Building a parallel bilingual syntactically annotated corpus

Jan Cuřín, Martin Čmejrek, Jiří Havelka
Center for Comp.Linguistics, Charles University,
Malostranské nám.25
Praha 1
{curin|cmejrek|havelka}@ckl.mff.cuni.cz

Vladislav Kuboň
UFAL, Charles University,
Malostranské nám.25
Praha 1
vk@ufal.mff.cuni.cz

Abstract

This paper describes a process of building a bilingual syntactically annotated corpus, the PCEDT (Prague Czech-English Dependency Treebank). The corpus is being created at Charles University, Prague, and the release of this corpus as Linguistic Data Consortium data collection is scheduled for the spring of 2004. The paper discusses important decisions made prior to the start of the project and gives an overview of all kinds of resources included in the PCEDT.

1 Introduction

Probably the most important trend in linguistics in the last decade is the massive use of large natural language corpora in many linguistic fields. The concept of collecting large amounts of written or spoken natural language data has become extremely important for several linguistic research fields.

The majority of large corpora used by linguists are monolingual, although there are several examples of bilingual corpora (e.g. Hansard corpus). The efforts of Czech computational linguists also concentrated in the past on creating large scale monolingual corpora as for example the Czech National Corpus (annotated on morphological level) or Prague Dependency Treebank (PDT). The PDT is annotated on three levels - morphological layer (lowest), analytic layer (middle) - superficial (surface) syntactic annotation, and tectogrammatical layer (highest) - level of linguistic meaning. Dependency trees, representing the sentence structure as concentrated around the verb and its valency, are used for the analytical and tectogrammatical layers of PDT

Only very few parallel bilingual corpora were available for the Czech-English language pair before the start of our project. A Reader's Digest corpus, which had been used in several smaller projects in the past, was one of the few exceptions. The situation with other language pairs was even worse, no reasonable resources were available at that moment.

2 The initial considerations

The experience gained in the process of building the above mentioned corpora indicated that collecting the data is much easier than annotating it. The deeper is the level of annotation, the longer and more expensive is the process of creating it. This fact is even more important for a parallel corpus, where every sentence is annotated twice, independently in each language.

Generally there are two possible strategies for building a parallel corpus. The first one is the parallel annotation of already existing parallel texts, the second one is the translation and annotation of already existing syntactically annotated corpus. The first approach has from our point of view two major drawbacks. In addition to the obvious problem of double annotation efforts there is also a problem of "relatedness" of parallel texts available. The up-to-now main parallel Czech-English resource, Reader's Digest corpus, contains extremely free translations, which have proved difficult in several machine-learning experiments (Al-Onaizan, et al., 1999).

The second approach, the human translation of an existing monolingual syntactically annotated corpus into the target language and its subsequent

syntactic annotation, not only seems to allow better control over the translation quality and reliability, but also reduces the necessary annotation efforts to annotation of a text in a single language. These initial considerations led to a decision to translate some already existing corpus.

2.1 Choosing the translation direction

When the choice of the general strategy had been made it remained to decide what kind of syntactic annotation to use and to choose a source language (and a source corpus, too). There were two natural candidates for the source text, namely the PennTreebank (for English) and the Prague Dependency Treebank (for Czech). The size of both corpora is approximately the same (more than 1 million words), both are syntactically annotated, both contain newspaper data (although PDT not exclusively, it also contains data from other sources). The choice of the PennTreebank as a source corpus was then pragmatically motivated - all the translators were native speakers of Czech and we have supposed that they should be able to provide higher quality of translation when translating into their native language.

2.2 Type of syntactic annotation

Another important issue was the choice of the annotation scheme. In fact this is the point where PennTreebank and PDT differ to a greatest extent. The syntactic annotations of the PennTreebank are relatively simple and transparent, they are based on constituent trees coded through a system of brackets accompanied by tags, while PDT is based on dependency trees. More precisely, Penn Treebank (version 3; LDC catalog no. LDC99T42, ISBN: 1-58563-163-9) comprises the surface syntactic structure, various types of null elements representing underlying positions for wh-movement, passive voice, infinitive constructions etc., and also predicate-argument structure markup.

