Prague Czech-English Dependency Treebank
Resource for Structure-based MT

Martin Čmejrek, Jan Cuřín, Jan Hajič, Jiří Havelka

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

Abstract. The Prague Czech-English Dependency Treebank (PCEDT) is a syntactically annotated Czech-English parallel corpus. The Penn Treebank has been translated to Czech, and its annotation automatically transformed into dependency annotation scheme. The dependency annotation of Czech is done from plain text by automatic procedures. A small 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 have been created at Charles University in Prague and released by Linguistic Data Consortium in 2004.

1. Introduction

The Prague Czech-English Dependency Treebank – PCEDT (Linguistic Data Consortium, 2004) is a project of creating a Czech-English syntactically annotated parallel corpus motivated by research in the field of machine translation. Parallel data are needed for designing, training, and evaluation of both statistical and rule-based machine translation systems.

When starting the PCEDT project, we decided to translate and annotate an existing syntactically annotated corpus, rather than to annotate in parallel an already existing parallel corpus of raw texts, since the latter option would have been more money and time consuming. The choice of the Penn Treebank as the source corpus was pragmatically motivated: firstly it is a widely recognized linguistic resource, and secondly the translators were native speakers of Czech, capable of high quality translation into their native language.

Since Czech is a language with a relatively high degree of word-order freedom, and its sentences contain certain syntactic phenomena, such as discontinuous constituents (non-projective constructions), which cannot be straightforwardly handled using the annotation scheme of Penn Treebank (Marcus et al., 1993; Linguistic Data Consortium, 1999), based on phrase-structure trees, we decided to adopt for the PCEDT the dependency-based annotation scheme of the Prague Dependency Treebank – PDT (Linguistic Data Consortium, 2001; Sgall et al., 1986), which is described in Section 3.

In Section 2., we describe the process of translating the Penn Treebank into Czech and reference retranslations. Section 4. presents manual tectogrammatical annotation for both languages. The automatic process of transformation of Penn Treebank annotation of English into both representations, analytical and tectogrammatical, is described in Section 5., the automatic annotation of Czech is described in Section 6. Section 7. gives an overview of additional resources included in the PCEDT corpus. Section 8. mentions two experiments that have been carried out on the data collection.

2. English to Czech Translation of Penn Treebank

Since the PCEDT is aimed as a resource for the purpose of MT, the translators were asked to translate each English sentence as a single Czech sentence and to avoid unnecessary stylistic changes of translated sentences. About half of the Penn Treebank has been translated so far (currently 21,628 sentences), the project aims at translating the whole Wall Street Journal part of the Penn Treebank.

2.1. English Retranslation

For the purpose of quantitative evaluation methods, such as NIST or BLEU, for measuring performance of 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 of them from the U.S.A.

This set might be also useful for a linguistic study of the variation between multiple translations. See Figure 1 for an example of reference translations of the sentence “Kaufman & Broad, a home building company, declined to identify the institutional investors.”

3. The Prague Dependency Treebank Annotation

The Prague Dependency Treebank is a manually annotated corpus of Czech. The corpus size is approx. 1.5 million words (tokens). In this section we briefly summarize the annotation scheme of PDT adopted by the PCEDT.

Original from PTB: *Kaufman & Broad, a home building company, declined to identify the institutional investors.*

Czech translation: *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.*

Figure 1: A sample English sentence from WSJ, its Czech translation, and four reference retranslations.

Three main groups ("layers") of annotation are used:

- the morphological layer, where lemmas and tags are being annotated based on their context,
- the analytical layer, which roughly corresponds to the surface syntax of the sentence,
- the tectogrammatical layer, or linguistic meaning of the sentence in its context.

3.1. The Morphological Layer

The annotation of Czech at the morphological layer is an unstructured classification of the individual tokens (words and punctuation) of the utterance into morphological classes (morphological tags) and lemmas. Since Czech is a highly inflective language, the tagset size used is 4257, with about 1100 different tags actually appearing in the PDT.

