Prague Czech-English Dependency Treebank

Syntactically Annotated Resources for Machine Translation

Martin Čmejrek, Jan Cuřín, Jiří Havelka, Jan Hajič, Vladislav Kuboň

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics, Charles University in Prague
Malostranské nám. 25, Praha 1, Czech Republic
{cmejrek,curin,havelka,hajic,vk}@ufal.mff.cuni.cz

Abstract

This paper introduces the Prague Czech-English Dependency Treebank (PCEDT), a new Czech-English parallel resource suitable for experiments in structural machine translation. We describe the process of building the core parts of the resources – a bilingual syntactically annotated corpus and translation dictionaries. A part of the Penn Treebank has been translated into Czech, the dependency annotation of the Czech translation has been done automatically from plain text. The annotation of Penn Treebank has been tranformed into dependency annotation scheme. A subset of corresponding Czech and English sentences has been annotated by humans. First experiments in Czech-English machine translation using these data have already been carried out. The resources being created at Charles University in Prague are scheduled for release as Linguistic Data Consortium data collection in 2004.

1. Introduction

The efforts of Czech computational linguists concentrated in the past on creating large-scale monolingual corpora, such as the Czech National Corpus (100 million words annotated on morphological level) and Prague Dependency Treebank (PDT; Linguistic Data Consortium, 2001). The PDT is manually annotated on three levels: morphological layer (lowest), analytical layer (middle) – 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 as proposed by Functional Generative Description (FGD; Sgall et al., 1986). The Prague Czech-English Dependency Treebank (PCEDT) uses the annotation style of PDT for Czech, and adapts it for English.

In Section 2, we describe the process of manual translation of the Penn Treebank into Czech. The phrase tree annotation of the English part is converted by two independent transformation procedures into analytical and tectogrammatical representations described in Sections 3 and 4, respectively. Section 5 is about the automatic parsing of the Czech part into analytical representation, and a subsequent rule-based conversion into tectogrammatical representation. Sections 6, 7, and 8 briefly summarize dictionary resources, additional corpora, and a set of tools included in the PCEDT distribution. The overview of resources is presented in Table 1. Section 9 mentions experiments that have been carried out on the data collection.

2. Czech Translation of Penn Treebank

When starting the PCEDT project, we were deciding between two possible strategies: either the parallel annotation of already existing parallel texts, or the translation and annotation of an existing syntactically annotated corpus. The up-to-now main parallel Czech-English resource, Reader's Digest corpus, contains extremely free translations, which has proved 'difficult' in several machine-learning experiments (Al-Onaizan et al., 1999). Therefore, we decided for the human translation of the Penn Treebank (Linguistic Data Consortium, 1999) into Czech and its subsequent syntactic annotation.

The translators were asked to translate each English sentence as a single Czech sentence and to avoid unnecessary stylistic changes of translated sentences. The translations are being revised on two levels, linguistic and factual. About one half (21,628 sentences) have been translated so far, and the project aims at translating the whole Wall Street Journal part of the Penn Treebank.

For the purpose of quantitative evaluation methods, such as NIST or BLEU, for measuring performance of automatic translation systems, we selected a test set of 515 sentences and had them retranslated from Czech into English by 4 different translator offices, two of them from the Czech Republic and two from the U.S.A.

3. English Analytical Trees

This section describes the automatic conversion of Penn Treebank annotation into analytical representation.

The general transformation algorithm from phrase-tree topology into dependency one works as follows:

- Terminal nodes of the phrase tree are converted to nodes of the dependency tree.
- Dependencies between nodes are established recursively: The root node of the dependency tree transformed from the head constituent of a phrase becomes the governing node. The root nodes of the dependency trees transformed from the right and left siblings of the head constituent are attached as the left and right children (dependent nodes) of the governing node, respectively.
- Nodes representing traces are removed and their children are reattached to the parent of the trace.

