An Application of the PDT-scheme to a Parallel Treebank

Jana Šindlerová, Lucie Mladová, Josef Toman, and Silvie Cinková
Charles University in Prague
Institute of Formal and Applied Linguistics

Abstract

Our paper comments on the divergences of the deep syntactic layers of Czech and English. We point out several phenomena potentially problematic for the syntactic alignment. We argue that at least some of the problems are caused by improper inferences from the deep syntactic layer of one language into the deep structures of the other one, and as such, they can be repaired on the theoretical level.

1 Motivation

In the recent past, syntax-based alignment models have gained the attention of many MT researchers. Such models make use of parallel treebanks, which are being built in different theoretical frameworks and based on different depths of sentence description.

The Prague Dependency Treebank (PDT) scheme is based on a multi-stratal annotation approach, the assumption being that the level of deep syntactic structures (the so-called "t-layer") might be beneficial for MT applications. This assumption is by no means new; it goes back to the well-known notion of interlingua (see (Hutchins, 1986) citing e.g. Vauquois or Mel'chuk), an "intermediate language" representing the underlying structures of different languages in a uniform manner, i.e. bearing as few differences as possible. In short, the core of this assumption is that machine translation systems profit from the simplicity and identity of the aligned syntactic structures. The less divergence there is between the structures, the more success we expect.

It should be noted that the deep-structure annotation in the PDT-scheme is basically meant to serve any language. Nevertheless it was developed first for the representation of Czech. The question then is whether there are any inferences from the deep structure of one language (in this case Czech) that would be inappropriate when applied to the deep structure of another one. In this paper, we would like to point out several aspects

De Smedt, K., Hajič, J. and Kübler, S. (Eds.)
Proceedings of the Sixth International Workshop
on Treebanks and Linguistic Theories (2007)

of using an annotation scheme developed for one particular language in the annotation of a multilingual treebank.

Our observations have been collected during the course of the work on the Prague Czech-English Dependency Treebank. As it is described below, this parallel treebank is still a sort of work-in-progress. No experiments with the human-annotated PCEDT t-layer which would confirm or refute the usefulness of the PDT-scheme for MT have been published yet. Therefore, we consider our findings to be a preliminary linguistic insight into the area of alignment. Such an investigation may lead to the successful repair of some of the problematic phenomena even before they can damage the outcome of alignment testing.

To illustrate the issues of particular interest, we will use reference data from the Prague Czech-English Dependency Treebank and the Prague Dependency Treebank.

2 Prague Czech-English Dependency Treebank

The Prague Czech-English Dependency Treebank (PCEDT) is a parallel, syntactically annotated corpus developed by Czech computational linguists since the 1990s mainly as a linguistic resource for the purposes of machine translation. The first version of PCEDT (PCEDT 1.0), publicly released by Linguistic Data Consortium in 2004 (see (Čmejrek et al., 2005)), contains in the first instance automatic, also partly manual, syntactic annotations of approximately 22 000 sentences from the Wall Street Journal part of the Penn Treebank (see (Marcus, Santorini, and Marcinkiewicz, 1994)) as well as annotations of their Czech translations. For the second release, the whole Wall Street Journal part of the Penn Treebank, 49 208 sentences (over 1.2 million words), has been translated to Czech by human translators, and we plan to include reference retranslations for Czech.

Originally used for the monolingual Czech corpus Prague Dependency Treebank (Mikulová et al., 2005) and adapted for English, the annotation scheme is based on Functional Generative Description (FGD; (Sgall, Hajičová, and Panevová, 1986)), the theoretical background for Prague treebanks, with its central idea of a dependency-based sentence structure handling the sentence structure as concentrated around the verb and its valency frame. According to the stratificational description of the language in FGD, there are principally 3 layers of annotation: morphological layer (m-layer) - full morphological annotation, analytical layer (a-layer) - superficial (surface) syntactic annotation, and tectogrammatical layer (t-layer) - deep or underlying syntactic annotation, it captures semantic relations, thus it is the level of linguistic meaning.

In the currently prepared PCEDT 2.0, however, no m-layer for the English texts is required, since the automatic conversion of the Penn Treebank

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

(PTB) phrasal trees to the PCEDT dependency trees (a-layer) already includes morphological information relevant for building the next layer. The a-layer then is conceptually the closest level to the syntactic annotation used in the PTB. Further, both the English and Czech parts of the corpus were automatically prepared for the subsequent manual annotation of the tectogrammatical layer by a) generating the t-layer for the English texts and b) automatic morphological tagging and subsequent automatic processing of the higher layers for the Czech texts.

