Building a Bilingual ValLex Using Treebank Token Alignment: First Observations

Jana Šindlerová, Ondřej Bojar

Charles University in Prague
Institute of Formal and Applied Linguistics (ÚFAL)
Malostranské nám. 25, Praha 1, CZ-118 00, Czech Republic
E-mail: {bojar,sindlerova}@ufal.mff.cuni.cz

Abstract

In this paper we explore the potential 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 bilingual treebank data and storing them in two already existing electronic valency lexicons, PDT-VALLEX and Engvallex. For this task a special annotation interface has been built upon the TrEd editor, allowing quick and easy collecting of frame equivalences in either of the source lexicons. The issues questioning the annotation practice encountered during the first months of annotation include limitations of technical character, theory-dependent limitations and limitations concerning the achievable degree of quality of human annotation. The issues of special interest for both linguists and MT specialists involved in the project include linguistically motivated non-balance between the frame equivalents, either in number or in type of valency participants. The first phases of annotation so far attest the assumption that there is a unique correspondence between the functors of the translation-equivalent frames. Also, hardly any linguistically significant non-balance between the frames has been found, which is partly promising considering the linguistic theory used and partly caused by little stylistic variety of the annotated corpus texts.

1. Introduction

In this paper we present our current experience with building a bilingual valency lexicon by collecting equivalences in valency structure of corresponding verbs from a bilingual treebank, the Prague Czech-English Dependency Treebank. By creating a bilingual valency lexicon, we hope to gain a multifunctional resource useful in many areas. Our bilingual lexicon is designed to provide an easy access to analogies and differences between the valency structures of Czech and English verbs, which is important for both theoretical linguistics research and MT applications development.

2. Building a Bilingual Valency Lexicon: Project Details

2.1. Source Data

Prague Czech-English Dependency Treebank (PCEDT, (Cuřín et al., 2004)) is a sentence-parallel manually annotated treebank in development. The annotation includes 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 (Hajič et al., 2003) has been developed as a resource for valency annotation in a large-scale syntactically annotated corpus, the Prague Dependency Treebank (Hajič et al., 2006). In PDT, verbal valency is embedded in the so-called tectogrammatical layer (the layer of 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 part 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 (Palmer et al., 2005), 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 (the so-called PEDT).

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 (Mareček et al., 2008). 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

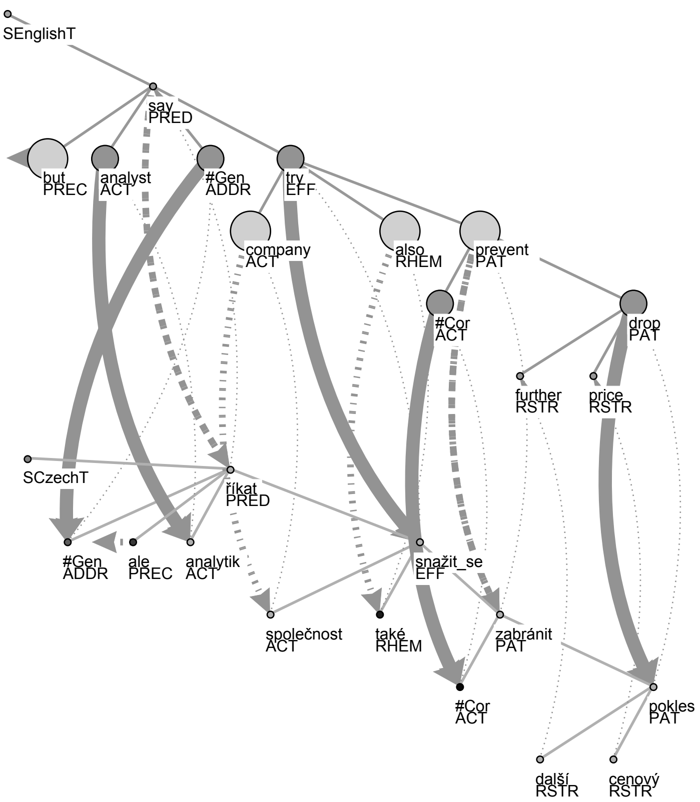
The project is divided into four phases, two of which, the preparation of source valency lexicons and preparation of the annotation interface, have already been completed.

The annotation interface is built on the tree editor TrEd[1] and the TectoMT[2] platform (Žabokrtský and Bojar, 2008). 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 the treebank, collect a typical representant of a frame mapping, and control and decide potential conflicting cases. Once collected, the frame mapping is automatically 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.



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.

