An MT System Recycled

Ondřej Bojar, Petr Homola, Vladislav Kuboň
Institute of Formal and Applied Linguistics
ÚFAL MFF UK, Malostranské náměstí 25, Praha 1, CZ-11800
Czech Republic
{bojar,homola,vk}@ufal.mff.cuni.cz

Abstract

This paper describes an attempt to recycle parts of the Czech-to-Russian machine translation system (MT) in the new Czech-to-English MT system. The paper descibes the overall architecture of the new system and the details of the modules which have been added. A special attention is paid to the problem of named entity recognition and to the method of automatic acquisition of lexico-syntactic information for the bilingual dictionary of the system. The paper also gives first results of an evaluation and it outlines the directions for further research.

1 Introduction

A huge effort has been invested into various MT systems in the past, but the results were not always satisfactory. There are of course exceptions, the history knows several examples of successful systems, both commercial or experimental ones, but the number of those systems, which consummated great amount of human labor and funds and which were silently abandoned when it turned out that the results are less than satisfactory, is much greater. This statement applies especially to classic rule-based MT systems, which require an enormous investment into grammar and dictionary development before they can provide results of acceptable quality. It is not a coincidence that the development of systems like Systran took decades - increasing the quality is a long process of enlarging the lexicons, debugging the grammar and tuning the system.

The last decade has witnessed several attempts to increase the quality of MT systems by introducing new methods. The strong stress on stochastic methods in the NLP in general and in the MT in particular, the attempts to develop hybrid systems, a wide acceptance of translation-memory based systems among the translation professionals, the aim at limited domain speech-to-speech translation systems, all these (and many other) trends have demonstrated encouraging results in recent years. Developing and using new methods definitely moves the whole MT field forward, but one should not forget about all the effort invested into the old systems. Reusing at least some parts of those systems may help to decrease the costs of new systems, especially when one of the languages is not a "big" language and therefore there is not such a wide range of tools, grammars, dictionaries available as for example for English, German, Japanese or Spanish. In this paper we would like to describe one such attempt to reuse the existing system for a new language pair.

2 RUSLAN

One of the systems which was silently abandoned in early nineties was the system for the translation from Czech to Russian called RUSLAN (Oliva, 1989). It was being developed in the second half of eighties as a joint project between the Faculty of Mathematics and Physics and the Research Institute of Mathematical Machines in Prague. It was oriented on a translation of a relatively closed thematic domain, the domain of operating systems of mainframes.

The system used transfer based architecture, although it claimed that for related languages (Czech and Russian) the size of the transfer module would be minimal due to the syntactic similarity of both languages. This claim turned out to be false, the testing and debugging of the system clearly showed that a number of differences between both languages had to be addressed by special transfer rules.

The implementation of the system was almost completely done in Q-systems, a formalism created by Alain Colmerauer (Colmerauer, 1969). Q-systems are a chart-parser-like formalism, it was for the first time successfully used in the machine translation system TAUM-METEO. Q-systems allow dividing the grammar into modules, where the output of a previous module serves immediately as the input for the following module. Each of the modules consists of a set of rules that in principle describe transformations of tree structures.Although considered to be outdated by many specialists, the Q-systems proved to be a suitable tool for our practically oriented project - the grammar interpreter was fast and reliable enough even for a grammar covering a large part of the syntax of both languages.

Apart from the grammar, the system also relied upon a set of dictionaries containing all data exploited by individual modules of the system. Each lexical item in the main (bilingual) dictionary contained not only lexico-syntactic data (valency frames etc.), but also a set of semantic features which were primarily used for resolving syntactic ambiguities in the process of syntactic analysis. Because the systems RUSLAN aimed at the translation of a limited domain, the size of its dictionaries reached only approximately 8000 items. Apart from that, it contained a special module called a transducing dictionary, which was capable to translate about 2000 additional technical terms of Greco-Latin origin by means of an algorithm for a direct transcription of Czech word-forms into Russian ones (Bémová and Kuboň, 1990).

The work on the system RUSLAN has been terminated in 1990, in the final phase of system testing and debugging. The reason was quite simple - after the political changes in 1989 there was no more any commercial demand for Czech to Russian MT system.

