Verb Argument Pairing in Czech-English Parallel Treebank

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

Abstract

We describe CzEngVallex, a bilingual Czech-English valency lexicon which aligns verbal valency frames and their arguments. It is based on a parallel Czech-English corpus, the Prague Czech-English Dependency Treebank, where for each occurrence of a verb a reference to the underlying Czech and English valency lexicons is explicitly recorded. CzEngVallex lexicon pairs the entries (verb senses) of these two lexicons, and allows for detailed studies of verb valency and argument structure in translation. While some related studies have already been published on certain phenomena, we concentrate here on basic statistics, showing that the variability of verb argument mapping between verbs in the two languages is richer than it might seem and than the perception from the studies published so far might have been.

Keywords: lexical resources, parallel corpus, treebank, valency, bilingual valency lexicon, Czech, English

1. Introduction

Valency, or verb argument structure, is an important phenomenon both in linguistic studies as well as in language technology applications, since verb is considered the core of a clause in (almost) every natural language utterance. Various lexicons have been built - from Propbank (Palmer et al., 2005) to Framenet (Baker et al., 1998). Various valency lexicons exist for several languages, such as Walenty (Przepiórkowski et al., 2014) for Polish, and several exist also for Czech: primarily VALLEX (Žabokrtský and Lopatková, 2007) and Verbalex (Horák, Aleš and Pala, Karel and Hlaváčková, Dana, 2013). However, there are no truly multilingual valency lexicons, and none link parallel corpora together through valency lexicons the way the CzEngVallex lexicon does, as described in (Urešová et al., 2015a) and analyzed in this paper. It thus offers an opportunity to learn not only about valency as generalized across languages, but also to study translation from a different perspective thanks to the explicit references between the parallel Czech-English corpus and the valency lexicons for the two languages.

In this paper, we briefly describe the resources and their interplay, and then analyze the CzEngVallex lexicon in more detail, showing also examples of the (mis)match of verb valency between the two languages.

2. The PCEDT parallel corpus

The Prague Czech-English Dependency Treebank (PCEDT 2.0) (Hajič et al., 2012) contains the WSJ part of the Penn Treebank (Marcus et al., 1993) and its manual professional translation to Czech, annotated manually using the tectogrammatical representation (Mikulová et al., 2005), first used for the Prague Dependency Treebank 2.0 (PDT) (Hajič et al., 2006).

2.1. PCEDT: the annotation scheme

The PCEDT contains 866,246 English tokens and 953,187 Czech tokens, aligned manually sentence-by-sentence and automatically word-by-word. It is annotated on all three annotation layers of the PDT: morphological, analytical (surface dependency syntax) and tectogrammatical (syntactic-semantic). However, as opposed to the PDT which is annotated fully manually,[1] PCEDT has been annotated for structure and valency at the tectogrammatical representation layer manually, but for POS and morphology and surface syntax only automatically.[2] Both language sides of the tectogrammatical representation have been enriched with valency annotation, using two valency lexicons: PDT-Vallex for Czech and EngVallex for English. Fig. 7 shows an example of an annotated pair of aligned sentences in the PCEDT (together with visualized CzEngVallex projection, see below Sect. 3.).

2.2. PDT-Vallex: Czech valency lexicon

The PDT-Vallex (Hajič et al., 2003; Urešová, 2011b; Urešová, 2011a) has been originally developed for the PDT annotation. It contains 12,000 verb frames for about 7,000 verbs, roughly corresponding to verb senses found during the annotation of the PDT and PCEDT treebanks. For each frame, verb arguments are listed together with the obligatoriness and constraints on surface morphosyntactic realization; examples and notes are given for each entry as well. Each occurrence of a verb in the PDT (and on the Czech side of the PCEDT) is linked to one verb frame in the PDT-Vallex lexicon. The same lexicon has also been used for the annotation of spoken Czech in the Prague Dependency Corpus of Spoken Czech, or PDTSC[3] (Hajic et al., 2009).

2.3. EngVallex: English valency lexicon

The EngVallex (Cinková, 2006) has been created for the English side of the PCEDT annotation. It is a semi-manual conversion of the Propbank frame files (Palmer et al., 2005) into the PDT style of capturing valency information in valency frames. The correspondence of the original Propbank

[1] With the exception of certain lexical node attributes.

[2] The surface dependency syntax on the English side has been derived from the Penn Treebank constituent syntax annotation, using head percolation rules, and thus can be considered semi-manual as well.

