Czech-English Dependency-based Machine Translation

Martin Čmejrek, Jan Cuřín, and Jiří Havelka
Institute of Formal and Applied Linguistics,
and Center for Computational Linguistics
Charles University in Prague
{cmejrek,curin,havelka}@ufal.mff.cuni.cz

Abstract

We present some preliminary results of a Czech-English translation system based on dependency trees. The fully automated process includes: morphological tagging, analytical and tectogrammatical parsing of Czech, tectogrammatical transfer based on lexical substitution using word-to-word translation dictionaries enhanced by the information from the English-Czech parallel corpus of WSJ, and a simple rule-based system for generation from English tectogrammatical representation. In the evaluation part, we compare results of the fully automated and the manually annotated processes of building the tectogrammatical representation.[1]

1 Introduction

The experiment described in this paper is an attempt to develop a full MT system based on dependency trees (DBMT). Dependency trees represent the sentence structure as concentrated around the verb and its valency. We use tectogrammatical dependency trees capturing the linguistic meaning of the sentence. In a tectogrammatical dependency tree, only autosemantic (lexical) words are represented as nodes, dependencies (edges) are labeled by tectogrammatical functors denoting the semantic roles, the information conveyed by auxiliary words is stored in attributes of the nodes. For details about the tectogrammatical representation see Hajičová et al. (2000), an example of a tectogrammatical tree can be found in Figure 3.

MAGENTA (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.

For processing the Czech part of the data, we reuse some modules of the MAGENTA system, but instead of MAGENTA's statistical tree-to-tree transducing module and subsequent language model, we implement a rule-based method for generating English output directly from the tectogrammatical representation.

First, we summarize resources available for the experiments (Section 2). Section 3 describes the automatic procedures used for the preparation of both training and testing data, including morphological tagging, and analytical and tectogrammatical parsing of Czech input. In Section 4 we describe the process of the filtering of dictionaries used in the transfer procedure (for its characterization, see Section 5). The generation process consisting mainly of word reordering and lexical insertions is explained in Section 6, an example illustrating the generation steps is presented in Sec-

[1] This research was supported by the following grants: MŠMT ČR Grant No. LN00A063 and NSF Grant No. IIS-0121285.

tion 7. For the evaluation of the results we use the BLEU score (Papineni et al., 2001). Section 8 compares translations generated from automatically built and manually annotated tectogrammatical representations. We also compare the results with the output generated by the statistical translation system GIZA++/ISI ReWrite Decoder (Al-Onaizan et al., 1999; Och and Ney, 2000; Germann et al., 2001), trained on the same parallel corpus.

2 Data Resources

2.1 The Prague Dependency Treebank

The Prague Dependency Treebank project (Böhmová et al., 2001) aims at complex annotation of a corpus containing about 1.8M word occurrences (about 80,000 running text sentences) in Czech. The annotation, which is based on dependency syntax, is carried out in three steps: morphological, analytical, and tectogrammatical. The first two have been finished so far, presently, there are about 18,000 sentences tectogrammatically annotated. See Hajič et al. (2001) and Hajičová et al. (2000) for details on analytical and on tectogrammatical annotation, respectively.

2.2 English to Czech translation of Penn Treebank

So far, there was no considerably large manually syntactically annotated English-Czech parallel corpus, so we decided to translate by human translators a part of an existing syntactically annotated English corpus (we chose articles from Wall Street Journal included in Penn Treebank 3), rather than to syntactically annotate existing English-Czech parallel texts. The translators were asked to translate each English sentence as a single Czech sentence and also to stick to the original sentence construction if possible. For the experiment, there were 11,189 WSJ sentences translated into Czech by human translators (see Table 1). This parallel corpus was split into three parts, namely *training*, *devtest* and *evaltest* parts.[2] The work on translations still continues, aiming at covering the whole Penn Treebank.

For both training and evaluation measured by BLEU metric, 490 sentences from devtest and evaltest data sets were retranslated back from Czech into English by 4 different translators (see an example of retranslations in Figure 2 and Section 8 for details on the evaluation).

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 the Czech sentences from devtest and evaltest data sets.

data category	#sentence pairs
training	10,699
devtest	242
evaltest	248

Table 1: Number of sentence pairs in English-Czech WSJ corpus

2.3 English Monolingual Corpus

The Penn Treebank data contain manually assigned morphological tags and this information substantially simplifies lemmatization. The lemmatization procedure searches a list of *triples* containing word form, morphological tag and lemma, extracted from a large corpus. It looks for a triple with a matching word form and morphological tag, and chooses the lemma from this triple. The large corpus of English[3] used in this experiment was automatically morphologically tagged by MXPOST tagger (Ratnaparkhi, 1996) and lemmatized by the *morpha* tool (Minnen et al., 2001), and contains 365 million words in 13 million sentences.