The concept of the head of a phrase is important during the tranformation described above. For marking head constituents in each phrase, we used Jason Eisner's scripts. In case of coordination, we consider the rightmost coordinating conjuction (`CC`) to be the head. The treatment of

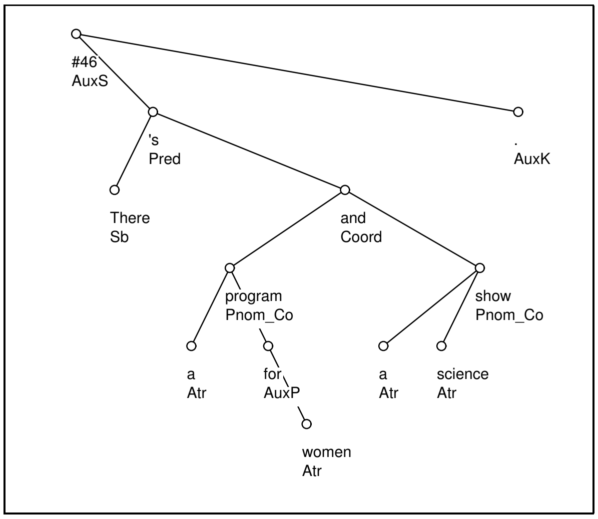

Figure 1: Analytical tree for the sentence "*There's a program for women and a science show.*"

apposition is a more difficult task, since there is no explicit annotation of this phenomenon in the Penn Treebank; constituents of a noun phrase enclosed in commas or other delimiters (and not containing `CC`) are considered to be in apposition and the rightmost delimiter becomes the head.

The information from both the phrase tree and the dependency tree is used for the **assignment of analytical functions**:

• Some function tags of a phrase tree almost unambiguously correspond to analytical functions in an analytical tree and can be mapped to them: `SBJ` $\rightarrow$ `Sb`, {`DTV`, `LGS`, `BNF`, `TPC`, `CLR`} $\rightarrow$ `Obj`, or {`ADV`, `DIR`, `EXT`, `LOC`, `MNR`, `PRP`, `TMP`, `PUT`} $\rightarrow$ `Adv`.

• For assigning analytical functions to the remaining nodes, we use rules querying local context (a node, its parent and grandparent) for POS and a phrase marker. For example, the rule `mPOS=MD` | `pPOS=VB` | `mAF=AuxV` assigns the analytical function tag `AuxV` to a modal verb headed by a verb.

Specifics of the PDT and Penn Treebank annotation schemes, mainly the markup of coordinations, appositions, and prepositional phrases are handled by these steps:

• The analytical function which was originally assigned to the head of a coordination or apposition is propagated to its child nodes by attaching the suffix `_Co` or `_Ap` to them and the head node gets the analytical function `Coord` or `Apos`, respectively.

• The analytical function originally assigned to a preposition node is propagated to its child and the preposition node is labeled `AuxP`.

• Sentences in the PDT annotation style always contain a technical root node labeled `AuxS`, which is inserted above the original root, and the final punctuation mark is moved under this new root.

• The dependency annotation scheme requires lemmatization – assigning base forms to words. This task was done automatically using the *morpha* tool (Minnen et al., 2001).

An analytical tree for a sample sentence automatically converted from the Penn Treebank can be found in Figure 1.

4. English Tectogrammatical Trees

The transformation of the Penn Treebank phrase trees into tectogrammatical representation consists of a structural transformation and an assignment of a tectogrammatical functor and a set of grammatemes to each node.

At the beginning of the structural transformation, an initial dependency tree is created by a general transformation procedure described in Section 3. However, functional (synsemantic) words, such as prepositions, punctuation marks, determiners, subordinating conjunctions, certain particles, auxiliary and modal verbs are handled differently. They are marked as "hidden" and information about them is stored in special attributes of their governing nodes (if they were to head a phrase, the head of the other constituent became the governing node in the dependency tree).