The tectogrammatical tree structures capture:

- syntactic dependency and coordination; these are represented by the edges of the tree;
- semantic relations between the parent and the child node or between the coordinated items; these are described by the semantic labels called tectogrammatical functors;
- valency of the verbs; the corresponding valency frame is assigned to the verbal node from the valency lexicon EngValLex through the linking with the treebank;
- topic-focus articulation and
- links to the lower layers, including the links to PTB phrasal trees in the English sentences.

At present, the manual tectogrammatical annotations of both parts of the treebank have proceeded. There is a separate annotator team for each language. We try to keep the annotation schemes as similar to each other as possible. The annotation also includes the manual checking of the correct linking to the lower, analytical layer.

3 Divergences Between the Czech and English T-layers

The traditionally mentioned translational divergences (e.g. (Hearne et al., 2007), (Dorr, 1994), especially for PCEDT also (Bojar and Prokopová, 2006)) fall largely within the surface structure phenomena, i.e. phenomena including grammatical words deletion/addition (articles, prepositions, verbal auxiliaries etc.), part of speech transition (like nominalizations), or those affecting the surface word order. These types of divergences, which are often referred to as a potential alignment challenge, do not appear divergent on the t-layer, and thus will be then treated on the t-layer parsing level within the individual languages. On the other hand, there are other, more complex syntactic phenomena which challenge the uniformity of the t-layer representation.

3.1 The Underspecified Prenominal Position

English tends to express qualification of nouns on a neutral prenominal position, whereas inflective Czech uses a scale of cases and postnominal prepositional phrases.

(1a) Upjohn spokesman

(1b) mluvčí společnosti Upjohn
spokesman company.Gen Upjohn

(2a) fourth-quarter charge

(2b) poplatek za čtvrté čtvrtletí
charge for fourth.Acc quarter.Acc

This may seem to be just a matter of the surface representation, but in fact, such a linguistic property affects the deep structure analysis as well.

The prenominal position is semantically less transparent than the direct case marking (or PP); its semantic function is underspecified. Without the deep understanding of the context, the annotator of the English t-layer is almost unable to decide which particular one of the accessible meanings lies in the underlying structure of the prenominal element. For instance, in (2a) we can select from a variety of temporal meanings (*from when*, *for how long*, *how long*...), all of them being equally accessible, unless a certain specific context is given.

Moreover, the fact that Czech tends to express a greater number of semantic nuances in the postnominal, semantically explicit position, and consequently the fact that it does not use the prenominal position for expressing a complex circumstantial modification as often as English does, leads also to a slightly more reluctant attitude to assigning primarily adverbial functors to the premodifications of non-action/process nouns. Therefore, in a noun phrase like *marble hall*, the premodification will get a different functor in Czech treebank when translated with a postmodifying PP (*hala z mramoru: origin*) than when translated with a premodification (*mramorová hala: adnominal_modification*).

The possibilities of paraphrasing a single meaning by filling in either the prenominal or the postnominal position vary in both languages, depending also on the choice of lexical items. Nevertheless, it appears that in the PDT-scheme the behaviour of the prenominal position is being perceived through the particular syntactic properties of Czech, and as such, it does not seem fully suitable for languages with looser restrictions on prenominal positions, like English.

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

3.2 The Insertion/Deletion of Words

We have already emphasized the role of the context in the interpretation. There are several ways in which context is important for the annotation. First, it is crucial for deciding dependencies in unclear cases. Second, it plays a great role in the interpretation of ellipses.

English is in a way more elliptical than Czech. This means that a certain amount of semantic underspecification is acceptable in English, but not in Czech (cf. the occupation of prenominal position in 3.1). Translating sentences from English to Czech sometimes requires inserting additional nodes into the structure, either nodes with a lexical reference to the previous context, or additional nodes required directly for the sake of the precision of the translation. Consider (3) showing an English PCEDT sentence and its Czech counterpart created by a human translator.

(3a) The company earlier this year adopted a shareholder-rights plan to ward off unwanted suitors.

(3b) Společnost dříve v letošním roce přijala plán
company earlier in this_year year adopted plan
zabývající se právy akcionářů, aby odvrátila nebezpečí
dealing_with_Refl rights shareholders.Gen in_order_to ward_off danger
případných nechtěných zájemců o koupi podniku.
possible.Gen unwanted.Gen suitors.Gen of buying company

The (3b) translation is indeed a legitimate one; a plain word-to-word conversion would sound quite unnatural in Czech. The multi-word expression *shareholder-rights* can by no means be preposed to a noun in Czech, neither is the relation expressible by means of a simple case assignment or prepositional phrase. Therefore, an additional semantic specification of the relation *dealing with* (zabývající se) must be inserted into the structure. *Odvrátit*, which is one of the adequate translations of *ward off*, is closely bound to words like *nebezpečí*, *hrozba* (*danger*, *threat*) and may require them overtly in the structure when the negative meaning is not marked explicitly in the lexical semantics of the direct object; here the structural difference is already influenced by the lexical choice of the verb. And finally, leaving the Czech equivalent of the word *suitor* (*zájemce*) on its own would cause reader's confusion, because its semantics is too broad to refer unequivocally to the intended meaning.