[2] training data, heldout data for running tests, and data for the final evaluation, respectively

[3] It consists of English part of French-English Canadian Hansards corpus, English part of English-Czech Readers' Digest corpus, English part of English-Czech IBM corpus, Wall Street Journal (years 95, 96), L.A. Times/Wash. Post (May 1994 – August 1997), Reuters General News (April 1994 – December 1996), Reuters Financial News (April 1994 – December 1996).

3 Czech Data Processing

3.1 Morphological Tagging and Lemmatization

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.

3.2 Analytical Parsing

The analytical parsing of Czech runs in two steps: the statistical dependency parser, which creates the structure of a dependency tree, and a classifier assigning analytical functors. We carried out two parallel experiments with two parsers available for Czech, parser I (Hajič et al., 1998) and parser II (Charniak, 1999). In the second step, we used a module for automatic analytical functor assignment (Žabokrtský et al., 2002).

3.3 Conversion into Tectogrammatical Representation

During the tectogrammatical parsing of Czech, the analytical tree structure is converted into the tectogrammatical one. These automatic transformations are based on linguistic rules (Böhmová, 2001). Subsequently, tectogrammatical functors are assigned by the C4.5 classifier (Žabokrtský et al., 2002).

4 Czech-English Word-to-Word Translation Dictionaries

4.1 Manual Dictionary Sources

There were three different sources of Czech-English manual dictionaries available, two of them were downloaded from the Web (WinGED, GNU/FDL), and one was extracted from the Czech and English EuroWordNet. See dictionary parameters in Table 2.

4.2 Dictionary Filtering

For a subsequent use of these dictionaries for a simple transfer from the Czech to the English tectogrammatical trees (see Section 5), a relatively huge number of possible translations for each entry[4] had to be filtered. The aim of the filtering is to exclude synonyms from the translation list, i.e. to choose one representative per meaning.

dictionary	*#entries*	*#transl*	*weight*
EuroWordNet	12,052	48,525	**3**
GNU/FDL	12,428	17,462	**3**
WinGED	16,296	39,769	**2**
merged	33,028	87,955	—

Table 2: Dictionary parameters and weights

First, all dictionaries are converted into a unified XML format and merged together preserving information about the source dictionary.

This merged dictionary consisting of entry/translation pairs (Czech entries and English translations in our case) is enriched by the following procedures:

- Frequencies of English word obtained from large English monolingual corpora are added to each translation. See description of the corpora in Section 2.3.
- Czech POS tag and stem are added to each entry using the Czech morphological analyzer (Hajič and Hladká, 1998).
- English POS tag is added to each translation. If there is more than one English POS tag obtained from the English morphological analyzer (Ratnaparkhi, 1996), the English POS tag is "disambiguated" according to the Czech POS in the appropriate entry/translation pair.

We select several relevant translations for each entry taking into account the sum of the weights of the source dictionaries (see dictionary weights in Table 2), the frequencies from English monolingual corpora, and the correspondence of the Czech and English POS tags.

4.3 Scoring Translations Using GIZA++

To make the dictionary more sensitive to a specific domain, which is in our case the domain of

[4] For example for WinGED dictionary it is 2.44 translations per entry in average, and excluding 1-1 entry/translation pairs even 4.51 translations/entry.

```
<e>zesílit<t>V

   [FSG]<tr>increase<trt>V<prob>0.327524
   [FSG]<tr>reinforce<trt>V<prob>0.280199
   [FSG]<tr>amplify<trt>V<prob>0.280198
   [G]<tr>re-enforce<trt>V<prob>0.0560397
   [G]<tr>reenforce<trt>V<prob>0.0560397

<e>výběr<t>N

   [FSG]<tr>choice<trt>N<prob>0.404815
   [FSG]<tr>selection<trt>N<prob>0.328721
   [G]<tr>option<trt>N<prob>0.0579416
   [G]<tr>digest<trt>N<prob>0.0547869
   [G]<tr>compilation<trt>N<prob>0.0547869
   []<tr>alternative<trt>N<prob>0.0519888
   []<tr>sample<trt>N<prob>0.0469601

<e>selekce<t>N

   [FSG]<tr>selection<trt>N<prob>0.542169
   [FSG]<tr>choice<trt>N<prob>0.457831

     [S] ... dictionary weight selection
     [G] ... GIZA++ selection
     [F] ... final selection
```

Figure 1: Sample of the Czech-English dictionary used for the transfer.

financial news, we created a probabilistic Czech-English dictionary by running GIZA++ training (translation models 1–4, see Och and Ney (2000)) on the training part of the English-Czech WSJ parallel corpus extended by the parallel corpus of entry/translation pairs from the manual dictionary. As a result, the entry/translation pairs seen in the parallel corpus of WSJ become more probable. For entry/translation pairs not seen in the parallel text, the probability distribution among translations is uniform. The translation is "GIZA++ selected" if its probability is higher than a threshold, which is in our case set to 0.10.