The well-formedness of a tectogrammatical tree structure requires the valency frames to be complete: apart from nodes that are realized on surface, there are several types of "restored" nodes representing the non-realized members of valency frames (cf. the pro-drop property of Czech and verbal condensations using gerunds or infinitives both in Czech and English). For the reconstruction of some of these nodes, we can use traces, which allow us to establish coreferential links and restore general participants in the valency frames.

For the assignment of tectogrammatical functors, we can use rules taking into consideration POS tags (e.g. `PRP` $\rightarrow$ `APP`), function tags (`JJ` $\rightarrow$ `RSTR`, `JJR` $\rightarrow$ `CPR`, etc.) and lemma ("not" $\rightarrow$ `RHEM`, "both" $\rightarrow$ `RSTR`).

Morphological grammatemes (e.g. tense, degree of comparison) are assigned to nodes of the tectogrammatical tree, based on PennTreebank POS tags and reflecting basic morphological properties of English.

In order to gain a "gold standard" annotation, 1,257 sentences (including the test set of 515 sentences) have been annotated manually. These data have been assigned morphological gramatemes (the full set of values), and the nodes have been reordered according to topic-focus articulation (information structure).

The automatic procedure briefly sketched above is described in detail in (Kučerová and Žabokrtský, 2002). The quality of such a transformation, based on comparison with manually annotated trees, is about 6% of wrongly aimed dependencies and 18% of wrongly assigned functors.

See Figure 2 with an example of the automatically converted tectogrammatical tree for the sample sentence.

5. Automatic Annotation of Czech

The Czech translations of Penn Treebank were automatically tokenized and morphologically tagged, each word form was assigned a basic form – *lemma* by (Hajič and Hladká, 1998) tagging tools.

Czech analytical parsing consists of a statistical dependency parser for Czech – either Collins parser (Hajič et al., 1998) or Charniak parser (Charniak, 1999), both adapted to dependency grammar – and a module for automatic analytical function assignment (Žabokrtský et al., 2002).

When building **the tectogrammatical structure**, the analytical tree structure is converted into the tectogram-

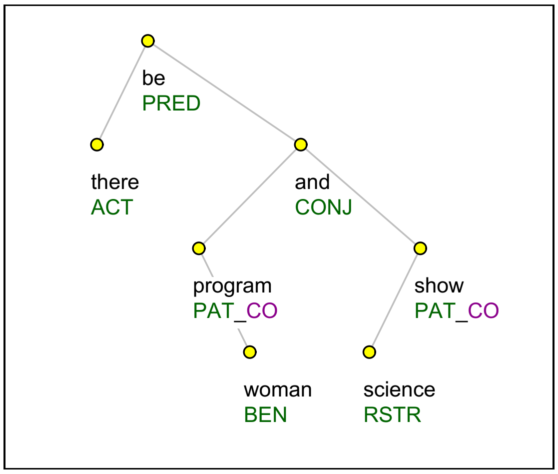

Figure 2: Tectogrammatical tree for the sentence "*There's a program for women and a science show.*"

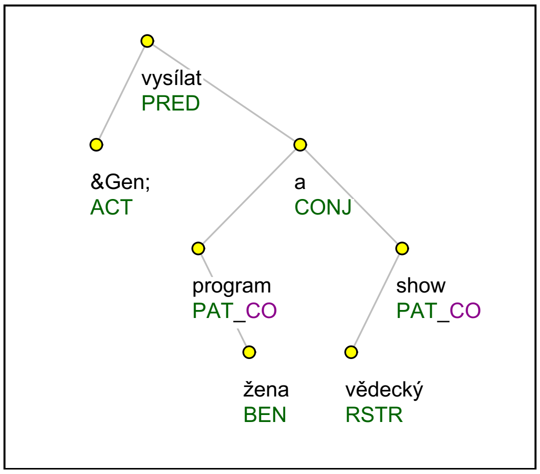

Figure 3: Tectogrammatical tree for the sentence "*Vysílá se program pro ženy a vědecká show.*"

matical one using linguistic rules (Böhmová, 2001). Then, tectogrammatical functors are assigned by a C4.5 classifier (Žabokrtský et al., 2002).