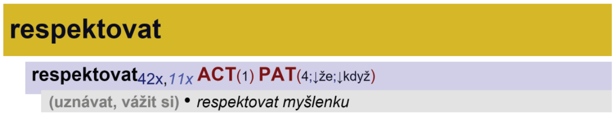
[3] http://ufal.mff.cuni.cz/pdtsc1.0/en/index.html

entries and valency frames in EngVallex is not necessarily 1:1 - entries have been occasionally merged or split. It contains over 7,000 frames for 4,300 verbs.

2.4. Treebank-lexicon links and lexicon entries

From the point of view of valency in general and this paper in particular, the most important part of the annotation of the corpus and its relation to the valency lexicons is the treatment of verb arguments and adjuncts. Every (non-auxiliary) verb node in the treebank refers to one particular sense of that verb in the respective valency lexicon (PDT-Vallex or EngVallex). The nodes dependent on the verb in the annotation are obligatory or optional complementations. All actants[4] and other obligatory complementations (we will call them collectively “arguments” for simplicity)[5] are also recorded in the valency lexion(s). In other words, the valency lexicon entry matches the verb-rooted subtree of the annotated tectogrammatical tree linked to it.

The “core” arguments (“actants” in the tectogrammatical terminology) are Actor (or deep subject, or first argument, ACT), Patient (deep object, or second argument, PAT), Addressee (ADDR), Effect (EFF) and Origin (in the transformational sense, such as *create a doll from wood*, labeled ORIG). Non-core arguments often deemed obligatory with certain verbs and their senses are Location (LOC), Direction-from (DIR1), Direction-to (DIR3), Manner (MANN), Beneficiary (BEN) and several others.



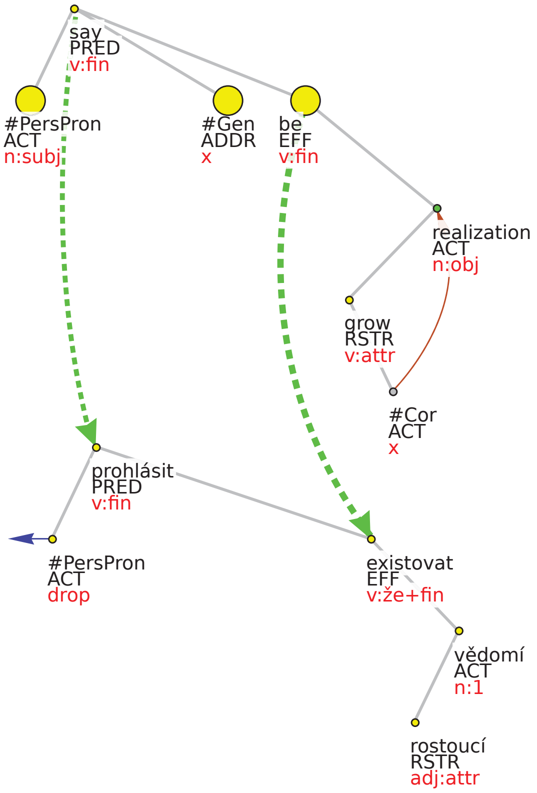
respektovat

respektovat$_{42x, 11x}$ **ACT**(1) **PAT**(4;↓že;↓když)

(uznávat, vážit si) • *respektovat myšlenku*

Figure 1: PDT-Vallex example entry of the valency frame for *respektovat* (lit. *respect, heed, honor*)

An example of a valency entry for the Czech verb *respektovat* is in Fig. 1. Since Czech is an inflective language and morphosyntactic features are essential for the description of verb arguments, they are listed in the lexicon entry as well, following the argument label (e.g., for the Patient argument in the figure, the number “4” means accusative case, and the arrows are used to specify that the argument can also be expressed as a subordinate clause, in this case using either the conjunction *“že”* or *“když”*).[6]



En: She said there is a “growing realization”...
Cz: Prohlásila, že ... existuje “rostoucí vědomí...