Unfortunately, this problem does not seem to be fully solvable by a simple modification of the PDT-scheme or improvement of the guidelines. Depending on which language we take as the source and which one as the target, we will once have to deal with the need for insertion or deletion of nodes on the t-layer to reach the naturalness of the machine translation.

3.3 Participial clauses

English participial constructions represent an issue deeply problematic from a structural point of view, both regarding their translation to Czech and because of their formal properties. English participles can express a wide range of surface and semantic functions. They can stand for a standard adjective modification of a noun, of a restrictive or a descriptive kind:

(4a) the haunted house

(4b) the man(,) wearing red suit

Or they can express different circumstantial meanings (e.g. temporal (5) or causal (6):

(5a) Having removed his coat, Jack rushed to the river.

(5b) Hned jak si svlékl kabát, se Jack rozběhl k vodě.

(6a) Having been a gymnast, Lynn knew the importance of regular exercise.

(6b) Jelikož bývala kdysi gymnastkou, znala Lynn důležitost pravidelného cvičení.

Traditionally, these constructions are being translated (by human translators) by means of a Czech participle or transgressive, both expressing the so called *dual semantic relation*.

(7a) transgressive: Tady ... vesnický student naráží na ... lhostejnost velkoměsta, poznávaje, že poctivost a odpovědnost jsou mnohým spíše k smíchu.

(7b) Here ... the country student encounters the ... indolence of the city, discovering that for many people, honesty and responsibility are just a laughing stock.

(8a) participle: Alexandr Makedonský, dotázán na smrtelném loži, komu že odkáže svou říši, lakonicky odpověděl: Tomu nejschopnějšímu.

(8b) Alexander of Macedon, asked on his deathbed, to whom he will bequeth his empire, answered in a laconic manner: To the most capable one.

In such cases, the participle/transgressive receives the functor for *predicative complement* relation in the Czech annotation.

In some cases, the participial construction is translated to Czech with a dependent clause, then it receives a correspondent functor according to the circumstantial meaning it expresses. This of course means a possible

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

divergence in the formal description. There are two formally and structurally different representations in Czech for a single representation in English.

The clash is caused by the simple fact that the Czech guidelines for the annotation of a predicative complement relation are based more on the structural than the semantic relation of a node to its father and nominal brother. Contrary to the Czech annotation practice (which prioritizes the fact that both the transgressives and the participles overtly express morphological agreement with the related nouns in Czech), it seems unfruitful to ignore the possible circumstantial semantic impact of a participle in favor of applying the semantically neutral *predicative complement* functor.

Looking at the English t-layer against the background of a theory developed for a language expressing syntactic relations morphologically, we can see that we do not have a clear formal or semantic criterion to decide the t-layer status of the English participial construction in one or the other way. Therefore, we can say that in the case of English participial constructions, the current theory is deficient for the explicit description.

Moreover, there are other instances of participle use, where its semantic relation to the modified verb/noun is clearly neutralized. Such use simply informs us about a vague simultaneity of the actions (where the simultaneity is not otherwise relevant in the overall semantic information).

(9) From the fee, the local phone company and the long-distance carrier **extract** their costs to carry the call, **passing** the rest of the money to the originator...

Consequently, following the current guidelines, the corresponding dependency would have the tree structure shown in Figure 1.

Nevertheless, there are almost no means to capture this dependency in the Czech translation (the existing means yielding only hardly acceptable archaic transgressive structures). Instead, the simultaneity is captured by means of a simple coordination of the two predicates (Figure 2).

A structure clash involving dual dependency is a repetitive problem in PCEDT. What seems as a dual dependency structure in English, can appear translated as a coordination/apposition structure, or it can be constructed as a simple dependency in Czech, depending on its intended meaning.

(10a) Quotron has had problems calculating the industrial average.

(10b) Quotron měl problémy s výpočtem průmyslového indexu.
Quotron had problems with calculating industrial average.

In part of the dual dependency problems, the unification of the guidelines is the preferable solution; others are basically irreparable on the t-layer and thus must be treated with great care.