3 Motivation

The demand for translation to and from English has grown dramatically during the years following the abandonment of the system RUSLAN. On the other hand, also the range of methods, tools and resources for MT has grown substantially. Several corpora were created for Czech, the most prominent ones being the morphologically annotated Czech National Corpus and syntactically annotated Prague Dependency Treebank. In 2002 we have started the work on the parallel bilingual Prague Czech English Dependency Treebank (PCEDT) (Cuřín et al., 2004), which contains about a half of the texts from PennTreebank 3 translated into Czech by native speakers. A large morphological dictionary of Czech has been developed (Hajič, 2001), allowing for a good quality morphological analysis of Czech, which has been tested in numerous commercial applications and scientific projects since then.

The last decade also witnessed the development of several stochastic parsers of Czech, which might be used for the source language analysis in the Czech-English MT projects. Although the quality of results of these parsers for Czech is way behind the results reported for English - the best results reported by Charniak and Collins (Collins et al., 1999) are close to 85% - it is possible to use these parsers for stochastic MT from Czech. One of the last stimuli for our endeavor to reuse parts of the RUSLAN were the experiments started in 2001 by M.Čmejrek a J.Cuřín. The project (described e.g. in (Čmejrek et al., 2003) aims at the fully automatic stochastic translation, including stochastic analysis of the source text to the tectogrammatical (deep syntactic) level. Although it originally aimed also at the stochastic transfer, the newer experiments replaced the stochastic transfer with a rule-based module. This move from the stochastic back to rule-based approach opened a question whether using a rule-based syntactic analysis created for an MT system would not be better than using a general purpose stochastic parser. The high number of correctly assigned edges may not necessarily mean bad analysis. For the MT it is more important to be able to cover continuous parts of input sentences with (sub)trees, allowing for a consistent translation of clauses. From our point of view, an MT system might serve as a certain testbed allowing the comparison of results of a manually created grammar with stochastic parsers. Such a comparison is difficult if it aims at really objective measuring due to the fact that the standard measures of the parser quality are very much tied to particular data and to a particular system of syntactic tagging used in treebanks.

The main motivation for our Czech-English MT experiment was to test several hypotheses. The most prominent of these hypotheses concerns the level, at which it is reasonable to perform the transfer. Due to the differences between both languages it is not sufficient to perform the transfer immediately after the morphological analysis or shallow parsing, as it has been done in the MT system eslko aiming at the translation between closely related (and similar) languages [cf (Hajič et al., 2003)]. On the other hand, it is a question whether the typological differences between Czech and English justify the transfer being performed at the tectogrammatical (deep syntactic) level, as it has been done by Cun and mejrek in their system. According to our opinion, the transfer at the surface syntactic level might be more appropriate. The main problem of the transfer at the tectogrammatical (deep syntactic) level is the low quality of results of parsing to the tectogrammatical level, which are much worse than the results of stochastic parsers parsing to the analytical (surface syntactic) level. All the results reported above belong to parsers aiming at the analytical (surface syntactic) level.

Last but not least, one of our aims was to develop a rule-based MT system with minimal possible costs, either reusing the existing modules or trying to use (semi)automatic methods whenever possible, concentrating on areas where using the human labor would be extremely expensive (for example building a large coverage bilingual dictionary, cf. the following paragraphs.)

4 Individual parts of the new system

The main goal of our project is to develop an experimental MT system for the translation of texts from the PCEDT from Czech to English. The system investigates the possibility of reusing the existing resources (grammar, dictionary) in order to decrease the development time. It also exploits the parallel bilingual corpus of syntactically annotated texts, although not as a direct learning material, more like an additional source of linguistic data especially for the dictionary development and for the testing of the system.

4.1 Morphological analysis

Apart from the original MT system RUSLAN and the bilingual PCEDT corpus we can also exploit the module of morphological analysis of Czech (Hajič, 2001). It covers almost the entire Czech language, with very few exceptions (it is estimated that it is able to handle about 800 000 lemmas). This module replaced the original limited coverage module of morphological analysis of Czech. It is very reliable, due to a really large coverage there are almost no unknown words in the whole PCEDT. The only problem was the incorporation of the new module into the system - the original module of syntactic analysis of Czech from the system RUSLAN was very closely bound to a dictionary lookup and to the morphological module. The new module also uses a different tagset.

4.2 Bilingual dictionary

We have already mentioned that the bilingual dictionary of the system RUSLAN contained approximately 8000 lexical items with a rich lexicon-syntactic information. We have originally assumed that the information contained in the dictionary might be transformed and reused in the new system, but this assumption turned to be false. Although the information contained in the original bilingual dictionary is extremely valuable for the module of syntactic analysis of Czech, we have decided to sacrifice this information. The mere 8000 lexical items constitute too small part of the new bilingual dictionary. We have decided to handle the dictionary in a uniform manner.