Figure 2: Verb and argument pairs suggested by the automatic preprocessing step (green arrows)

3. The CzEngVallex lexicon

The CzEngVallex lexicon (Urešová et al., 2015a; Urešová et al., 2015b)[7] is a bilingual valency lexicon with explicit pairing of verb senses (corresponding to valency frames) and their arguments, built upon the Prague Czech-English Dependency Treebank (PCEDT), as described in the previous section. It contains 20,835 frame pairs. It should be noted that not all verbs from the PCEDT can be found in the CzEngVallex: some verbs have not at all been translated as verbs, and vice versa, and some verb translations have been so structurally different that even if translated as verbs, they have not been included in the CzEngVallex. According to (Urešová et al., 2015a), 71% of English verb tokens found in the corpus have been aligned and can be found in the CzEnVallex (for Czech verb occurrences, it is 77%). Also, due to the fact that the CzEngVallex is restricted to the parallel corpus only, it also covers only about 2/3rd of the underlying valency lexicons, i.e., PDT-Vallex and EngVallex. Exacts statistics are given in Table 1 (Urešová et al., 2015a).

	Verb	Frame	PCEDT Tokens	
Language	types	types	verbs	aligned
English	3,292	5,010	130,514	92,747
Czech	4,218	6,930	118,189	91,656

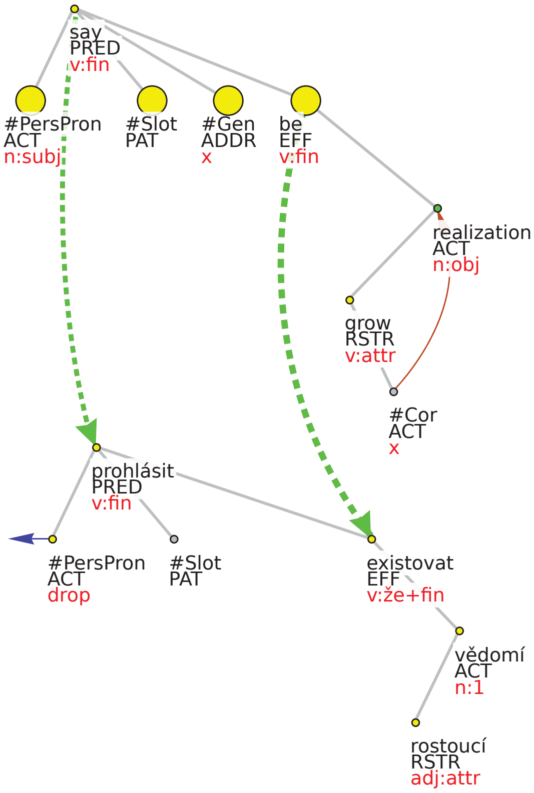
Table 1: Alignment coverage - CzEngVallex/PCEDT

[4] Sometimes called “core” arguments, see below for a list.

[5] The distinction between arguments and adjuncts is often understood differently by different authors, but that is not the important point; here, our use of “argument” is wider than usual, as it gets clearer later.

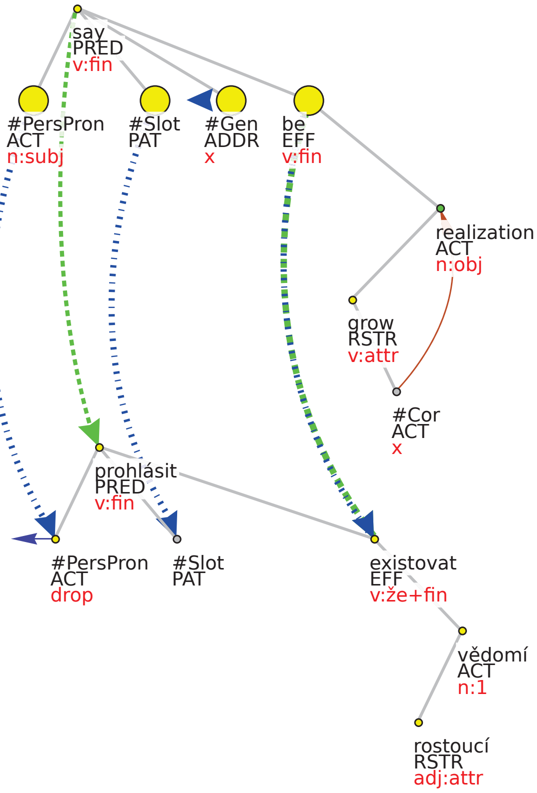
[6] Frequencies in the PDT and PCEDT treebank are included as well, and so are synonyms and a human-readable description or definition of the particular verb sense, especially to distinguish entries of polysemous verbs.

