Czech-English Machine Translation Dictionary

Final report for GAUK 351/2005 (years 2005-2006)

Ondřej Bojar and Magdalena Prokopová

April 17, 2007

Abstract

We are proposing a format for translation dictionaries suitable for machine translation. The dictionary format is concise and generalizes phrases by introducing rules for morphological generation instead of using simple phrase to phrase mapping.

We describe a simple way how to automatically construct our compact entries from a machine-readable dictionary originally intended for human users using parallel corpora. We further describe how to expand the compact dictionary entries to phrase table dictionary that can be used further on by machine translation systems (until the systems will support morphological generation from a translation dictionary natively). We performed manual annotation of a small set of entries to analyze problems of this approach.

1 Introduction

This report summarizes research results of grant GAUK 351/2005 conducted in the years 2005 and 2006. The aim of the grant was to design a format for Czech-English translation dictionary suitable for machine translation systems and to examine methods of reusing existing translation dictionaries to populate the format with translation entries.

Traditional translation dictionaries usually do not contain detailed morphological and syntactic information necessary in linguistically motivated approaches to MT. Full-featured MT dictionaries that are or were built mainly for rule-based MT systems (such as Hajič (1987) for Czech and Russian) are typically limited to a narrow domain and face low coverage problems.

Our approach is to enrich available Czech-English machine-readable translation dictionaries (such as WinGED (Win, 2003) or Svoboda (2001)) with the necessary information by scanning large automatically-annotated parallel text corpora.

2 Czech-English MT Dictionary Format

This section describes the format of our Czech-English MT dictionary.

The format of our dictionary was designed with knowledge of internal limitations of current top-performing MT systems. As documented by the NIST evaluation[1] for Chinese and Japanese translated to English, best results are currently obtained by statistical systems that make use of little or no linguistic information. Therefore, we decided to keep our dictionary reasonably close to what current systems can actually utilize. In future, as top-performing systems will explicitly handle of deeper linguistic information, we foresee deepening the annotation of our lexicon, along the lines proposed by the ISLE/MILE project (Calzolari et al., 2001).

2.1 Dictionary Format and Its Semantics

The Czech-English MT dictionary is a set of TRANSLATION PAIRS. Each TRANSLATION PAIR consists of Czech and English ENTRIES and a set of binary morphological constraints on the entries. Each ENTRY consists of a sequence of WORDS and binary morphological constraints over the words. WORDS are represented by the base form (lemma) and a set of allowed values for morphological features expressed by means of unary constraints. A UNARY CONSTRAINT is used to assign a fixed value to morphological feature of a word, see Section 2.4 for the details. A BINARY CONSTRAINT is used to bind together morphological features of two words, either both from the same language or one from the source and one from the target language (cross-lingual binary constraints). Binary constraints do not necessary assign a fixed value to a feature, they can merely express the requirement of some agreement between the features, see Section 2.5 for the details.

The semantics of the dictionary is as follows: given an input "phrase" (sequence of words in a sentence in the source language), the phrase can be translated to the target language using a matching translation pair from the dictionary. A translation pair is said to MATCH the input phrase, if the lemmas in the input phrase are equal to the lemmas in the source language entry in the specified order with no intervening words and if all unary and binary morphological constraints of the entry are satisfied by the actual values of morphological features of input words. Cross-lingual binary constraints transfer morphological properties from the source words to target words.

In theory, all the entries in a translation dictionary should be complete with respect to the set of constraints. Given a translation pair and any matching occurrence of the source entry, all phrases matching the target entry should be a valid translation of the source words (provided that all constraints expressed in the translation pair are satisfied). Situations where such a translation is not desired should be ruled out by additional constraints of the translation pair, e.g. by adding some more words from the neighbouring context or by adding further unary or binary constraints.

In practice, insisting on this hard interpretation is not feasible. First, the constraints would need some more expressive power in order to test for words from close or more distant neighbourhood and not just words participating in the entry. Second and more importantly, by making an entry "bullet-proof" against bad usage, we would inevitably make it too specific. For most sentences where a similar term occurs, this entry would not be applicable and we would need to add an enormous amount of entries to cover a reasonable percentage of input phrases. This difficulty is usually solved by relaxing the completeness

[1] http://www.nist.gov/speech/tests/mt/

requirement and keeping the set of constraints too permissive. Some relatively crude means is then used to decide which of the many possible translation options should be used. Typically, co-occurrence counts from domain-specific training corpus serve as a reasonable estimation.

