Building a Bilingual ValLex Using Treebank Token Alignment: First Observations

Ondřej Bojar and Jana Šindlerová

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

1 Introduction

In this paper we explore the possibilities and limitations of a concept of building a bilingual valency lexicon based on the alignment of nodes in a parallel treebank. Our aim is to build an electronic Czech↔English Valency Lexicon by collecting equivalences from treebank data.

2 Building a Bilingual Valency Lexicon: Project Details

2.1 Source Data

Prague Czech-English Dependency Treebank (PCEDT, [1]) is a sentence-parallel manually annotated treebank in development. The annotation includes also links to two valency lexicons, PDT-VALLEX for Czech and Engvallex for English. We utilize the annotation to add explicit links between the lexicon entries, thus raising the interlinking of verb tokens to a formally represented interlinking of verb types.

PDT-VALLEX [2] has been developed as a resource for valency annotation in a large-scale syntactically annotated corpus, the Prague Dependency Treebank [3]. In PDT, verbal valency is embedded in the so-called tectogrammatical layer (deep syntactic dependency relations), therefore PDT-VALLEX contains information about 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, each slot is marked as obligatory or optional.

By now, PDT-VALLEX contains 10593 valency frames for 6667 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 further annotations, including the annotation of the Czech side of PCEDT.

PDT-VALLEX has been developed in close relation to the annotation works on PDT. The frames have been created during the process of syntactic annotation, with great respect to the authentic linguistic material available. The theory of tectogrammatical representation, though aspiring to a high degree of universality, has been primarily developed on Czech language data. Thus, an attempt of creating a parallel treebank and parallel valency lexicon is a challenge to the whole theory.

Engvallex was created by a (largely manual) adaptation of an already existing resource of English verbs valency characteristics, the PropBank [5], to PDT labeling standards. First, all slots have been renamed using functors, second, the non-obligatory 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 English part of PCEDT.

Engvallex only contains verbs so far. Currently, it contains 6213 valency frames for 3823 verbs. As in case of PDT-VALLEX, it is being constantly expanded and refined in the course of further annotations.

2.2 Annotation Goal

To summarize the whole structure of manual data available in PCEDT, there is a corpus of parallel sentences, each of which is annotated at the tectogrammatical layer and each of which links verb occurrences to entries in PDT-VALLEX and Engvallex, respectively. There is no manual alignment between the two trees but an automatic one can be created e.g. using the tool by [4]. What we add are manual links between frame entries and slots of the frames.

The information about translation frames and functor (slot) equivalences is stored right within the frame entry, as a list of valency slot mappings. The mappings simply consist of tuples <Czech slot functor, English slot functor>. The format permits also 1-0 mapping (no counterpart slot in the target frame) and 0-1 mapping (unspecified mapping in the source representation). For the final version of the lexicon we plan to include mapping information into both PDT-VALLEX and Engvallex part, but in practice we start in English-Czech direction, storing the information in Engvallex only.

2.3 Progress of the Project

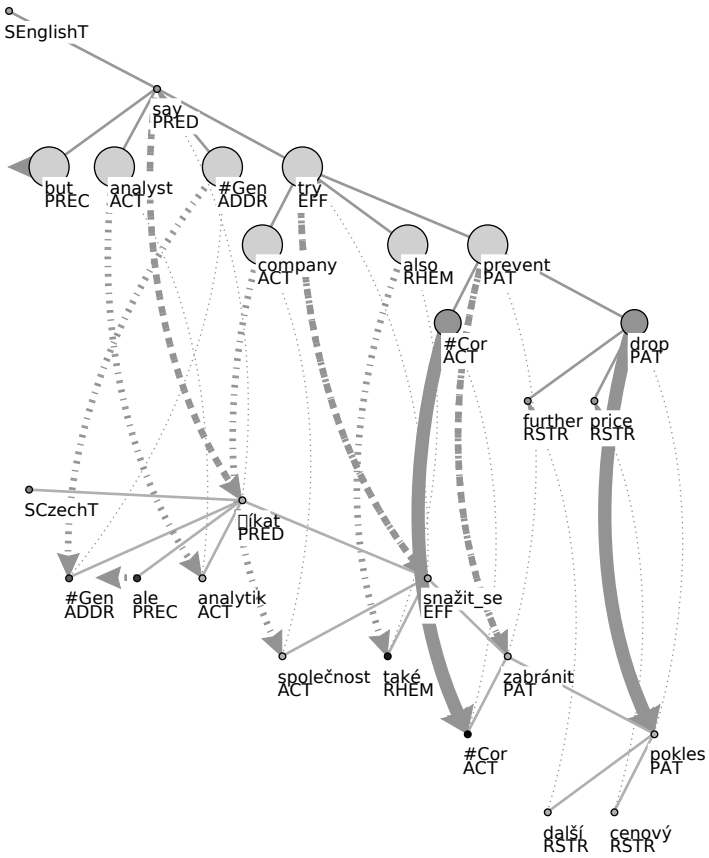
The project is divided into four phases, two of which, the preparation of source valency lexicons and preparation of the annotation interface, have already been completed.

