) adaptation of an already existing resource for English with similar aim, namely the PropBank Lexicon (PropBank "frame files") [18, 10], to PDT labeling standards (see also [3]).

During the adaptation process, arguments were re-labeled, obligatoriness was marked 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 whenever 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 Prague Czech-English Dependency Treebank (PCEDT).

The EngVallex currently contains 7,148 valency frames for 4,337 verbs. As in case of the PDT-Vallex, it is being constantly expanded and refined in the course of further annotation.

4 Building CzEngVallex

4.1 The annotation goal

For fulfilling the goal stated in the Sec. 2.1.4, an explicit linking between valency frames of Czech and English verbs based on a parallel corpus is needed. This has been accomplished by creating a bilingual Czech-English Valency Lexicon (CzEngVallex).

The CzEngVallex stores alignments between Czech and English valency frames and their arguments. The resulting alignments are captured in a standoff 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.

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.[9] Aligned pairs of verb frames are grouped by the English verb frame (<en_frame>), 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). In this particular example, for the pair *abandon*[10] - *opustit* (Lit. *leave [alone]*), we can observe a match of the first two arguments (ACT:ACT, PAT:PAT) and a zero alignment of the third frame element: EFF does not match any verb argument in this particular Czech counterpart.

[9] Similar scheme is used in [9].

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

```
<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 the `CzEngVallex` (part of *abandon* pairing)

On the other hand, for the pair *abandon* - *zříci se* (Lit. *get rid of [for sth]*), the third argument is involved in functor mismatch: `EFF` in English maps onto the Czech adjunct `SUBS` (substitution).

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 this 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 interface is an extension of the tree editor `TrEd` [16][11] environment.[12]

`TrEd` is a fully customizable and programmable graphical editor and viewer for tree-like structures (though it also can be used for annotating constituent trees). Among other projects, it was used as the main annotation tool for the tectogrammatical annotation of both source treebanks. It allows displaying and annotating sentential tree structures annotated on multiple linguistic layers with

[11] `http://ufal.mff.cuni.cz/tred`

[12] There exist also other environments for manual alignment, such as [11, 24, 1] and others; usually, they work with plain text or phrases, not dependency trees.

a variety of tags using either the Prague Markup Language (PML) format[13] or the `Treex` format.[14]

`Treex` (formerly TectoMT) [36, 22] is a development framework for general as well as specialized NLP tasks (such as machine translation) working with 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). The annotation interface, called (perhaps a bit confusingly also) `CzEngVallex`, has been built as an extension of the `TrEd` environment. It provides an annotation mode for valency frames alignment between the `PDT-Vallex` and the `EngVallex`. This extension builds on previously used `TrEd` extensions: pdt2.0 (for the annotation of the PDT 2.0), the `PDT-Vallex` extension, pedt (extension for annotating the English side of the PCEDT); they enable the functions necessary for browsing Czech and English treebanks and their valency lexicons, while the `CzEngVallex` extension provides the proper interlinking annotation environment.

The annotation is based on the data from the parallel Czech-English corpus PCEDT 2.0[15] which contains, aside from the bilingual data, also both (Czech and English) valency lexicons.

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 Prague Czech-English Dependency Treebank 2.0 (PCEDT);
- 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 filelists grouping together all verb-sense pairs for every English verb as collected within the previous step.

The GIZA++[16] algorithm was used for the word alignment of the PCEDT data, and subsequently, this alignment was mapped to the nodes of the corresponding (deep/tectogrammatical) dependency trees of the original and translated sentence.

Then, the process of 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; 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. 5.5). Thus, the annotator is able to inspect the same verb occurrences together in a single data block, and similarly, the individual pairs for the same source verb sense are sorted in succession within the groups.

[13]`http://ufal.mff.cuni.cz/jazz/PML`
[14]`http://ufal.mff.cuni.cz/treex`
[15]http://ufal.mff.cuni.cz/pcedt2.0/en/index.html
[16]`https://code.google.com/p/giza-pp`

5.3 The filelists

As described in the previous section, corresponding pairs of Czech and English verbs were looked up in the PCEDT, using a `btred`[17] 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 are 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 are sorted based on the English verb lemma and organized alphabetically into folders according to the first letter of the source verb. If a single English verb corresponds to more than one Czech verb, those verbs are located 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 have been 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 one 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 verbs and the number of occurrences of this pair in the parallel corpus (e.g., *abide_by.1v.2p.dodržovat.2.fl*).

The annotator is handed 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 size 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[18] 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 *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. will be 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).

5.4 Annotation setup

The PCEDT data are kept in a separate folder; each annotator works with her/his own copy. It is not allowed to change the core data, potential mistakes are marked (as notes) for further corrections.

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

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

The vallexes are part of the `CzEngVallex` extension and as such they are updated together with this extension. If the annotator wants to make changes to the vallexes, s/he should make her/his own working copy and change the path appropriately in the *ResourcePath* setting in the *Config File* (Menu-Setup-Edit Config File).[19]

Since the paths to the nodes in the filelists are relative, it is important that the filelists folder (containing a-z filelists subfolders) be placed in the same folder as the treex_files folder (containing PCEDT data).

5.5 The annotation process

During the actual annotation process, English and Czech verbs and their arguments are manually aligned, and after checking carefully all the occurrences of any given pair in the PCEDT data, the corresponding arguments are captured in the `CzEngVallex` dictionary, using the structure described in Sec. 4.2, which is in turn based on [25, 2].[20]

Even though all PCEDT occurrences of all verb–verb pairs are to be inspected manually, the process is helped substantially by the existence of the corpus and the valency lexicons themselves, as well as by several automatic preprocessing steps, as described in Sec. 5.2.

5.6 The capabilities of the annotation tool in detail

The nodes of the trees representing the translation-equivalent sentences of the PCEDT are automatically pre-aligned (Sec. 5.2). The children slots of the verb under inspection, as well as their automatically suggested inter-lingual English-Czech alignment, are highlighted to the annotator (for details, see Fig. 2 and Sec. 6.3).

The annotator may choose to see either all links in the given tree, or just links for the annotated verb pair. The annotator operates the annotation environment mostly with macros from the `CzEngVallex` extension. Their list is given in Table 2. Macros usually change values of individual attributes, or they add or delete whole nodes from the structure. Links, which lead from the source (English) to the target sentence (Czech) are manipulated using a drag-and-drop function.

The annotator-decided frame and valency slots alignments are stored in a separate file, called `frames_pairs.xml` (see Sec. 4.2), which is interlinked with the treebank data, as well as with the original Czech and English valency lexicons. This file must be saved continuously during the annotation process to avoid losing the work, using the `Save file FramesPairs` macro.

For saving the collected alignment of the given verb pair, use `P`.

Also, mind the necessity to save the changes in the trees (copy of the PCEDT) using the `Save current file` button in the `TrEd` interface. This

[19] For further reference, see the following link
`http://ufal.mff.cuni.cz/tred/documentation/ar01s13.html`

[20] In [25] and [2], only a pilot experiment has been described; the current process differs from the suggestions in these papers in several substantial respects.

!	Add Note
S	Add all artificial sons
s	Add artificial sons
ALT+c	Add or Remove valalign coref to remembered node
F	Browse FramesPairs file
CTRL+SHIFT+Return	Browse valency frame lexicon
ALT+l	Change not_collect (only for English nodes)
ALT+r	Change slot_remove (only for English nodes)
c	Collect slot links to FramesPairs
C	Collect slot links to FramesPairs for all nodes
r	Debug: Redraw automatic slot links
R	Debug: Redraw automatic slot links for all nodes
ALT+R	Delete slot links from FramesPairs
n	Edit Note
SHIFT+space	Forget remembered node
H	Handle all coordinations, appositions and SM
h	Handle coordination, apposition and SM
L	Reload file FramesPairs
space	Remember current node
P	Save file FramesPairs
CTRL+Return	Select and assign valency frame
f	Set functor for `CzEngVallex` purposes
a	Toggle display all nodes
A	Toggle display suggested arrows for all nodes

Table 2: `TrEd` Macros for CzEngVallex annotation

is important for keeping the changes in the alignment or functor correction etc., for the purposes of further processing and corrections in the original treebank.

The `TrEd` 's CzEngVallex extension offers the following edit options to the annotator:

1. align two nodes in between the sentences (i.e., add alignment(s) or correct the computer-suggested alignments);
2. delete alignments from the data;
3. mark nodes for not to be collected into the `frames_pairs.xml`;
4. add missing arguments nodes as direct verb dependents (missing frame nodes or nodes hanging higher or lower in the structure, due to coordination resolution etc.), in order to be able to add their alignment;
5. assign different `CzEngVallex` functor to arguments in the tree (the `CzEngVallex` functor will gain precedence over the PCEDT functor in the process of links collecting);
6. change frame assignment for a Czech or English verb in the data;
7. collect the slot alignments into the main file, `frames_pairs.xml`;

8. make changes in the `frames_pairs.xml` entries (add to or delete alignments from the entry);
9. delete entries from `frames_pairs.xml`;
10. write a note.

For each of the ten actions mentioned above, a simple macro (or a few macros) is provided within the `CzEngVallex` extension. For a complete list of macro functions see Table 2.

On the other hand, the following options are not allowed, due to the general requirement of keeping the annotation well defined:

1. align a source node to more than one target node;
2. align nodes with no functor assigned;
3. make changes in the structure of the sentence (nodes "re-hanging" or shifting);
4. collect slot alignments marked with non-argument functors (such as effective root functors, functors for rhematizers and linking adverbials, modals, adnominals, coordination, etc.; see Table 1 for allowed functors).

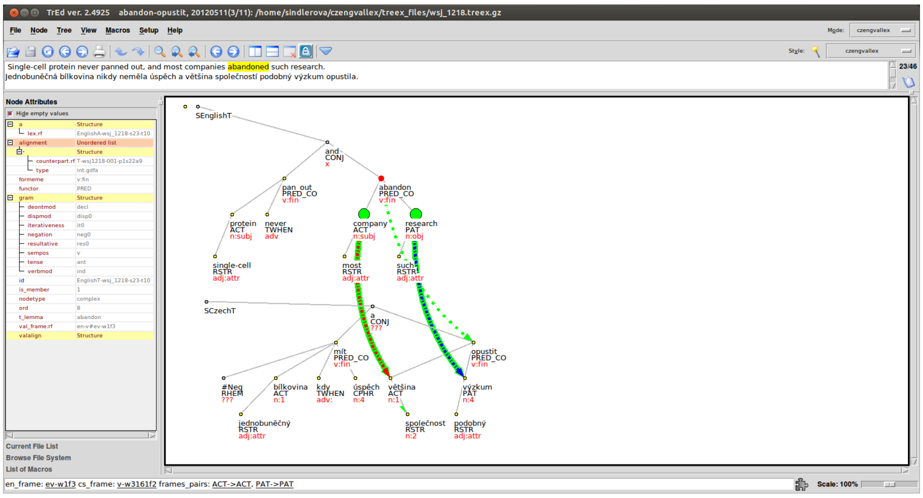


Figure 2: Annotation environment at work

The structure of the `frames_pairs.xml` file is a simple XML structure (cf. Fig. 1 in Sec. 4.2), 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). Collecting "zero alignment", i.e., alignment from a functor-labeled slot in the source language to "no slot" in the target language, and vice-versa, is possible.

Any problems encountered which are outside of the allowed changes, such as bad translation or annotation errors, can be described by the annotator in the appropriate "note" attribute of the governing verb, in order to be able to correct the lexicons and treebank data later.

The CzEngVallex extension offers the following pre-defined "note" attributes to the annotator, which can be extended by free text:

1. Frame: there is a mistake in the frame elements labeling in the lexicon, or the appropriate frame is missing completely, or a wrong frame is assigned;
2. Functor: there is a mistake in the functor assignment in the data;
3. Structure: the tree is ill-construed, or there is something in the structure that does not allow proper `CzEngVallex` annotation;
4. Translation: the translation is inappropriate, incorrect or too loose;
5. Question: space for storing theoretical or methodological questions of the annotators;
6. Other: residual issues.

6 Manual annotation workflow, functions and operations

6.1 Manual alignment - the starting point

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

Each annotator has her/his own copy of the `PDT-Vallex`, the `EngVallex` and the PCEDT and the filelists to work on (Sec. 5.2).[21]

S/he is expected to go through all verb occurrences in the filelist and build a typical valency frame alignment for each verb sense by collecting the frame alignments. S/he is 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 is automatically extended to all occurrences of the pair of the valency frames; it is 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` (i.e., in the `frames_pairs.xml` file).

Direct changes (changing the tree structure) in the treebank are disallowed, but the annotator reports problems through a note system for later corrections, and s/he is allowed to change the valency frame link if considered inappropriate. The changes made by the annotators over the separate copies of the valency lexicons and the pairing files are to be merged in the later stages of the project.

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

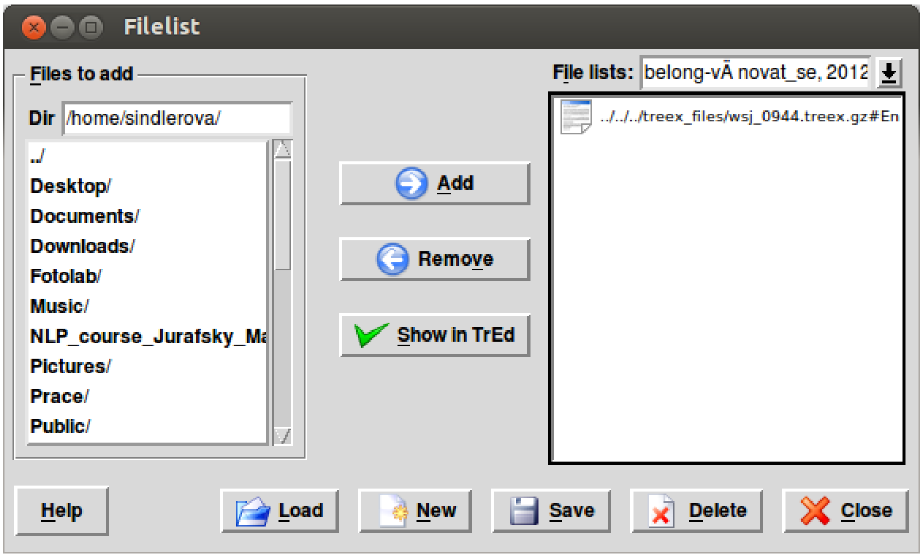


Figure 3: Manage Filelists Menu

6.2 Accessing the data - working with filelists

Filelists are being browsed, loaded and opened via the *File - File Lists - Manage Filelists* menu (see Fig. 3). After opening the *File Lists* dialogue the annotator can load the filelist s/he wants to annotate. By clicking the *Load* button (at the bottom of the *File Lists* window), or using `Alt-l`, the *Load Filelist* window opens and the annotator is asked to select the appropriate directory and file from the list of prepared filelists.

The *Show in TrEd* button (or using `Alt-s`) then shows the selected filelist in the editing environment and displays the first sentence pair on the screen (cf. Fig. 2).

The sentence pairs of an open filelist are passed through either by using arrow-to-stop buttons in the editor (the *Visit the next file* icon), or by pressing `F12` (forward) and `F11` (backward).

