Prague Czech-English Dependency Treebank
Any Hopes for a Common Annotation Scheme?

Martin Čmejrek, Jan Cuřín, Jiří Havelka
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics, Charles University in Prague
Malostranské nám. 25, Praha 1, Czech
{cmejrek,curin,havelka}@ufal.mff.cuni.cz

Abstract

The Prague Czech-English Dependency Treebank (PCEDT) is a new syntactically annotated Czech-English parallel resource. 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. We discuss some of the problems we have experienced during the automatic transformation between annotation schemes and hint at some of the difficulties to be tackled by potential guidelines for dependency annotation of English.

1 Introduction

The Prague Czech-English Dependency Treebank (PCEDT) 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.

Since Czech is a language with 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). The PDT is annotated on three levels: morphological layer (lowest), analytic 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 levels, as proposed by Functional Generative Description (Sgall et al., 1986).

In Section 2, we describe the process of translating the Penn Treebank into Czech. Section 3 sketches the general procedure for transforming phrase topology of Penn Treebank into dependency structure and describes the specific conversions into analytical and tectogrammatical representations. The following Section 4 describes the automatic process of parsing of Czech into analytical representation and its automatic conversion into tectogrammatical representation. Section 5 briefly discusses some of the problems of annotation from the point of view of mutual compatibility of annotation schemes. Section 6 gives an overview of additional resources included in the PCEDT.

2 English to Czech Translation of Penn Treebank

When starting the PCEDT project, we chose the latter of two possible strategies: either the parallel annotation of already existing parallel texts, or the translation and annotation of an existing syntactically annotated corpus. The choice of the Penn Treebank as the source corpus was also 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.

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 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.

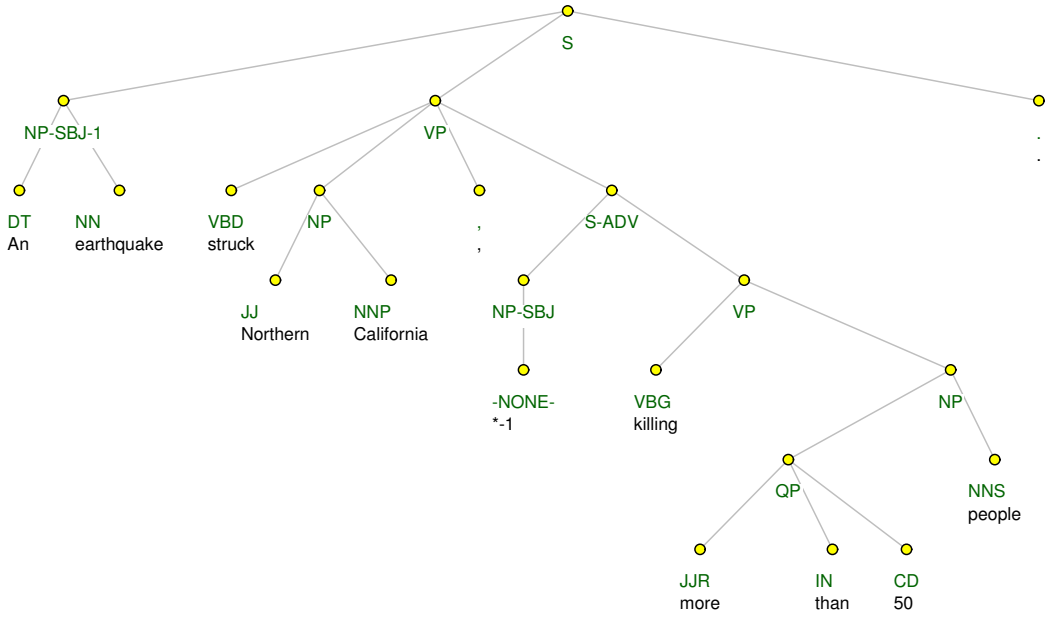

Figure 1: Penn Treebank annotation of the sentence "*An earthquake struck Northern California, killing more than 50 people.*"

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.

3 Transformation of Penn Treebank Phrase Trees into Dependency Structure

The transformation algorithm from phrase-structure topology into dependency one, similar to transformations described by Xia and Palmer (2001), works as follows:

- Terminal nodes of the phrase 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.

3.1 Preprocessing of Penn Treebank

Several preprocessing steps preceded the transformation into both analytical and tectogrammatical representations.