There are 13 categories used for morphological annotation of Czech: Part of speech, Detailed part of speech, Gender, Number, Case, Possessor's Gender and Number, Person, Tense, Voice, Degree of Comparison, Negation and Variant.

For English we adopted the Penn Treebank POS annotation.

3.2. The Analytical Layer

At the analytical layer, two attributes are being annotated:

- (surface) sentence structure,
- analytical function.

A rooted dependency tree is being built for every sentence as a result of the annotation. Every item (token) from the morphological layer becomes (exactly) one node in the tree, and no nodes (except for the single "technical" root of the tree) are added. Analytical functions, despite being kept at nodes, are in fact names of the dependency relations between a dependent (child) node and its governor (parent) node.

Coordination and apposition is handled using "technical" dependencies: the conjunction is the head and the members are its "dependent" nodes. Common modifiers of the coordinated structure are also dependents of the coordinating conjunction, but they are not marked as coordinated structure members. This additional "coordinated structure member" markup (`_Co`, `_Ap`) gives an added flexibility for handling such constructions.

Ellipsis is not annotated at this level (no traces, no empty nodes etc.), but a special analytical function (`ExD`) is used at nodes that are lacking their governor, even though they (technically) do have a governor node in the annotation.

There are 24 analytical functions used, such as `Sb` (Subject), `Obj` (Object, regardless of whether the direct, indirect, etc.), `Adv` (Adverbial, regardless of type), `Pred`,`Pnom` (Predicate / Nominal part of a predicate for the (verbal) root of a sentence), `Atr` (Attribute in noun phrases), `Atv`, `AtvV` (Verbal attribute / Complement), `AuxV` (auxiliary verb – similarly for many other auxiliary-type words, such as prepositions (`AuxP`), subordinate conjunctions (`AuxC`), etc.), `Coord`, `Apos` (coordination/apposition "head"), `Par` (Parenthesis head), etc.

3.3. The Tectogrammatical Layer

The tectogrammatical layer is the most elaborated, complicated, but also the most theoretically grounded layer of syntactico-semantic (or "deep syntactic") representation. For the purposes of the annotation of PCEDT, we will sketch only the core components of the tectogrammatical annotation.

The tectogrammatical layer goes beyond the surface structure of the sentence, replacing notions such as "subject" and "object" by notions like "actor" (`ACT`), "patient" (`PAT`), "addressee" (`ADDR`) etc., but the representation still relies upon the language structure itself rather than on world knowledge. The nodes in the tectogrammatical tree are autosemantic (content) words only. Dependencies between nodes represent the relations between the (autosemantic) words in a sentence, the dependencies are labeled by functors, which describe the dependency relations. Every sentence is thus represented as a dependency tree, the nodes of which are autosemantic words, and the (labeled) edges name the dependencies between a dependent and its governor. Coordination and apposition is handled in the same way as on the analytical level.

Many nodes found at the morphological and analytical layers disappear (such as function words, prepositions, subordinate conjunctions, etc.). The information carried by the deleted nodes is not lost, of course: the relevant attributes of the autosemantic nodes they belong to now contain enough information (at least theoretically) to reconstruct them.

Ellipsis is being resolved at this layer. Insertion of (surface-)deleted nodes is driven by the notion of *valency* and completeness: if a word is deemed to be used in a context in which some of its valency frames applies, then all the frame's obligatory slots are "filled" (using regular dependency relations between nodes) by either existing nodes or by newly created nodes, and these nodes are annotated accordingly.

4. Manual Tectogrammatical Annotation of Czech and English

Since there are no guidelines for tectogrammatical annotation of English yet, and in order to acquire some initial experience before the work on the guidelines begins, a "gold standard" tectogrammatical annotation of 1,257 sentences has been done. These data are assigned morphological grammatemes (the full set of values), and the nodes are reordered according to topic-focus articulation (information structure). The manually annotated sentences comprise the whole development and evaluation test sets. Also the Czech counterpart of the test set (515 sentences) has been manually annotated according to the guidelines for tectogrammatical annotation of Czech.