If there is a need for looking at the adjacent sentences within a file for a document context, i.e., if one only wants to move between sentences within one file, s/he just uses the arrow buttons (the *Next tree* icon).

If one wants to get to a specific sentence pair from the filelist, s/he can use a macro for jumping to a specific sentence, i.e., `Alt-Shift-g`, and enter the number of the desired sentence in a *Give a File Number* pop-up window. If one wants to view the list of sentences of the whole file, s/he opens a dialog in the right-hand corner of the `TrEd` editor (the “book” icon). This dialog window is also accessible via the *View - List of Sentences* menu.

For annotating the next filelist, the whole procedure must be repeated. Pressing the sequence of macros `Alt-f-l-m-l` retrieves a list of filelists, double clicking on an item selects a new filelist from the list, and pressing the *Show in TrEd* button (or using `Alt-s)` opens the first sentence pair of the new filelist in the `TrEd` editor for annotation.

6.3 Correcting alignments and data annotation mistakes

After the first sentence of the filelist is loaded and displayed, the annotator can start the `CzEngVallex` annotation proper, making the amendments and other actions as described in Sec. 5.6.

In each new sentence displayed by `TrEd` , the pre-aligned source verb is activated automatically (visually marked with the red color of the node). The automatic pre-alignment of the verb pair to be worked on is marked with a green dash-and-dot arrow. The direct children of the verb are highlighted with big yellow dots, their suggested alignment is marked with blue dash-and-dot arrows (see Fig. 4).

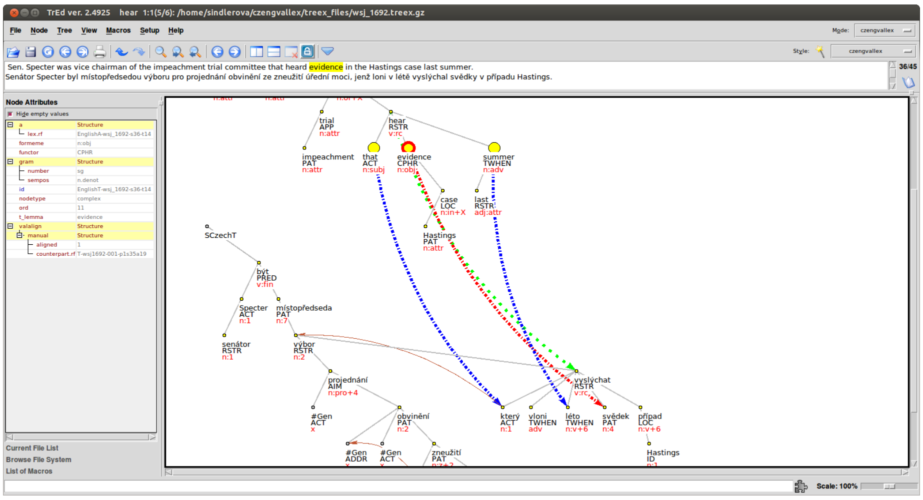


Figure 4: Highlighted alignments in `TrEd`; color-coding: green for verbs, blue for arguments, red for manually corrected alignment of arguments

6.3.1 Correction of the automatic pre-alignment

In case there is a mistake in the automatic verb pre-alignment (e.g., one of the verbs should be aligned with a different one than the one displayed, e.g., to a missing modal verb, or the automatic procedure has simply chosen a wrong word of the target sentence etc.), the annotator corrects the erroneous automatic pre-alignment, using the drag-and-drop function, dragging in the direction from the English verb node to the correct Czech verb node. Note that the arrow is redrawn immediately, and it is colored red to signal manual correction.

The annotator must always indicate that the verb alignment has been changed. S/he has to mark it in the note attribute of the English verb by using the macro `Shift-!` which opens the `Comment type` window. After choosing the type *Other* from the Menu, s/he inscribes the appropriate note into the *Comment text window*.

If the source verb cannot be easily aligned to any verb of the target sentence, the annotator should not annotate the frames-pair and s/he should also make an appropriate note (using note type *Other*) in the same way as described above.

Blue arrows connect arguments according to the automatic pre-alignment of the data. The annotator can re-align the arrows by clicking on the source node and dragging it to the target node. The manually-created argument-aligning arrows will appear marked in red. Note that if the drag-and-drop is performed on a verbal node, the displayed argument alignments will be switched to the alignments of the (now active) verbal node pair.

In order to change the displayed alignments back to the original verb pair (or in order to switch between alignments of different verb pairs in general), click on the source verb and use the macro `r`.

For redrawing the alignment of the given verb pair, use `r`.

An alignment between two nodes can also be deleted. This is to be done if an argument node is in fact not aligned to any argument on the other side. An unaligned argument will be marked by `---` on the side with "missing" node in the `frames_pairs.xml` file. There is no macro for deleting an alignment; it has to be done manually by deleting the alignment reference (the *counterpart.rf* and *type* attribute values) from the `attribute-value matrix` of the source node, see Fig. 5. It is accomplished by clicking the minus sign next to the `Structure` container within the `alignment` attribute. The resulting arrow displayed in the data will then visually lead from the source node horizontally into a free space; see e.g. the `MANN` node in Fig. 14.

Usually, not all the arguments captured in the valency frame of the annotated verb are actually present in the data as immediately dependent on the verb node. Some of the modifications may appear raised to a higher position in the tree as a consequence of identity of arguments of two coordinated verbs (the so-called common modification of coordinated nodes), see Fig. 6,[22] the node *price* belonging to the valency of both the coordinated verbs *continue* and *reverse*.

Special treatment has been chosen for arguments shared between verbs and coordinated arguments. In the case of shared arguments, the argument appears higher in the tree (as if dependent on the coordination node), and in the case of coordinated arguments, the arguments appear lower. Both examples can be seen at Fig. 6, where the coordinated verbs *continue* and *reverse* represent the `PAT` modification of the verb *believe*, and the *price*.`ACT` node represents a shared argument for these two verbs. For "pulling" the functor up the tree through intervening coordination/apposition nodes, the macro the `H` (`Shift-h`), is used, whereas the case of a shared argument is resolved by the macro `s`. Macro `s` is also responsible for virtually adding all the other missing arguments of the frame into the data for subsequent collection. After invoking these macros, the added arguments appear dependent on the verb directly as `#Slot` nodes.[23] This extra node can be in case of need removed by the *Node – Remove Active Node* option.

In order to solve both the coordination related issues properly, macro `H` (`Shift-h`) must be used before macro `s`, not vice versa.

[22]The images have been cropped or otherwise adjusted for the sake of clarity.

[23]This is only a simplification of the `CzEngVallex` extension, to make the visual representation of the alignment more transparent. In fact, this could have been solved by using the "effective parent" and "effective child" functions of `TrEd`, but the visual representation would then be confusing.

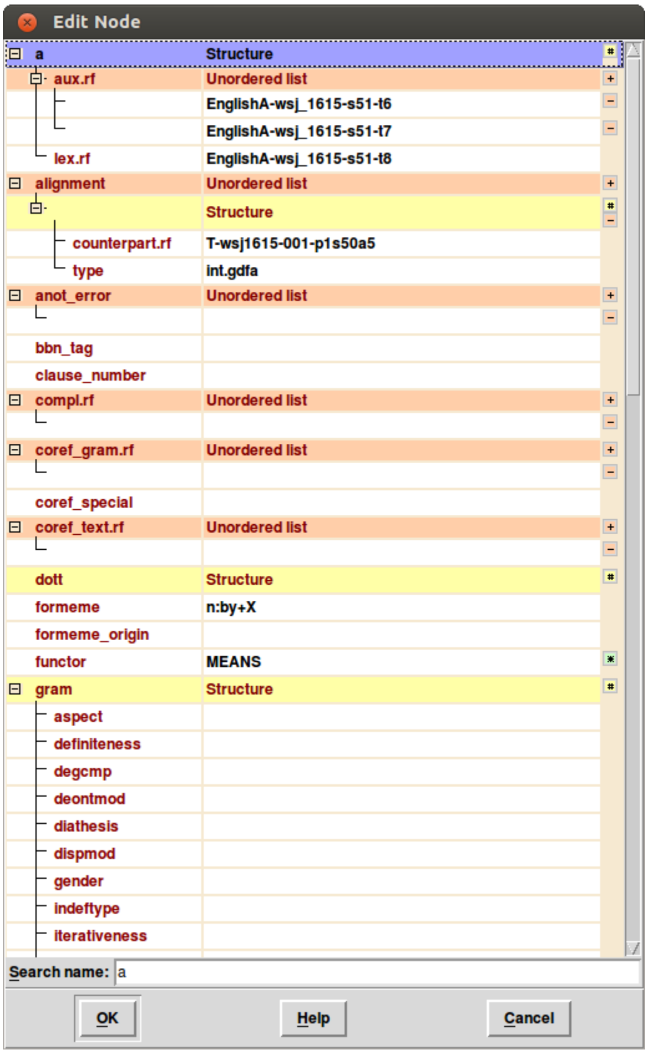


Figure 5: Attribute-value matrix

For solving the coordination and getting all relevant modifications under the verb node, use `H` (`Shift-h`) and `s`, in this order.

Note that the manual alignment (macro `s`) draws an alignment link for each daughter node of the verb, i.e., even for nodes that do not constitute valency slots of the frame. Alignment links that are not to be collected can be "forgotten" using the macro `Alt-l` which operates on the active source node of the unwanted alignment. After using this macro, the value of the attribute *not_collect* will be set to 1 and the blue/red arrow will disappear from the screen.

To toggle the (dis)allow flag for a particular node, use `Alt-l`.

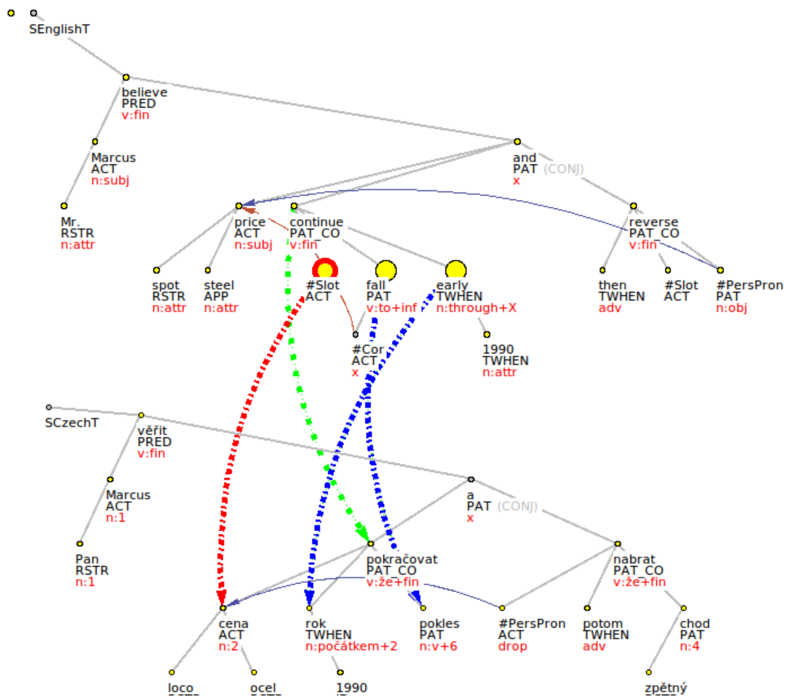


Figure 6: Coordination resolution

The annotator is expected to align all valency frame members (this includes possible zero alignments, on either side), and *exceptionally*, s/he may also introduce and alignment for other (non-argument) dependents of the verb, if there is a suspicion that their meaning might be important for the interpretation of the verb(-pair)'s meaning.

6.3.2 Dealing with mistakes from the previous data annotation

The annotator is generally not asked to check or correct previous tectogrammatical[24] annotation of the treebank (the PCEDT). The initial default approach of the annotator should be:

> **I believe there was a reason for choosing the particular translation/dependency/functor/frame. I trust the judgment of the translators and treebank annotators.**

Nevertheless, if the annotator comes across an evident and harmful mistake s/he is asked to make a note, and, if possible and in case it involves an allowed data change, also to correct it. Note that it is not possible to make changes in translation or dependencies (that would be a too massive change in the data, for which the current workflow is not ready), but it is possible

[24]Obviously, the same applies to the syntactic (analytic) and morphological annotation.

1. to change a functor (the new value will not overwrite the functor value in the data but will appear as a new `CzEngVallex` functor);
2. or to assign a different frame.

A different (`CzEngVallex`) functor value is assigned via the macro `f`. The annotator is asked to make a note (Comment type *Functor*) for postprocessing purposes.

For changing a functor, use `f`.

A corrected valency frame of the annotated verbs can be assigned directly in the `EngVallex` or the `PDT-Vallex` interface, using the macro `Ctrl-Enter` which opens the appropriate lexicon. S/he is however asked to always mark the change into the appropriate note (Comment type *Frame*) for postprocessing purposes.

It is not allowed to correct even obvious mistakes in a particular valency frame (using the `EngVallex` or the `PDT-Vallex` editor). In case the annotator sees the need of making changes in the valency frame itself, s/he is asked to suggest the change in an appropriate note (Comment type *Frame*) for a possible change later.

6.4 Collecting the node alignment

After the node alignments show the correct alignment of verb arguments, i.e., the annotator is satisfied with these alignments, the alignments are collected into the `frames_pairs.xml` file via the macro `c`. Such a *collect* operation means that the alignment between the valency frames and the alignments between the argument slots corresponding to the arguments and alignments in the displayed sentence pair are recorded into the `frames_pairs.xml` file.

The annotator is expected to collect the alignment of all valency frame members (including the zero alignment), and *exceptionally* also for other daughters of the verb, if there is a suspicion that their meaning might be important for the interpretation of the verb's meaning.

For collecting the alignment, use `c`.

Even after the *collect*, changes may be made when necessary. Changes in the stored alignment, or additions to the collected pairs are made by simply making a new *collect* (in the same or in a different tree).

Partial deletion of one or more alignments within the frames-pair is done using macro `Alt-r` while standing on the active source node of the unwanted alignment (this macro sets the value of the *slot_remove* attribute to 1, the arrow becomes marked in purple, see Fig. 7) and making a new *collect* afterwards. The whole frames-pair alignment is deleted from the `frames_pairs.xml` file through using macro `Alt-Shift-r` while standing on the active source verb node.

For deleting the collected alignment of two complementations, use `Alt-r`.

For deleting the collected alignment of the overall frames-pair, use `Alt-Shift-r`.

