Towards English-Czech Parallel Valency Lexicon via Treebank Examples

Jana Šindlerová and Ondřej Bojar

Charles University in Prague
Institute of Formal and Applied Linguistics (ÚFAL)
E-mail: {sindlerova,bojar}@ufal.mff.cuni.cz

Abstract

The paper describes an ongoing project of building a bilingual valency lexicon in the framework of Functional Generative Description. The bilingual lexicon is designed as a result of interlinking frames and frame elements of two already existing valency lexicons.

First, we give an overall account of the character of the lexicons to be linked, second, the process of frame linking is explained, and third, a case study is presented to exemplify what the information contained in frame links tells us about crosslinguistic differences in general and the linguistic theory applied.

1 Introduction

Bilingual and multilingual lexicons have been arising quite commonly on the computational linguistics scene in the past decade. Besides the simple fact that electronic bi- and multilingual dictionaries are necessary tools for NLP projects concerned with machine translation, there is also a strong assumption that bi- and multilingual valency lexicons can be useful in the MT area as well.

Many researchers in the field of lexicography underline the need to support the lexicographic data with the evidence from linguistic corpora, e.g. [2]. Dictionaries and valency lexicons thus often arise directly in the process of corpora annotation (see e.g. [4]).

Contrary to the idea of building a valency lexicon as a resource for treebank annotation, we present an ongoing project of building a bilingual valency lexicon using a treebank as an annotation tool. The project takes advantage of two already existing valency lexicons: PDT-VALLEX and Engvallex, which have been developed during the annotation of the Prague Dependency Treebank [3] and Prague Czech-English Dependency Treebank (PCEDT, [1]), and of the parallel treebank PCEDT itself.

PCEDT is a syntactically annotated parallel corpus of approximately 50,000 sentences originally from the Penn Treebank (Wall Street Journal section), translated into Czech. The merit of PCEDT lies in the fact that the core annotation takes place on the tectogrammatical layer (t-layer), i.e. on the layer of deep syntactic relations with an overrun into the area of semantic relations. The deep syntactic annotation of PCEDT is still in progress. The annotation works on the Czech part (PCEDT_CZ) and the English part (PEDT) proceed independently albeit synchronized. Currently, about 11,500 mutually corresponding sentences are finished, which amounts to about 23% of the whole corpus (though the percentage of sentences already finished on each individual side of PCEDT reaches higher, to about 40% and 60%). By the time our multilingual valency lexicon is concluded, we expect the PCEDT corpus to have been completed.

By creating a bilingual valency lexicon, we hope to gain a multifunctional resource useful in many areas. First it will provide linguistic information about the behaviour of verbal valencies in a crosslinguistic perspective. Second, the resulting multilingual valency lexicon created in a specific linguistic framework (FGD) may serve as an interesting test for the usability and appropriateness of the framework itself. Fourth, since there is an assumption of a certain degree of universal behaviour across languages, comparing the frames in the two lexicons can be used as a test of the accuracy of the lexicons. And last but not least, with respect to the fact that verbal valencies serve as the core of syntactic structure in most languages, it will provide an interesting resource for MT applications.

2 Lexicographic Process in a Parallel Treebank

2.1 Construction of Source Valency Lexicons

PDT-VALLEX[1] has been developed as a resource for valency annotation in a large-scale syntactically annotated corpus, the Prague Dependency Treebank. Information about verbal valency is embedded into the tectogrammatical layer of annotation, i.e. the layer of deep syntactic dependency relations, therefore it does not specify any surface requirements but rather syntactico-semantic requirements of the verbs. Each headword contains one or more valency frames corresponding (mostly) to the individual senses of the headword. Valency frames contain participant slots represented by tectogrammatical functors, i.e. labels from the layer of syntactico-semantic representation. Only the so called "inner participants" and obligatory "free modifiers" are included in the frame, information about possible typical background elements is not stored except for some short notes in the example area. Each slot is marked as obligatory or optional.