[7] Available publicly for download from the `http://lindat.cz` repository, together with the monolingual valency lexicons and the PCEDT corpus.



En: She said there is a "growing realization"...
Cz: Prohlásila, že ... existuje "rostoucí vědomí...

Figure 3: Verb and argument pairs after insertion of elided valency slots



En: She said there is a "growing realization"...
Cz: Prohlásila, že ... existuje "rostoucí vědomí...

Figure 4: Verb and argument pairs as marked by the annotator (blue arrows) for entering them into CzEngVallex

3.1. Building CzEngVallex

CzEngVallex has been built, as it has been mentioned above, on top of the Prague Czech-English Dependency Treebank. The corpus, annotated manually for (monolingual) valency, has been first (automatically) pre-processed to align all nodes of the tectogrammatically-annotated trees, and all trees which contained at least one verb–verb pair have been extracted, re-sorted to show all pairs of trees with the same sense of the English verb in one group, and passed to the CzEngVallex annotators. Fig. 2 shows an example (fragment) of a sentence pair (Eng: *She said there is a "growing realization"* ...) containing the verb *say* and its translation (*prohlásit*, in this sentence), as displayed for the annotator, with green arrows showing pre-aligned verb and argument pairs. The main task of the annotators has been to check the pairings of both verbs and their arguments, and to add or correct them if necessary. The underlying hypothesis which has determined the design of the valency frame pairing scheme was that for each verb sense pair, the alignment of their arguments is the same (otherwise, the verb sense on one or both sides would have to be refined). This was the key point of the annotation, apart from corrections of the errors of the original automatic node alignment or corrections of the treebank annotation itself.[8] In our example sentence, the annotator fills in missing (non-overt) arguments for both *say* and its Czech translation, namely, the deep object (PAT, with the lexeme represented only as `#Slot`, see Fig. 3). After filling in all of the elided valency slots, the annotator adds alignment links for the newly introduced arguments and for those that have not been identified by the automatic preprocessing step. In the displayed case, the nodes with ACT and PAT have been aligned and the ADDR node has been marked as non-corresponding to any Czech argument (Fig. 4, blue arrows).

Only after a careful review of the whole group of *all* PCEDT examples for the given pair of verb senses and their valency frames the alignment of the arguments has been confirmed by the annotator and the valency frame pair entered into CzEngVallex.

3.2. Annotation rules in specific cases

Due to slight inconsistencies in the handling of verb arguments and adjuncts on the two sides of the PCEDT, the annotation rules had to be gradually extended to contain con-

[8] To keep the annotation consistent, corrections in the treebank have only been suggested and passed to the treebank maintainers to include them in the next version, i.e., the underlying treebanks have not been corrected immediately.

ventions for such cases, in order to keep the CzEngVallex pairings consistent. For example, EngVallex (used for the valency annotation of the English side of the PCEDT) often includes certain adjuncts (i.e., optional free modifications in the PDT terminology) in the valency frame, while PDT-Vallex strictly does not. This is, of course, not a cause for a "true" argument mismatch, but the treatment for these had to be unified so that these cases are easily identifiable afterwards.

Similarly, certain types of verb constructions using more than one verb (typically, catenative verb or a modal) might have structurally different annotation, if only for the fact that one one side of the translation only one verb is used carrying the same meaning. In these cases, the "semantic" annotation rule takes effect, i.e., the modal or catenative verb is left out and the alignment is made between the more semantically "full" verb and its single-word counterpart in the other language (node in the annotated tree). For example, *keep* and *riding (up)* are represented as two nodes in the English tree annotation, while their translation is only *klouzat* in Czech (albeit complemented by and adverbial *stále*, meaning *lit. still*); in such a case, *keep* is not considered part of the pair and alignment is made for *ride (up)* and *klouzat* and their arguments only. In addition to *keep*, *need* or *get* (when complemented by a non-finite verb) also appear often translated in the same way.

In some cases, the translation itself could be plain wrong (however unlikely it might seem after professional translation editing and fully manual tectogrammatical annotation took place on the data prior to this alignment effort). In these cases, the corpus pairing is excluded from consideration and the error reported to the treebank maintainers.

3.3. CzEngVallex format

The resulting CzEngVallex is represented as a simple standoff file which refers back to the PDT-Vallex and EngVallex lexicons, or more precisely, to the individual valency frames in them. In other words, the underlying two lexicons are not modified at all, which makes it easier to maintain them in the future (Fig. 5). The valency frames are referred to by their respective IDs, while the arguments are identified by their labels (since they are for each frame unique). Technically, all Czech frame pairs are listed for every English verb, but the relations are symmetric.