4.2.1 Reusing existing dictionaries

A lot of manual labour has been already invested to building large coverage translation dictionaries, so we wish to make use of it. However, there are no such dictionaries between Czech and English aimed at supporting MT systems. As a result, the available machine-readable dictionaries built mainly for a human user (such as WinGED[1] or Svoboda (2001))) suffer from important limitations:

- Sometimes, several variants of translation are combined in one entry[2].
- No clear annotation of meta-language is present, although the entries contain valuable morphological or syntactic information to some extent. (E.g. valency frames are encoded by means of rather inconsistent abbreviations in plain text: *accession to = vstoupení do* or *adjudge sb. to be guilty = uznat vinným koho.*)
- Usually, no morphological information is given along the entries, although the morphological information can be vital for correctly recognizing an occurrence of the entry in a text. See Figure 1 for an illustration.
- No syntactic information is available and no consitent rules have been adopted by the

[1] http://www.rewin.cz/

[2] Throughout the text, we use the term ENTRY as a synonym to translation pair, i.e. a pair of Czech and English expressions.

Noun and Noun/Adjective	English Translation
husa divoká	grey goose
kniha účetní†	account book
napětí dovolené†	permissible stress
→ disambiguate to Noun Adjective	
chyba měření	measurement error
plán prací†	schedule of operation
rozsah měření	range of measurement
→ disambiguate to Noun Noun	
Numeral/Verb and Noun	**English Translation**
tři prdele†	shitloads
pět švestek	one's duds*
→ disambiguate to Numeral Noun	
pět chválu	sing someone's praises
→ disambiguate to Verb Noun	

† These expressions allow for the other interpretation, too, mostly kind of funny.
* Part of the idiom *pick up one's duds*.

Figure 1: Examples of morphological ambiguity in translation dictionaries.

lexicographers to annotate syntactic properties in plain text (such as putting the head of the clause as the first word).

From the point of view of structural machine translation, the lack of syntactic information in the translation dictionary is crucial. In the course of translation, the input sentence is syntactically analyzed before searching for foreign language equivalents. In order to check for presence of multi-word expressions in the input, the dictionary must encode the structural shape of such entries, otherwise the system does not know how to traverse the relevant part of the tree. Similarly, some expressions require some constraints to be met (such as an agreement in case or number) in the input text. If these constraints are not fulfilled, the proposed foreign language equivalent is not applicable.

The importance of valency (subcategorization) frames and their equivalents should be stressed, too. In the described system, already the syntactic analyzer requires verb and adjective valency frames in order to allow for specific syntactic constructions. In general, knowledge of translation equivalents of valencies is important to preserve the meaning (*přijít na nějaký nápad = come at an idea*, literal translation: *come on an idea*; *chodit na housle = attend violin lessons*, lit. *walk on violin*) or to handle auxiliary words properly (*čekat na něhoko = wait for somebody*, lit. *wait on sb.*; *říci něco = tell something* but *přejet něco = run over something*).

4.2.2 Cleaning the dictionary up

In order to handle the problems mentioned above, we performed an extensive cleanup of the data from available machine-readable dictionaries. The core steps of the cleanup are as follows:

Identifying meta-information.

We manually processed all the entries and searched for frequent words that typically encode some meta-information, such as *sth.*, *st.*, *oneself*. We also checked all entries ending with a word that is potentially a preposition. Based on the expression in the other language, we were able to recognize the meaning and identify, whether the suspicious word expresses a "slot" in the expression or whether it is a fixed part of the expression. (E.g. *mít o sobě vysoké mínění = think something of oneself*, only the word *oneself* encodes a slot, the word *something* is a fixed part of the expression.)

During this phase, entries encoding several translation variants at once were disassembled into separate translation pairs, too.

Part-of-speech disambiguation.

We processed the Czech part of each entry with a morphological analyzer ((Hajič, 2001)) and we performed manual part-of-speech disambiguation of expressions with ambiguity. It should be noted that automatic tagging would not provide us with satisfactory results due to the lack of sentential context around the expressions.

The manual disambiguation was carried out in blocks, each block contained all the expressions bearing the same kind of POS ambiguity. This allowed us to proceed the most of the expressions relatively quickly. However, as we demonstrated in Figure 1, it does happen that different annotations are required for expressions in the same block, so there is no way to completely avoid reading the entries.