[1]Not to be confused with VALLEX [5], a general lexicon with very similar formal background but not tailored to any corpus.

By now, PDT-VALLEX contains 10,593 valency frames for 6,667 verbs. The verbs and frames come mostly from the data appearing in the PDT, version 2.0, the lexicon is being constantly enlarged by data gained from PCEDT annotation.

The origin of Engvallex is different, though the motivation (gaining a resource for syntactic annotation of a treebank) is similar. At the time PCEDT begun to be annotated on the tectogrammatical layer, a reliable version of PDT-VALLEX had already been finished, fully checked and published. A similar resource was needed to be available for the English annotation in a reasonable time, therefore, instead of creating Engvallex manually on the basis of pure data, we decided to adapt an already existing resource of English verbs valency characteristics, the PropBank [7].

The PropBank lexicon has been adapted to the Functional Generative Description scheme in several ways. First, all slots have been renamed using functors, second, the non-obligatory (according to FGD) free modifiers have been deleted and optional elements marked. Third, frames corresponding to the same verb sense have been merged. Fourth, the lexicon has been refined in the process of treebank annotation by addition of other frames, whole verb lemmas, and also, the PropBank adapted frames were corrected manually with respect to the language data available in the PCEDT corpus.

Engvallex only contains verbs so far. Currently, it contains 6,213 valency frames for 3,823 verbs. As in case of PDT-VALLEX, it is being constantly expanded and refined in the course of PCEDT annotation.

In the process of PropBank adaptation to FGD theory, some core differences of the two valency theories came alight that are supposed to affect also the intended linking process. For example, it appeared that the PropBank argument range is much broader than the one usually admitted by FGD approach. It results from stricter criteria for "argumentness" used in FGD (the famous *dialogue test* [8], disallowance of non-obligatory free modifiers in verb frames etc.). Such frame arguments were usually deleted during the adaptation process, though we kept them in sporadic cases where the resulting frame would have been otherwise divested of adjuncts too typical. The deleted frame arguments were typically benefactives, non-obligatory attributes of arguments, commitatives or locatives.

2.2 Annotating Types While Seeing Tokens

Our aim of aligning two existing valency lexicons is considerably easier than the lexicographic process carried out at the time the individual lexicons were built. Still, we face the problem of formally describing verb (or frame) *types* while observable items are verb *tokens*.

Traditionally, lexicographers collected corpus evidence, organized the tokens into groups of examples with similar syntactic and/or semantic properties and derived a single description of the given type. Little or no effort was

spent in checking whether the description well matches the "training" tokens or even an independent set of "test" tokens. In our opinion, this is the root of troubles faced when trying to apply a traditional lexicon in NLP applications. Fortunately, recent projects (e.g. FrameNet [9]) try hard to provide enough real-world example sentences coupled with lexicon entries.

We design our annotation process to carefully separate the annotation attributed to types (i.e. lexicon entries) from the annotation attributed to tokens (i.e. verb occurrences in a treebank), but we require the annotator to see and provide both annotations simultaneously. In order to simplify the annotation process, we implement automatic procedures to project type annotation to an observed token and vice versa.

With the automatic procedures at hand, the annotator usually constructs the type annotation at the first token of the given type. Subsequent tokens of the same type will automatically reveal how the type annotation projects in the particular case. We can easily highlight any conflicts between the projection and the token annotation.

We feel that this design of the lexicographic process has several advantages:

- The annotation of types is presented not in an abstract form but rather naturally projected on a given example, i.e. verb frames are displayed in example sentences, but not available in the form of written lists of slots.
- While building the lexicon, we get an annotated corpus as a by-product, including explicit links between the two resources.
- Automatic highlighting of conflicts between the annotation of tokens and the projected annotation of types serves as quality assurance for all three components in question: the lexicon, the corpus of lexicon examples and the automatic procedures that apply lexicon entries to (unseen) sentences.