The final selection contains translations selected by both the dictionary and GIZA++ selectors. In addition, translations not covered by the original dictionary can be included into the final selection, if they were newly discovered in the parallel corpus by GIZA++ training and their probability is significant (higher than the most probable translation so far).

The translations from the final selection are used in the transfer. See sample of the dictionary in Figure 1.

5 Czech-English Lexical Transfer

In this step, tectogrammatical trees automatically created from Czech input text are transfered into "English" tectogrammatical trees. The transfer procedure itself is a lexical replacement of the tectogrammatical base form (*trlemma*) attribute of autosemantic nodes by its English equivalent found in the Czech-English probabilistic dictionary.

For practical reasons such as time efficiency, a simplified version, taking into account only the most probable translation, was used. Also 1–2 translations were handled as 1–1 — two words in one trlemma attribute.

Compare an example of a Czech tectogrammatical tree after the lexical transfer step (Figure 3), with the original English sentence in Figure 2.

6 Generating English Output

When generating from the tectogrammatical representation, two kinds of operations (although often interfering) have to be performed: lexical insertions and transformations modifying word order.

Since only autosemantic (lexical) words are represented in the tectogrammatical structure of the sentence, for a successful generation of English plain-text output, insertion of synsemantic (functional) words (such as prepositions, auxiliary verbs, and articles) is needed. Unlike in Czech, where different semantic roles are expressed by different cases, in English, it is both prepositions and word order that are used to convey their meaning.

In our implementation, the generation process consists of the following five consecutive groups of generation tasks:

1. determining contextual boundness

2. reordering of constituents

3. generation of verb forms

4. insertion of prepositions and articles

5. morphology

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.

R1: Kaufman & Broad, a company specializing in housing development, refused to give the names of their corporate investors.

R2: Kaufman & Broad, a firm specializing in apartment building, refused to list institutional investors.

R3: Kaufman & Broad, a firm specializing in housing construction, refused to name the institutional investors.

R4: Residential construction company Kaufman & Broad refused to name the institutional investors.

Figure 2: A sample English sentence from WSJ, its Czech translation, and four reference retranslations.