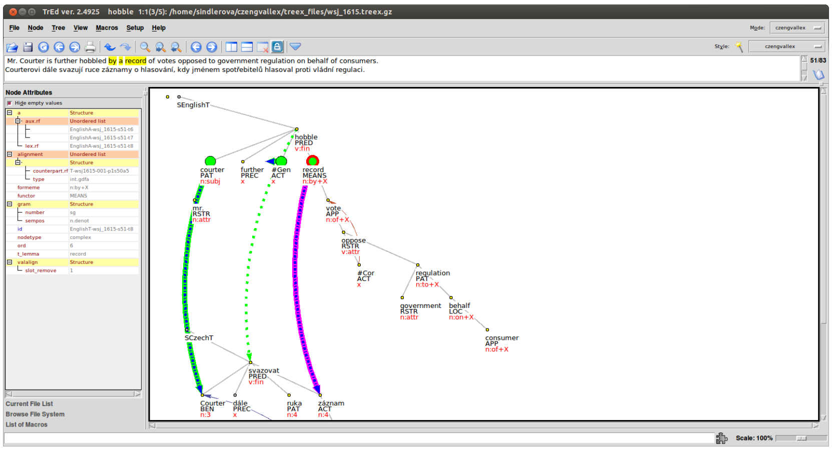


Figure 7: Highlighting of the alignment to be deleted

6.5 Annotation operations summarized

1. Coordination/apposition resolution

   - macro `H` (`Shift-h`) resolves coordination and apposition of the given verb by "pulling up" the arguments' nodes from below coordination and apposition nodes (`CONJ` or `APPS`), so that they end up directly dependent on the verb and thus ready for the collect operation (to work properly).

2. Completion of missing valency nodes according to the appropriate valency frame for the given verb sense

   - macro `s` adds nodes for arguments from the appropriate valency frames of the source and target language not contained in the particular sentence pair yet. To see the assigned valency frame, either drag your mouse over the verb and the associated frame shows in the yellow pop-up window, or use the `Ctrl-Enter` macro to open the Czech or English valency lexicon interface.

3. Possible removal of the automatic pre-alignment of some nodes

   - after activating the appropriate English node, use `Alt-l`.

4. Possible change in node alignment and deletion of an alignment

   - after activating the English node, use the drag-and-drop function to the target Czech node, i.e., hold the mouse button and move the node close enough to the Czech node which you wish to be newly aligned with the English node. In the correct position, when the Czech node turns green (originally being yellow), the mouse button might be released. Doing so, the automatically prepared blue arrow is being redrawn and turns red. Red color indicates the manually changed alignment. In case there is a need for zero alignment, delete the alignment identifier from the `attribute-value matrix` of the source node (open it by `Enter` when the node is activated).

5. Possible change of functor value of some valency nodes

   - for changing a functor, the macro `f` is used.

6. Recording the alignment into the `frames_pairs.xml` file

   - for this operation, the macro `c` (for `c`*ollect*) is used. The collected alignments, originally blue or red arrows, are overdrawn with a full green arrow. In case of need to change the recorded mapping, first change the alignment manually, subsequently use `c` for a new `c`*ollect*. The content of the `frames_pairs.xml` file for the given verb senses will be changed and the *collect* arrow redrawn. Note that the collected alignment is displayed in the bottom bar of the `TrEd`. For saving the `frames_pairs.xml` file, macro `P` is used.

7. Removal of node alignment from the frame alignment

   - to change an alignment in the already collected mapping, activate the English source node and use `Alt-r`. This macro is toggling the `vallalign–slot remove` attribute value to 0 or 1 and allows deleting the specific alignment from the `frames_pairs.xml` file when the next collect operation is performed.

8. Removal of a whole frame alignment from the `frames_pairs.xml` file

   - to delete a wrongly performed collect operation, use `Alt-Shift-r`. This will delete the alignment for the displayed frame pair.

9. Making notes

   - notes are added using the macro `Shift-!`; they can be deleted or amended using the macro `n`.

7 Basic Annotation Guidelines

7.1 When not to annotate (ignoring a pair)

Sometimes, all occurrences (one or more) of the same frame pair, as pre-aligned during the preprocessing stage, align such a diverging structures that a reasonable alignment of the verbs and their arguments is not possible.

These cases include:

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.

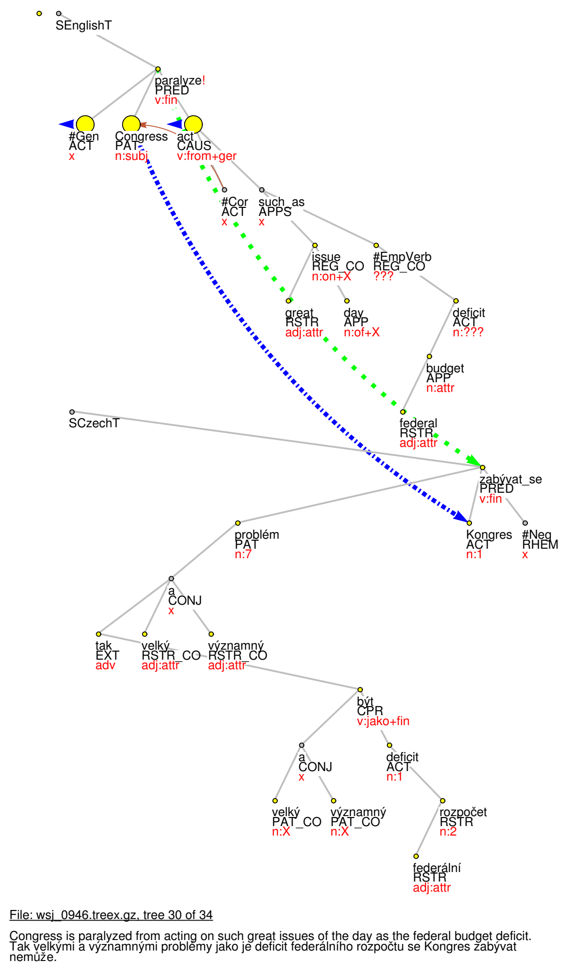


Figure 8: Too loose translation

Judging the degree of syntactic diversity is fully up to the annotator. In case of complex and rare syntactic differences, the annotator is required not to include the sentence (or more sentences for a given frame pair) in the annotation. The reason should be described in the note attribute (`Comment type` should be selected according to the type of the problem, e.g., often as *Other*). For example, if the translation is substantially inaccurate or if the translation is too loose (see Fig. 8), the sentences remain manually "unannotated," i.e., there is no attempt to correct alignments in the data or to make other data adjustments.

In case the annotator decides not to annotate, s/he also does not invoke the *collect* operation on the frame pair on that particular sentence, effectively not including the frame pair in the resulting `CzEngVallex`.

This is a different situation than a mere conflict in one (or a small number of) occurrence(s) of the frame pair among otherwise reasonably aligned occurrences of that frame pair in the treebank (for the annotation guidelines applicable to such cases, see Sec. 7.2.3).

7.2 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 correspond 1:1 and that their functors match.

Nevertheless, this is often not the case, thus the annotator will come across discrepancies and conflicts of different kinds in the data and s/he will have to deal with them.

By discrepancies, we refer either to the so-called zero alignment (see Sec. 7.2.1), 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 functor mismatch (7.2.2), i.e., two aligned corresponding nodes have different tectogrammatical functor labels.

By conflicts in annotation (Sec. 7.2.3), we refer to cases where the annotator is unable to align nodes representing the translation of the verb and its arguments for a given frame pair occurring somewhere in the data because the arguments were previously collected (for the same frame pair) in a different way. In other words, for that frame pair, such an alignment would be in conflict with the alignments observed elsewhere in the data.[25]

The most common types of alignment discrepancies are described in detail in Sec. 7.3.

7.2.1 Zero alignment

By zero alignment we mean such structural configurations that involve different number of arguments in 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 non-presence of a lexical or structural counterpart in the translation, or deeper embedding of an argument counterpart in a subtree.[26]

In Fig. 9, the reason is that in English the word *plan* is treated as an argument of the light verb *have*, whereas in Czech its counterpart (*plán*) depends on the nominal part of the light verb constructions (the word *připomínka* - lit. *comment*).

Similar case is depicted at Fig. 10 where the Czech equivalent (*propagace*) of the English `PAT` argument *advertising* is embedded lower in the structure.

Slightly different case appears for the verb pair *call/volat*, En: *...this calls into question the validity of the Rowland-Molina theory* / Cz: *...to volá po otázce po správnosti ... teorie*: the Czech equivalent *správnost* to the English *validity*.`PAT`ient 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 collecting the alignment of the appropriate "superfluous" node to "no specific node", see also Sec. 6.3.1.

[25] 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.

[26] For more details about zero alignment from the linguistic point of view see also [26].

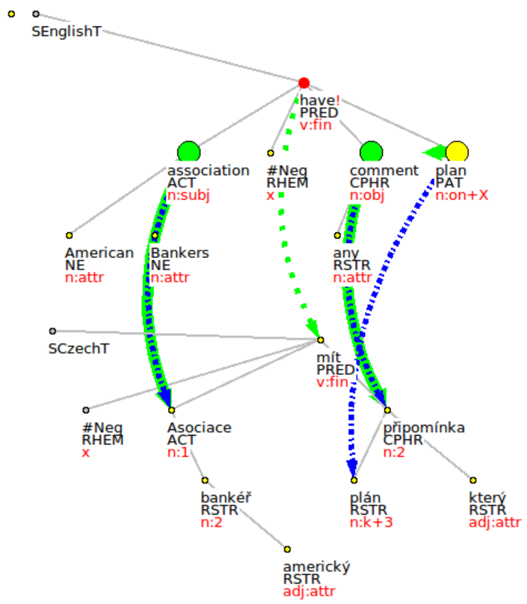


Figure 9: Zero alignment (embedded argument) `PAT`→`---`

7.2.2 Functor mismatch

By functor mismatch, we mean alignment of nodes with different functor labels, see Fig. 11. These alignments can involve either (proper) argument-argument mapping, or even an argument-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.

The collect operation with mismatching functors is technically unproblematic, but the annotator should always carefully consider correctness of the treebank annotation or the appropriateness of the assigned frame and in case of doubts, make a note. In the following paragraphs, we present some reasons for functor mismatch, as a possible guidance for the annotator what could be reasonable cases of a functor mismatch which can be treated by simply keeping such an alignment.

The data show that it is quite often the case that the alignment connects an argument (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. The linguistic reasons for this type of mismatch are usually grounded in different morphosyntactic forms of the given modifications, which were perhaps a bit overstressed when assigning the functor(s) to slots in the valency frames.

The alignment for individual functor pairs seems to be quite consistent throughout certain verb pairs or even verb classes. 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*

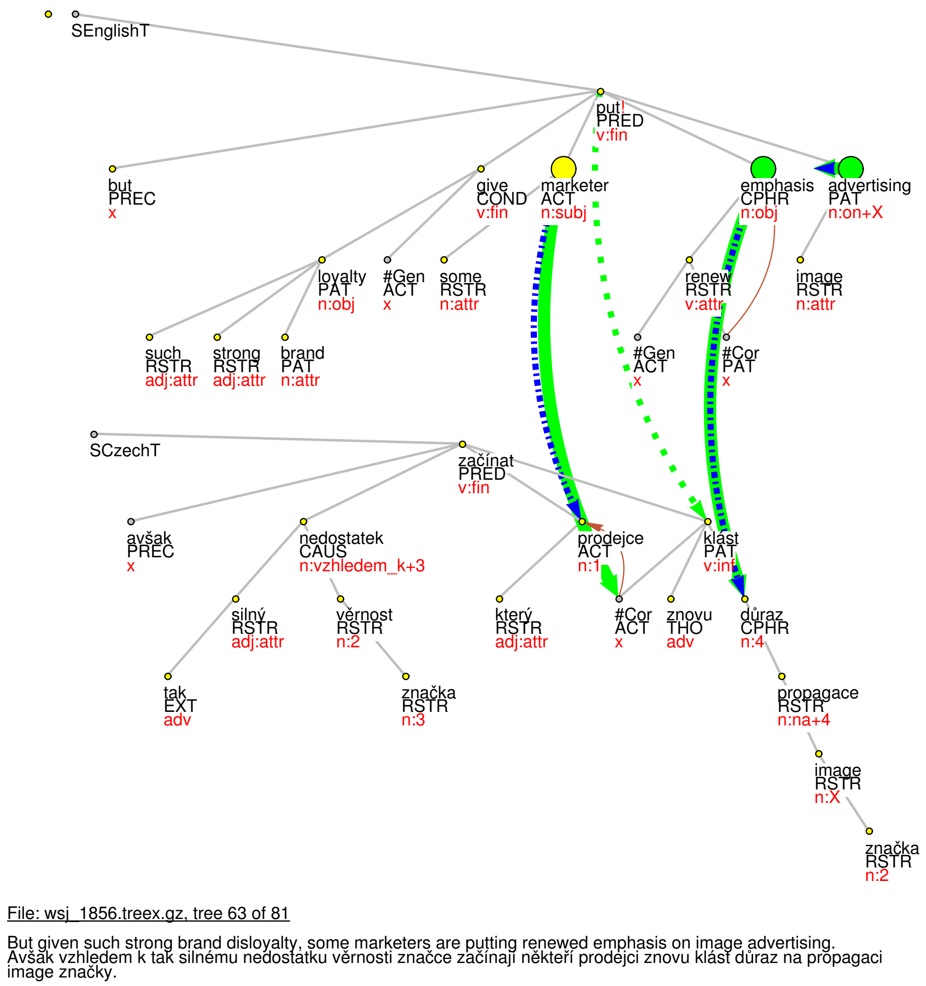


Figure 10: Zero alignment (embedded argument) PAT→---

than Sunday / Cz: *...budou posouzeny jako včas podané nabídky*.COMPL).

This kind of functor mismatch can happen with any argument label, even with the ACT. For example, the case of ACT aligning to MEANS as 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" argument in the structure, this third (or higher-numbered) argument also bears a different label in English and Czech, even in cases where the 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 argu-

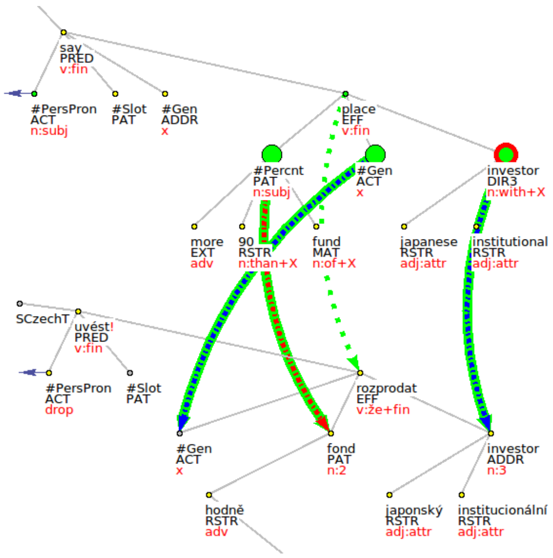


Figure 11: Functor mismatch `DIR3`→`ADDR` in the data

ments in the Czech and English guidelines for valency entries. This was also found for adjuncts, 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 arguments (`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`.

The annotators also have to deal with a "dynamic versus static expression of location", i.e., the `DIR3/LOC` mismatch: for example, for the verb pair *include/zahrnout*, we found the following illustration of the problem: En: *...real-estate assets are included 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`.

Another example is shown in Fig. 12, where instead of having the matching alignment `DIR3`→`DIR3` the asymmetrical alignment `DIR3`→`LOC` is captured in the following parallel sentences: En: *If positioned over the brain's motor-control area.* `DIR3`, ... / Cz: *Pokud se umístí nad částí.*`LOC` *mozku.* This is because the English verb *position* has just one valency frame: `ACT`, `PAT` and `DIR3`, while the Czech verb *umístit* has two valency frames: `ACT`, `PAT` and `DIR3` and `ACT`, `PAT` and `LOC`.