We believe that this explicit type-token link is vital for future applicability of the constructed resource. For instance, if a lexicon entry is doubted by a human user, he or she can use the treebank examples to understand better the generalization captured in the lexicon. For NLP tools, the set of annotated examples can serve as a test set or as a training set for machine-learning algorithms.

2.3 Description of Annotation Environment

2.3.1 Tools Used

The annotation tool we developed builds on two large software projects: TectoMT [11] framework for various NLP tasks (including MT) and tree editor TrEd[2].

[2] `http://ufal.mff.cuni.cz/~pajas/tred`

TectoMT is a modular programming environment aimed at linguistically rich processing of text. The two features of TectoMT we exploit are: automatic alignment of Czech and English t-nodes [6] and TectoMT native file format TMT, capable of storing dependency analyses of two languages at all three layers of linguistic description. As our source examples of verb usage are already manually annotated at the t-layer for both English and Czech, we do not need the automatic analyses implemented in TectoMT, but this option would be clearly very useful for a potential future annotation of a different text type.[3]

TrEd is a highly customizable and extensible editor of dependency trees. TrEd was used for manual annotation of all the above-mentioned Prague treebanks and an extension of TrEd allows to edit TMT files, i.e. to work with several trees of a given sentence pair at once.

2.3.2 Design of User Environment

The design of the user environment for the annotation follows the principles outlined in Section 2.2. The user is presented with a pair of t-trees and aligned verbs. We use several types of arrows to indicate *token annotation*, i.e. links between dependents of the verbs in this particular sentence, and *type annotation*, i.e. links indicating the correspondence between the slots of the two frames in question, see Figure 1.

The user environment facilitates the following annotation actions:

Token annotation: Correction of automatic node alignment.

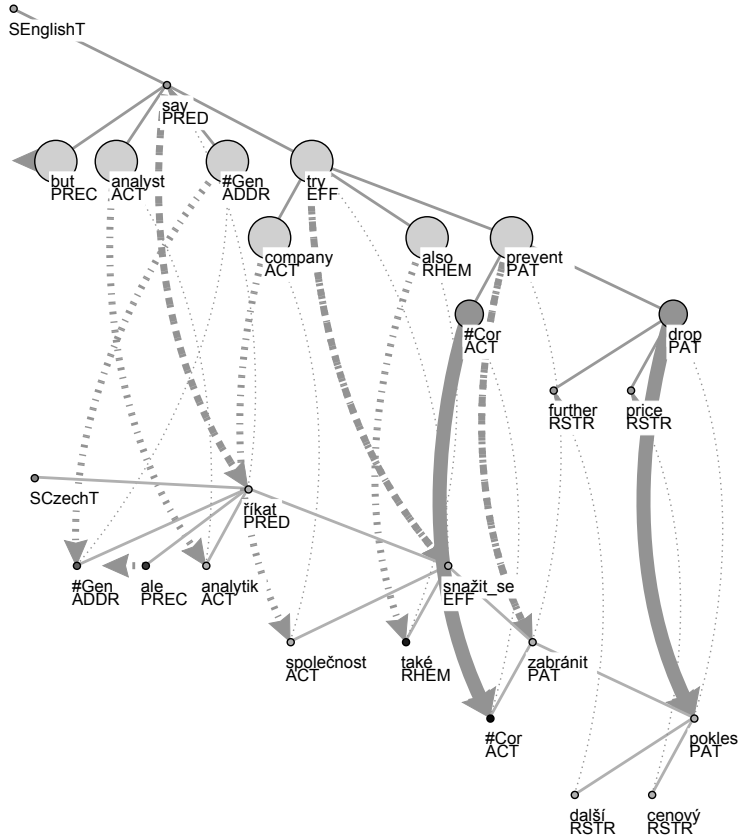
The pair of t-trees has already been automatically node-aligned, so most English t-nodes have a Czech t-node counterpart (one, at most). We use the node alignment to find both: pairs of matching verbs as well as pairs of verbs' immediate dependents.