Marking of Heads in English

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.

Lemmatization of English

Czech is an inflective language, rich in morphology, therefore lemmatization (assigning base forms) is indispensable in almost any linguistic application. Mostly for reasons of symmetry with Czech data and compatibility with the dependency annotation scheme, the English part was automatically lemmatized by the *morpha* tool (Minnen et al., 2001) using manually assigned POS tags of the Penn Treebank.

Unique Identification

For technical reasons, a unique identifier is assigned to each sentence and to each token of Penn Treebank.

3.2 English Analytical Dependency Trees

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

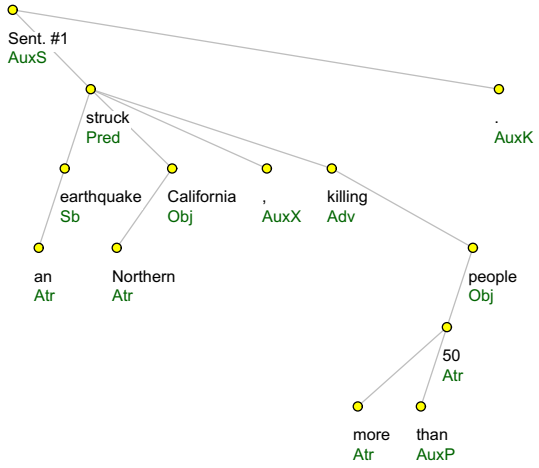

Figure 2: Analytical tree for the sentence "*An earthquake struck Northern California, killing more than 50 people.*"

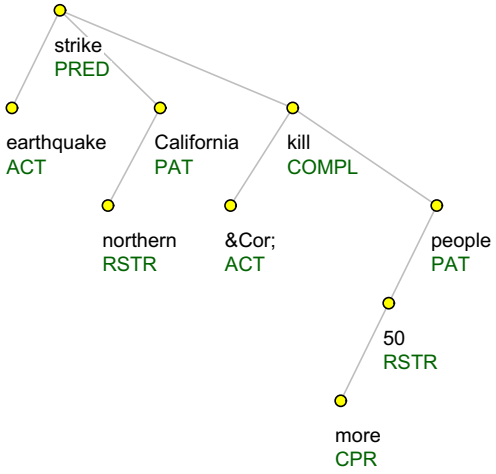

Figure 3: Tectogrammatical tree for the sentence "*An earthquake struck Northern California, killing more than 50 people.*"

The **structural transformation** works as described above. Because the handling of coordination in PDT is different from the Penn Treebank annotation style and the output of Jason Eisner's head assigning scripts, in the case of a phrase containing a coordinating conjunction (CC), we consider the rightmost CC as the head. The treatment of 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**:

- Penn Treebank function tag to analytical function mapping: some function tags of a phrase tree correspond to analytic functions in an analytical tree and can be mapped to them:

  $\texttt{SBJ} \rightarrow \texttt{Sb}$,

  $\{\texttt{DTV}, \texttt{LGS}, \texttt{BNF}, \texttt{TPC}, \texttt{CLR}\} \rightarrow \texttt{Obj}$,

  $\{\texttt{ADV}, \texttt{DIR}, \texttt{EXT}, \texttt{LOC}, \texttt{MNR}, \texttt{PRP}, \texttt{TMP}, \texttt{PUT}\} \rightarrow \texttt{Adv}$.

- Assignment of analytical functions using local context of a node: for assigning analytical functions to the remaining nodes, we use rules looking at the current node, its parent and grandparent, taking into account POS and the phrase marker of the constituent in the original phrase tree headed by the node. For example, the rule

  $$\texttt{mPOS} = \texttt{DT} | \texttt{mAF} = \texttt{Atr}$$

  assigns the analytical function Atr to every determiner, the rule

  $$\texttt{mPOS} = \texttt{MD} | \texttt{pPOS} = \texttt{VB} | \texttt{mAF} = \texttt{AuxV}$$

  assigns the function tag AuxV to a modal verb headed by a verb, etc. The attribute mPOS representing the POS of a node is obligatory for every rule. The rules are examined primarily in the order of the longest prefix of the POS of the given node and secondarily in the order as they are listed in the rule file. The ordering of rules is important, since the first matching rule found assigns the analytical function and the search is finished.

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

- Coordinations and appositions: the analytical function that 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.