These findings lead us to perhaps 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.

Once again, annotation-wise, all the above cases should be considered "normal" and collected into `CzEngVallex`.

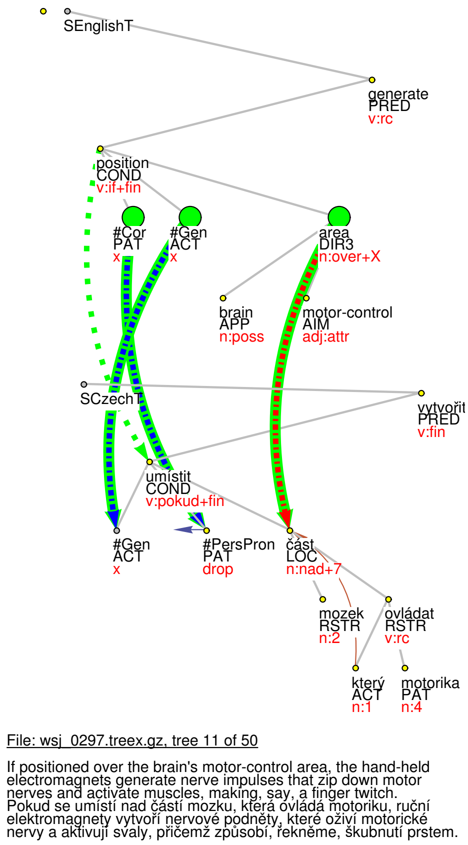


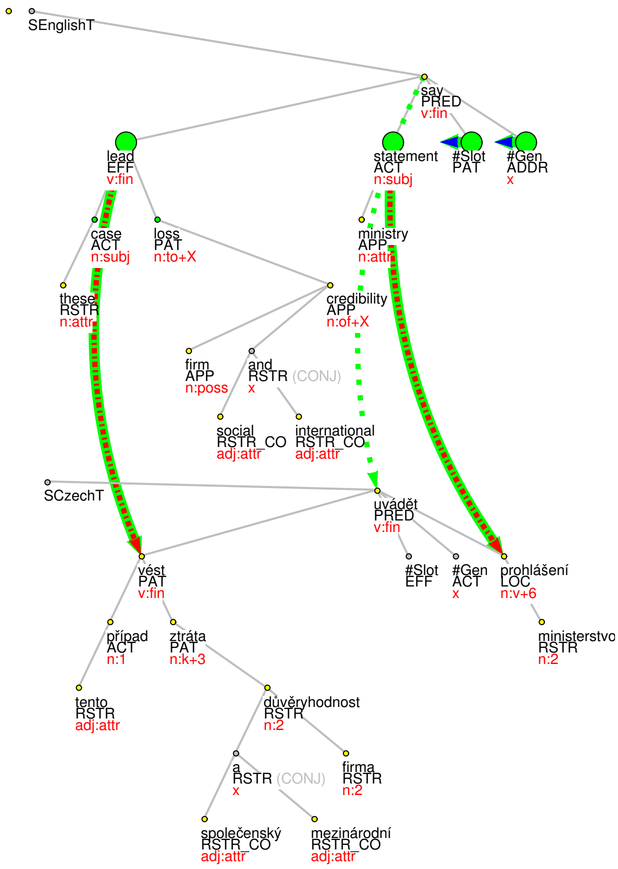
Figure 12: Functor mismatch `DIR3`→`LOC` due to labeling standards

7.2.3 Conflicts

Conflicts in the data annotation (i.e., inconsistent argument alignment for the same verb frame pair at an occurrence of the pair in the data vs. [many] other occurrences) arise mainly from two reasons:

1. 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 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



File: wsj_0051.treex.gz, tree 10 of 34

``These cases lead to the loss of the firms' social and international credibility,'' a ministry statement said.
"Tyto případy vedou ke ztrátě společenské a mezinárodní důvěryhodnosti firem," uvádělo se v prohlášení ministerstva.

Figure 13: Conflicting occurrence of an `ACT`→`LOC` alignment (vs. `ACT`→`ACT`)

English `PAT` with a Czech `DIR3` functor. In another occurrence, arguably the more general one, the `PAT` arguments 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.

2. 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.

   An example documenting this case is shown in Fig. 13, 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. 13: En.: *"These cases lead to the loss of the firms' social and international credibility," a ministry state-*

ment said. – Cz: *"Tyto případy vedou ke ztrátě společenské a mezinárodní důvěryhodnosti firem," uvádělo se v prohlášení ministerstva.*, the same frame pair would lead to a 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 syntactically typical 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, as it would any other subject.

3. Third, conflicts can be lexically motivated, given the translation 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.

   An example is in Fig. 14. Most occurrences of the verb pair *learn* → *poučit se* suit the standard alignment `ACT`→`ACT`, `PAT`→`PAT`, `ORIG`→`ORIG`. However, an occurrence of the same frame pair has been found with an untypical alignment (Fig. 14), where instead of the `ORIG`→`ORIG` alignment, that particular occurrence offers the non-compliant `MANN`→`ORIG` alignment based on the (perhaps a bit rough) lexical correspondence *the hard way* – *z vlastních chyb*, lit. *"from one's mistakes"*.

The annotators should decide the way of handling the conflicts found in the data according to the underlying reason for the conflict. They can suggest (in a note) a change/correction of either of the two valency lexicons, they can suggest (in a note) a change in treebank annotation guidelines (i.e., for consistently assigning a different functor to such a modification in a future version of the treebank) or, in case the reason for conflict is a wrong tectogrammatical annotation, or is supposed to be statistically infrequent, they may decide to ignore the conflict and choose the statistically more probable mapping (as found in other occurrences of the frame pair in the treebank) for collection into `CzEngVallex`, leaving also a note in the data.

7.3 Specific Annotation Issues

7.3.1 Automatic argument post-alignment

If the Czech valency frame contains more complementations than the paired English valency frame (i.e, there is a zero alignment in the reverse, target-source direction), the "superfluous" Czech arguments cannot be manually captured in the frames-pair alignment using the *collect* operation. Such pairs (e.g., `---`→`DIR1`, `---`→`LOC`, `---`→`MANN`) are recorded into the `frames_pairs.xml` file automatically after the manual annotation is finished.

In Fig. 15, the Czech verb *přejít* has three valency frame members: `ACT`, `DIR1` and `DIR3`, as opposed to the English counterpart *go*, which has (in the given sense) only two valency frame members: `ACT` and `PAT`. After the manual annotation, the `frames_pairs.xml` file contains only the `ACT`→`ACT` and the `PAT`→`DIR3` alignments. The remaining alignment `---`→`DIR1` will be added later, through the automatic post-alignment.

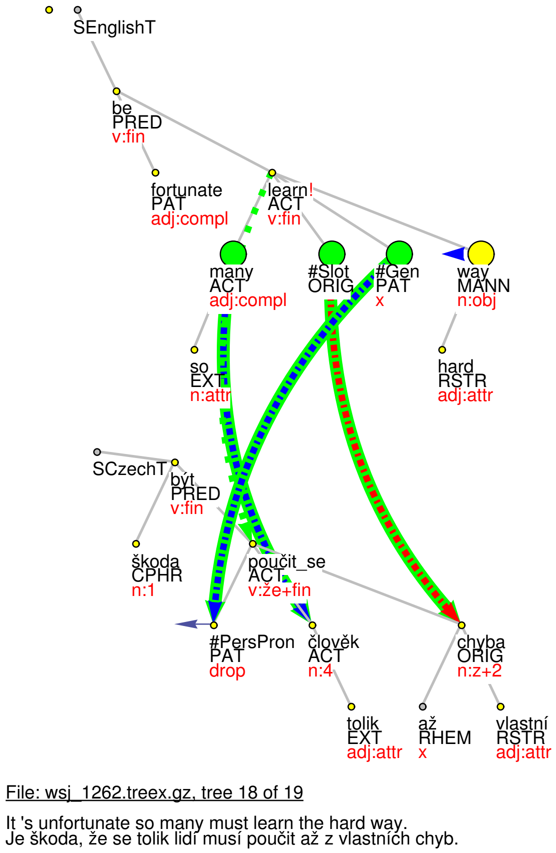


Figure 14: Conflicting occurrence of an MANN→ORIG alignment (vs. ORIG→ORIG)

7.3.2 Treatment of automatic pre-alignment for an already collected frame pair

If the annotator has already performed a *collect* for a given frame pair (at some occurrence of that frame pair in the treebank data), s/he does not need to make any changes (repair, remove etc.) in the visualized automatic alignment of next occurrences of the same frame pair (by the blue arrows), even if those are wrong. The main task of the annotation is to record the correct alignments in the `frames_pairs.xml` file, rather than repair automatic alignment directly in the treebank data at every occurrence. The automatic alignment serves mainly as a hint to the annotator.

Fig. 16 shows the difference in the automatic vs. collected alignment for the verb pair *save*→*ušetřit*. The collected pairing ACT→ACT, PAT→PAT has already been manually assigned at some previous occurrence(s). As can be seen here, the collected alignment (full green arrows) fits the given occurrence correctly. The redundant alignments (blue arrows) do not need to be removed.

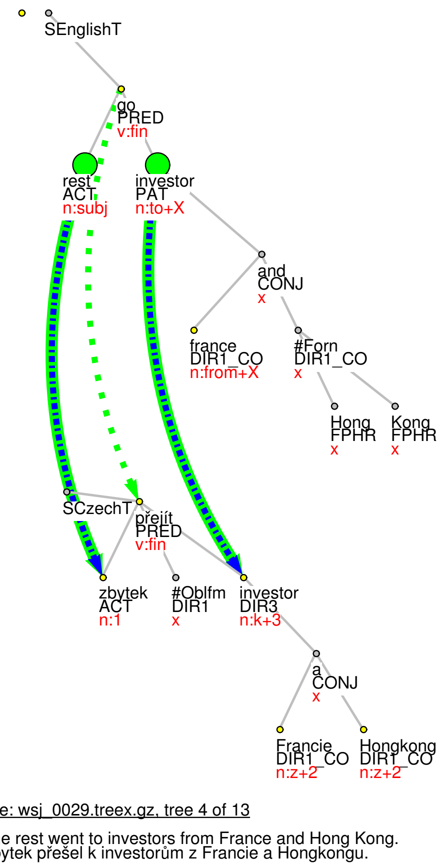


Figure 15: Automatic argument post-alignment `---`→`DIR1`

7.3.3 Automatic warnings

During the process of annotation, the annotators have an automatic warning system at their disposal. There are two types of automatic warnings:

(i) The annotator gets a warning in case s/he wants to align disallowed functors, i.e., specific functors that never label any argument or adjunct node of the sentence, such as `RHEM`, `PREC`, `ATT`, `DENOM`, `PARTL`, `VOCAT` etc. In such cases, the annotator gets the following warning called *Wrong functor*:

Cannot collect t_lemma (En verb) and t_lemma (Cz verb). Node t_lemma (id) has not allowed functor.

(ii) The annotator gets a warning in case one of the functors to be collected is a `RSTR` functor (e.g., when the desired alignment for collect is `RSTR`→`PAT`). The `RSTR` label is reserved for dependents of nominal nodes and it is not

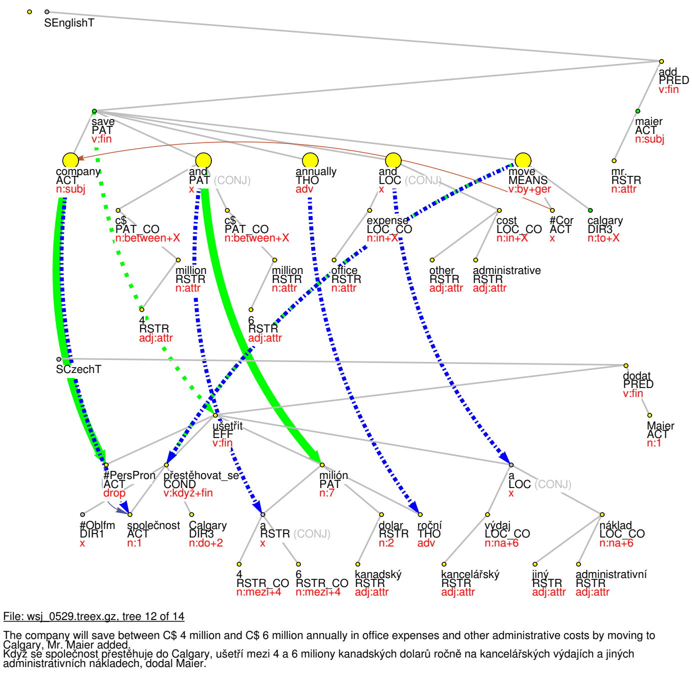


Figure 16: Treatment of automatic pre-alignment for an already collected frame pair

normally expected to appear in any verb frame. In such cases, the annotator gets the following warning called *RSTR functor*:

> **Note "t_lemma" (node_id) has `RSTR` functor. Still want to collect?**

As opposed to the *Wrong functor* warning, the *RSTR functor* warning can be overridden by the annotator.

7.3.4 Erroneous automatic verb pre-alignment

If the automatic pre-alignment was wrong and non-corresponding verbs were aligned, the annotator corrects the wrong alignment (using the drag-and-drop operation) to get the right one. S/he is not required to make any notes about this change.

For example, in the following translational sentence pairs: *In addition, Ms. Consolo says, tenants usually can negotiate to pay rents that are about one-quarter lower than landlords' initial asking price.* → *Paní Consolo říká, že nájemci navíc obvykle dokáží usmlouvat cenu nájmu o zhruba čtvrtinu nižší, než majitel původně požadoval*, the verb *negotiate* was wrongly pre-aligned with the Czech verb *dokázat*. The annotator corrects the wrong alignment by aligning *negotiate* with *usmlouvat* (and then possibly corrects its arguments etc.). Than s/he collects this verb pair into the `frames_pairs.xml` file, if needed.

Wrong verb alignment usually arises with quasimodals and process verbs in either language.

7.3.5 Erroneous functor

In case of an erroneous functor label that is to be used in the alignment to be collected, the annotator corrects the wrongly assigned functor using the macro `f`, which sets a new functor for the `CzEngVallex` purposes, and marks this change in the note attribute (Comment type *Functor*).

For example, Fig. 17 shows that the word *akcie* in the Czech sentence is labeled with an erroneous `REG` functor instead of the correct `EFF` functor. The `REG` functor must be corrected in order to get the correct alignment, as shown in Fig. 18. Moreover, after this correction, the redundant artificial node `#Slot.EFF` must be deleted in order not to have two nodes of the same functor (here `EFF`) in the structure.

Erroneous functors should be corrected at every relevant occurrence in the treebank.