If the automatic alignment does not provide a link, or there is an error in the alignment, the user can provide manual node alignment links simply by dragging an English node onto a Czech node. If a node is aligned both manually and automatically, the manual alignment takes precedence and the automatic alignment is not displayed at all.

Type annotation: Collection of node alignment to Engvallex.

When the manual or automatic node alignment correctly represents the alignment of verb dependents, the annotator uses a single keystroke command to collect and store it as the slot alignment in the correspoding Engvallex entry.

[3] The only step performed in our source manual trees with no automatic counterpart in TectoMT is the selection of frame ID of a given verb occurrence, i.e. the verb-frame disambiguation task. However, the task itself has already been explored for Czech [10].



But analysts say the company is also trying to prevent further price drops.
Ale analytici říkají, že společnost se také snaží zabránit dalším cenovým poklesům.

Figure 1: Sample pair of sentences with manual and automatic alignment of verb dependents and projected alignment of frame slots (thick arrows). In practice, the arrows are color-coded.

A feedback to the user visually combines both type and token annotation. For every pair of aligned verbs (indicated by dashed green arrows) we highlight immediate dependents and their alignment:

- **Manual and automatic node alignments** are displayed as dotted red and blue arrows. (The complete automatic node alignment is indicated by very thin dotted lines.)
- If the frame entry contains a **slot alignment** specification, the slot alignment is projected on the pair of verbs and indicated by thick green arrows.
- All English verb dependents with a missing or mismatching slot alignment are displayed as large (yellow) nodes.

- When the type and token alignment matches (as illustrated for the verb *prevent–zabránit* in Figure 1), the nodes are smaller (and green).

In order to simplify the access to individual verb examples, we use TrEd "filelists". A filelist contains a list of corpus positions, i.e. filenames and node IDs. Filelists allow to browse the parallel treebank data in various ways. For the time being, we prepared a filelist for each pair <English verb, its Czech translation> and we organize the filelists based on the number of corpus examples. With a different filelist, the same corpus could be browsed from the most complex verb frames or from frames with most conflicts in the (automatic) token and type annotation.

2.3.3 Implementation Details

Both parts of the parallel treebank we build upon use their respective file formats to store Czech and English t-layers. We identify sections of data annotated in both PEDT and PCEDT_CZ and merge them into TMT files. In the subsequent annotation, we use only the combined TMT files and never modify the original independent treebank files.[4]

Engvallex and PDT-VALLEX are stored in XML files with a similar but not exactly identical structure. Both lexicons are still under development. In order to avoid conflicts, we detach from their development and preserve some fixed versions of the lexicons for our purposes.[5]

Technically, manual node alignments are stored directly in the TMT files. The slot alignment information should belong to both valency dictionaries, but for the time being we prefer to store it in Engvallex only.

We extend the representation of Engvallex to include a set of frame counterparts for each frame of an English verb. Each of the frame counterparts specifies the ID of the target frame in PDT-VALLEX accompanied by a mapping of slots. As slots in both valency dictionaries are uniquely identified by their functors, the mapping simply consists of tuples <Czech slot functor, English slot functor>. The format currently permits also 1-0 mapping (no counterpart slot in the Czech frame) and we will soon also store the list of unaligned English slots, i.e. 0-1 mapping, to differentiate between no mapping and still unspecified mapping in the representation.

[4] We preserve sentence and node IDs (and do not modify the t-layer annotation apart from a few corrections in functor values), so all our annotations can be transfered back to the treebanks, if desired.

[5] Frame IDs are usually preserved, so later our alignment should be easily transferable to fresh versions of the lexicons.

3 Preliminary Observations

The most frequent problem with the annotation environment is the lack of support for coordinated verb dependents or, even worse, coordinated verbs. While this limitation does not completely block the annotation process (all problematic examples can be simply skipped), it requires the annotator to walk the filelist searching for a suitable example. The solution for examples with coordinated verb dependents is rather simple: a conjunction node should serve as a representative for both coordinated members and it should be understood as bearing the functor common to the dependents instead of a technical functor CONJ.