The annotation interface is built on the tree editor TrEd[1] and the TectoMT[2] platform [6]. TrEd is being used for the annotation of both source treebanks while TectoMT adds a unified file format capable of storing trees in two languages in the same file and also tools for automatic processing of the data, including the alignment of the trees.

Figure 1 illustrates the core of our annotation user interface. The annotator is provided with both Czech and English tectogrammatical trees with automatic node alignment (very thin lines). The automatic node alignment is used to suggest alignment between verb tokens (dashed lines) and verb dependents (dotted lines). These suggestions can be manually corrected (we use colors to indicate which links are manual and which are automatic). Once the alignment of the dependents is finished, the annotator uses a single keystroke to "collect" the token alignment and store it as alignment of verb types and their slots in the dictionary. The alignment of slots in the dictionary is then projected back onto the sentence (very thick arrows) and previously unseen sentences as well to allow for a quick visual confirmation and validity of the alignment for other instances.

The third phase, links collection, has been started in September 2009 and is expected to finish in June 2010. Due to the fact that we work with corpus data already annotated for syntactic relations including verbal valency attribution, we decided to keep only one annotator. Her task is to go through the verb occurrences in a treebank, collect a typical representant of a frame mapping, and control and decide potential conflicting cases. Once collected, the frame mapping is automatically

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



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.

applied to all its other potential representants. The annotator is asked not to change the tree structure, but she is allowed to change frame attribution if considered inappropriate.

The fourth phase will include control and amendment works, adaptation of user interface for external users, and further extraction and exploration of linguistically important and interesting issues.

3 Issues Encountered during the Annotation Process

3.1 Different Set of Slots in Frames

We notice three cases of asymmetry in the set of slots of equivalent frames. First, the frames include the same number of slots but different labeling, i.e. there is a difference in linguistic structuring of the situation described by the verb. Such cases in fact justify the need for a bilingual valency lexicon in MT applications with a deep-syntactic transfer. Second, one of the lexicons includes an obligatory slot for a dependent while the other does not (the dependent is considered a non-obligatory free modifier). Our annotation process thus has to decide whether to include links if only one side of

the link is a valid slot in a frame. Third, one of the lexicons includes an obligatory actant slot while the other includes only a facultative actant slot. These cases are solved in the annotation process by allowing 1-0 or 0-1 mapping and inserting "phantom" slots (slots for non-expressed facultative actants of the frame) into the tree representation.

3.2 Conflicting Mappings

The annotation process is designed to collect frame-to-frame relations. It is believed that there exists a unique functor-to-functor mapping within this relation (coming from the assumption that each frame describes a verbal situation generally and the slots the individual participants of the situation take do not differ in different uses of the verb frame). Therefore, it is possible to store a list of target frames for each source frame, but in each of these relations only a single functor mapping is available.

Nevertheless, it appeared during the annotation process that certain syntactic constructions behave contra this assumption, i.e. if the construction is applied to the verb frame use in either source or target utterance, whereas the translation counterpart uses a different syntactic configuration, the lexical alignment results in different slot alignment than desired.

3.2.1 Unspecified Agent: Said

An example of such a construction is a typical construction with unspecified agent, shown in (1).

(1) a. The documents also said that although the 64-year-old Mr. Cray has been working on the project for more than six years, the Cray-3 machine is at least another year away from a fully operational prototype. *(PCEDT English sentence)*

b. V dokumentech se tak řeklo, že ačkoliv 64-letý pan Cray pracuje na projektu více než šest let, je počítač Cray-3 nejméně další rok vzdálen od plně funkčního prototypu. *(PCEDT Czech sentence)*

c. It was said in the documents that although the 64-year-old Mr. Cray has been working on the project for more than six years, the Cray-3 machine is at least another year away from a fully operational prototype. *(Strict translation of the Czech sentence in b.)*

English sentence uses *documents* in actor position to the verb *say*. On the contrary, Czech sentence uses passive voice with actor position not overtly expressed (and unspecified), and *documents* are constructed as locatives. (Note that a Czech sentence with *documents* in an overt actor position would hardly sound natural.)

Such cases of conflicting functor-mappings are of great importance to us. If we only concentrated on mapping asymmetries in the lexicon, we would lose the part of the story that lies in corpus data. This is the grate "pro" of the approach we chose.

4 Conclusion

We describe our ongoing efforts in aligning two valency lexicons on the basis of a parallel treebank. The projects serves not only the purpose of creating the resource but also the purpose of a compatibility check between the two lexicons and theory validation. We notice and document some issues of the lexicon alignment: different sets of slots included in the two lexicons and conflicting slot alignment for some verb occurrences.

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, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová, and Petr Pajas. PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9 of *Mathematical Modeling in Physics, Engineering and Cognitive Sciences*, pages 57–68. Växjö University Press, November 14–15, 2003 2003.

[3] Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, and Magda Ševčíková Razímová. Prague Dependency Treebank 2.0. LDC2006T01, ISBN: 1-58563-370-4, 2006.

[4] 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.

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

[6] Zdeněk Žabokrtský and Ondřej Bojar. TectoMT, Developer's Guide. Technical Report TR-2008-39, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague, December 2008.