CzEngVallex is also publicly available online for quick browsing and search.[9] This interface allows for searching for particular argument pairs aligned by CzEngVallex, resulting in a list of verbs (and their particular valency frames) where this pairing occurs. Individual verb and verb pairs can also be browsed alphabetically, in both directions (English->Czech as well as Czech->English). Moreover, each pair of valency frames displayed is complemented with all the real-usage examples from the parallel PCEDT corpus (Fučíková et al., 2015). All the displayed material (verb entry heading, valency frames, etc.) are linked through HTML links to the monolingual entries in PDT-Vallex and EngVallex, to display additional information and, in the case of PDT-Vallex, additional examples from the monolingual Czech PDT corpus.

```
<frames_pairs owner="...">
 <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 5: Structure of the CzEngVallex (part of *abandon* pairing)

Number of argument pairs	Number of frame pairs	Percent of all pairs
0	9	0.04%
1	593	2.85%
2	8746	41.98%
3	7939	38.10%
4	2613	12.54%
5	813	3.90%
6	103	0.49%
7	19	0.09%

Table 2: Argument pairing statistics

4. Argument matching in the CzEngVallex / PCEDT

Out of the 20,835 frame pairs recorded in the CzEngVallex lexicon, Table 2 summarizes argument alignment diversity in these frame pairs: it shows how many times a certain number of argument pairs appears in the CzEngVallex lexicon.

It should be noted that not necessarily the number of arguments on both sides is equal to the number of pairs; some pairs might in effect pair an argument with "nothing" on the other side. A study on such a "zero" alignment can be found in (Šindlerová et al., 2015).

One of the reasons for creating CzEngVallex was to have explicitly annotated corpus material for the study of translation differences in Czech and English valency, or verb argument (and in some cases, also adjunct) use. Overall statistics are given in Table 3.

An example of a well-behaved verb pair is in Fig. 6, where all three arguments match between the two languages for

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

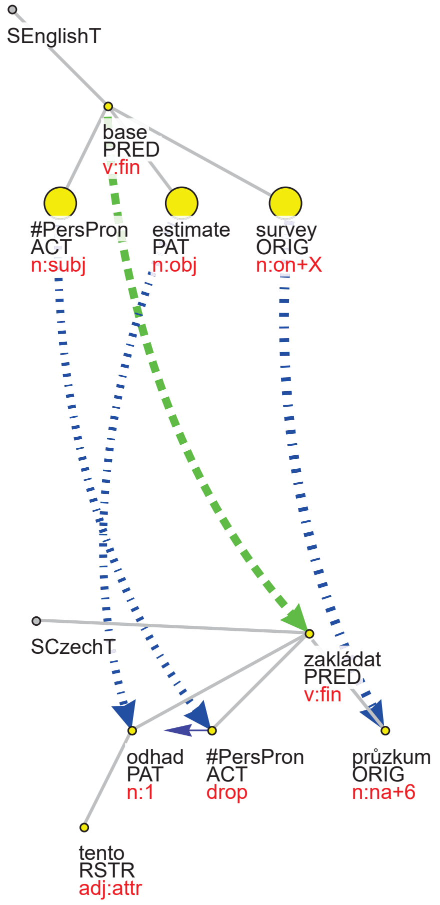
No. of argument pair differences	Number of frame pairs	Percent of all pairs
0	9302	44.646%
1	6737	32.335%
2	3313	15.901%
3	1157	5.553%
4	267	1.281%
5	49	0.235%
6	9	0.043%
7	1	0.005%

Table 3: Argument pair differences in numbers

Number of argument pairs	Number of unique pairing types	Percentage
0	1	0.04%
1	4	0.15%
2	238	9.20%
3	980	37.88%
4	935	36.14%
5	338	13.07%
6	76	2.94%
7	15	0.58%

Table 4: Argument pair differences in numbers

the verb sense pair *base–zakládat*.



En: He bases the estimate on a survey...
Cz: Tento odhad zakládá na průzkumu...

Figure 6: Matching arguments in verb pair base–zakládat; verb pair in green, argument links in blue.

However, quite clearly as the table shows, there are more differing pairs (over 55%) than those which match in all argument pairings.
An example of an aligned sentence with five differences in argument mapping is captured in Fig. 7.
The example with seven differences comes from the translation of the English verb "to sell" to Czech as "vyvážet" (lit. *export*) in

- En: For example, Nissho Iwai Corp., one of the biggest Japanese trading houses, now buys almost twice as many goods from China as it.ACT sells to that country.ADDR
- Cz: Společnost Nissho Iwai Corp., jedna z největších japonských obchodních firem, dnes například kupuje dvakrát tolik zboží z Číny, než kolik.PAT do této země.DIR3 vyváží

In this case, the English entry has five argument slots, labeled ACT, PAT, ADDR, EFF, BEN and the Czech entry ACT, PAT and DIR1;[10] ACT maps to PAT, ADDR to DIR3 (not included as an argument in the valency frame), and all others are unaligned (in either direction), accounting for the seven pairing differences.
Out of the frame pairs with just one argument pair, four different cases have been found. While it is not surprising that by far the most frequent pair is the expected ACT:ACT labeled argument pair, three other differing pairs have been found:[11]