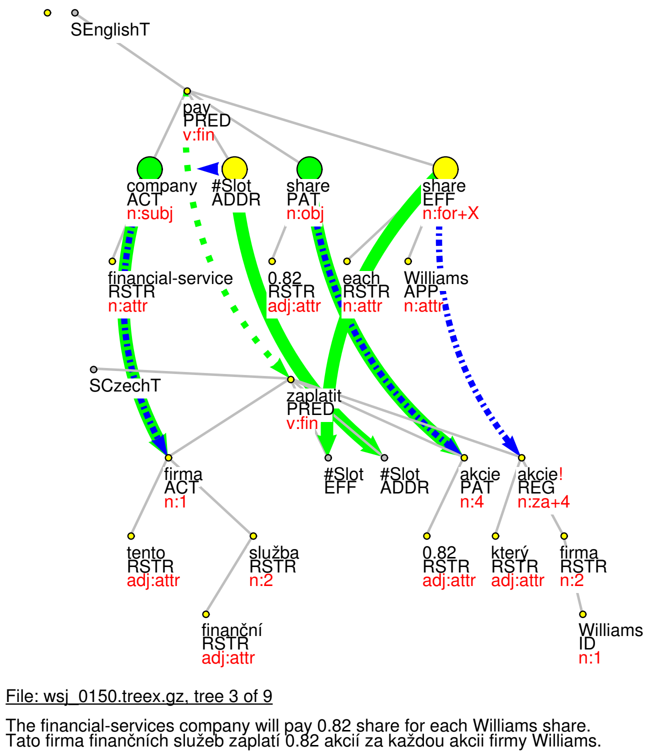


Figure 17: Erroneous functor

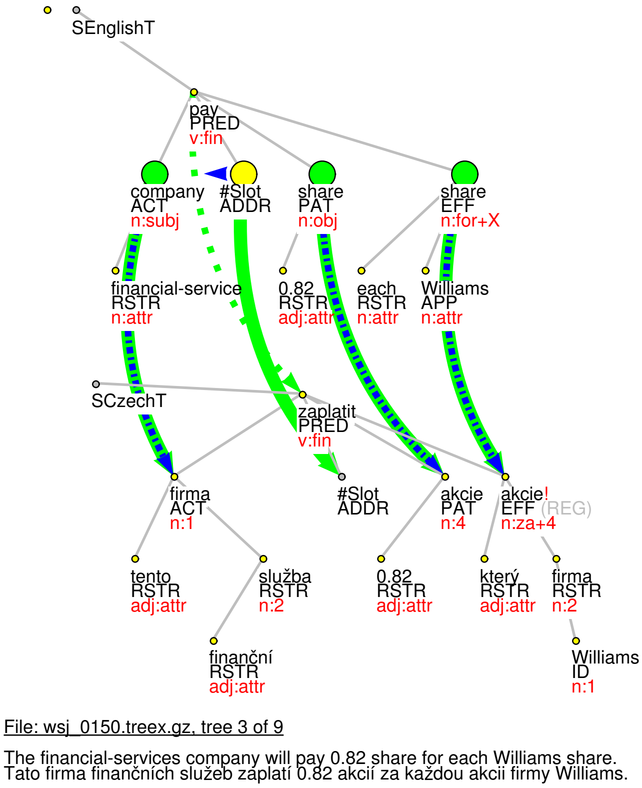


Figure 18: Erroneous functor, corrected

7.3.6 Arguments competing for the same position in the valency frame

With certain verbs, one of the valency positions may be alternatively occupied by arguments bearing different functors, while the meaning of the verb is preserved (or it only changes a little). Such arguments are called "competing arguments" or "alternating functors", mirroring the fact that only one of them can appear at any particular occurrence of the frame in an actual utterance. For such cases, the concept of "modification alternatives" was introduced [12]. This approach has so far been adopted mainly for different types of manner adjuncts, where the obligatory manner adjunct position may be taken up by modifications with the following list of functors: `MANN`, `CRIT`, `ACMP`, `BEN`, `MEANS` or `CPR`.

However, in our data, the issue of alternating functors poses a problem if the English valency frame contains such an alternative and the aligned Czech one does not, or vice versa. In these cases, we have adopted an interim solution: since technically all the alternatives appear as separate arguments (e.g., using the macro `Alt-s`), the annotator aligns the (most frequently found, or the most typical) alternative, leaving the others unaligned (i.e., aligned to `---`).

For example, see Fig. 19, where the frame pair consists of the following alignments: `ACT`→ `ACT`, `MANN` → `MANN`, `CRIT` → `---`, `CPR` → `---`. This phenomenon needs further research and possibly it might be necessary to revisit the relevant frame pairs.

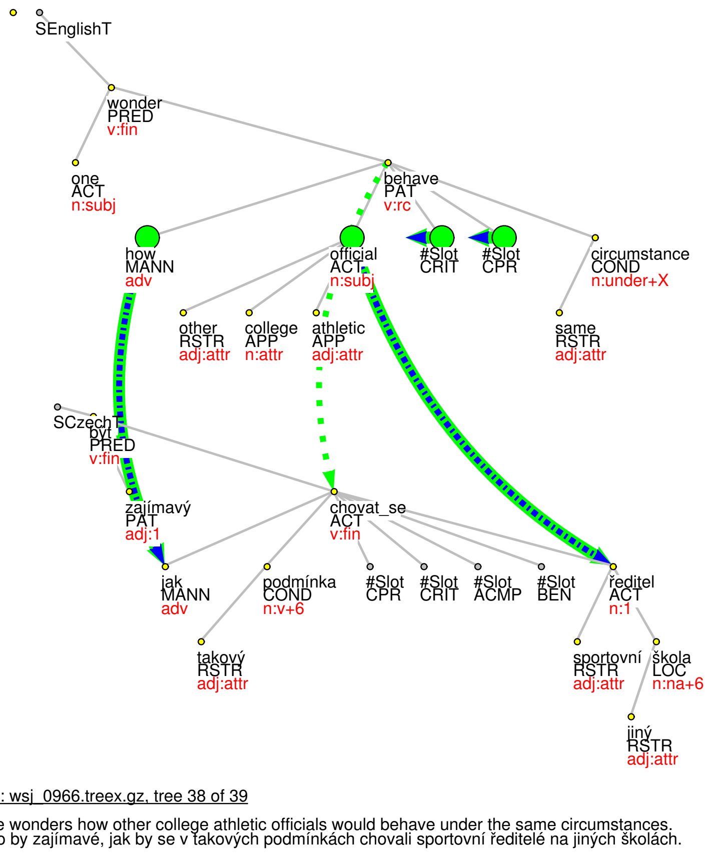


Figure 19: Competing valency modifications (alternating functors)

7.3.7 Problematic alignment

For various reasons, the alignment of the English and Czech verb arguments might not make a clear sense in the structure. The most frequent reason is the presence of language-specific syntactic structure or too different syntactic structures in both languages. In such cases the annotator does not build new frame pairs, nor does s/he collect. S/he just writes the reason for not collecting the frames into a note (Comment type *Other*).

This covers cases which do not fall into the categories of mismatch or conflict as described in Sec. 7.2, i.e., which are “more complicated” and thus beyond a possibility to establish a reasonable alignment.

Here are some possible cases which constitute such problematic alignments:

1. idiomatic utterances (as shown in Fig. 20);
2. wrongly annotated structure (see Fig. 21);
3. inaccurate or too loose translation, cf. Sec. 7.3.8, Fig. 8.

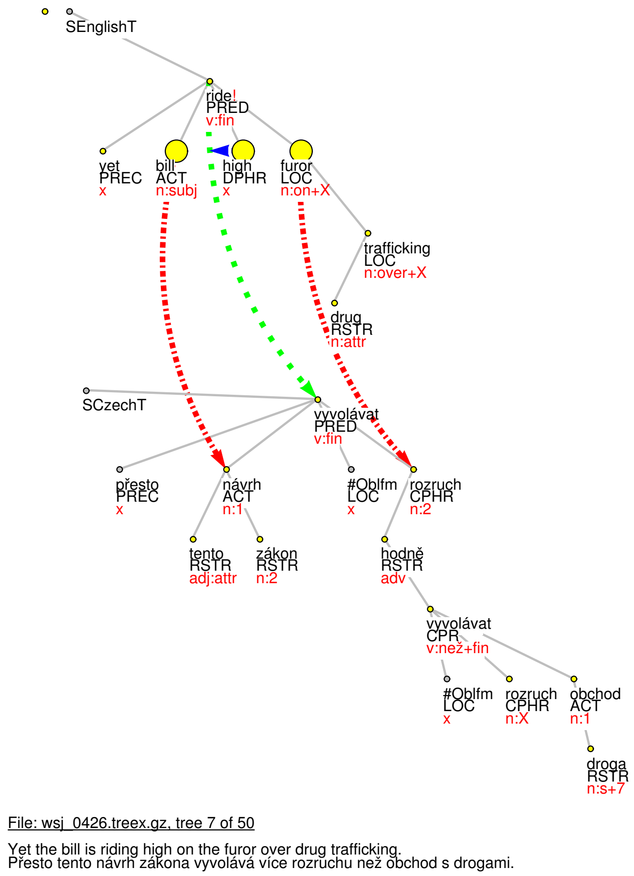


Figure 20: Idiomatic structure

7.3.8 Adequate vs. inadequate translation

In general, the annotator should only annotate sentences containing adequate translation pairs and should not annotate sentences with an inadequate translation. The decision whether the translation is adequate or inadequate is left to the annotators.

Sentences with a too loose translation can be classified as either adequate or inadequate, depending mostly on their syntactic complexity:

(i) Sentences with a too loose translation having too different syntactic structures, where the proper alignment is either impossible or very difficult and misleading, are not annotated, see Fig. 22;

(ii) In contrast, syntactic structures of too loosely translated sentences that still allow more or less straightforward alignment can be annotated as usual, see Fig. 23.

While loose translation sometimes keeps the corresponding arguments "in sync" (labeled with the same functor), it often (predictably) leads to non-corresponding valency realizations.

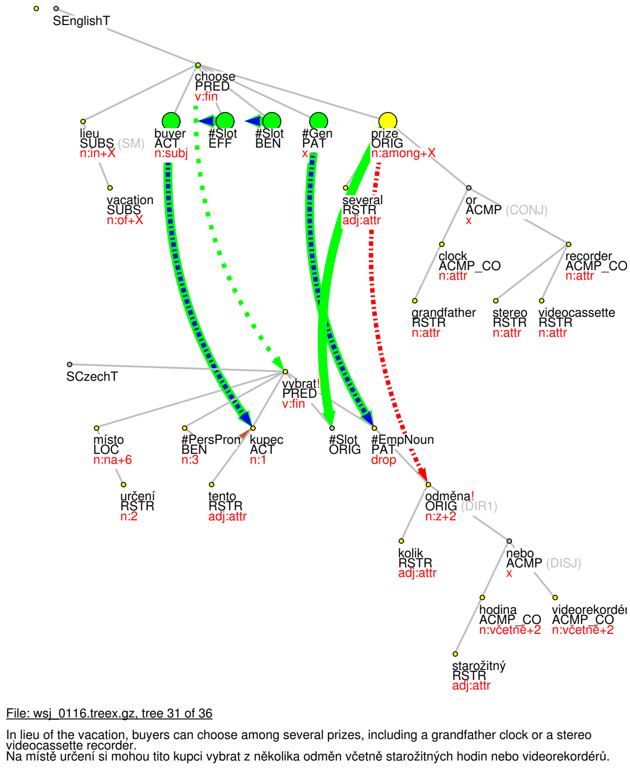


Figure 21: Wrong structure in Czech

For example, in the pair En: *The company.*ACT*, which went public.*PAT / Cz: *Společnost.*ACT*, která se stala veřejně obchodovatelnou.*PAT show both ACT and PAT in the corresponding positions. However, in En: *Nicaragua.*ACT *had gone communist.*PAT / Cz: *Z Nikaraguy.*PAT *se stala komunistická země.*ACT the arguments have been swapped due to the use of the verb *stát se* (lit. *become*) with a different syntactic configuration of modifications . This also happens if there is a choice between a more literal translation and a looser one (*infuriate* translated as *rozzuřit* or more loosely as *dohnat k zuřivosti*): En: *...other snags that.*ACT *infuriated some fund investors.*PAT *in October 1987* / Cz: *...jiným zádrhelům, které.*ACT *v říjnu roku 1987 doháněly některé investory.*ADDR *fondů k zuřivosti.*PAT).

8 Advanced annotation guidelines (more difficult cases)

8.1 Catenative and modal verbs

Special attention in the annotation must be paid to verbs that are involved, together with another verb, in 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

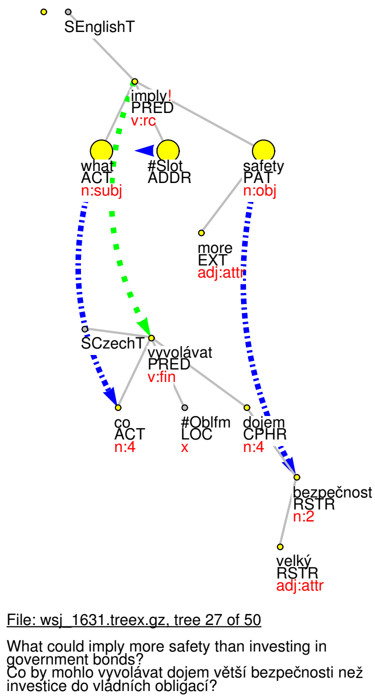


Figure 22: Inadequate translation

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 [17, 14] 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.[27]

Auxiliaries, or modals, do not appear in the tectogrammatical annotation, though there are certain verbs in both English and Czech that have a similar function and behavior (e.g., *dokázat*, or the so-called quasimodal verbs) and are therefore often translated with a proper modal verb. On the other hand, catenatives are usually displayed as regular nodes in the annotation. Nevertheless, the complex phrase they are involved in is often translated with a single lexical unit. In case of such verbs appearing just in one side of the translation, the annotator should not align the quasimodal or the catenative verb, but their dependent verbs according to their semantic correspondence.

For example, the annotator should not align the English catenative verb (such as *keep*, *need* or *get*) with its Czech equivalent (which can be an adverbial, for example) but aligns the dependent verb of the catenative verb and its Czech translation. S/he makes an appropriate note (Comment type *Other*) with the catenative verb in question (see the exclamation mark sign at the node *keep* in Fig. 24, denoting a comment filled in the `note` attribute of that node).

[27] 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. Some discrepancies have already been described in Sec. 7.2.

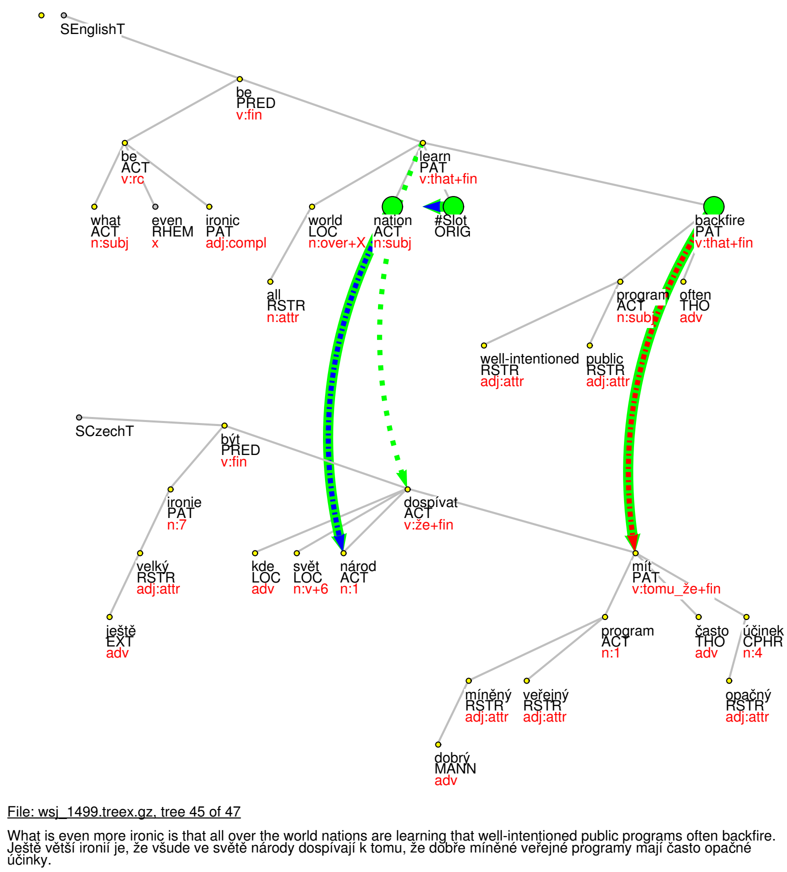


Figure 23: Adequate, though loose, translation

In Fig. 24, the catenative verb *keep* is not aligned to any Czech verb. The annotator aligned its dependent verb *ride up* with its Czech translation *klouzat*. This alignment is then collected to `CzEngVallex` as a frame pair.