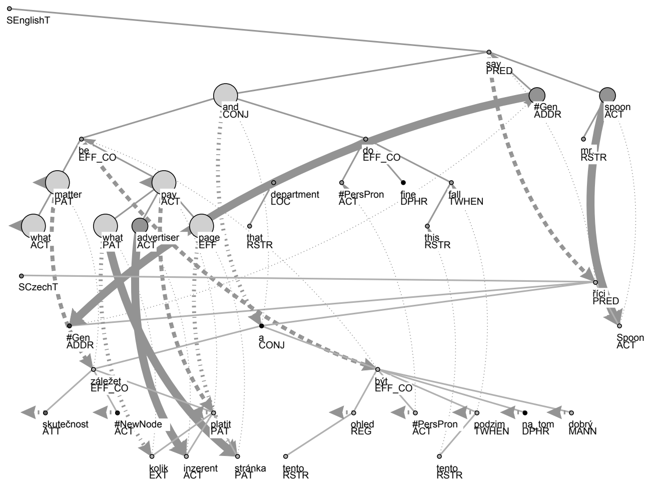
3. Issues Encountered during the Annotation Process

There are several issues we found problematic that are not related to a different linguistic structuring of a situation, but to the technical, theoretical or human factor limitations of the annotation practice.

3.1. Technical Limitations

Vast majority of corpus sentences includes either coordinated valency positions or coordination of predicates. Both these cases represent a minor technical problem considering slot identification. With coordinated slot positions, slots bearing the relevant functor are positioned one level lower than other, non-coordinated slots (there is an intervening technical level for coordination marking lexically represented by the conjunction), see Figure 2. In order to gain maximum available sentences for the annotator to work with, it was decided to copy the relevant functor from the first coordinated slot (since it might be the exceptional case of two different functors being coordinated on the slot

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



"What matters is what advertisers are paying per page, and in that department we are doing fine this fall," said Mr. Spoon. "Ve skutečnosti záleží na tom, kolik inzerenti platí za stránku, a v tomto ohledu jsme na tom tento podzim dobře," řekl pan Spoon.

Figure 2: Sentence illustrating the intervention of slot coordination. The EFF participant of the verb *say* is split, the relevant nodes are positoned under the technical conjunction node and therefore unrecognized by the procedure identifying slot sons of the predicate.

position) to the conjuction position. With coordination of predicates, it is usually the case that one of the relevant valency slots is positioned one level higher, i.e. as a *common dependent* it is displayed as a sister of the verbs (see Figure 3). For keeping the annotation trasparent it appears as the most convenient solution to copy the common node as a "phantom" node to its usual position of a daughter of the verb. More complicated cases (combination of both coordination types etc.) will probably be ignored and the corresponding sentences will be skipped in the annotation.

3.2. Theory-based Limitations

Some cases of Czech-English frame asymmetries are due to different linguistic backgrounds used for analysis of Czech and English sentences. An example of such a theoretical mismatch are the so-called "raising verbs". The accusative element following the object raising verb in a raising construction is traditionally being analysed as a deep subject of the infinitive part of the construction. Following the fact that PEDT t-layer was originally converted from Penn Treebank syntactic annotation, it has been decided to stick to this kind of analysis. On the other hand, Czech annotation growing from Formal Generative Description formalism does not recognize raising as an instance of an unexpected case realized on a deep subject, but identifies the accusative element as a semantic patient of the higher verb and the infinitive as its complement. Thus we gain one more valency slot in the Czech frame than in the English frame, see Figure 4.

Though the case clearly ends up in a slot mismatch, this type of asymmetry is not interesting from the point of view of collecting evidence for deep-layer asymmetries between Czech and English as different languages. It is rather important as a theory-checker. One of the inevitable effects of building a bilingual treebank is the challenge of general applicability of the theory. Here we have to reconsider the question whether raising is a valid construction for Czech in the same way as it is for English and what consequences it might have for the representation of the structure in the theoretical framework we use.