1. five frame pairs with PAT:ACT argument pair; this is apparently the relict of not shifting the English valency slot label PAT to ACT, due to its origins in Propbank which often uses Arg1 alone (such as in *the glass.Arg1 broke*, and Engvallex typically used PAT for Arg1;
2. four times no English frame argument corresponding to ACT in the Czech frame, and
3. one case of an ACT on the English side corresponding to no argument on the Czech side.

With the increasing number of arguments, there are more and more different pairings of arguments, as the combinatorics also suggest. The numbers are given in Table 4. The percentages are computed from the total number of 2,587 different (unique) pairs found in the CzEngVallex lexicon across all argument pair counts.

[10] Not all of them are present in the (surface form of the) example, but the alignment is not affected by argument ellipsis.
[11] More examples and their breakout (including possible annotation errors) will be presented in the full version of the paper.

5. Mismatch classification

While complete breakout and classification of the 2,500+ mismatch types apparently needs further study, we can already provide (a coarse grained) classification. The "zero alignment" has already been mentioned and studied (Šindlerová et al., 2015), since it accounts for a large proportion of argument alignment discrepancies. However, when we step up from the investigation of individual argument alignments to the level of the whole valency frame, the situation is far richer. Nevertheless, there are certain common reasons for various types of mismatches:

- verb translation choice often combined with differing argument expression and/or representation, which can further be subdivided into several types (plain argument expression (*to drive a car.PAT* vs. *jezdit v autě.MEANS*, lit. *go in a car*), light verb constructions translated as a single verb or vice versa, such as *uzavřít smlouvu s ...*, lit. *close a contract with ...* → *(to) contract sb*, "cross-language" alternation (cf. also Fig. 7 and below), other structural differences)

- treebank annotation convention and guidelines (e.g., choice of direction vs. origin), cf. Fig. 7: *is derived from the U.S.*.ORIG vs. *pochází z USA*.DIR1

- valency frame composition convention mismatch (for example: En: *spread* ACT DIR1 DIR2 DIR3 → Cz: *rozšířit se* ACT, where the direction(s) of spreading are not included in the Czech valency frame, being considered optional complementations).

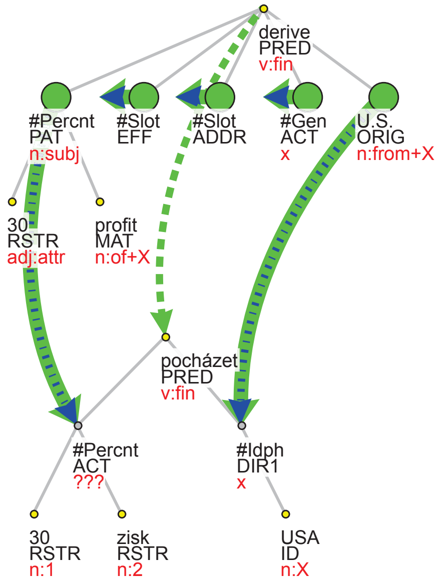
As an illustration,[12] consider the following translation:

- En: ... the change ended [the series]

- Cz: ... série skončila změnou (lit. *[the] series ended by-change*)

where the (deep) object (*series*) has moved to (deep) subject position in Czech (this alternation process applies to both languages; it was the translator's choice to do so in Czech). In a slightly more complex example, we refer to another case of "cross-language" alternation (Fig. 7): the passive form "is derived" has been translated as intransitive "pocházet" (more literally translated as *come from*), where the deep subject (ACT) represents the theme, while in English this is the deep object (PAT) of "derive": someone derives something.PAT from ... This example suggests that in translation, the choice of the translation is often not done at the more syntactically-oriented valency (or "propbanking") level, but at a much deeper, FrameNet-like more semantically-oriented level (Baker et al., 1998); while this might not be surprising for human translators, it confirms that it has to be taken into account for MT. Interlinking all the valency/propbanking/semantic role lexicons, similarly to (Bonial et al., 2013), would give us more insight, but it must be complemented with multilingual annotation in a similar way that we have attempted here with CzEngVallex in the bilingual case.