For example, the sentence *UAL's decision to remain independent company sent share prices tumbling.* is annotated in PennTreebank as follows:

Example 1:

```
( (S
    (NP-SBJ
      (NP (NNP UAL) (POS 's) )
      (NN decision)
      (S
        (NP-SBJ (-NONE- *) )
        (VP (TO to)
          (VP (VB remain)
            (NP-PRD (DT an)
(JJ independent) (NN company)
)))))
    (VP (VBD sent)
      (S
        (NP-SBJ (NN share)
(NNS prices) )
        (VP (VBG tumbling)
)))
    (. .) ))
```

Due to the fact that Czech is a language with relatively high degree of word-order freedom its sentences relatively often contain some phenomena (discontinues constituents etc.) which cannot be coded by a simple bracketing system. The annotation scheme of the PDT is therefore more complicated and less transparent than that of the PennTreebank.

For example, the annotation of the first three words in the sentence *Smlouvy o debetu však KB poskytuje pouze omezenému počtu vybraných klientů*. [The KB provides debit agreements only to a limited number of selected clients.] is the following in PDT (similarly as in the previous case the annotation contains both morphological and syntactic tags):

Example 2:

```
<s id="ln95047:001-p5s4">
<f cap>Smlouvy<l>smlouva<t>
NNFP4-----A----<MDl src="a">
smlouva<MDt src="a"> NNFP1---
--A----<MDl src="b">smlouva
<MDt src="b">NNFP4-----A----
<A>Obj<r>1<g>6
<f>o<l>o-1<t> RR--6----------
<MDl src="a"> o-1<MDt
src="a">RR--6----------<MDl
src="b">o-1 <MDt src="b"> RR-
-6----------<A>AuxP<r>2<g>1
<f>debetu<l> debet<t>NNIS6---
--A----<MDl src="a">debet<MDt
```

```
src="a"> NNIS6-----A----<MDl
src="b"> debet<MDt
src="b">NNIS6-----A----
<A>Atr<r>3<g>2
```

3 The translation process

In order to achieve maximal quality of the translation we have divided the process of translation of PennTreebank data into several steps.

3.1 Filtering the tags from the treebank

Although the CD containing PennTreebank 3 contains not only fully morphologically and syntactically annotated data, but also various other levels of annotation (including text files with the data in the plain text format), we have decided to take as a basis for translation the files containing the fully annotated data (files having the *.mrg* extension in the PennTreebank 3) which will be included in the PCEDT. This decision was motivated by the endeavor to maintain a closest possible relationship between the annotated English and annotated Czech data. We have found several examples where the data in the plain text format were not included in the annotated part of PennTreebank, therefore we have decided to apply simple filters removing all tags and assigning each sentence its unique number consisting of the file name and the sequential number of a sentence in the file, starting from 0. The sentence from the Example 1 then looks like this:

<wsj_1102.mrg:3::>UAL's decision to remain independent company sent share prices tumbling.

3.2 Preparing the glossary

Due to the fact that the translation of Wall Street Journal texts from PennTreebank is extremely difficult, we have decided to provide the translators with a glossary of most frequent terms. The glossary should help to maintain the consistency of translation even when multiple translators do the job.

Originally we have considered to use a translation tool DèjaVue for the extraction of terms, but it turned out that it is not able to handle more than 1 million words of PennTreebank, so we were forced to create our own simple extraction tool in Perl. The extraction tool made a list of frequently co-occurring word sequences of various length. This list of course contained multiple random word sequences, so we have applied manual filtering in order to get a list of real terms. This list was then translated into Czech and distributed to all translators.