Similarly, the annotator should not align the English verb *want*, the translation of which in Czech (*chtít*) is sometimes considered to be a modal, as it is the case in Fig. 25. In such a case, the annotator should simply ignore the pair of sentences (i.e., "not to annotate", Sec. 7.1).

8.1.1 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". In short, raising and ECM are generally considered a marginal phenomenon in Czech and are not being treated conceptually [21], except for several attempts to describe agreement issues, e.g., the morphological behavior of predicative complements described in a phrase structure grammar formalism [23].

The reason for this negligent approach to ECM is probably rooted in the

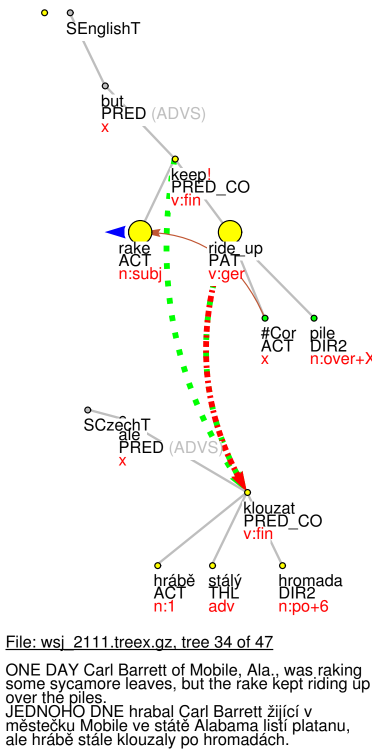


Figure 24: Catenative and modal verbs – *keep*

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. 26, or they include a nominalization, thus keeping the structures parallel and the annotation is unproblematic.

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.ACC is coming.

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

In this type of an accusative-infinitive sequence, the accusative element is in the FGD analyzed consistently as the direct object of the matrix verb (the `PAT` argument) and the non-finite verb form then as the predicative complement of the verb (the `EFF` argument in the annotation).

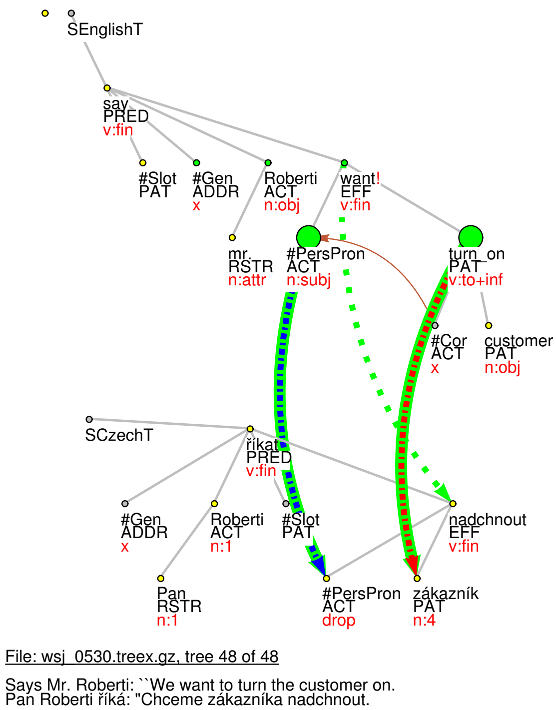


Figure 25: Catenative and modal verbs – *want*

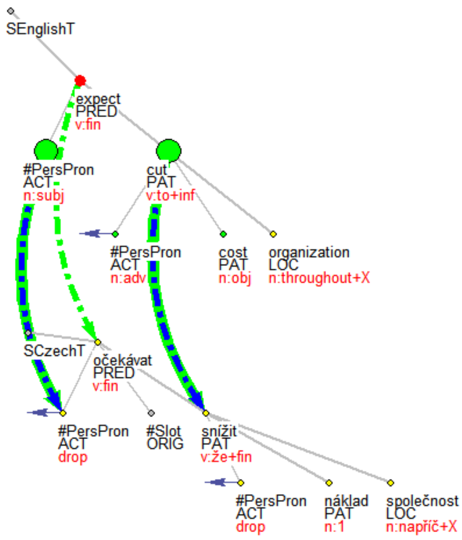
The PCEDT annotation of verbs of perception is shown in Fig. 27, with frame arguments aligned in the following way: `ACT`→`ACT`; `PAT`→`EFF`; `---`→`PAT`.

The tectogrammatical annotation of the catenative verb constructions in the English data is a relatively "deeply structured" (this is sometimes called "layered annotation", not to be confused with the PDT/PCEDT major annotation layers). On the other hand, the Czech part of the PCEDT data uses a relatively "flat" annotation, partly because the catenative construction with raising structure is fairly uncommon in Czech. The flat structure is easier to interpret, or synthesize 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 causes alignment mismatches in the parallel data.

The question of treating ECM constructions is a conceptual theoretical question to be further researched. For the time being, the annotators are asked to align the verb dependent modifications according to their semantic correspondence where possible, usually leaving an unaligned modification in the flat structure, and to make an appropriate note (Comment type *Structure*).

8.1.2 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-



En: They expect him to cut costs throughout the organization.
Cz: Očekávají, že sníží náklady napříč celou společností.

Figure 26: Alignment of the ECM construction

object in both languages, i.e., the intervening NP is dependent on the matrix verb (and licensed by it) and there is usually a co-referential empty element of some kind in the valency structure of the dependent verb form. OCV constructions are similarly frequent in Czech and English and their alignment in the PCEDT data is balanced, see Fig. 28.[28]

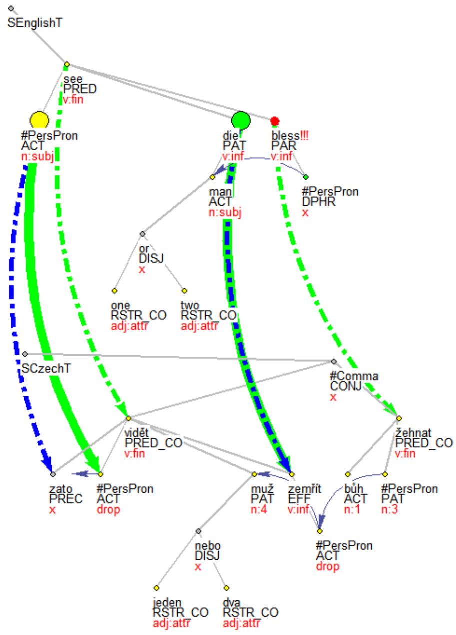
Interestingly, it is sometimes the case that English control verbs in the treebank are translated with non-control, non-catenative verbs on the Czech side, and the intervening NP is transformed to a dependent of the lower verb of the dependent clause (see Fig. 29), or even a more complex nominalization of the dependent structure is used (Fig. 30).

The verb involved in this kind of translation shift may be either a more remote synonym, or a conversive verb.[29] 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*).[30] Nevertheless, this type of semantic shift does not prevent the use of the structure as a sufficiently equivalent expression of the semantic content. We approach this as an inherent property of (any) language to suppress certain aspects of meaning without losing the general sense of synonymity.

[28]In Fig. 28, 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*).

[29]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.

[30]Note that the de-causativisation process is possible without objections whereas the reverse shift, from non-control verb to a control verb, is rare if it at all exists.



En: I have seen one or two men die, bless them.
Cz: Zato jsem viděla jednoho nebo dva muže zemřít, bůh jim žehnej.

Figure 27: Alignment of the perception verbs' arguments

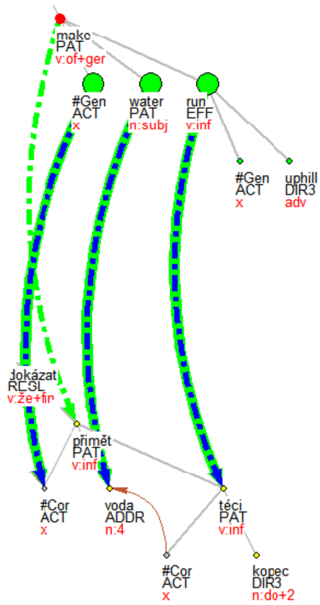
The annotators are asked to treat such occurrences as typical examples of zero alignment (see Sec. 7.2.1).

8.2 Complex predication - light verb constructions

Next issue deserving special attention is the issue of 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 tectogrammatical annotation), 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 syntactic annotation of the parallel data.

A complex predication in one language can often 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 (zero alignment). There are essentially 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.

If there is a "third" valency complementation within the complex predication structure, e.g., En: *placed weight on retailing* / Cz: *klást důraz na prodej*, we



En: But is he so clever that he has achieved the political equivalent of making water run uphill?
Cz: Je ale tak chytrý, že by v politice dokázal přimět vodu téct do kopce?

Figure 28: Alignment of the control verbs' arguments

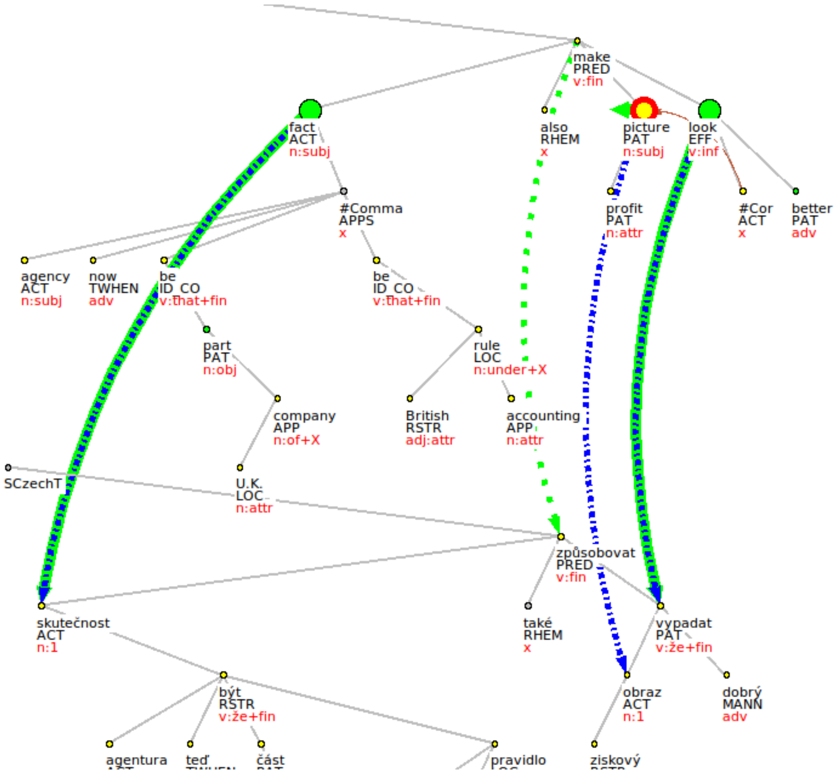
may get another instance of zero alignment in the data, see Fig. 31.

Complex predicates have been annotated according to quite a complicated set of rules on the Czech side of the PCEDT data (for details, see [12]). Those rules include also the so-called dual function of a valency modification. There are two possible dependency positions for the "third" valency argument of the complex predicate: either it is modeled 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 structure of the noun in other contexts, behavior of synonymous verbs etc. On the Czech side, the "third" valency argument was strongly preferred to be a dependent of the nominal component.

On the English side of 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 a unified way. The annotators are therefore asked to respect the structure of the existing tectogrammatical annotation and in case of discrepancy, treat them as instances of a zero alignment (see Sec. 7.2.1).

Dealing with `CPHR`-labelled (light-verb-)constructions involves also a different issue: for example, for the verb pair *offer/poskytnout (důkazy)* we encountered light-verb constructions explicitly annotated just in Czech, but for the verb pairs *make (complaint)/podat (stížnost)* and *give (help)/poskytnout (pomoc)*, we have



En: The fact that the agency will now be part of a U.K. company, under British accounting rules, will also make the profit picture look better.
Cz: Skutečnost, že agentura nyní bude částí společnosti Spojeného Království, pod britskými účetními pravidly, také způsobuje, že ziskový obraz vypadá lépe.

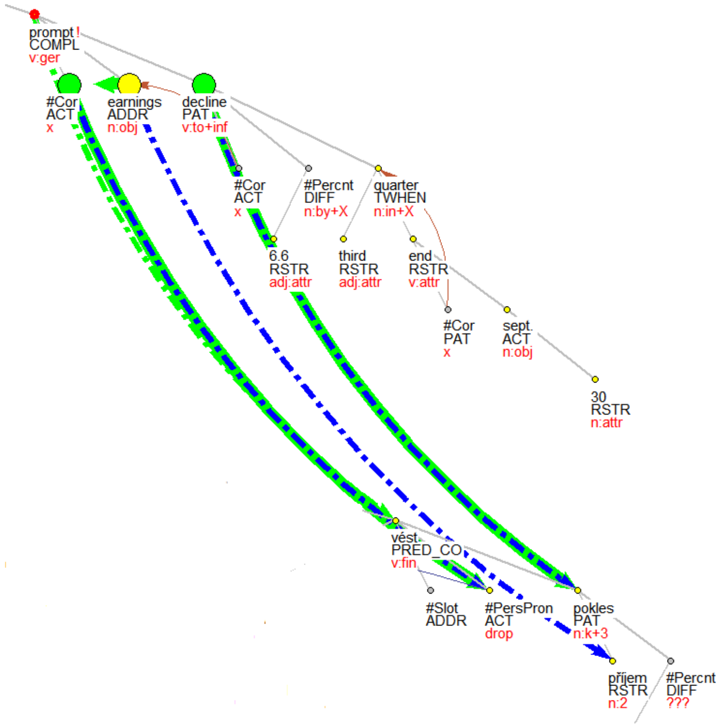
Figure 29: Alignment of English OCV with Czech non-OCV construction

encountered light-verb constructions in both languages: En: *The report offered new evidence*.`PAT` / Cz: *Uvedená zpráva poskytla nové důkazy*.`CPHR`; En: *...a complaint*.`CPHR` *was made* / Cz: *...stížnost*.`CPHR` *byla podána*; En: *...she gave them similar help*.`CPHR` / Cz: *...poskytla jim obdobnou pomoc*.`CPHR`.