2.2 Czech Morphology and Lemmatization

For Czech, a positional morphological tagset (Hajič, 2004b) has been well established in projects like the Czech National Corpus (Kocek et al., 2000) or the Prague Czech Dependency Treebank (Hajič, 2004a). Several tools for automatic tagging and lemmatization using this tagset exist, with Hajič and Hladká (1998) being the one most widely used. We adopt this format and the cited tool for automatic annotation.

2.3 English Morphology and Lemmatization

For English, the most widely used morphological (part-of-speech, POS) tagset was defined for the Penn Treebank (PTB, Marcus et al. (1993)) and there are many taggers that can automatically add tags to English plain text. In our case, we used the tagger MXPOST by Ratnaparkhi (1996).

For the purposes of unified formulation of morphological constraints in Czech and English (Sections 2.4 and 2.5 below), we defined a positional English tagset that maps 1-to-1 to the PTB tagset.

Table 1 lists the complete PTB tagset with our positional equivalent and explanation.

Due to a limited morphological variance, less attention has been paid to automatic lemmatization of English and English taggers in general do not provide this information. Therefore, we used a separate tool for English lemmatization, the Morpha tool by Minnen et al. (2001).

2.4 Unary Constraints

A unary constraint on a word in an entry expresses that the entry can be meaningfully and correctly used only if the constrained morphological feature of the input word bears one of the allowed values. Currently, only one type of unary constraints can be used: the constrained feature must have a constant value.

The use of a positional tag system allows us to encode all unary constraints on a word as a simple wildcard expression. For constrained features, the required value is expressed as a single character, for unconstrained features, we use the character '*'.

For instance the Czech entry for *black cat* could be encoded as a two-word entry:

Lemmas	**Unary Constraints**
černý kočka	`AAF** NNF**`

The morphological constraints express that the word *černý* must be used as an adjective in feminine gender to be a valid component of this entry, and similarly for the word *kočka* (feminine noun). This rules out many word forms of the lemma *černý* that would be used for different genders.

PTB	Positional	Description	PTB	Positional	Description
#	Z:-----	Punctuation	NNS	NNXPX--	Noun, plural
$	Z:-----	Punctuation	PDT	Td-----	Predeterminer
''	Z:-----	Punctuation	POS	Ts-----	Possessive ending
(	Z:-----	Punctuation	PRP	PPXXX--	Personal pronoun*
)	Z:-----	Punctuation	PRP$	PSXXX--	Possessive pronoun*
,	Z:-----	Punctuation	RB	DD----1	Adverb
-LRB-	Z:-----	Punctuation	RBR	DD----2	Adverb, comparative
-RRB-	Z:-----	Punctuation	RBS	DD----3	Adverb, superlative
.	Z:-----	Punctuation	RP	TT-----	Particle
:	Z:-----	Punctuation	SYM	Z:-----	Symbol
CC	J^-----	Coordinating conjunction	TO	To-----	to
CD	C=-----	Cardinal number	UH	II-----	Interjection
DT	Th-X---	Determiner	VB	Vf-X-X-	Verb, base form
EX	Tt-----	Existential there	VBD	Ve-X-X-	Verb, past tense
FW	X@-----	Foreign word	VBG	Vg-X-X-	Verb, gerund or present participle
IN	Ti-----	Preposition or subordinating conjunction	VBN	Vp-X-X-	Verb, past participle
JJ	AAX-X-1	Adjective	VBP	VB-X-X-	Verb, non-3rd person singular present
JJR	AAX-X-2	Adjective, comparative	VBZ	VB-S-3-	Verb, 3rd person singular present
JJS	AAX-X-3	Adjective, superlative	WDT	Tw-X---	Wh-determiner
LS	Z,-----	List item marker	WP	PWXXX--	Wh-pronoun
MD	Vm-----	Modal	WP$	PxXXX--	Possessive wh-pronoun
NN	NNXSX--	Noun, singular or mass	WRB	Dv-----	Wh-adverb
NNP	NCXSX--	Proper noun, singular	‘‘	Z:-----	Punctuation
NNPS	NCXPX--	Proper noun, plural			

* For English pronouns we use the lemma information to replace the three undefined values (XXX) with gender (M, F, N or X), number (S, P or X) and case information (1, 4 or X) information.

Table 1: Penn Treebank POS tagset as produced by MXPOST, with our positional notation.

The Czech morphological system as defined by Hajič (2004b) defines aggregate values for some features. For instance the letter Y used as the value of gender means either masculine (M) or inanimatum (I). All the aggregate values defined in Zeman et al. (2005) are allowed in our unary constraints.