- Prepositional phrases: the analytical function originally assigned to the preposition node is propagated to its child and the preposition node is labeled AuxP.

- Sentences in the PDT annotation style always contain a root node labeled AuxS, which, as the only one in the dependency tree, does not correspond to any terminal of the phrase tree; the root node is inserted

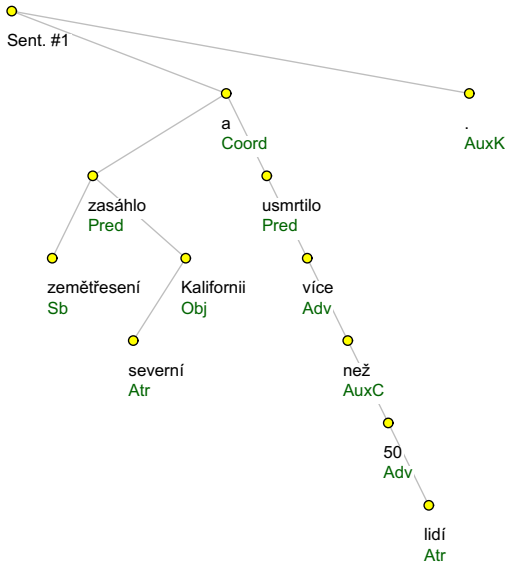

Figure 4: Analytical tree for the Czech translation "*Zemětřesení zasáhlo severní Kalifornii a usmrtilo více než 50 lidí.*"

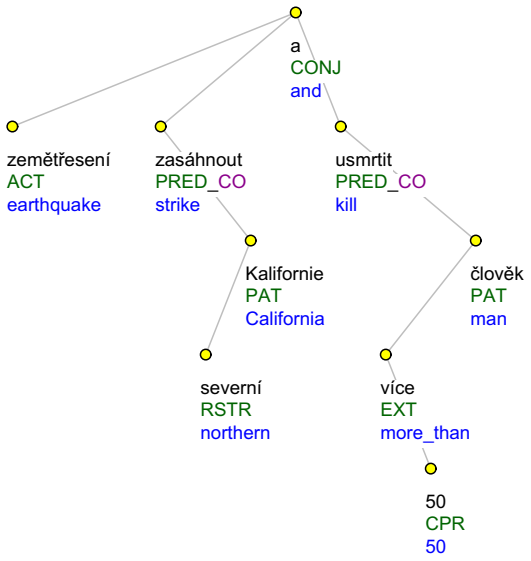

Figure 5: Tectogrammatical tree for the Czech translation "*Zemětřesení zasáhlo severní Kalifornii a usmrtilo více než 50 lidí.*"

above the original root. While in the Penn Treebank the final punctuation is a constituent of the sentence phrase, in the analytical tree it is moved under the technical sentence root node.

Compare the phrase structure and the analytical representation of a sample sentence from the Penn Treebank in Figures 1 and 2.

3.3 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 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 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. pro-drop property of Czech and verbal condensations using gerunds and infinitives both in Czech and English). For a partial reconstruction of such nodes, we can use traces, which allow us to establish coreferential links, or 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` → `APP`), function tags (`JJ` → `RSTR`, `JJR` → `CPR`, etc.) and lemma ("not" → `RHEM`, "both" → `RSTR`).

Grammateme Assignment – morphological grammatemes (e.g. tense, degree of comparison) are assigned to each node of the tectogrammatical tree. The assignment of the morphological attributes is based on PennTreebank tags and reflects basic morphological properties of the language. At the moment, there are no automatic tools for the assignment of syntactic grammatemes, which are designed to capture detailed information about deep syntactic structure.

The whole procedure is described in detail in Kučerová and Žabokrtský (2002).