3.3. Human-factor Limitations

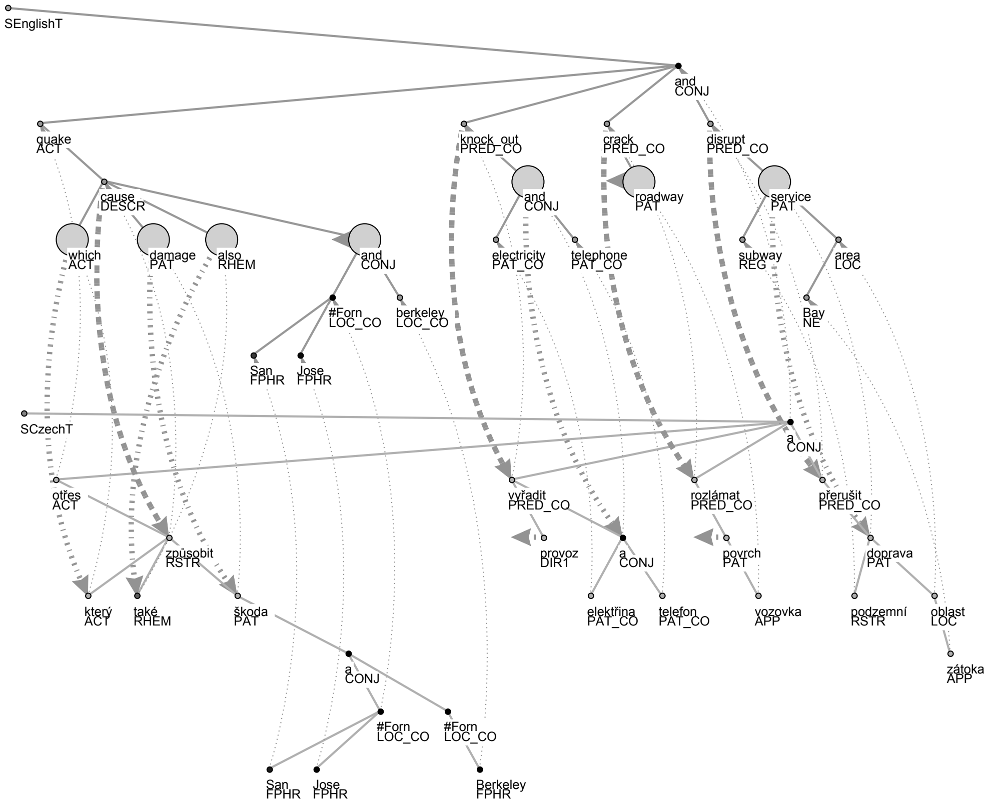
In the project we make use of a previous annotation of valency frames attribution done by PEDT annotators. In some cases it appears that the annotators are unable to keep consistent in attributing two frames of a single verb with mutually close meanings, i.e. they are unable to distinguish properly between the two frames. This surprisingly takes place both on the Czech and on the English part of annotation, so it cannot be blamed on translation difficulties but perhaps rather on too fine-grainedness of the lexicon entry. This is the case of the verb *expect*. The lexicon entry for Czech *očekávat* (expect) contains two frames, both containing ACT (actor) and PAT (patient) participants which are morphologically realized in a similar way in the example sentences. The two frames are only distinguished by the presence of facultative ORIG (source) participant in one of them (cf. *expect an advice from somebody* and *expect further declines *from somebody*) For some reason, the annotators were unable to keep this difference in mind when annotating, which is a signal for us to reconsider whether the splitted meaning is really justified. (Questions about the actual nature and place of ORIG's dependence and the possibility of semantic inclusion of ORIG in other components of the construction come to one's mind immediately.)

3.4. Different Set of Slots in Frames

We basically expect three cases of asymmetry in the set of slots of equivalent frames. First, the frames may 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. After first months of annotation we must admit not to have encountered a clear example of a mismatch of this type in PCEDT. Nevertheless, this can be easily explained. It is caused by the nature of texts used for PCEDT. We have previously suggested in (Šindlerová, 2010) that these mismatches appear quite frequently in a large parallel corpus of fiction texts. In largely economic and financially-aimed texts of Wall Street Journal we are limited by a restricted range of verb meanings present in the corpus, and naturally, even the translations tend to be more literal than in case of fiction.

Second, one of the lexicons may include 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 may include an obligatory actant slot while the other includes only a facultative actant slot. These cases are solved in the annotation process by allow-



The quake, which also caused damage in San Jose and Berkeley, knocked out electricity and telephones, cracked roadways and disrupted subway service in the Bay Area.
Otřesy, které také způsobily škody v San Jose a v Berkeley, vyřadily z provozu elektřinu a telefony, rozlámaly povrchy vozovek a přerušily podzemní dopravu v oblasti zátoky.

Figure 3: Sentence illustrating the intervention of predicate coordination. The ACT participant *quake* is common to the predicates *knock out*, *crack* and *disrupt*, therefore displayed as their sister and unrecognized by the procedure.

ing 1-0 or 0-1 mapping and inserting "phantom" slots (slots for non-expressed facultative actants of the frame) into the tree representation.

3.5. 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 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 for 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.