2.5 Binary Constraints

Binary constraints express agreement requirements between two words in the same language or one word in the source and one word in the target language.

We define the following binary constraints:

cCASE:X=Y	Czech words X and Y have to agree in CASE.
cNUM:X=Y	Czech words X and Y have to agree in NUMBER.
cGEND:X=Y	Czech words X and Y have to agree in GENDER.
cCNG:X=Y	Czech words X and Y have to agree in CASE, NUMBER and GENDER.
cPERS:X=Y	Czech words X and Y have to agree in PERSON.
ceNUM:X=Y	Czech word X and English word Y have to agree in NUMBER.

Please note that there are more ways to express equivalent set of constraints on a translation pair. For example, the constraint cCNG can be equivalently replaced by the three individual constraints `cCASE`, `cNUM` and `cGEND` spanning the same words. Similarly, if there are three words to agree in a feature value, it is not significant which pairwise agreements are used to express the requirement, as long as all the words are covered.

2.6 Compact Representation of the Dictionary

For the sake of simplicity and ease of use, the format of our dictionary is rather minimalistic.

The dictionary is stored in a plain-text file (UTF-8 encoded). Each line of the file represents one translation pair and contains five tab-delimited columns:

- Czech lemmas,
- Czech unary constraints,
- English lemmas,
- English unary constraints,
- both monolingual and cross-lingual binary constraints

The number of Czech lemmas must match the number of (space-delimited) wildcard patterns representing Czech unary constraints, and likewise for English. A few sample entries are given in Figure 1.

In principle, more fancy notation styles might be considered, e.g. each word represented as a feature structure expressing both the lemma and unary constraints.[2] The feature structures should be then stored in XML conforming the respective TEI guidelines[3]. However, we believe that before investing effort in more verbose notation, experiments and applications have to prove the utility of the concept as such.

3 Enriching Machine-Readable Dictionaries with Constraints

3.1 Available Dictionaries

A lot of manual labour has been invested in developing translation dictionaries for human users. Some of such Czech-English dictionaries are available in a machine-readable form, e.g. WinGED (Win, 2003) or Svoboda (2001) and possibly EuroWordNet[4].

Taking the dictionaries as a source of translation pairs lemmas, we can automatically add the missing constrains on morphology. (None of the dictionaries contains all necessary morphological information explicitly encoded.)

[2] For Czech, this data format was used by Bojar (2002) and conversion tools to and from the positional tagset are ready.

[3] `http://www.tei-c.org/P4X/FD.html#FDFS`

[4] EuroWordNet (Pala and Smrž, 2004) can be used as a translation dictionary thanks to the interlingual links it contains. Unfortunately, the version available to us contained only about 10,000 nominal expressions and about 3,000 verbal expressions, an order of magnitude less than other dictionaries.

absence	NNF**
absence	NNX*X--
ceNUM:1=1	
akciový společnost	AAF** NNF**
company	NNX*X-
cCNG:1=2 ceNUM:1=1 ceNUM:2=1	
akt-1 násilí	NNI** NNNS2
act of violence	NNX*X-- Ti----- NNXSX-
ceNUM:1=1 ceNUM:2=3	
automatický převodovka	AAF** NNF**
automatic transmission	AAX-X-1 NNX*X-
cCNG:1=2 ceNUM:1=2 ceNUM:2=2	
betonový blok	AAIS7 NNIS7
concrete block	AAX-X-1 NNXSX-
cCNG:1=2 ceNUM:1=2 ceNUM:2=2	

Figure 1: Sample translation pairs in compact format.

The missing morphological information is not the only problem of machine-readable dictionaries. We noticed that many entries contain not only the translated phrases, but also additional remarks on usage, in a very irregular fashion. Before exploiting a dictionary for the phrases, one has to go through a tedious process of semi-manual clean-up of the entries. We described the clean-up in (Bojar, 2005).

3.2 Available Parallel Corpora

A good source of sample usage of translation pairs is a sentence-parallel bilingual corpus.

For Czech and English, the Prague Czech-English Dependency Treebank (PCEDT, (Čmejrek et al., 2004)) was already available but its domain is too specific. The core part of PCEDT are only economical texts originally from Wall Street Journal. PCEDT contains also a section of generic stories from Reader's Digest, but the size in total is still rather limited.

Therefore, we collected additional Czech-English parallel data and made them publicly available for research purposes as the CzEng parallel corpus (Bojar and Žabokrtský, 2006). CzEng contains approximately 1 million 1-to-1 aligned Czech and English sentences, which is 40-times bigger than the WSJ section of PCEDT.