3.3 Translation and revisions

The translators (there were about 20 human translators involved in various stages of translation) were asked to translate each English sentence as a single Czech sentence and to avoid unnecessary stylistic changes of translated sentences. The translations are revised on two levels, linguistic and factual. It turned out that especially the factual revision is extremely difficult due to the nature of source texts. The Wall Street Journal articles are written in a style which is very far from the style of Czech newspaper articles. The text is full of economic slang, it is extremely compact, packed with references to institutions, people, and events which are not generally known outside of Wall Street circles. This is the main reason why the revisions are proceeding slower than we have expected – only about one fifth of translated texts is fully revised at the moment.

4 The annotation of the PCEDT

As mentioned above, the annotation scheme of PDT has been chosen for the annotation of the PCEDT. Apart from the linguistic reasons for this decision there was very strong technical one – due to the long experience with annotations of the PDT we had at our disposal several skilled annotators involved in annotations of the PDT.

Let us now briefly summarize basic facts about the annotation scheme of PDT (and PCEDT). In PDT, a single-rooted dependency tree is being built for every sentence as a result of the annotation both at the analytical (surface-syntactic) and tectogrammatical (deep syntactic) level. Every item (token) from the morphological layer becomes (exactly) one node in the analytical tree, and no nodes (except for the single "technical" root of the tree) are added. The order of nodes in the original

sentence is being preserved in an additional attribute, but non-projective constructions are allowed. Analytical functions, despite being kept in nodes, are in fact names of the dependency relations between a dependent (child) node and its governor (parent) node. Only a single (manually assigned) analytical annotation (dependency tree) is allowed per sentence. There are 24 analytical functions used, such as Sb (Subject), Obj (Object), Adv (Adverbial), Atr (Attribute in noun phrases) etc.

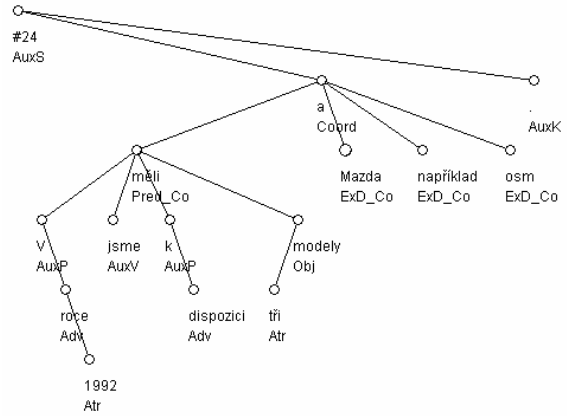

Figure 1: Analytical annotation of the sentence *"V roce 1992 jsme měli k dispozici tři modely a Mazda například osm.*" [In the year 1992 we had at our disposal three models and Mazda (had) for example eight (models).]

The tectogrammatical level is the most elaborated, complicated but also the most theoretically based layer of syntactico-semantic (or "deep syntactic") representation. The tectogrammatical layer annotation scheme is divided into four sublayers:

– dependencies and functional annotation,

– the topic/focus annotation including reordering according to the deep word order,

– coreference,

– the fully specified tectogrammatical annotation (including the necessary grammatical information).

As an additional data structure we use a syntactic lexicon, mainly capturing the notion of valency. The lexicon is not needed for the interpretation of the tectogrammatical representation itself, but it is helpful when working on the annotation since it defines when a particular node should be created that is missing on the surface. In other words, the notion of (valency-based) elipsis is defined by the dictionary.

The tectogrammatical layer goes beyond the surface structure of the sentence, replacing notions such as "subject" and "object" by notions like "actor", "patient", "addressee" etc. The representation itself still relies upon the language structure itself rather than on world knowledge. The nodes in the tectogrammatical tree are autosemantic words only. Dependencies between nodes represent the relations between the (autosemantic) words in a sentence, for the predicate as well as any other node in the sentence. The dependencies are labeled by functors, which describe the dependency relations. Every node of the tree is furthermore annotated by such a set of grammatical features that enables to fully capture the meaning of the sentence.