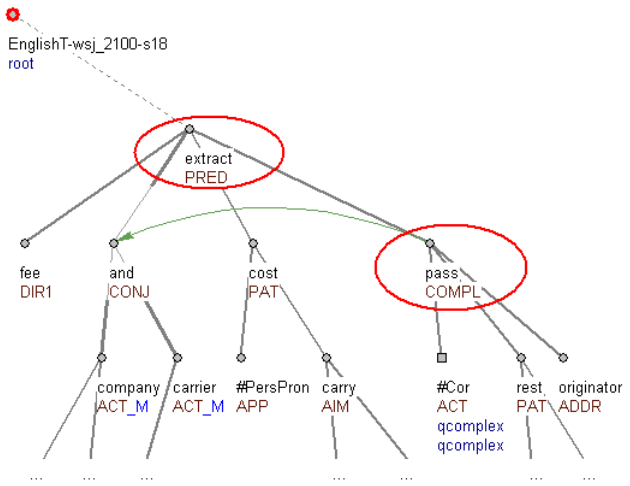

Figure 1: Part of the tree capturing the sentence in (9)

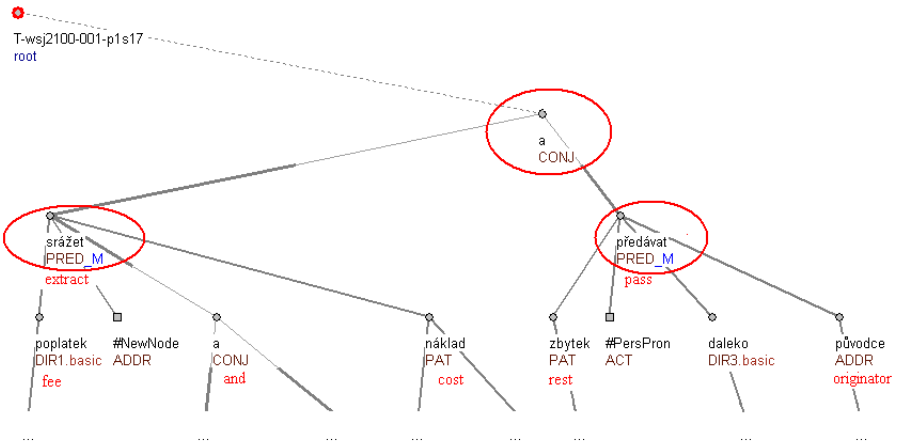

Figure 2: Part of the tree capturing the Czech translation of the sentence in (9): *Z tohoto poplatku sráží místní telefonní společnost a dálková telekomunikační společnost své náklady na přenos hovoru a zbytek peněz předávají dál původci...*

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

3.4 Valency

A considerable number of divergences appear in the area of valency. For both the English and the Czech treebank there is a valency lexicon available. *PDT-ValLex 2.0* is a result of several years of research and covers the valency properties of 2730 Czech verbs, or 6460 verb senses, and the most frequent verbal nouns and adjectives. *EngValLex* (Cinková, 2006) has been created by manual conversion of PropBank semantic roles (see (Palmer, Gildea, and Kingsbury, 2005)) to the traditional PDT-scheme. Currently, it covers 3687 English verbs, or 5877 verb senses/valency frames, but nouns and adjectives are still to be included. The two valency lexicons are to be interrelated to one another soon.

Basically, the differences in the description of verbal participants have more sources. Either the translated equivalent of the original verb requires a different number or different type of participants or the overall translation is improper, or there is primarily an inconsistency in the lexicon entry, which should be corrected at the theoretical level. Interrelating the two lexicons and the consequent revision of the problematic semantic roles would definitely help.

4 Interannotator Agreement as the Indicator of Divergences

In the course of the annotation of the Prague English Dependency Treebank, we have found a rather surprising thing. Most of the inconsistencies measured between the annotators of the English t-layer have grounds in its divergence from the standard Czech t-layer treatment. The greatest disagreement between the annotators appeared in annotating complex participial constructions and in the functor assignment to the prenominal position. However, this finding is not as surprising if we consider the nature of the annotation process. Only then we can see that it is not by chance, that the big problematic issues overlap.

4.1 Interannotator Agreement in PEDT

The interannotator disagreement is basically caused by a) human error (lack of concentration, distress, or simply lack of necessary knowledge or native-speaker intuition), or b) diversity of annotators' judgment (different evaluation of semantic phenomena or different interpretation of dependencies/constituents in underspecified cases). The latter has proved a useful source of information about the possible points of t-layers divergence.