Adding morphological constraints.

Morphological constraints on word entries describe which values of morphological features are required to hold for each word of the entry or have to be shared among a few words of the entry. Once identified, morphological constraints can be used to check whether a word group in input text represents an entry or not. With respect to our final task (translation from Czech to English), we aim at discovering Czech constraints only.

We decided to induce morphological constraints automatically, based on corpus examples of the entries. For each entry, we look up sentences that contain all the lemmas of the entry in a close neighbourhood (but irrespective to the word order and some extra words inserted). We weight the instances to promote those with no intervening words and those with connected dependency graph[3]. The list of weighted instances is scanned for pre-defined constraints both unary (such as "case is accusative", "number is singular") and binary ("the case of the first and second words match") selecting those constraints that are satisfied by at least 75% of total weight.

Despite the obvious simplicity of the algorithm and known limitations of accuracy of parsers and taggers used for corpus preparation, we are quite satistifed with the results. Most of the expressions with at least 10 corpus instances obtain a valid set of constraints, see Figure 2 for an illustration. Only expressions containing very common words (so that the words do appear quite often close together without actually forming the expression) obtain too weak constraints. For instance, no case and gender agreement constraints are selected for the expression *bohatý člověk (wealthy man).*

Adding syntactic information.

Syntactic information (dependency relations among words in the expression) is needed mainly during the analysis of input sentences, therefore we focussed on adding the information to the Czech part of entries first. For most of the entries, it was possible to add the dependency structure manually, based on the part-of-speech pattern of the entry. For instance all the entries containing an adjective followed by a noun get the same structure: the noun governs the preceeding adjective. For the remaining entries (with very varied POS patterns), we employ a corpus based search similar to the automatic procedure of identifying morphological constraints.

4.3 Named entity recognition module

A relatively independent module handles idiomatic constructions, named entities and terminological units. No such module existed in the original MT system, the nature of texts was very different from the nature of texts from PCEDT and didn't contain that many named

[3]When using a treebank as the input corpus, dependency graphs are available, when using plain corpus, we first employ a Czech adaptation of a parser by (Charniak, 2000).

Expression	Unary Constraints Binary Constraints
za nízkou cenu at low cost/price	RR–4 AAF** NNF*4 cng:2=3
v jediném dnu on one day	RR–6 AA*S* NNIS* num:2=3
v jiném směru to other direction	RR–6 AAI** NNIS6 case:1=3 num:2=3 gend:2=3
v jiném stavu in the family way	RR–6 AA*S* NNIS*
v jistém smyslu in certain sense	RR–6 AAIS6 NNIS6 case:1=2 case:1=3 cng:2=3
bohatý člověk wealthy man	AA*** NNM** num:1=2
získané informace gained information	AAFP* NNFP* cng:1=2
zkušební provoz trial operation	AAIS* NNIS* cng:1=2
první světová válka first world war	CrFS* AAFS* NNFS* cng:1=2 cng:1=3 cng:2=3

Unary constraints are expressed as Czech positional morphological tags with a wildcard character "*" at places where the value is context-dependent (i.e. where no unary constraint holds).
Binary constraints indicate which attributes have to be shared between some words of the expression. E.g., "cng:2=3" denotes that the case, number and gender have to match between the word 2 and word 3.

Figure 2: Examples of automatically selected morphological constraints.

entities or idioms. As mentioned above, the PCEDT contains a translation of texts from the Wall Street Journal section of the PennTreebank, in which the named entities, terminological units and idioms occur rather very often.

Named entities (NE) are atomic units such as proper names, temporal expressions (e.g., dates) and quantities (e.g., monetary expressions). They occur quite often in various texts and carry important information. Hence, proper analysis of NEs and their translation has an enormous impact on MT quality (Babych and Hartley, 2004).

NE translation involves both semantic translation and phonetic transliteration. Each type of NE is handled in a different way. For instance, person names do not undergo semantic translation (only transliteration is required), while certain titles and part of names do (e.g., *první dáma Laura Bushová* → *first lady Laura Bush*). In case of organizations, application of regular transfer rules for NPs seems to be sufficient (e.g., *Ústav formální a aplikované lingvistiky* → *Institute of formal and applied linguistics*), although an idiomatic translation may be probably preferable sometimes. With respect to geographical places we apply bilingual glossaries and a set of regular transfer rules as well.