See (Hajič et al. 2001) and (Hajičová et al. 2000) for details on analytical and tectogrammatical annotations of PDT, respectively.

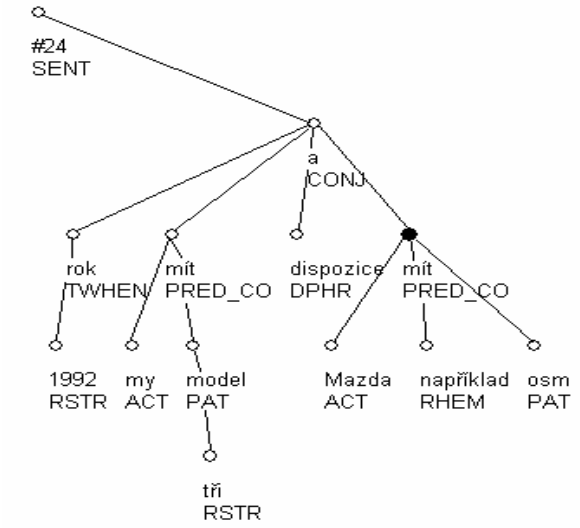

Figure 2.: Tectogrammatical tree for the sentence from the Fig.1

5 The annotation tools used in PCEDT

While the morphological annotation of the English part is simply taken over from the Penn Treebank, the analytical and tectogrammatical markups of the English part of the corpus are obtained by two in-

dependent procedures transforming the phrase trees into dependency ones. The Penn Treebank phrase trees had to be automatically transformed into dependency trees: only terminal nodes of the phrase tree are converted to nodes of the dependency tree and the dependency structure is built recursively so that the node representing a head constituent governs the nodes representing its sibling constituents. The transformation procedure is based on rules taking into account the information from the phrase tree (POS, functional markup, traces etc.), resulting into two different structures -- analytical and tectogrammatical ones.

The annotation of Czech at the morphological level is an unstructured classification of individual tokens (words and punctuation) of the utterance into morphological classes (morphological tags) and lemmas. The original word forms are preserved, too. In fact, every token has gotten its unique ID within the corpus for reference reasons. Sentence boundaries are preserved and/or corrected if found wrong (the errors in original texts contained in the Czech National Corpus have been preserved in the corpus). The number of tags actually appearing in the PDT is about 1100 out of 4257 theoretically possible. The data has been double annotated fully manually, the annotators selected a correct tag out of a set provided by a module of an automatic morphological analysis (cf. Hajič et al., 2001).

The Czech part is automatically annotated by the BH tagging tools (Hajič and Hladká, 1998) on the morphological level. The analytical representation is obtained by a statistical dependency parser (Charniak, 1999) and a C4.5 classifier assigning syntactic functions to nodes of the dependency tree. The tectogrammatical markup is a result of an automatic, rule-based transformation of analytical trees according to linguistic rules (Böhmová, 2001) and a C4.5 classifier assigning tectogrammatical functions (Žabokrtský et al. 2002).

6 Additional resources included in PCEDT

For both development testing and evaluation measured by BLEU metrics (Papiniemi et al., 2001), a test set of about 500 sentences was retranslated back from Czech into English by 4 different translator offices, two of them from the Czech Republic and two of them from the U.S. Example 3 illustrates the differences between retranslated sentences and an original sentence from the Penn Treebank.

Example 3:
Original:
Kaufman & Broad, a home building company, declined to identify the institutional investors.
Czech:
Kaufman & Broad, firma specializující se na bytovou výstavbu, odmítla institucionální investory jmenovat.
Reference 1:
Kaufman & Broad, a company specializing in housing development, refused to give the names of their corporate investors.
Reference 2:
Kaufman & Broad, a firm specializing in apartment building, refused to list institutional investors.
Reference 3:
Kaufman & Broad, a firm specializing in housing construction, refused to name the institutional investors.
Reference 4:
Residential construction company Kaufman & Broad refused to name the institutional investors.