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

The quality of the automatic transformation procedure described above, based on comparison with manually an-

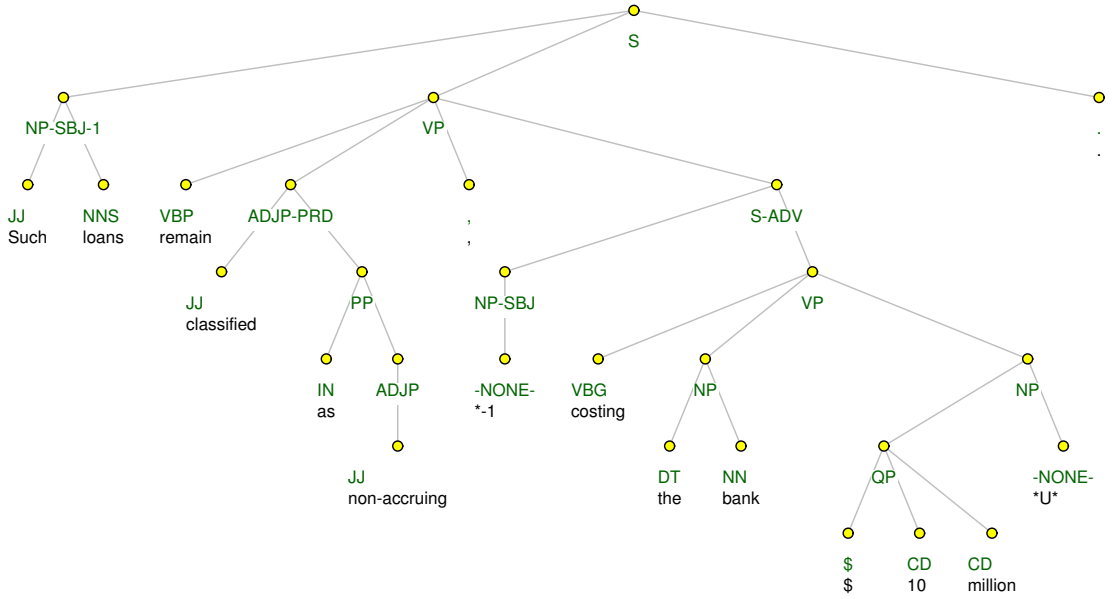

Figure 6: Penn Treebank annotation of the sentence "*Such loans remain classified as non-accruing, costing the bank $10 million.*"

notated trees, is about 6% of wrongly aimed dependencies and 18% of wrongly assigned functors.

See Figure 3 for the manually annotated tectogrammatical representation of the sample sentence.

4 Automatic 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 – 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 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).

The test set of 515 sentences (which have been retranslated into English) has been also manually annotated on tectogrammatical level.

See Figures 4 and 5 for automatic analytical and manual tectogrammatical annotation of the Czech translation of the sample sentence.

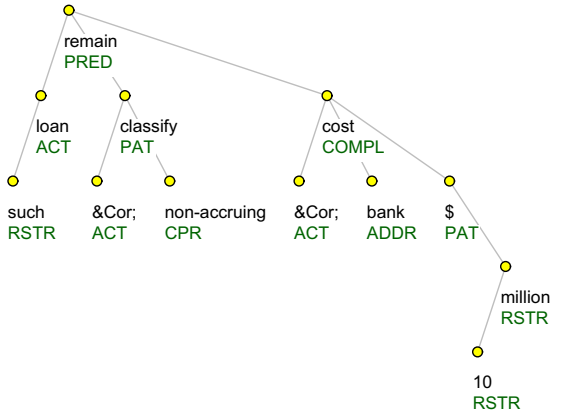

Figure 7: Tectogrammatical tree for the sentence "*Such loans remain classified as non-accruing, costing the bank $10 million.*"

5 Problems of Dependency Annotation of English

The manual annotation of 1,257 English sentences on tectogrammatical level was, to our knowledge, the first attempt of its kind, and was based especially on the instructions for tectogrammatical annotation of Czech. During the process of annotation, we have experienced both phenomena that do not occur in Czech at all, and phenomena

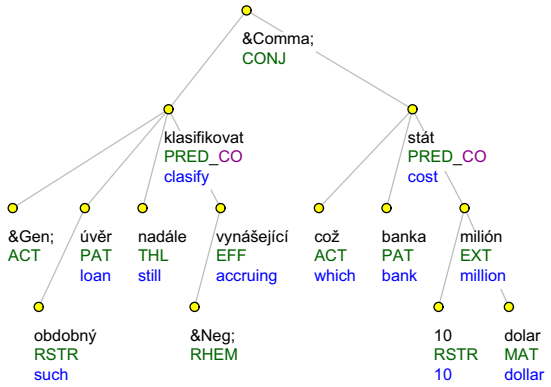

Figure 8: Tectogrammatical tree for the Czech translation "*Obdobné úvěry jsou nadále klasifikovány jako nevynášející, což banku stálo 10 milionů dolarů.*"