For NE-recognition, we have developed a grammar based on regular expressions that processes typed feature structures. The grammar framework, similarly as the formally a bit weaker platform SProUT (Bering et al., 2003), uses finite-state techniques and unification, i.e., a grammar consists of pattern/action rules, where the left-hand side is a regular expression over typed feature structures (TFS) with variables, representing the recognition pattern, and the right-hand side is a TFS specification of the output structure. The type hierarchy is defined globally, a simple example is given in Figure 3.

The NE grammar is based on the experiment described in (Piskorski et al., 2004). An example of a simple rule is:

```
#subst[LEMMA: ministerstvo]$s1
+ #top[CASE: gen, PHRASE: $phr]$s2
== $s1#ministry[ATTR: $s2,
PHRASE: &('ministerstvo ' + phr)]
```
(1)

The first TFS matches any morphological variant of the word *ministerstvo* (ministry), followed by a genitive NP. The variables $s1, $s2 and $phr create dynamic value assignments and allow to transport these values to the slots in the output structure of type *ministry*. The output structure contains a new attribute called PHRASE with the lemmatized value of the whole phrase.

If the input phrase is

informace ministerstva zahraničí
o cestování do ohrožených oblastí (2)

then the phrase "ministerstva zahraničí" will be recognized as a NE and handled as an atomic unit in the whole MT process:

$$\begin{bmatrix} \textit{ministry} & \\ \text{LEMMA} & \text{ministerstvo} \\ \text{FORM} & \text{ministerstva} \\ \text{PHRASE} & \text{ministerstvo zahraničí} \\ \text{ATTR} & \begin{bmatrix} \textit{subst} & \\ \text{LEMMA} & \text{zahraničí} \\ \text{PHRASE} & \text{zahraničí} \\ \text{FORM} & \text{zahraničí} \\ \text{CASE} & \text{gen} \\ \text{NUMBER} & \text{sg} \\ \text{GENDER} & \text{n} \end{bmatrix} \\ \text{CASE} & \text{gen} \\ \text{NUMBER} & \text{sg} \\ \text{GENDER} & \text{n} \end{bmatrix} \quad (3)$$

Lemmatization of NEs is crucial in the context of MT. However, it might pose a serious problem in case of languages with rich inflection due to structural ambiguities, e.g., internal bracketing of complex noun phrases might be difficult to analyze. The core of the framework is based on grammars that have been developed for the MT system Česílko (Hajič et al., 2003).

4.4 Syntactic analysis of Czech

Although we have originally assumed that the module of syntactic analysis of Czech will require only small modifications, it was necessary to include new grammar rules, for example the rules handling the huge number of numerals appearing in the source texts. The syntactic analysis is nevertheles the only module which has been fully incorporated into the new system.

4.5 Transfer

The main task of this module is to transform the syntactic structure (syntactic tree) of the input Czech sentence into the syntactic structure (tree) of the corresponding English sentence. The transfer module does not handle the translation of regularly translated lexical units, it is handled by the bilingual dictionary in the earlier phases of the system. The transfer concentrates on three main tasks:

- The transformation of the Czech syntactic tree into the English one reflecting the differences in the word order between both languages.
- The identification and translation of those constructions in Czech, which require specific (irregular) translation into English.
- The insertion of articles (which do not exist in Czech) into the target language sentences.

The delopmenet of this module still continues, the initial tests confirmed that a substantial improvement can be achieved in the future.

4.6 Syntactic synthesis of English

The syntactic synthesis of Russian in RUSLAN is very closely bound to transfer, therefore we have tried to use as big portion of the grammar as possible, but of course, substantial modifications of the grammar were necessary. As well as the work on the transfer module, also the work on this modules still continues.

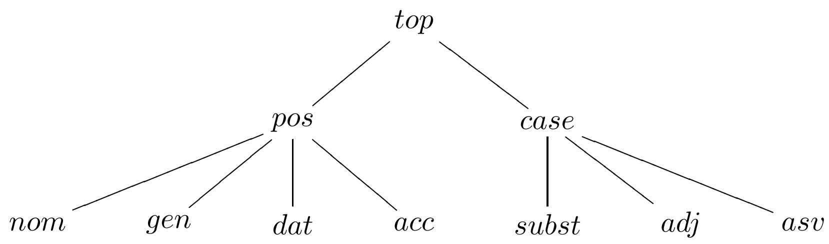

Figure 3: Example of a simple type hierarchy

4.7 Morphological synthesis of English

Due to the simplicity of English morphology this module has a very limited role in our system. It handles plurals, 3rd persons and irregular words.