3.3 A Method for Automatic Constrain Induction Outlined

Given a set of translation pair lemmas from a machine-readable dictionary and a parallel corpus, we can try to automatically induce unary and binary constraints to create a full-featured translation pair.

We make use of the Manatee corpus search engine (Rychlý and Smrž, 2004) to search our parallel corpus for sentences where all Czech lemmas occur in the

Czech side and all English lemmas occur in the English side. We do not require the lemmas to occur in the particular order specified by the translation pair and we allow intervening words. In order to increase accuracy of the method (at the expense of samples found), one could enforce such restrictions or at least increase the weight for occurrences where the lemmas are close to each other measured by linear distance or by the number of edges in an automatic dependency analysis of the sentence.

We collect all occurrences of the translation pair and check, how often is a unary or a binary constraint satisfied by the occurrence. We use a simple thresholding technique to determine the validity of a constraint: if more than say 70% of occurences satisfy a constraint, we include it to the translation pair representation.

Despite the simplicity of this approach and the amount of errors in the automatic tagging of our corpus, the set of induced constraints seems reasonable for translation pairs with more than e.g. 10 occurrences. The most problematic translation pairs are those containing very common words. Common words can often co-occur in a sentence pair in the parallel corpus by chance and do not exemplify the translation pair. Such occurrences naturally need not (and do not) satisfy any constraints. Confused by these false occurrences, our simple method tends to induce that no constraints does not need to be satisfied by the translation pair.

3.4 Comparing Manual and Automatically Induced Constraints

In order to analyze the quality of computer-generated constraints, we first manually annotated a sample of eighty translation pairs and then compared the manual and automatic constraints.

The sample translation pairs were selected randomly from our cleaned version of the dictionary to cover translation pairs of 30 different types with a high number of occurrences in total. By “translation pair type” we mean main part of speech of the words and also an estimated frequency rank of the translation pair. The selection thus contained sample translation pairs for various parts of speech and also of a varied number of occurrences (both low and high), provided that either some of the translation pairs of a given type are highly frequent or there are many different (less frequent) translation pairs of the given type. The selection thus represents mainly the kinds of translation pairs that would be used most frequently when translating real texts.

Two independent annotators were asked to provide all sample translation pairs with both unary and binary constraints. For cases where already the set of lemmas covered by the translation pair was not appropriate, annotators were given the possibility to mark the whole translation pair as invalid. To speed up manual annotation, a preliminary set of unary constraints was automatically suggested, the annotators thus had to make sure the unary constraints are valid. Binary constraints had to be constructed from scratch, no automatic suggestion was provided.

By comparing the two independent manual annotation we identified potential problems of our dictionary format. Inconsistencies between the two annotations can be divided into 3 categories – inconsistence in unary restrictions, inconsistence in binary restrictions and in marking whether the whole entry is valid. Since each entry was marked in several ways, only 17% of items were annotated in exactly the same way by the two annotators.

Annotators were given some freedom in using unary and binary constraints. For example, indicating that two words should be in plural form can be done either by using a unary constraint on each of the words or one unary and one binary constraint (to mark e.g. the first word as plural and to require an agreement between the number of the first and the second). All possible notations leading to the same set of possible word forms were taken into account when comparing annotation to ensure that entry is counted as correct even though each of annotators used different notation.

The manually annotated data sets served two goals: first, to identify possible problems of the dictionary format as such, and second to compare manual and automatically generated constraints. We report our observations for both the tasks together.

Unary Restrictions The main problem in this category was a different extent of generalization used by the annotators. For the given data set we found 11 differences in generalization for category detailed part-of-speech and similar amount of differences in other categories such as gender, number and case.

Binary Restrictions The most significant problem when looking at the two data sets from our two annotators was caused by verbs. One of our annotators decided to introduce new binary restrictions whilst the other one marked all verb entries as incorrect. We explore the problem of verbs in a greater detail below.

Entries Validity We have already mentioned problem of annotating verbs, also some other issues appeared.

Several prepositional entries showed disagreement in validity. Every preposition can be translated in several different ways which entirely depend on the context. Thus when an annotator is asked to add constraints to an entry consisting only of a preposition (and no context), he or she cannot tell whether the entry is right or wrong. We conclude that translation pairs should never cover a preposition only, prepositions need to be accompanied by words bearing some meaning to make a reasonable translation pair.