A test set of 515 sentences has been manually annotated on tectogrammatical level. See Figure 3 with the manual annotation of the translation of our sample sentence.

6. Czech-English Translation Dictionaries

The **Czech-English probabilistic dictionary** was compiled as the translation of the words occurring in the Czech translation of the Penn Treebank extended by words that occur more than 100 times in the Czech National Corpus (455M words). For the translation of this set of words we used three different Czech-English manual dictionaries: two of them were available on the Web (WinGED and GNU/FDL) and one was extracted from Czech and English EuroWordNets. We included only translations that occurred in at least two of the three dictionaries or the frequency of which is significant in the English North American News Text Collection (310M words).

POS tag and lemma were added to each Czech entry. If possible, we selected the same POS for the English translation, otherwise the most frequent one.

By training GIZA++ translation model (Och and Ney, 2003) on the training part of the PCEDT extended by the obtained entry-translation pairs, we created a probabilistic Czech-English dictionary more sensitive to the domain of financial news specific for the Wall Street Journal.

The resulting Czech-English probabilistic dictionary contains 46,150 entry-translation pairs.

Since Czech is highly inflective, the PCEDT also comprises a **Czech-English translation dictionary of word forms** containing pairs of Czech and English word forms agreeing in appropriate morphological categories (such as number and person). This dictionary contains 496,673 entry-translation pairs.

We have incorporated also an **English-Czech Dictionary** downloaded from the web under GNU/FDL licence (Svoboda, 2004). The dictionary was created from the probabilistic dictionary, and contains 115,929 entry-translation pairs, and unlike the dictionaries mentioned above, it contains also multi-word translations.

7. Additional Resources

Reader's Digest parallel corpus contains raw text in 53,000 automatically aligned segments in 450 articles, years 1993–1996. The Czech part is a free translation of the English original text.

A large **corpus of Czech** contains 39M words in news articles published in the newspaper Lidové Noviny, years 1994–1995.

8. Tools

SMT Quick Run is a package of scripts and instructions for building a statistical machine translation system using the PCEDT or any other parallel corpus. The system uses translation models GIZA++ and ISI ReWrite decoder (Germann et al., 2001).

TrEd is a graphical editor and viewer of tree structures. Its modular architecture allows easy handling of diverse annotation schemes, it has been used as the principal annotation environment for the PDT and PCEDT.

Netgraph is a multi-platform client-server application for browsing, querying and viewing analytical and tectogrammatical dependency trees, either over the Internet or locally.

9. Experiments in Structural MT

Two experiments in structural Czech-English machine translation have been carried out on the PCEDT.

The first one – MAGENTA system (Hajič et al., 2002) – is an experimental framework for machine translation implemented during 2002 NLP Workshop at CLSP, Johns Hopkins University in Baltimore. Modules for parsing of Czech, lexical transfer, a prototype of a statistical tree-to-tree transducer for structural transformations used during transfer and generation, and a language model for English based on dependency syntax are integrated in one pipeline.

The second experiment – Dependency-based Machine Translation, described in (Čmejrek et al., 2003) – uses a

Description of Data	*Size*
PTB Corpus: English part (# sentences)	
– manually annotated on tectogrammatical level	1,257
– automatically transformed into analytical & tectogrammatical levels	49,208
– retranslated by 4 different human translators for the purposes of quantitative evaluation	515
PTB Corpus: Czech part (# sentences)	
– manually annotated on tectogrammatical level	515
– automatically parsed into analytical & tectogrammatical levels	21,628
Reader's Digest corpus (# aligned segments)	58,656
Czech monolingual corpus (# sentences)	2,385,000
Dictionaries (# entry-translation pairs)	
Czech-English probabilistic dictionary	46,150
Czech-English dictionary of word forms	496,673
GNU/FDL English-Czech dictionary	115,929

Table 1: Data Sizes

rule-based method for generating English output directly from the tectogrammatical representation. DBMT comprises the whole way from the Czech plain-text sentence to the English one using the state-of-the-art parsers into analytical and tectogrammatical representation for Czech and a word-to-word probabilistic dictionary built from manual dictionaries and dictionaries automatically obtained from the parallel corpus.