5 The results of the translation

For the evaluation of results we are using a method presented in (Hajič et al., 2003). The method measures the similarity of an original text and the retranslated text. The similarity is measured by the TRADOS Translation Workbench. TRADOS evaluates the percentage of match between the original English text from the PennTreebank and the text translated by our system from Czech to English in the same manner as it normally evaluates the percentage of match of source text with sentences stored in a translation memory.

The paper (Hajič et al., 2003) mentioned that the commercial MT system PC Translator tested on the 256 sentences from the PCEDT has achieved a weighted average of 30% match. Our tests on a different (and smaller) set of sentences from the same source indicate slightly better results (close to 35%). It is nevertheles too early to draw any conlusions from the first tests, the system is currently still undergoing extensive testing and debugging.

6 Conclusion

Although the first tests have shown encouraging results, there are still many possible directions how to improve its behavior. Apart from the work carried on the existing modules there are at least three directions for future research. One of the big problems is the morphological ambiguity of individual Czech word forms. In average, there are more than four morphological tags per single word form. This can, of course, be solved by exploitation of a stochastic tagger, but we are intentionally trying to avoid this direction. The best taggers for Czech are currently reporting accuracy slightly over 95%, in other words, almost every single input sentence would contain a wrongly assigned tag. This fact may have dire consequences on the quality of the key module, the module of syntactic analysis of Czech. The direction we have decided to test goes towards partial, but error-free disambiguation of the results of morphological analysis of Czech.

Another way how to decrease the ambiguity is the exploitation of a special module resolving the lexical ambiguity in those cases when the bilingual dictionary provides more than one lexical equivalent. This stochastic module would exploit the context and would suggest the best translation.

The third direction of research in fact means that we might abandon the legacy of RUSLAN and to exploit one of the existing stochastic parsers of Czech instead of the rule-based grammar. Such an experiment might provide an answer to the crucial question - does it really pay off to recycle the old system or not?

References

B. Babych and A. Hartley. 2004. Selecting translation strategies in MT using automatic named entity recognition. In *Proceedings of the Ninth EAMT Workshop, Valetta, Malta.*

Alevtina Bémová and Vladislav Kuboň. 1990. Czech-to-Russian Transducing Dictionary. In *Proceedings of the 13th International Conference COLING, Vol. 3*, pages 314–316.

C. Bering, W. Drożdżyński, G. Erbach, C. Guasch, P. Homola, S. Lehmann, H. Li, H.-U. Krieger, J. Piskorski, U. Schaefer, A. Shimada, M. Siegel, F. Xu, and D. Ziegler-Eisele. 2003. Corpora and evaluation tools for multilingual named entity grammar development.

Eugene Charniak. 2000. A Maximum-Entropy-Inspired Parser. In *Proceedings of NAACL-2000*, pages 132–139, Seattle, Washington, USA, April.

Michael Collins, Jan Hajič, Eric Brill, Lance Ramshaw, and Christoph Tillmann. 1999. A

Statistical Parser of Czech. In *Proceedings of 37th ACL Conference*, pages 505–512, University of Maryland, College Park, USA.

Alain Colmerauer. 1969. Les Systemes Q ou un formalisme pour analyser et synthetiser des phrases sur ordinateur.

Jan Cuřín, Martin Čmejrek, Jiří Havelka, and Vladislav Kuboň. 2004. Building a Parallel Bilingual Syntactically Annotated Corpus. In *Proceedings of the 1st International Joint Conference on NLP*.

Jan Hajič. 2001. *Disambiguation of Rich Inflection - Computational Morphology of Czech*, volume I. Prague Karolinum, Charles University Press. 334 pp.

J. Hajič, P. Homola, and V. Kuboň. 2003. A simple multilingual machine translation system. In *In: Proceedings of the MT Summit IX*, New Orleans.

Karel Oliva. 1989. A Parser for Czech Implemented in Systems Q. *Explizite Beschreibung der Sprache und automatische Textbearbeitung*.

J. Piskorski, P. Homola, M. Marciniak, A. Mykowiecka, A. Przepiórkowski, and M. Woliński. 2004. Information extraction for Polish using the SProUT platform. In *Proceedings of the International IIS:IIP WM'04 Conference, Zakopane, Poland*.

Milan Svoboda. 2001. GNU/FDL English-Czech Dictionary. `http://slovnik.zcu.cz/`.

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.