The interannotator agreement in Prague English Dependency Treebank (PEDT, the English subpart of PCEDT) is being regularly measured in

several respects. The most telling numbers result from measuring the agreement in semantic labeling, i.e. the functor assignment, and the agreement in structure, i.e. the dependencies.

In the measuring process, the annotated sentences are perceived as a system of "attribute-value" pairs. For every two annotators we sum up the identical pairs and compute their percentual agreement. A pair is matching when both annotators assigned the same value to the particular attribute. First, the tree nodes are paired based mostly on their links to the analytical layer. Then the values contained in the paired nodes are compared.

The computation of the agreement percentage goes as follows: Let us have two numbers of the "attribute-value" pairs A1 and A2 (one for each annotator). Consider a subset consisting of the matching pairs of the size M. A pair is matching when both annotators assigned the same value to the particular attribute. The number M/A1 tells us how much of his data the first annotator annotated in the same way as the second annotator. The number M/A2 can be interpreted as the amount of the data produced by the second annotator, on which the first annotator agreed with him. In other words, it is almost a precision/recall evaluation, but instead of having the correct and the test data, we have something like two sets of test data, none of them fully correct. We can also compute F-measure as 2M/(A1+A2).

At present, the interannotator agreement in PEDT is measured around every other month and the results of each survey are used to improve the annotation guidelines, so that the most common errors could be successfully avoided in future annotations. The progress seems to be greater in terms of structure than in terms of functors, which is obvious from the nature of measured facts; the decisions about structure are more apt to be liable to a particular prescription rule than the semantic judgment.

The "English" annotators have already reached a comparable agreement level with the "Czech" group in terms of structure annotation, whereas the average accuracy of semantic evaluation is still about 3 percentage points lower. The fact that all the annotators proved an excellent level of language knowledge and the nature of the points of disagreement indicate that the true reason for the lower effectivity of the English annotation lies in the annotators' effort to apply the PDT-scheme guidelines even to those phenomena for which they do not appear suitable. From the linguistic point of view, the interannotator (dis)agreement is the major indicator of problematic issues to be solved either on the linguistic description level or within the organization of the annotation scheme.

5 Conclusion and Future Work

In this paper, we have discussed the issue of divergence of deep syntactic layers of different languages, using the example of the Czech-English par-

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

allel corpus annotation. We have argued that the PDT-scheme for parallel treebank annotation is generally applicable to the multilingual treebank annotation, but that there are specific issues, which cannot be generalized due to their language-specific character. Some of them are reparable within the specification of the annotation guidelines, some by the improvement of the general scheme; others remain to be dealt with within the alignment rules.

The current semantic annotation of PCEDT seems to stick quite a lot to the surface representations, which may be the cause of considerable differences in translation. Dealing with cross linguistic research, we should disengage from such an approach which makes us safe in terms of monolingual treebanks, but brings unnecesary divergences into the multilingual approach.

Our future task lies in elaborating on current annotation guidelines, interconnecting the two available valency lexicons into a consistent whole, and reconsidering the possibilities and limits of the unification of the Czech and English annotation guidelines in terms of the still unresolved structural questions.

6 Acknowledgements

This work was funded in part by the Companions project (www.companions-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-034434, and in part by the following grants: ME838, GA405/06/0589.

References

Bojar, Ondřej and Magdalena Prokopová. 2006. Czech-English Word Alignment. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1236–1239. ELRA.

Cinková, Silvie. 2006. From Propbank to Engvallex: Adapting the Propbank-Lexicon to the Valency Theory of the Functional Generative Description. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*. ELRA.

Dorr, Bonnie J. 1994. Machine Translation Divergences: A Formal Description and Proposed Solution. *Computational Linguistics*, 20(4):597–633.

Hearne, Mary, John Tinsley, Ventsislav Zhechev, and Andy Way. 2007. Capturing Translational Divergences with a Statistical Tree-to-Tree Aligner. In *Proceedings of TMI 2007*, pages 85–94.

Hutchins, John. 1986. *Machine Translation: past, present, future.* Ellis Horwood.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1994. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Mikulová, Marie, Allevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štepánek, Zdeňka Urešová, Kateřina Veselá, and Zdeněk Žabokrtský. 2005. Anotace Pražského závislostního korpusu na tektogramatické rovině: pokyny pro anotátory. Technical report, UFAL MFF UK, Prague, Czech Republic.

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

Sgall, Petr, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht:Reidel Publishing Company and Prague:Academia.

Čmejrek, Martin, Jan Cuřín, Jiří Havelka, Jan Hajič, and Vladislav Kuboň. 2005. Prague Czech-English Dependency Treebank. In *EAMT 2005 Conference Proceedings*, pages 73–78.

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476