The annotators are expected to align and collect the mismatching functors as usual (see Sec. 7.2). Only in case they are strongly convinced that the construction with a non-`CPHR` functor fulfills the requirements for being treated as a complex predication, they might suggest the change of frame in the appropriate note (Comment type *Other*) for further processing of the PCEDT data.

8.3 Conversive verbs

A considerable number of unaligned modifications 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 in language, topic-focus articulation etc.), the translator decides not to use the syntactically most similar lexical unit, but uses a conversive one (for a similar



En: ... demand for Nekoosa's commodity paper has weakened, prompting earnings to decline by 6.6 in the third quarter ended Sept. 30.
Cz: ... poptávka ... vedla k poklesu příjmu o 6.6.

Figure 30: Alignment of English OCV with Czech complex nominalization

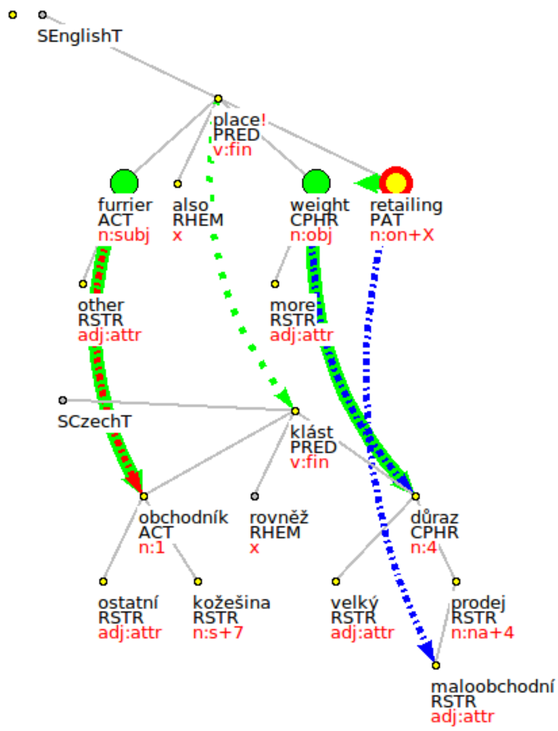
process, see also Sec. 8.1.2), thus causing the arguments to relocate in the deep syntactic structure, see Fig. 32.

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 modification embedded even lower in the structure.

Prevalent morphosyntactic realization of `ACT` is nominative case, but certain exceptions are recognized (verbs of feeling etc.). Also, the `ACT` position (first actant) is subject to the process called "shifting of cognitive roles" [19], i.e., other semantic roles can take 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, see the English side of Figs. 33 and 34.

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

It is often unclear whether such verb instances correspond to different meanings of the verb (as represented by different verb frames), or whether they correspond to a single meaning (as represented by a single valency frame). It is



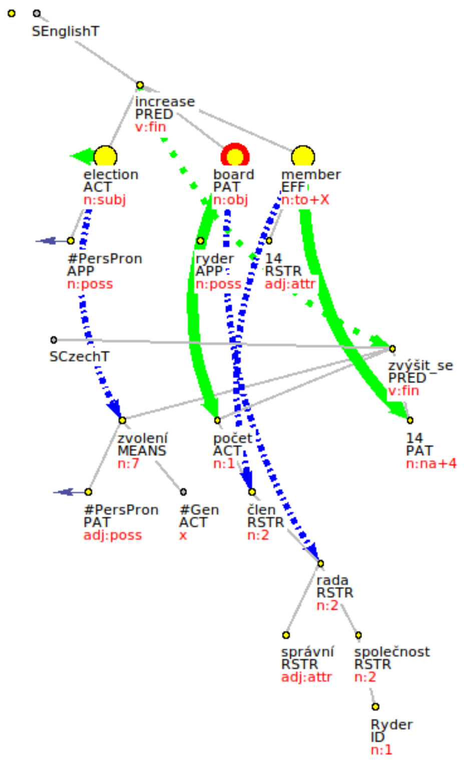
En: Other furriers have also placed more weight on retailing.
Cz: Ostatní obchodníci s kožešinami rovněž kladou větší důraz na maloobchodní prodej.

Figure 31: Mismatch due to complex predication solution

often the case, that the Czech data tend to overgeneralize the valency frames by 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. 33.

As we can see, the conflict arose due to the fact that the alignment of the verb pair was previously collected on a different occurrence of the same verb pair with a semantically different realization of the Czech `ACT` position, see Fig 34. The valency structure for both instances of *base* is identical; only in the first case, the verb is used in active voice, whereas in the second case, it is used in the passive voice.

Moreover, there are seemingly only two modifications in the structure of the Czech sentences, which means that in both cases, one semantic modification (a different one for each of the Czech sentences) was sort of backgrounded from the semantic structure. This modification would then, in case there is a need of its being made overt, appear as a non-argument, either as a locative adjunct, or a dependent modification of the noun, see the xample (2).



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

Figure 32: Mismatch due to the the use of conversive verbs

(2) a. Wertheimer se ve svém názoru opírá o prohlášení Keatinga.
Wertheimer REFL in his opinion leans to the statement by Keating

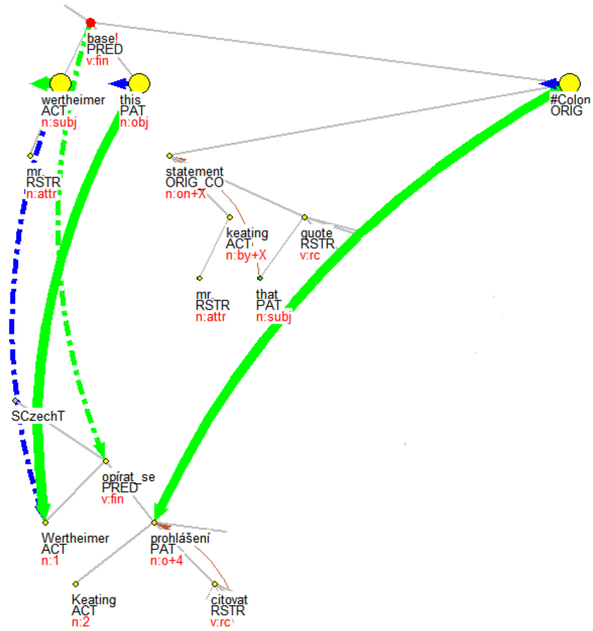
b. Wertheimerův názor se opírá o prohlášení Keatinga.
Wertheimer's opinion REFL leans to the statement by Keating

c. Wertheimer opírá svůj názor o prohlášení Keatinga.
Wertheimer leans his opinion to the statement by Keating

Note that such conflicts often involve Czech verbs with an adjoining *se* particle.

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 between the languages. Czech often uses the *se*-reflexivization in order to preserve the topic focus articulation (information) structure, whereas English does not have such an operation to work with. Therefore, it often uses simple passivization or middle construction.

Moreover, the first valency position in Czech is often overgeneralized, allowing a multitude of semantically different modifications, which is, due to "econ-



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

Figure 33: Conflict due to the underspecification of the ACT position

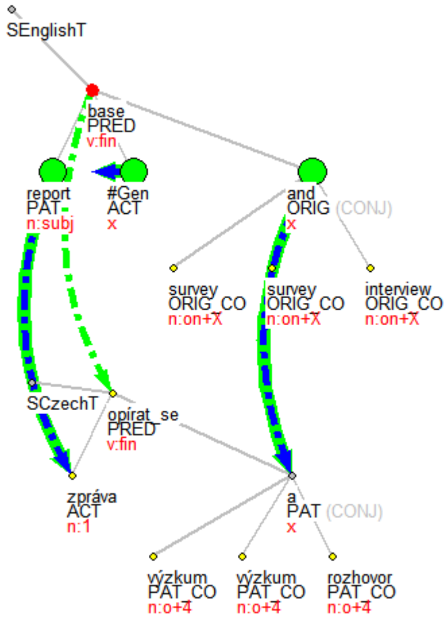
omy of description", sometimes not reflected in the linguistic theory.

When encountering this type of conflict, the annotator should choose to annotate and collect one of the possible alignments (usually the one with greater similarity of aligned functors, i.e., `ACT`→`ACT` rather than `ACT`→`PAT`) and add an appropriate note suggesting multiplication of frames in the valency lexicon. Simple functor mismatches due to conversiveness of the verbs are to be annotated as usual.

8.4 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 valency 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 valency argument) but in its Czech translation *řada ekonomů*, the word *řada* is 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 annotator is asked to align head to head (i.e., align both direct daughters



En: The report was based on a telephone survey of 1,250 low-income households across the state...
Cz: Zpráva se opírá o telefonický výzkum prováděný ve 1250 domácnostech s nízkými příjmy po celém státě...

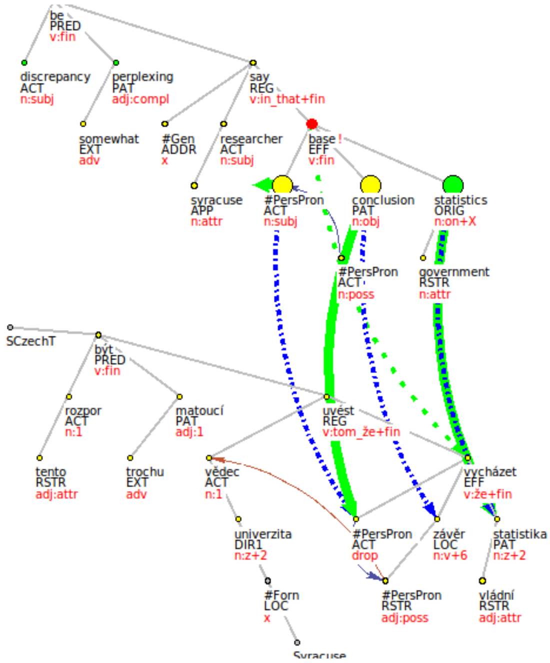
Figure 34: Original collect for the verbs *base* and *opírat se*

of the verb and valency arguments). In the above example, *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. 36.

Names of companies (e.g., *IBM*) are usually preceded with a generic name *společnost* (*company*) in the Czech translation, 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.

8.5 Direct speech

According to the annotation guidelines, the annotation rules for direct speech in English [4] and Czech [12] on the tectogrammatical level are similar. Both languages add a new node representing the gerund (transgressive) of a verb of saying to the tectogrammatical annotaation in cases where 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?”*)



En: The discrepancy is somewhat perplexing in that the Syracuse researchers said they based their conclusions on government statistics.
Cz: Tento rozpor je trochu matoucí v tom, že vědci z univerzity v Syracuse uvedli, že ve svých závěrech vycházejí z vládních statistik.

Figure 35: Original collect for the verbs base and vycházet with LOC modification linked to PAT

Due to the same instructions, mismatches are not expected in collecting direct speech utterances. Nevertheless, the annotation process reveals some discrepancies, as shown in Fig. 37, where the collected frame pair is as follows: `ACT`→`ACT PAT`→`---`, `---`→`PAT`.

The mismatch occurs due to different practical annotation approach to direct speech in the individual languages, most notably, the English annotation often differs from the common guidelines. While in Czech the use of `#EmpVerb` and the functor `COMPL` is common and according to the guidelines, 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. 7.2.1). Such structures are left for future treatment within possible tectogrammatical annotation revisions.

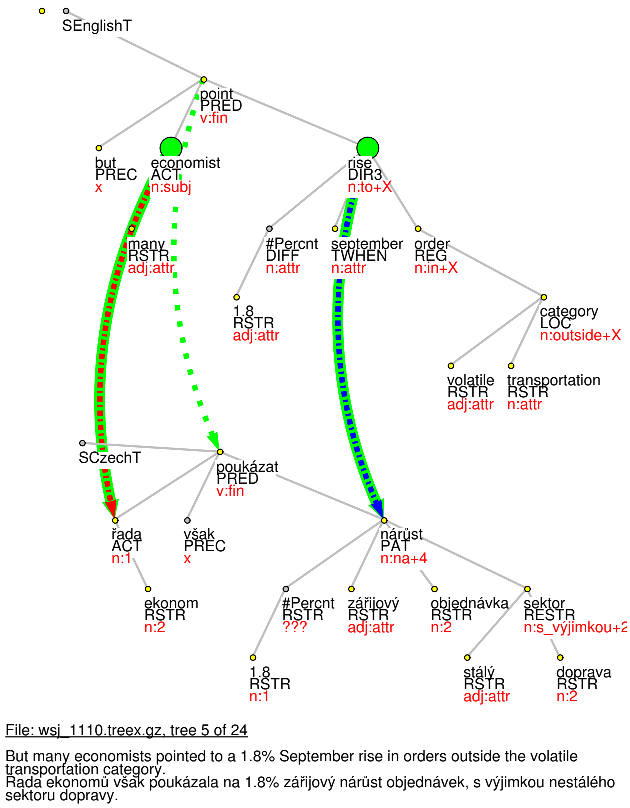


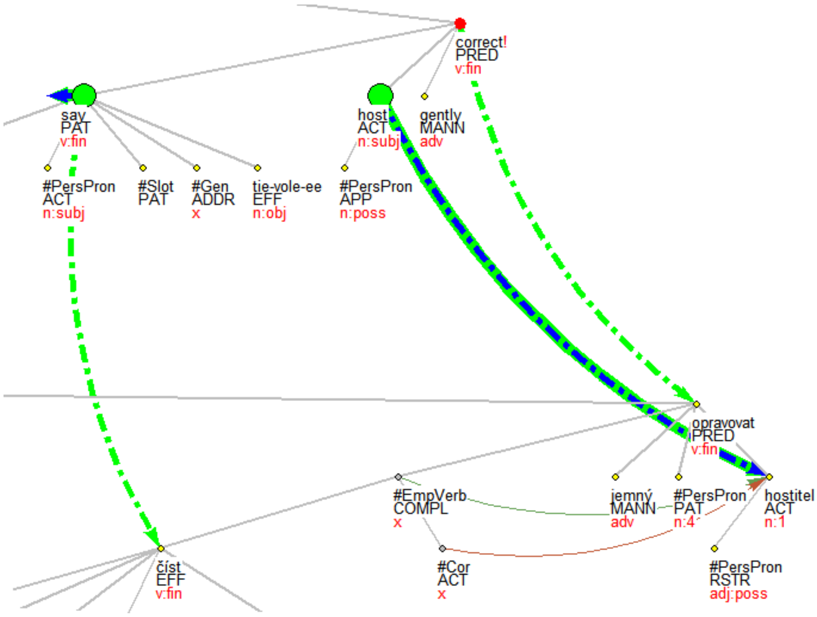
Figure 36: Head-dependent switch

9 Using CzEngVallex: linguistic theory and NLP experiments

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 a innovative use in various NLP tasks.

In both of these areas, the `CzEngVallex` has proved to be a valid resource. Our publications [13, 27, 32, 28, 33, 31, 26] show some interesting and important results concerning verbal valency from the Czech-English comparison perspective, while [6, 5] showes 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 [35, 34, 15].

We 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.



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

Figure 37: Direct speech alignment

References

[1] Lars Ahrenberg, Mikael Andersson, and Magnus Merkel. 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, 2002.