5. Automatic Transformation of Penn Treebank Annotation

This section gives an overview of the automatic procedures used in obtaining the automatic dependency annotation of the Penn Treebank part of PCEDT.

For illustration, different annotations of example sentence "An earthquake struck Northern California, killing more than 50 people." are shown in Figures 2 and 3. The Czech translation of this sentence is "Zemětřesení zasáhlo severní Kalifornii a usmrtilo více než 50 lidí.", which can be literally translated as "An earthquake struck Northern California and killed more than 50 people."

5.1. English Tectogrammatical Dependency Trees

The transformation of 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, the initial dependency tree is created by a general transformation procedure analogous to the one as described above. 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 whole procedure is described in detail in (Kučerová and Žabokrtský, 2002).

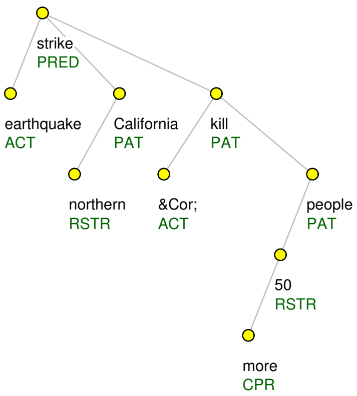

Figure 3: Output of automatic conversion into tectogrammatical representation for the sentence "*An earthquake struck Northern California, killing more than 50 people.*"

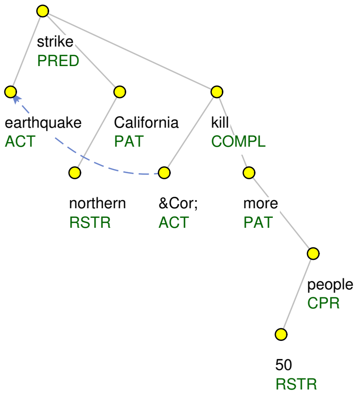

Figure 4: Manually annotated tectogrammatical tree for the sentence "*An earthquake struck Northern California, killing more than 50 people.*"

The quality of the automatic transformation procedure described above, based on comparison with manually annotated trees, is about 6% of wrongly aimed dependencies and 18% of wrongly assigned functors.

See Figures 3 and 4 for a comparison of the manually annotated tectogrammatical tree and the output of the automatic conversion into tectogrammatical representation for the sample sentence.

6. Automatic Annotation of Czech

6.1. Analytical Annotation of Czech

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

Czech **analytical parsing** consists of a statistical dependency parser for Czech – either Collins parser (Collins et al., 1999) or Charniak parser (Charniak, 1999), both adapted to dependency grammar

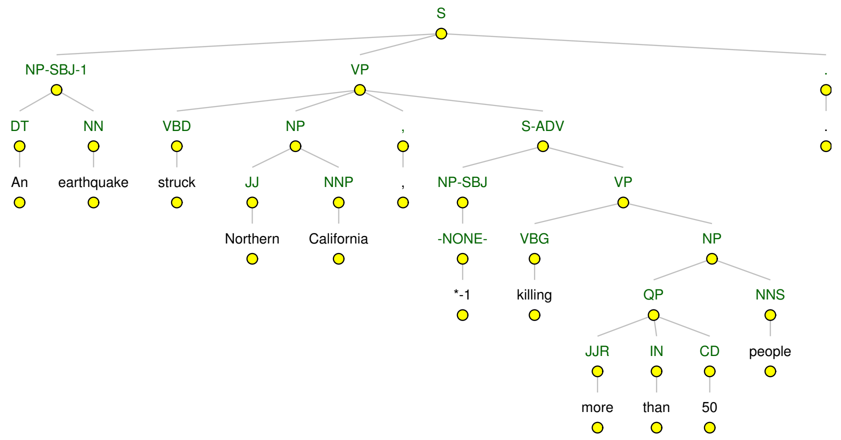

Figure 2: Original Penn Treebank annotation for the sentence “*An earthquake struck Northern California, killing more than 50 people.*”