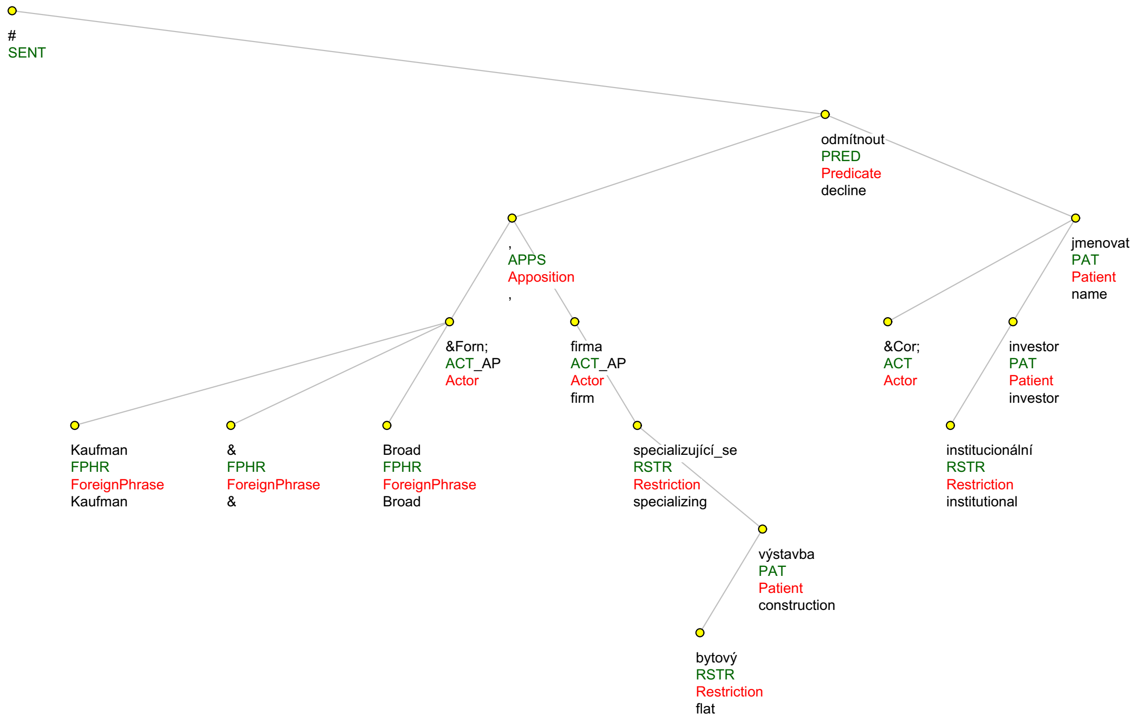

Figure 3: An example of a manually annotated Czech tectogrammatical tree with Czech lemmas, tectogrammatical functors, their glosses, and automatic word-to-word translations to English.

Cz	Kaufman & Broad		firma	specializující_se		bytový	výstavba	odmítnout		instit.	investor		jmenovat
0.	Kaufman & Broad		firm	specializing		flat	construction	decline		instit.	investor		name
1.	Kaufman & Broad		firm	specializing		flat	*construction*	decline		instit.	investor		*name*
2.	Kaufman & Broad		firm	specializing		flat	construction	decline		name		instit.	investor
3.	Kaufman & Broad		firm	specializing		flat	construction	decline	to	name		instit.	investor
4.	Kaufman & Broad	DEF	firm	specializing	INDEF	flat	construction	decline	to	name	DEF	instit.	investor
5.	Kaufman & Broad	the	firm	specializing	a	flat	construction	declined	to	name	the	instit.	investors

Figure 4: An illustration of the generation process for the resulting English sentence:
Kaufman & Broad, the firm specializing a flat construction declined to name the institutional investors.

In each of these steps, the whole tectogrammatical tree is traversed and rules pertaining to a particular group are applied. Considering the nature of the selected data, our system is limited to declarative sentences only.

Contextual boundness

Since neither the automatically created nor the manually annotated tectogrammatical trees capture topic–focus articulation (information structure), we make use of the fact that Czech is a language with a relatively high degree of word order freedom and uses mainly the left to right ordering to express the information structure. In written text, given (contextually bound) information tends to be placed at the beginning of the sentence, while new (contextually non-bound) information is expressed towards the end of the sentence. The degree of communicative dynamism increases from left to right, and the boundary between the contextually bound nodes on the left-hand side and the contextually non-bound nodes on the right-hand side is the verb. We consider information structure to be recursive in the dependency tree, and use it both for the reordering of constituents in the English counterpart of the Czech sentence, and for determining the definiteness of noun phrases in English.

Reordering of constituents

Unlike Czech, English is a language with quite a rigid SVO word order, therefore verb complements and adjuncts have to be rearranged in order to conform with the constraints of English grammar, according to the sentence modality. In the basic case of a simple declarative sentence, we place first the contextually bound adjuncts, then the subject, the verb, verb complements (such as direct and indirect objects), and contextually non-bound adjuncts, preserving the relative order of constituents in all these groups. The functors in a tectogrammatical tree denote the semantic roles of nodes. So we can use the contextual boundness/non-boundness of ACTor (deep subject), PATient (deep object), or ADDRessee, and realize the most contextually bound node as the surface subject.

Generation of verb forms

According to the semantic role selected as the subject of the verb, the active or passive voice of the verb is chosen. Categories such as tense and mood are taken over from the information stored in the Czech tectogrammatical node. Person is determined by agreement with the subject. Auxiliary verbs needed to create a complex verb form are inserted as separate children nodes of the lexical verb.

Insertion of prepositions and articles

The correspondence between tectogrammatical functors and auxiliary words is a complex task. In some cases, there is one predominant surface realization of the functor, but, unfortunately, in other cases, there are several possible surface realizations, none of them significantly dominant (mostly in cases of spatial and temporal adjuncts). For deciding on the appropriate surface realization of a preposition, both the original Czech preposition and the English lexical word being generated should be taken into account.

The task of generating articles in English is nontrivial and challenging due to the absence of articles in Czech. The first hint about what article should be used is the contextual boundness/nonboundness of a noun phrase. The definite article is inserted when the noun phrase is either contextually bound, postmodified, or is premodified by a superlative adjective or ordinal numeral. Otherwise, the indefinite article is used.

An article may be prevented from being inserted altogether in cases where uncountable or proper nouns are concerned, or the noun phrase is predetermined by some other means (such as possessive and demonstrative pronouns).