10. Conclusion

Building a large-scale parallel treebank is a demanding challenge. We have created a parallel corpus for a pair of languages with a relatively different typology, Czech and English, and made an attempt to bridge between two linguistic theories commonly used for their description.

We are convinced that the PCEDT will be usefull for further experiments in Czech-English machine translation. A certain disproportion between the English part converted from a manual annotation and the Czech part automatically parsed from plain text corresponds to the real situation in Czech-English machine translation, where modules for transfer and generation have to adapt to errors caused by automatic analysis of the input language. Several input options for Czech (plain text, analytical and tectogrammatical representations–both automatic and manual) and a test set for quantitative evaluation can be used in various experimental settings, allowing to identify insufficiencies in analysis, transfer, and generation.

Acknowledgements

This research was supported by the following grants: MŠMT ČR Grants No. LN00A063 and No. MSM113200006, and NSF Grant No. IIS-0121285. Travel expenses were partially supported by the Czech Society for Cybernetics and Informatics.

References

Al-Onaizan, Yaser, Jan Cuřín, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz-Josef Och, David Purdy, Noah A. Smith, and David Yarowsky, 1999. The Statistical Machine Translation. Technical report. NLP WS'99 Final Report.

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

Charniak, Eugene, 1999. A Maximum-Entropy-Inspired Parser. Technical Report CS-99-12.

Germann, Ulrich, Michael Jahr, Kevin Knight, Daniel Marcu, and Kenji Yamada, 2001. Fast decoding and optimal decoding for machine translation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*.

Hajič, Jan, 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.

Hajič, Jan and Barbora Hladká, 1998. Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In *Proceedings of COLING-ACL Conference*. Montreal, Canada.

Hajič, Jan, Martin Čmejrek, Bonnie Dorr, Yuan Ding, Jason Eisner, Daniel Gildea, Terry Koo, Kristen Parton, Gerald Penn, Dragomir Radev, and Owen Rambow, 2002. Natural Language Generation in the Context of Machine Translation. Technical report. NLP WS'02 Final Report.

Kučerová, Ivona and Zdeněk Žabokrtský, 2002. Transforming Penn Treebank Phrase Trees into (Praguian) Tectogrammatical Dependency Trees. *Prague Bulletin of Mathematical Linguistics*, 78:77–94.

Linguistic Data Consortium, 1999. Penn Treebank 3. LDC99T42.

Linguistic Data Consortium, 2001. Prague Dependency Treebank 1. LDC2001T10.

Minnen, G., J. Carroll, and D. Pearce, 2001. Applied Morphological Processing of English. *Natural Language Engineering*, 7(3):207–223.

Och, Franz Josef and Hermann Ney, 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

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

Svoboda, Milan, 2004. Anglicko-český slovník. http://slovnik.zcu.cz.

Čmejrek, Martin, Jan Cuřín, and Jiří Havelka, 2003. Czech-English Dependency-based Machine Translation. In *Proceedings of the 10th Conference of The European Chapter of the Association for Computational Linguistics*. Budapest, Hungary.

Žabokrtský, Zdeněk, Petr Sgall, and Džeroski Sašo, 2002. Machine Learning Approach to Automatic Functor Assignment in the Prague Dependency Treebank. In *Proceedings of LREC 2002*, volume V. Las Palmas de Gran Canaria, Spain.