3.5.1. Unspecified Agent: Said

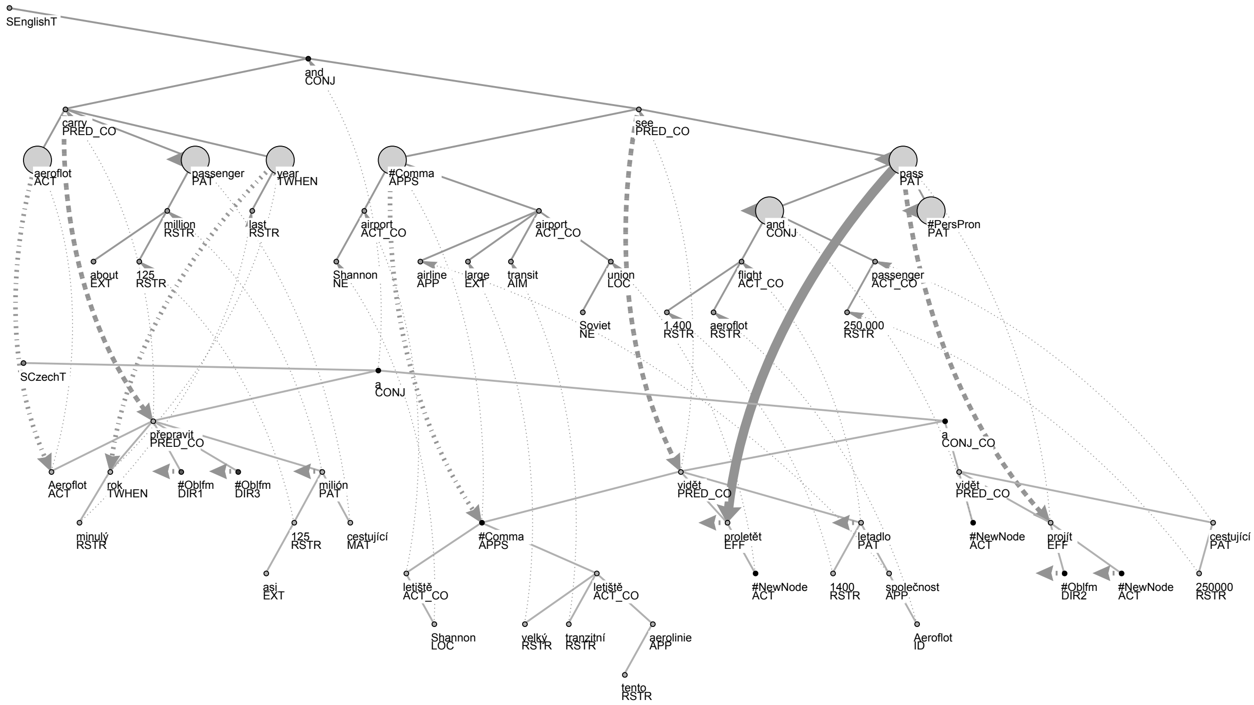
An example of such a construction is a typical construction with unspecified agent, shown in (1).

(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 locative. (Note that a Czech sentence with *documents* in an overt actor position would hardly sound natural.)



Aeroflot carried about 125 million passengers last year, and Shannon Airport, the airline's largest transit airport outside the Soviet Union, saw 1,400 Aeroflot flights and 250,000 passengers pass through.
Aeroflot minulý rok přepravil asi 125 milionů cestujících a letiště v Shannonu, největší tranzitní letiště těchto aerolinií, vidělo proletět 1400 letadel společnosti Aeroflot a projít 250 000 cestujících.

Figure 4: Sentence illustrating unbalanced mapping due to different syntactic analysing of raising verb predicate *see – vidět*. The ACT participant of the verb *pass* ("flight") is recognized as a PAT participant of the verb *vidět* in Czech and therefore has no direct slot mapping counterpart. ACT participants of *see – vidět* are unrecognized due to the intervention of apposition structure.

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 great "pro" of the token annotation approach we chose.

4. Conclusion

We describe our ongoing efforts in aligning two valency lexicons on the basis of a parallel treebank. The project serves not only the purpose of creating an important computational and linguistic resource but also the purpose of a compatibility check between the two source lexicons and theory validation. We notice and document some issues of lexicon alignment problems and asymmetries. Technical limitations of the chosen approach can be easily overcome by making changes to the procedure, human-factor and theory-dependent limitations require better training or theory adjustment. After finalizing the annotation phase we plan to investigate deeper into the question of non-balanced slot-mappings as a crosslinguistic phenomenon. Being aware of the limitations stemming from a specific corpus type we use, we hope to be able to use also some other Czech-English parallel data in the future and increase thus the reliability a relevance of the bilingual lexicon.

5. Acknowledgment

The work on this project was supported by the following grants: EuroMatrixPlus (FP7-ICT-2007-3-231720 of the EU and 7E09003 of the Czech Republic), GAUK 19008/2008, GAUK 52408/2008, and MSM 0021620838. 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.

6. References

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

Jan Hajič, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová, and Petr Pajas. 2003. 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.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, and Magda Ševčíková Razímová.

2006. Prague Dependency Treebank 2.0. LDC2006T01, ISBN: 1-58563-370-4.

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

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

Jana Šindlerová. 2010. (A)symetrie valenčních vlastností českých a anglických sloves pohybu. In *Sborník konference Intercorp*. In print.

Zdeněk Žabokrtský and Ondřej Bojar. 2008. 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.