Another issue is caused by ellipsis: many examples do not contain dependents to fill all the slots of a frame. Currently, the annotator has to wait for an example explicitly mentioning a dependent of a given functor to be able to annotate the slot link. We plan to add artificial nodes for all slots not expressed at the t-layer so that the annotator would be able to align them.

The last issue is less important in our case but should be taken as a caveat for similar annotation enterprises. In our case, each example is a pair of t-trees, occupying a large portion of screen and requiring a short but observable time to render. If the lexicographer should be provided with many examples at once, e.g. for the purposes of comparison, the t-layer would be a too rich representation.

4 Case Study: Verbs of Commercial Transaction

Verbs of Commercial Transaction are due to the character of the corpus (WSJ texts, economic focus) one of the most common verb classes in PCEDT. What is more, they are characteristic by a great number of hypothetic arguments, which often fail to be realized in a surface syntactic structure. As such, they represent a verb class highly attractive as verbal valency investigation issue.

Due to the lack of space we will limit our observations to one member of the class only, the verb *sell*.

4.1 Sell

Sell is a typical representative of the verb class in question. The representations of its valency frames in the individual valency lexicons are in Table 1.

PropBank provides a single verb meaning with a single set of participants exemplified by several surface argument layouts. Engvallex, on the other hand, provides three different frames, though representing the same meaning of the verb (which is quite an unusual situation in FGD framework). Those three frames are exemplified further in (1)–(3) respectively.

Propbank entry:	**Engvallex entries:**			**PDT-VALLEX entry for** *prodat/prodávat***:**
Arg0: Seller	ACT	ACT		ACT
Arg1: Thing Sold	PAT	PAT	ACT	PAT
Arg2: Buyer	?ADDR			ADDR
Arg3: Price Paid	(EXT)	(EXT)	(EXT)	(EXT)
Arg4: Benefactive	—			—

Table 1: Comparison of valency entries for *sell–prodat/prodávat*. The question mark "?" sign indicates that the frame element is optional only and the brackets "(. . .)" around the functor label represent the information that the element is considered a free modifier, and as such it is not included in the frame.

(1) a. At last count, Candela had sold $4 million of its medical devices in Japan.

b. Celkem prodala Candela v Japonsku své lékařské přístroje za 4 miliony dolarů.

Example (1) is an instance of the most common positive sentence constellation of the arguments. It can be seen that the three lexicons do not substantially differ in how they capture the valency properties of such uses of the verb.[6]

(2) a. A more recent novel , "Norwegian Wood" (every Japanese under 40 seems to be fluent in Beatles lyrics), has sold more than four million copies since Kodansha published it in 1987.

b. Novějšího románu "Norské dřevo" (snad každý Japonec pod 40 zná texty Beatles) se prodalo od jeho vydání v nakladatelství Kodansha roku 1987 více než čtyři miliony výtisků.

c. Four million copies of a more recent novel, "Norwegian Wood" . . . , have been sold since Kadansha published it in 1987.

Example (2) is an instance of a nonstandard shift of arguments (as PropBank interprets it). The sold item moves into the position of seller and its place is taken by an expression of transaction proportion. The Czech translation (2b) uses correctly a type of passive voice, which does not require an additional valency frame. Nevertheless, if we were to consider (2b) an instance of the frame constellation used in (1), we would have to consider an underlying structure such as (2c), which would have exactly the same tree representation as (2b).

Nevertheless, there is no imaginable way of justifying a transformation of this kind, for there is (almost for sure) no leading case for such a shift between a deep representation and a surface structure in the whole treebank. For this

[6] We decided to keep ADDRessee optional due to the dialogue test [8] results and due to the fact that in vast majority of examples in PCEDT it is semantically suppressed and not realized on the surface.

reason, we decided to keep a separate frame (despite the fact that the same verb sense is employed). Keeping a separate frame in this case also minimizes the risk of linkage conflict in the bilingual valency lexicon.