[2] Ondřej Bojar 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.

[3] Silvie Cinková. 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.

[4] Silvie Cinková, Jan Hajič, Marie Mikulová, Lucie Mladová, Anja Nedoložko, 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, 2006.

[5] Ondřej Dušek, Eva Fučíková, Jan Hajič, Martin Popel, Jana Šindlerová, and Zdeňka Urešová. Using parallel texts and lexicons for verbal word sense

disambiguation. In *Proceedings of the Third International Conference on Dependency Linguistics, Depling 2015, To appear.*

[6] Ondřej Dušek, 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.

[7] 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. Prague Czech-English Dependency Treebank 2.0, 2011.

[8] Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, Magda Ševčíková-Razímová, and Zdeňka Urešová. Prague Dependency Treebank 2.0, 2006.

[9] Silvia Hansen-Schirra, Stella Neumann, and Mihaela Vela. 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, 2006.

[10] P. Kingsbury 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.

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

[12] 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. Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual. Technical Report 30, Prague, Czech Rep., 2006.

[13] Marie Mikulová, Jan Štěpánek, and Zdeňka Urešová. Liší se mluvené a psané texty ve valenci? *Korpus – gramatika – axiologie*, 8:36–46, 2013.

[14] Dieter Mindt. 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.

[15] Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Silvie Cinková, Dan Flickinger, Jan Hajič, and Zdeňka Urešová. Semeval 2015 task 18: Broad-coverage semantic dependency parsing. 2015.

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

[17] F. R. Palmer. *The English verb / F. R. Palmer*. Longman London, 2d ed. edition, 1974.

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

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

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

[21] Jarmila Panevová. More remarks on control. *Prague Linguistic Circle Papers*, 2(1):101–120, 1996.

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

[23] Adam Przepiórkowski and Alexandr Rosen. Czech and Polish raising/control with or without structure sharing. 3:33–66, 2005.

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

[25] Jana Šindlerová 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.

[26] Jana Šindlerová, Eva Fučíková, and Zdeňka Urešová. Zero alignment of verb arguments in a parallel treebank. In *Proceedings of the Third International Conference on Dependency Linguistics, Depling 2015, To appear.*

[27] Jana Šindlerová, Zdeňka Urešová, and Eva Fučíková. Verb valency and argument non-correspondence in a bilingual treebank. In Katarína Gajdošová 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.

[28] Jana Šindlerová, Zdeňka Urešová, and Eva Fučíková. Resources in conflict: A bilingual valency lexicon vs. a bilingual treebank vs. a linguistic theory. In Nicoletta Calzolari, 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.

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

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

[31] Zdeňka Urešová, 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.

[32] Zdeňka Urešová, 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, Association for Computational Linguistics.

[33] Zdeňka Urešová, Eva Fučíková, Jan Hajič, and Jana Šindlerová. Verb-noun idiomatic combinations in a czech-english dependency corpus. In *PARSEME 2nd general meeting*, Athens, Greece, 2014. Institute for Language and Speech Processing of the Athena Research Center.

[34] Zdeňka Urešová, 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, 2014. Dublin City University, Association for Computational Linguistics and Dublin City University.

[35] Nianwen Xue, 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 Nicoletta Calzolari, 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.

[36] Zdeněk Žabokrtský. 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, 2011. Univerzita Pavla Jozefa Šafárika v Košiciach.

The ÚFAL/CKL Technical Report Series

ÚFAL

ÚFAL (Ústav formální a aplikované lingvistiky; http://ufal.mff.cuni.cz) is the Institute of Formal and Applied linguistics, at the Faculty of Mathematics and Physics of Charles University, Prague, Czech Republic. The Institute was established in 1990 after the political changes as a continuation of the research work and teaching carried out by the former Laboratory of Algebraic Linguistics since the early 60s at the Faculty of Philosophy and later the Faculty of Mathematics and Physics. Together with the "sister" Institute of Theoretical and Computational Linguistics (Faculty of Arts) we aim at the development of teaching programs and research in the domain of theoretical and computational linguistics at the respective Faculties, collaborating closely with other departments such as the Institute of the Czech National Corpus at the Faculty of Philosophy and the Department of Computer Science at the Faculty of Mathematics and Physics.

CKL

As of 1 June 2000 the Center for Computational Linguistics (Centrum komputační lingvistiky; http://ckl.mff.cuni.cz) was established as one of the centers of excellence within the governmental program for support of research in the Czech Republic. The center is attached to the Faculty of Mathematics and Physics of Charles University in Prague.

Technical Reports

The ÚFAL/CKL technical report series has been established with the aim of disseminate topical results of research currently pursued by members, cooperators, or visitors of the Institute. The technical reports published in this Series are results of the research carried out in the research projects supported by the Grant Agency of the Czech Republic, GAČR 405/96/K214 ("Komplexní program"), GAČR 405/96/0198 (Treebank project), grant of the Ministry of Education of the Czech Republic VS 96151, and project of the Ministry of Education of the Czech Republic LN00A063 (Center for Computational Linguistics). Since November 1996, the following reports have been published.

ÚFAL TR-1996-01 Eva Hajičová, *The Past and Present of Computational Linguistics at Charles University*
Jan Hajič and Barbora Hladká, *Probabilistic and Rule-Based Tagging of an Inflective Language – A Comparison*

ÚFAL TR-1997-02 Vladislav Kuboň, Tomáš Holan and Martin Plátek, *A Grammar-Checker for Czech*

ÚFAL TR-1997-03 Alla Bémová at al., Anotace na analytické rovině, *Návod pro anotátory (in Czech)*

ÚFAL TR-1997-04 Jan Hajič and Barbora Hladká, *Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structural Tagset*

ÚFAL TR-1998-05 Geert-Jan M. Kruijff, *Basic Dependency-Based Logical Grammar*

ÚFAL TR-1999-06 Vladislav Kuboň, *A Robust Parser for Czech*

ÚFAL TR-1999-07 Eva Hajičová, Jarmila Panevová and Petr Sgall, *Manuál pro tektogramatické značkování (in Czech)*

ÚFAL TR-2000-08 Tomáš Holan, Vladislav Kuboň, Karel Oliva, Martin Plátek, *On Complexity of Word Order*

ÚFAL/CKL TR-2000-09 Eva Hajičová, Jarmila Panevová and Petr Sgall, *A Manual for Tectogrammatical Tagging of the Prague Dependency Treebank*

ÚFAL/CKL TR-2001-10 Zdeněk Žabokrtský, *Automatic Functor Assignment in the Prague Dependency Treebank*

ÚFAL/CKL TR-2001-11 Markéta Straňáková, *Homonymie předložkových skupin v češtině a možnost jejich automatického zpracování*

ÚFAL/CKL TR-2001-12 Eva Hajičová, Jarmila Panevová and Petr Sgall, *Manuál pro tektogramatické značkování (III. verze)*

ÚFAL/CKL TR-2002-13 Pavel Pecina and Martin Holub, *Sémanticky signifikantní kolokace*

ÚFAL/CKL TR-2002-14 Jiří Hana, Hana Hanová, *Manual for Morphological Annotation*

ÚFAL/CKL TR-2002-15 Markéta Lopatková, Zdeněk Žabokrtský, Karolína Skwarská and Vendula Benešová, *Tektogramaticky anotovaný valenční slovník českých sloves*

ÚFAL/CKL TR-2002-16 Radu Gramatovici and Martin Plátek, *D-trivial Dependency Grammars with Global Word-Order Restrictions*

ÚFAL/CKL TR-2003-17 Pavel Květoň, *Language for Grammatical Rules*

ÚFAL/CKL TR-2003-18 Markéta Lopatková, Zdeněk Žabokrtský, Karolina Skwarska, Václava Benešová, *Valency Lexicon of Czech Verbs VALLEX 1.0*

ÚFAL/CKL TR-2003-19 Lucie Kučová, Veronika Kolářová, Zdeněk Žabokrtský, Petr Pajas, Oliver Čulo, *Anotování koreference v Pražském závislostním korpusu*

ÚFAL/CKL TR-2003-20 Kateřina Veselá, Jiří Havelka, *Anotování aktuálního členění věty v Pražském závislostním korpusu*

ÚFAL/CKL TR-2004-21 Silvie Cinková, *Manuál pro tektogramatickou anotaci angličtiny*

ÚFAL/CKL TR-2004-22 Daniel Zeman, *Neprojektivity v Pražském závislostním korpusu (PDT)*

ÚFAL/CKL TR-2004-23 Jan Hajič a kol., *Anotace na analytické rovině, návod pro anotátory*

ÚFAL/CKL TR-2004-24 Jan Hajič, Zdeňka Urešová, Alevtina Bémová, Marie Kaplanová, *Anotace na tektogramatické rovině (úroveň 3)*

ÚFAL/CKL TR-2004-25 Jan Hajič, Zdeňka Urešová, Alevtina Bémová, Marie Kaplanová, *The Prague Dependency Treebank, Annotation on tectogrammatical level*

ÚFAL/CKL TR-2005-27 Jiří Hana, Daniel Zeman, *Manual for Morphological Annotation (Revision for PDT 2.0)*

ÚFAL/CKL TR-2005-28 Marie Mikulová a kol., *Pražský závislostní korpus (The Prague Dependency Treebank) Anotace na tektogramatické rovině (úroveň 3)*

ÚFAL/CKL TR-2005-29 Petr Pajas, Jan Štěpánek, *A Generic XML-Based Format for Structured Linguistic Annotation and Its application to the Prague Dependency Treebank 2.0*

ÚFAL/CKL TR-2006-30 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Annotation on the tectogrammatical level in the Prague Dependency Treebank (Annotation manual)*

ÚFAL/CKL TR-2006-31 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Petr Sgall, Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Anotace na tektogramatické rovině Pražského závislostního korpusu (Referenční příručka)*

ÚFAL/CKL TR-2006-32 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Petr Sgall,Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Annotation on the tectogrammatical level in the Prague Dependency Treebank (Reference book)*

ÚFAL/CKL TR-2006-33 Jan Hajič, Marie Mikulová, Martina Otradovcová, Petr Pajas, Petr Podveský, Zdeňka Urešová, *Pražský závislostní korpus mluvené češtiny. Rekonstrukce standardizovaného textu z mluvené řeči*

ÚFAL/CKL TR-2006-34 Markéta Lopatková, Zdeněk Žabokrtský, Václava Benešová (in cooperation with Karolína Skwarska, Klára Hrstková, Michaela Nová, Eduard Bejček, Miroslav Tichý) *Valency Lexicon of Czech Verbs. VALLEX 2.0*

ÚFAL/CKL TR-2006-35 Silvie Cinková, Jan Hajič, Marie Mikulová, Lucie Mladová, Anja Nedolužko, Petr Pajas, Jarmila Panevová, Jiří Semecký, Jana Šindlerová, Josef Toman, Zdeňka Urešová, Zdeněk Žabokrtský, *Annotation of English on the tectogrammatical level*

ÚFAL/CKL TR-2007-36 Magda Ševčíková, Zdeněk Žabokrtský, Oldřich Krůza, *Zpracování pojmenovaných entit v českých textech*

ÚFAL/CKL TR-2008-37 Silvie Cinková, Marie Mikulová, *Spontaneous speech reconstruction for the syntactic and semantic analysis of the NAP corpus*

ÚFAL/CKL TR-2008-38 Marie Mikulová, *Rekonstrukce standardizovaného textu z mluvené řeči v Pražském závislostním korpusu mluvené češtiny. Manuál pro anotátory*

ÚFAL/CKL TR-2008-39 Zdeněk Žabokrtský, Ondřej Bojar, *TectoMT, Developer's Guide*

ÚFAL/CKL TR-2008-40 Lucie Mladová, *Diskurzní vztahy v češtině a jejich zachycení v Pražském závislostním korpusu 2.0*

ÚFAL/CKL TR-2009-41 Marie Mikulová, *Pokyny k překladu určené překladatelům, revizorům a korektorům textů z Wall Street Journal pro projekt PCEDT*

ÚFAL/CKL TR-2011-42 Loganathan Ramasamy, Zdeněk Žabokrtský, *Tamil Dependency Treebank (TamilTB) – 0.1 Annotation Manual*

ÚFAL/CKL TR-2011-43 Nguy Giang Linh, Michal Novák, Anna Nedoluzhko, *Coreference Resolution in the Prague Dependency Treebank*

ÚFAL/CKL TR-2011-44 Anna Nedoluzhko, Jiří Mírovský, *Annotating Extended Textual Coreference and Bridging Relations in the Prague Dependency Treebank*

ÚFAL/CKL TR-2011-45 David Mareček, Zdeněk Žabokrtský, *Unsupervised Dependency Parsing*

ÚFAL/CKL TR-2011-46 Martin Majliš, Zdeněk Žabokrtský, *W2C – Large Multilingual Corpus*

ÚFAL TR-2012-47 Lucie Poláková, Pavlína Jínová, Šárka Zikánová, Zuzanna Bedřichová, Jiří Mírovský, Magdaléna Rysová, Jana Zdeňková, Veronika Pavlíková, Eva Hajičová, *Manual for annotation of discourse relations in the Prague Dependency Treebank*

ÚFAL TR-2012-48 Nathan Green, Zdeněk Žabokrtský, *Ensemble Parsing and its Effect on Machine Translation*

ÚFAL TR-2013-49 David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Daniel Zemana, Zdeněk Žabokrtský, Jan Hajič *Cross-language Study on Influence of Coordination Style on Dependency Parsing Performance*

ÚFAL TR-2013-50 Jan Berka, Ondřej Bojar, Mark Fishel, Maja Popović, Daniel Zeman, *Tools for Machine Translation Quality Inspection*

ÚFAL TR-2013-51 Marie Mikulová, *Anotace na tektogramatické rovině. Dodatky k anotátorské příručce (s ohledem na anotování PDTSC a PCEDT)*

ÚFAL TR-2013-52 Marie Mikulová, *Annotation on the tectogrammatical level. Additions to annotation manual (with respect to PDTSC and PCEDT)*

ÚFAL TR-2013-53 Marie Mikulová, Eduard Bejček, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Pavel Straňák, Magda Ševčíková, Zdeněk Žabokrtský, *Úpravy a doplňky Pražského závislostního korpusu (Od PDT 2.0 k PDT 3.0)*

ÚFAL TR-2013-54 Marie Mikulová, Eduard Bejček, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Pavel Straňák, Magda Ševčíková, Zdeněk Žabokrtský, *From PDT 2.0 to PDT 3.0 (Modifications and Complements)*

ÚFAL TR-2014-55 Rudolf Rosa, *Depfix Manual*

ÚFAL TR-2014-56 Veronika Kolářová, *Valence vybraných typů deverbativních substantiv ve valenčním slovníku PDT-Vallex*

ÚFAL TR-2014-57 Anna Nedoluzhko, Eva Fučíková, Jiří Mírovský, Jiří Pergler, Lenka Šíková, *Annotation of coreference in Prague Czech-English Dependency Treebank*

ÚFAL TR-2015-58 Zdeňka Urešová, Eva Fučíková, Jana Šindlerová, *CzEngVallex: Mapping Valency between Languages*