For completeness, we should also mention our previous work on investigating how verb-noun phrasal and verb idiomatic constructions are translated (Urešová et al., 2013). We have found that only a minority of such constructions are translated as idiomatic or phrasal constructions (from English to Czech), and perhaps even more surprisingly, it also holds in the other direction, namely that idioms (in the Czech translation) are often coming from non-idiomatic constructions. The findings about translations of verb-noun idiomatic constructions has led to more focus on the representation of such constructions themselves in valency dictionaries in different languages; comparison between Czech and Polish with suggestions for improvement in representation of verb-based idiomatic constructions has been described in (Przepiórkowski et al., 2016 in print).



En: ... 30% of ... profit ... is derived from the U.S..
Cz: .. 30 % zisků ... pochází z USA.

Figure 7: Functor mismatch in 5 argument pairs

6. Related work

The predecessor to CzEngVallex, which has used machine learning methods based on a parallel corpus, has been described in (Šindlerová and Bojar, 2009), but it did not produce a manually checked and corrected resource. Another preliminary attempt at a comparison of English and Czech Valency has been using several resources (PDEV on the English side and VerbaLex on the Czech side), but it has not used a parallel corpus for linking and checking the actual usage (Pala et al., 2014). Obviously, multilingual dictionaries like FrameNet (Fillmore et al., 2003; Baker et al., 1998; Materna and Pala, 2010) inherently contain links between verb sense equivalents, but we are not aware of any work that would start from a parallel corpus, use the same methodology of valency description for both languages and that has underwent a thorough manual check.

[12] Due to the limited space in the abstract - more examples and finer grained classes in the full version of the paper.

7. Conclusions

We have described some basic statistics derived from the CzEngVallex lexicon, a bilingual valency lexicon created over the Prague Czech-English Dependency Treebank, a parallel corpus of over 50,000 sentences. Perhaps it should not be surprising that there is a large number of differences in the use of verb arguments across the two languages. The 2,587 different valency frame pairs (in the alignment of their arguments) offer a large amount of material for further studies.

Apart from studying the properties of the lexical entries themeselves, we have already used the lexicon in various NLP applications, such as in word sense disambiguation using the argument and verb pairings coming from the parallel corpus as an additional features, getting an improvement over the (monolingual) baseline (Dušek et al., 2015).

Since the CzEngVallex lexicon, both underlying valency lexicons (PDT-Vallex for Czech and EngVallex for English) are now publicly available online,[13] we believe that it will be possible to get more insight into the use of verb arguments in translation, benefiting both linguistic studies as well as language technology, especially machine translation.

8. Acknowledgments

This work described herein has been supported by the grant GP13-03351P of the Grant Agency of the Czech Republic and by the LINDAT/CLARIN Research Infrastructure projects, LM2010013 and LM2015071 funded by the MEYS of the Czech Republic. It has also been using language resources developed and distributed by the LINDAT/CLARIN project. The results described in this article have been contributed to the LINDAT/CLARIN open access repository, as a item `http://hdl.handle.net/11234/1-1512` and as an user-accessible application at `http://lindat.mff.cuni.cz/services/CzEngVallex`.

9. References

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bonial, C., Stowe, K., and Palmer, M., (2013). *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, chapter Renewing and Revising SemLink, pages 9 – 17. Association for Computational Linguistics.

Cinková, S. (2006). From Propbank to Engvallex: Adapting the PropBank-Lexicon to the Valency Theory of the Functional Generative Description. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006), Genova, Italy*.

Dušek, O., Fučíková, E., Hajič, J., Popel, M., Šindlerová, J., and Urešová, Z. (2015). Using parallel texts and lexicons for verbal word sense disambiguation. In *Proceedings of the Third International Conference on Dependency Linguistics, Depling 2015*, Uppsala, Sweden. Uppsala universitet, Uppsala universitet.