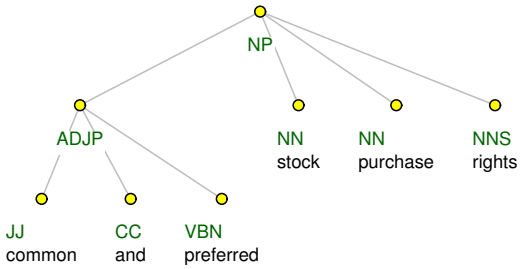

Figure 9: Penn Treebank annotation of the noun phrase "*common and preferred stock purchase rights*".

whose counterparts in Czech occur rarely, and therefore are not handeled thoroughly by the guidelines for tectogrammatical annotation designed for Czech. To mention just a few, among the former belongs the annotation of articles, certain aspects of the system of verbal tenses, and phrasal verbs. A specimen of a roughly corresponding phenomenon occurring both in Czech and English is the gerund. It is a very common means of condensation in English, but its counterpart in Czech (usually called transgressive) has fallen out of use and is nowadays considered rather obsolete.

The guidelines for Czech require the transgressive to be annotated with the functor `COMPL`. The reason why it is highly problematic to apply them straightforwardly also to the annotation of English, is that the English gerund has a much wider range of functions than the Czech transgressive. The gerund can be seen as a means of condensing subordinated clauses with in principle adverbial meaning (as it is analyzed in the phrase-structure annotation of Penn Treebank). Since the range of functors with adverbial meaning is much more fine-grained, we deem it inappropriate to mark the gerund clauses in such a simple way on the tectogrammatical level.

From the point of view of machine translation, the gerund constructions pose considerable difficulties because of the many syntactic constructions suitable as their translations corresponding to their varied syntactic functions.

We present two examples illustrating the issues mentioned above. Each example consists of three figures, the first one presenting the Penn Treebank annotation of a (in the second case simplified) sentence from the Penn Treebank, the second one giving its tentative tectogrammatic representation (according to the guidelines for Czech applied to English), and the third one containing the tectogrammatical representation of its translation into Czech (see Figures 1, 3, 5, and Figures 6, 7, 8). Note that in neither of the two examples the Czech transgressive is used as the translation of the English gerund; a coordination structure is used instead.

On the other hand, we have also experienced phenomena in English whose Penn Treebank style of annotation is insufficient for a successfull conversion into dependency representation.

In English, the usage of constructions with nominal premodification is very frequent, and the annotation of such noun phrases in the Penn Treebank is often flat, grouping together several constituents without reflecting finer syntactic and semantic relations among them (see Figure 9 for an example of such a noun phrase). In fact, the possible syntactic and especially semantic relations between the members of the noun phrase can be highly ambiguous, but when translating such a noun phrase into Czech, we are not usually able to preserve the ambiguity and are forced to resolve it by choosing one of the readings (see Figure 10).

Sometimes we even may be forced to insert new words explicitly expressing the semantic relations within the nominal group. An example of an English noun phrase and the tectogrammatical representation of its Czech translation with an inserted word "podnikající" ('operating') can be found in Figures 11 and 12.

6 Other Resources Included in PCEDT

6.1 Reader's Digest Parallel Corpus

Reader's Digest parallel corpus contains raw text in 53,000 aligned segments in 450 articles from the Reader's Digest, years 1993–1996. The Czech part is a free translation of the English version. The final selection of data has been done manually, excluding articles whose translations significantly differ (in length, culture-specific facts, etc.). Parallel segments on sentential level have been aligned by Dan Melamed's aligning tool (Melamed,

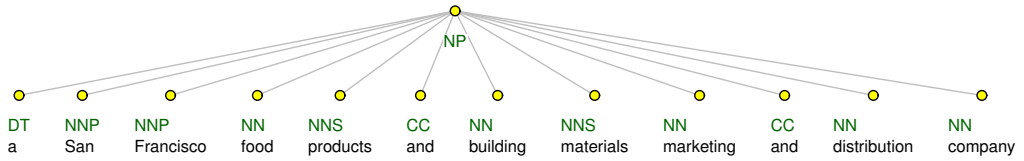

Figure 11: Penn Treebank annotation of the noun phrase "*a San Francisco food products and building materials marketing and distribution company*".