Morphology

When generating the surface word form, we are searching through the table of triples [word form, morphological tag, lemma] (see Section 2.3) for the word form corresponding to the given lemma and morphological tag. Should we fail in finding it, we generate the form using simple rules. Also, the appropriate form of the indefinite article is selected according to the immediately following word.

MT system	*BLEU – devtest*	*BLEU – evaltest*
DBMT with parser I	0.1857	0.1634
DBMT with parser II	0.1916	0.1705
DBMT on manually annotated trees	0.1974	0.1704
GIZA++ & ReWrite – plain text	0.0971	0.0590
GIZA++ & ReWrite – lemmatized	0.2222	0.2017
MAGENTA WS'02	0.0640	0.0420
Avg. BLEU score of human retranslations	—	0.5560

Table 3: BLEU score of different MT systems

7 An Example

Figure 4 illustrates the whole process of translating a sample Czech sentence, starting from its manually annotated tectogrammatical representation (Figure 3). The first line contains lemmas of the autosemantic words of the sample sentence from Figure 2. The next line, labeled 0, shows their word-to-word translations. The remaining lines correspond to the generation steps described in Section 6.

The order of nodes is used to determine their contextual boundness (line 1, contextually nonbound nodes are in italics). In line 2, the constituents are reordered according to contextual boundness and their tectogrammatical functors. The form of the complex verb is handled in step 3. In the next step, prepositions and articles are inserted. However, not every functor's realization can be reconstructed easily, as can be seen in the case of the missing preposition "in". It is also hard to decide whether a particular word was used in an uncountable sense (see the wrongly inserted indefinite article). The last line contains the final morphological realization of the sentence.

8 Evaluation of Results

We evaluated our translations with IBM's BLEU evaluation metric (Papineni et al., 2001), using the same evaluation method and reference retranslations that were used for evaluation at HLT Workshop 2002 at CLSP (Hajič et al., 2002). We used four reference retranslations of 490 sentences selected from the WSJ sections 22, 23, and 24, which were themselves used as the fifth reference. The evaluation method used is to hold out each reference in turn and evaluate it against the remaining four, averaging the five BLEU scores.

Table 3 shows final results of our system compared with GIZA++ and MAGENTA's results.

The DBMT with parser I and parser II experiments represent a fully automated translation, while the DBMT experiment on manually annotated trees generates from the Czech tectogrammatical trees prepared by human annotators.

For the purposes of comparison, GIZA++ statistical machine translation toolkit with the ReWrite decoder were customized to translate from Czech to English and two experiments with different configurations were performed. The first one takes the Czech plain text as the input, the second one translates from lemmatized Czech. In addition, the word-to-word dictionary described in Section 4 was added to the training data (every entry-translation pair as one sentence pair). The language model was trained on a large monolingual corpus of Wall Street Journal containing about 52M words. The corpus was selected from the corpus mentioned in Section 2.3.

We also present the score reached by the MAGENTA system.

All systems were evaluated against the same sets of references.

Both our experiments show a considerable improvement over MAGENTA's performance, they also score better than GIZA++/ReWrite trained on word forms. We were still outperformed by GIZA++/ReWrite trained on lemmas, but it makes use of a large language model.

9 Conclusion and Further Development

The system described comprises the whole way from the Czech plain-text sentence to the English one. It integrates the latest results in analytical and tectogrammatical parsing of Czech, experiments with existing word-to-word dictionaries combined with those automatically obtained from a parallel corpus, lexical transfer, and simple rule-based generation from the tectogrammatical representation.

In spite of certain known shortcomings of state-of-the-art parsers of Czech, we are convinced that the most significant improvement of our system can be achieved by further refining and broadening the coverage of structural transformations and lexical insertions. We consider allowing multiple translation possibilities and using additional sources of information relevant for surface realization of tectogrammatical functors. Finally, an integrated language model would discriminate the best of the hypotheses.

References

Yaser Al-Onaizan, 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. WS'99, Johns Hopkins University.

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2001. The Prague Dependency Treebank: Three-Level Annotation Scenario. In Anne Abeillé, editor, *Treebanks: Building and Using Syntactically Annotated Corpora*. Kluwer Academic Publishers.

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.

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, 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.

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. English translation of the original Czech version, `http://shadow.ms.mff.cuni.cz/pdt/Corpora/PDT_1.0/References/aman_en.pdf%`.

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. WS'02, Johns Hopkins University — in preparation.

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.

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

F. J. Och and H. Ney. 2000. Improved statistical alignment models. In *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447, Hongkong, China, October.

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

Adwait Ratnaparkhi. 1996. A maximum entropy part-of-speech tagger. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 133–142, University of Pennsylvania, May. ACL.

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