(3) a. At Christie's, a folio of 21 prints from Alfred Stieglitz's "Equivalents" series sold for $396,000, a single-lot record.

b. Na Christie's bylo 21 fotografií ve foliovém formátu z řady "Ekvivalenty" od Alfreda Stieglitze prodáno za 396 000 dolarů, rekordní částku za jedinou položku.

For similar reasons, we decided to keep a separate frame for (3a), though the construction evokes alternations of the type which is considered a mere derivation of the basic frame in the FGD application to English data.

Another issue connected to the verb *sell* is the issue of the element named EXT and standing for the *price* of the goods in the commercial transaction. FGD considers EXT a free modifier, not allowing it any role in the valency frame. Though with other verbs of commercial transaction, such as *pay*, the *price* argument has its place in the frame (being considered obligatory), here it falls out. This is a disadvantageous property of the linguistic framework we use and it, of course, has limiting consequences for the task of our interlinking the frame elements.

5 Conclusion and Future Work

Despite the fact that the process of our creating a bilingual valency lexicon is still at its beginnings, we have, thanks to it, already gained some important insight into the theoretical issues of crosslinguistic comparison of verbal valencies. By accessing the linguistic core of verbal valency via its treebank manifestations we are in hope of gaining a valuable, reliable and useful resource of linguistic information. The methodological solution we have chosen turns out as easy, user-friendly and effective in practical use.

We expect to continue annotation works and complete the linking process in approximately a year horizon. Further, we plan to utilize the bilingual valency lexicon in a forthcoming linguistic research in verbal valency and its impact on the verb semantic classes, and also, we would like to use the lexicon in future MT experiments.

6 Acknowledgment

The work on this project was supported by the following grants: GAUK 19008/2008, GAUK 52408/2008, MSM 0021620838, and FP7-ICT-2007-3-231720 (EuroMatrix Plus). We are grateful to Josef Toman for significant help with the implementation of our extension to TrEd and to Miroslav Týnovský for preliminary merging of PCEDT and PDT files into the TMT format.

References

[1] Jan Cuřín, Martin Čmejrek, Jiří Havelka, Jan Hajič, Vladislav Kuboň, and Zdeněk Žabokrtský. Prague Czech-English Dependency Treebank, Version 1.0. Linguistics Data Consortium, LDC2004T25, 2004.

[2] Jan Hajič and Zdeňka Urešová. Linguistic annotation: from links to cross-layer lexicons. In *Proc. of TLT 2*, pages 69–80, 2003.

[3] Jan Hajič. Complex Corpus Annotation: The Prague Dependency Treebank. In *Insight into the Slovak and Czech Corpus Linguistics*, pages 54–73, 2006.

[4] Erhard W. Hinrichs and Heike Telljohann. Constructing a Valence Lexicon for a Treebank of German. In *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories*, pages 41–52, 2009.

[5] Markéta Lopatková, Zdeněk Žabokrtský, and Václava Kettnerová. *Valenční slovník českých sloves*. Nakladatelství Karolinum, Praha, 2008.

[6] David Mareček, Zdeněk Žabokrtský, and Václav Novák. Automatic Alignment of Czech and English Deep Syntactic Dependency Trees. In *Proc. of EAMT 2008*, Hamburg, Germany, 2008.

[7] Martha Palmer, Daniel Gildea, and Paul Kingsbury. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106, 2005.

[8] Jarmila Panevová. *Formy a funkce ve stavbě české věty [Forms and functions in the structure of the Czech sentence]*. Academia, Prague, Czech Republic, 1980.

[9] Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. FrameNet II: Extended Theory and Practice. Technical report, ICSI, 2005. `http://framenet.icsi.berkeley.edu/book/book.pdf`.

[10] Jiří Semecký. *Verb Valency Frames Disambiguation*. PhD thesis, Charles University, Prague, 2007.

[11] Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. TectoMT: Highly Modular Hybrid MT System with Tectogrammatics Used as Transfer Layer. In *Proc. of ACL Workshop on Statistical Machine Translation*, pages 167–170, 2008.