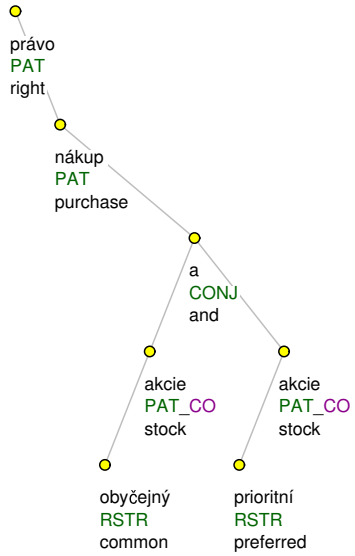

Figure 10: Tectogrammatical tree for the Czech translation "*právo na nákup obyčejných a prioritních akcií*".

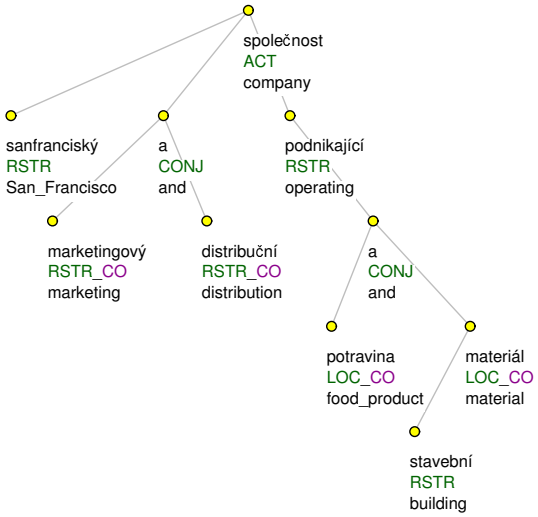

Figure 12: Tectogrammatical tree for the Czech translation "*sanfranciská marketingová a distribuční společnost podnikající v potravinách a stavebních materiálech*".

1996). The topology is 1–1 (81%), 0–1 or 1–0 (2%), 1–2 or 2–1 (15%), 2–2 (1%), and others (1%).

6.2 Dictionaries

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 Czech and English monolingual corpora, and the correspondence of the Czech and English POS tags. Furthermore, by training GIZA++ (Och and Ney, 2003) 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 domain of financial news specific for the Wall Street Journal.

The resulting Czech-English probabilistic dictionary contains 46,150 entry-translation pairs in its lemmatized version and 496,673 pairs of word forms in the version where for each entry-translation pair all the corresponding word form pairs have been generated.

6.3 Tools

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 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.

7 Conclusion

We have described the process of building the first version of a parallel treebank for two relatively distant languages, Czech and English, during which we have also attempted to reconcile two fairly incompatible linguistic theories used for their description.

The resulting data collection contains data syntactically annotated on several layers of analysis. There have already been experimental machine translation systems MAGENTA (Hajič et al., 2002) and DBMT (Čmejrek et al., 2003) confirming the exploitability of the corpus and showing that we are capable of performing automatic transformations from phrase structures to dependency representation with an acceptable, though still not impeccable quality.

However, for both languages, we have presented examples of phenomena, for which the "native" annotation scheme does not provide a sufficiently fine-grained analysis. In such cases, automatic conversion between annotation schemes is not possible, and the less we can hope for successfull machine translation.

The question of enhancing the annotation schemes to allow for a lossless transformation between them remains still open, and its difficulty presents a yet unfathomed depth.

8 Acknowledgements

This research was supported by the following grants: MŠMT ČR Grants No. LN00A063, No. MSM113200006, and NSF Grant No. IIS-0121285.

References

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

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

Michael Collins, 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.

Ulrich Germann, 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*, pages 228–235.

Jan Hajič and Barbora Hladká. 1998. Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In *Proceedings of COLING-ACL Conference*, pages 483–490, Montreal, Canada.

Jan Hajič, 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.

Ivona Kučerová 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.

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

I. Dan Melamed. 1996. A geometric approach to mapping bitext correspondence. In *Proceedings of the First Conference on Empirical Methods in Natural Language Processing*.

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

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

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

Martin Čmejrek, 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*, pages 83–90, Budapest, Hungary, April.

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*, volume V, pages 1513–1520, Las Palmas de Gran Canaria, Spain.

Fei Xia and Martha Palmer. 2001. Converting dependency structures to phrase structures. In *Proceedings of HLT 2001, First International Conference on Human Language Technology Research*, San Francisco.