Fillmore, C. J., Johnson, C. R., and L.Petruck, M. R. (2003). Background to framenet: Framenet and frame semantics. *International Journal of Lexicography*, 16(3):235–250.

Fučíková, E., Hajič, J., Šindlerová, J., and Urešová, Z. (2015). Czech-english bilingual valency lexicon online. In *14th International Workshop on Treebanks and Linguistic Theories (TLT 2015)*, pages 61–71, Warszawa, Poland. IPIPAN, IPIPAN.

Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V., and Pajas, P. (2003). PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In Nivre, Joakim//Hinrichs, E., editor, *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9 of *Mathematical Modeling in Physics, Engineering and Cognitive Sciences*, pages 57—68, Vaxjo, Sweden. Vaxjo University Press.

Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., Žabokrtský, Z., Ševčíková-Razímová, M., and Urešová, Z. (2006). Prague Dependency Treebank 2.0, LDC Catalog No. LDC2006T01.

Hajic, J., Pajas, P., Marecek, D., Mikulova, M., Uresova, Z., and Podvesky, P. (2009). Prague dependency treebank of spoken language (PDTSL) 0.5. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Bojar, O., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z., and Žabokrtský, Z. (2012). Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3153–3160, Istanbul, Turkey. European Language Resources Association.

Horák, Aleš and Pala, Karel and Hlaváčková, Dana. (2013). Preparing VerbaLex Printed Edition. In *Seventh Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2013*, pages 3–11.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of english: The penn treebank. *COMPUTATIONAL LINGUISTICS*, 19(2):313–330.

Materna, J. and Pala, K. (2010). Using Ontologies for Semi-automatic Linking VerbaLex with FrameNet. In *LREC*, pages 3331–3337.

Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová, V., Lopatková, M., Pajas, P., Panevová, J., Razímová, M., Sgall, P., Štěpánek, J., Urešová, Z., Veselá, K., Žabokrtský, Z., and Kučová, L. (2005). Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka. Technical

[13] `http://hdl.handle.net/11234/1-1512` and `http://lindat.mff.cuni.cz/services/CzEngVallex`

Report TR-2005-28, ÚFAL MFF UK, Prague, Prague.

Pala, K., Baisa, V., Sitová, Z., and Vonšovský, J. (2014). Mapping czech and english valency lexicons: Preliminary report. In *Eighth Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 139–145, Brno. Tribun EU.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

Przepiórkowski, A., Hajič, J., Hajnicz, E., and Urešová, Z. (2016, in print). Phraseology in two slavic valency dictionaries: limitations and perspectives. *International Journal of Lexicography*.

Przepiórkowski, A., Hajnicz, E., Patejuk, A., Woliński, M., Skwarski, F., and Świdziński, M. (2014). Walenty: Towards a comprehensive valence dictionary of Polish. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 2785–2792, Reykjavík, Iceland. ELRA.

Šindlerová, J. and Bojar, O. (2009). Towards English-Czech Parallel Valency Lexicon via Treebank Examples. In *Eighth International Workshop on Treebanks and Linguistic Theories*, pages 185–195.

Šindlerová, J., Fučíková, E., and Urešová, Z. (2015). Zero alignment of verb arguments in a parallel treebank. In Hajičová, E. and Nivre, J., editors, *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 330–339, Uppsala, Sweden. Uppsala University, Uppsala University.

Urešová, Z., Fučíková, E., Hajič, J., and Šindlerová, J. (2013). An Analysis of Annotation of Verb-Noun Idiomatic Combinations in a Parallel Dependency Corpus. Proceedings from The 9th Workshop on Multiword Expressions, Workshop at NAACL 2013.

Urešová, Z., Dušek, O., Fučíková, E., Hajič, J., and Šindlerová, J. (2015a). 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. Association for Computational Linguistics.

Urešová, Z., Fučíková, E., and Šindlerová, J. (2015b). Czengvallex: Mapping valency between languages. Technical Report TR-2015-58.

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

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

Žabokrtský, Z. and Lopatková, M. (2007). Valency information in VALLEX 2.0: Logical structure of the lexicon. *The Prague Bulletin of Mathematical Linguistics*, (87):41–60.