To be able to observe the relationship between the tectogrammatical structure of a Czech sentence and its English translation (without distortions caused by automatic parsing), we have manually annotated on the tectogrammatical level both Czech and English sentences from the test set.

The PCEDT comprises also a translation dictionary, compiled from three different Czech-English manual dictionaries: two of them were downloaded form the Web and one was extracted from Czech and English EuroWordNets. Entry-translation pairs were filtered and weighed taking into account the reliability of the source dictionary, the frequencies of the translations in the English monolingual corpus, and the correspondence of the Czech and English POS tags. Furthermore, by training GIZA++ (Och and Ney, 2000) translation

model on the training part of the PCEDT extended by the manual dictionaries, we obtained a probabilistic Czech-English dictionary, more sensitive to the specific domain of financial news.

7 Conclusion

Building a large scale syntactically annotated parallel bilingual corpus is an extremely difficult endeavor, even if both languages are typologically similar and the syntactic annotation is based on similar linguistic tradition. This paper describes a method developed for the situation when both languages are typologically different as well as the data types traditionally used for the description of syntax. We do not think that our method is the only method possible, but nevertheless, we hope that the description of our method may help other researchers to avoid some of our mistakes when developing their own parallel syntactically annotated corpora.

The exploitation of the PCEDT for the stochastic machine translation is only most obvious application of this new parallel bilingual corpus. We hope that after the publication of currently available data (slightly more than one half of the Wall Street Journal section of Penn Treebank has been translated up to now) and especially after the completion of the whole project the PCEDT will prove to be a valuable source of data for various applications.

Acknowledgement

The work described in this paper has been supported by the grant of the MŠMT ČR No. ME642, No. LN00A063 and partially supported by the grant of the GAČR No. 405/03/0914.

References

Yaser Al-Onaizan, Jan Cuřín, Michael Jahr, Kevin Knight, John Lafferty, et al., 1999, The statistical machine translation. Technical report. WS'99, Johns Hopkins University, 1999

Jan Hajič, Barbora Hladká, 1998, Tagging Inflective Languages: Predicting Morphological Categories for a Rich, Structured Tagset. In Proceedings of ACL-Coling'98, Montreal, Canada. pp. 483-490, 1998

Alena Böhmová, 2001, Automatic procedures in tectogrammatical tagging. The Prague Bulletin of Mathematical Linguistics, 76, 2001.

Eugene Charniak, 1999, A maximum-entropy-inspired parser. Technical Report CS-99-12, 1999.

Jan Hajič, Eric Brill, Michael Collins, Barbora Hladká, Douglas Jones, Cynthia Kuo, Lance Ramshaw, Oren Schwartz, Christopher Tillmann, and Daniel Zeman, 1998, Core Natural Language Processing Technology Applicable to Multiple Languages}.Technical Report Research Note 37, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, 1998.

Jan Hajič, Jarmila Panevová, Eva Buráňová, Zdeňka Urešová, Alla Bémová, Jan Štěpánek, Petr Pajas and Jiří Kárník, 2001, A Manual for Analytic Layer Tagging of the Prague Dependency Treebank, Prague, Czech Republic, 2001.

Eva Hajičová, Jarmila Panevová and Petr Sgall, 2000, A Manual for Tectogrammatic Tagging of the Prague Dependency Treebank, Technical report TR-2000-09, ÚFAL MFF UK, Prague, Czech Republic, 2000.

Franz Josef Och and Hermann Ney, 2000, Improved statistical alignment models. Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, pp. 440--447, Hongkong, China, October 2000.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, 2001, Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176, IBM, 2001.

Zdeněk Žabokrtský, Petr Sgall and Džeroski Sašo, 2002, Machine learning approach to automatic functor assignment in the Prague Dependency Treebank. In Proceedings of LREC 2002 (Third International Conference on Language Resources and Evaluation), volume V, pp. 1513-1520, Las Palmas de Gran Canaria, Spain, 2002.