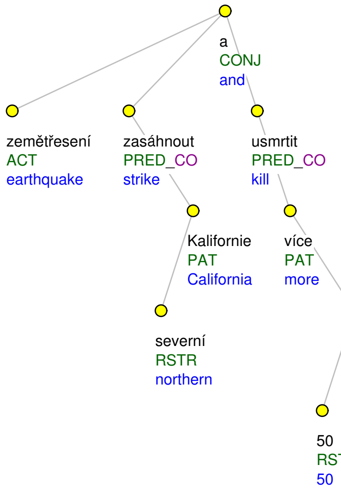

Figure 5: Manually annotated tectogrammatical tree for the Czech translation “*Zemětřesení zasáhlo severní Kalifornii a usmrtilo více než 50 lidí.*”

and trained on Prague Dependency Treebank (Linguistic Data Consortium, 2001) – and a module for automatic analytical function assignment (Žabokrtský et al., 2002).

6.2. Tectogrammatical Annotation of Czech

When building the **tectogrammatical structure**, the analytical tree structure is converted into the tectogrammatical one. These transformations are described by linguistic rules (Böhmová, 2001). Then, tectogrammatical functors are assigned by a C4.5 classifier (Žabokrtský et al., 2002).

7. Other Resources

7.1. Reader’s Digest Corpus

This corpus contains parallel raw text of 450 articles from the Reader’s Digest, years 1993–1996. The Czech part is a translation of the English one. Sentence pairs were aligned automatically by Dan Melamed’s SIMR/GMA tool. Since the translations in this corpus are relatively free, only 43969 of 54091 aligned segments contain one-to-one sentence alignments.

7.2. Czech Monolingual Corpus

The electronic text sources have been provided by the Institute of Czech National Corpus. Originally, all data come from news articles which were published in the daily newspaper Lidove Noviny, 1994–1995. The total amount of data is more than 39M tokens (words proper + punctuation) in about 2,385K sentences.

7.3. Czech-English Translation Dictionaries

Czech-English Probabilistic Dictionary

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 probabilistic dictionary contains 46,150 entry-translation pairs.

Czech-English Dictionary of Word Forms

Since Czech is highly inflective, the PCEDT also comprises a 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 was created from the probabilistic dictionary and contains 496,673 entry-translation pairs.

English-Czech Dictionary under GNU/FDL

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

7.4. Tools

SMT Quick Run

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

Tree Editor TrEd

TrEd (Pajas, 2005) 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. TrEd has a modular architecture allowing custom input/output modules to be created in order to support other data formats.

NetGraph

Netgraph is a multi-platform client-server application allowing you to browse, select and view analytical and tectogrammatical dependency trees. It can either view Czech trees from Prague Dependency Treebank (PDT) on the remote server located at the Institute of Formal and Applied Linguistics in Prague, or you can install your own server for viewing trees from PCEDT.

8. 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 rule-based method for generating English output directly from the tectogrammatical representation. DBMT comprises the whole way from the Czech plaintext 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.

9. 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 useful 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.

10. Acknowledgements

This research was supported by the following grants: MŠMT ČR Project No. MSM0021620838, MŠMT ČR Project No. ME642, and NSF Grant No. IIS-0121285.

References

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.

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.

Collins, Michael, Jan Hajič, Lance Ramshaw, and Christoph Tillmann, 1999. A Statistical Parser for Czech. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. College Park, Maryland.

Germann, Ulrich, Michael Jahr, Kevin Knight, Daniel Marcu, and Kenji Yamada, 2001. Fast Decoding and Optimal Decoding for Machine Translation. In *Meeting of the Association for Computational Linguistics*.

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.

Linguistic Data Consortium, 2004. Prague Czech-English Dependency Treebank 1.0. LDC2004T25.

Marcus, M. P., B. Santorini, and M. A. Marcinkiewicz, 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

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

Pajas, Petr, 2005. Tree Editor TrEd. http://ckl.mff.cuni.cz/˜pajas/tred/.

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.

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.