In total, it happened in 30% of cases that one annotator marked translation pair as incorrect and the other one as correct.

Verbs Comments from both of our annotators indicate that more appropriate rules or a different coding standard need to be defined for verb entries.

Verb forms and their usage are significantly different in Czech and English and therefore to make an entry valid more restrictions would be needed. Unlike English, Czech does not require any pronoun as a subject to a verb in a sentence. When translating from Czech to English using our format of dictionary, the translation entry for the Czech verb has to include a pronoun as a part of the corresponding English side otherwise the English output would not be grammatical.

Due to usage of auxiliary verbs translation pairs intended to represent verbs used in past or past perfect tenses would need to contain also a “have” for English and/or “být” for Czech.

With the current dictionary format, the downside of adding more specificity to translation pairs by adding all necessary constraints is striking. The size of the dictionary is growing without any information gain. In future research we plan to design a specific format for verbal entries.

3.5 Summary of Problems Identified

The format of the dictionary suits well for nominal or adjectival (multi-word) entries. There is not much added value for adverbial entries over plain pairs of strings. For verbs and idiomatic expressions, an extension to our format is to be searched for in order to achieve the same level of conciseness.

4 Expanding Entries

The majority of current statistical MT systems are phrase-based, they rely on a very simple notion of a “phrase table”. Phrase table consists of pairs of sequences of word forms, the left-hand side in the source language, the right-hand side in the target language. Phrase tables are extracted by automatic methods from parallel corpora. See e.g. Koehn (2004) or the documentation of the MT system Moses[5] for more examples and description of methods used to extract phrase tables.

4.1 Expansion Process

We developed a tool to transform our dictionary format into the phrase table format used by machine translation systems.

Expanding each entry has tree main steps:

1. Czech entry expansion: For expanding Czech entries we use the tool provided by Hajič (2004b) as part of the Czech “Free” Morphology package[6]. Along with the basic form of word this tool lets us specify all unary constraints by using wild card characters (asterisk and dot) in the morphological tag. By expanding each word from the Czech phrase and applying Czech-only binary constraints, we obtain the set of possible Czech phrases that can be generated from the Czech entry.

2. English entry expansion: In order to generate English word forms from the lemmas and morphological constraints contained in an English entry, we could use the complementary tool to Morpha by Minnen et al. (2000). Due to the introduction of our positional version of the English morphological tagset, we decided to use a simple morphological dictionary collected from the English side of our corpus instead. The dictionary is stored as a simple file containing all observed English word forms accompanied by their tag and lemma.

[5] http://www.statmt.org/moses/

[6] http://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Morphology/index.html

Given the following sample translation pair:

černý kočka   AAF** NNF**   black cat   AAX-X-1 NNX*X--   cCNG:1=2 ceNUM:1=2

the following items will be produced as part of the translation table:

černá kočka	AAFS1---------- NNFS1----------	black cat	AAX-X-1 NNXSX--
černé kočky	AAFS2---------- NNFS2----------	black cat	AAX-X-1 NNXSX--
černé kočce	. . .		
černé kočky	AAFP1---------- NNFP1----------	black cats	AAX-X-1 NNXPX--
černých koček	AAFP2---------- NNFP2----------	black cats	AAX-X-1 NNXPX--
černým kočkám	. . .		
. . .			

Figure 2: An example of automatic entry expansion.

Given an English entry, we retrieve all possible word forms for all the lemmas in the entry and then apply all unary and binary constraints to remove inappropriate forms.

3. Applying cross language binary constraints: As the last step, all possible combinations of Czech and English phrases generated from a translation pair are checked for cross language binary constraints and only those satisfying all the constraints are produced to the final phrase table.

In place of unary constraints the produced phrase table contains valid morphological tags for both Czech and English. An example is given in Figure 2.

5 Conclusion

This report summarizes the format of a Czech-English translation dictionary aimed at supporting machine translation.

We outlined and implemented a process of automatic enriching of data available in machine-readable dictionaries by information necessary for MT. On a sample of entries, we compared the information as provided by two human annotators and the information obtained automatically to learn entries of which types are represented and constructed satisfactorily and entries of which type will require specific treatment.

We implemented a tool that converts our dictionary format to the phrase-table format that can be directly used in statistical phrase-based systems.

We hope that the suggested format is simple but powerful enough to express entries useful for current MT systems. The real utility of our dictionary and the tools developed still has to be confirmed by extensive employment in MT systems.

6 Acknowledgement

The work on this project was supported by the grants GAUK 351/2005, Collegium Informaticum GAČR 201/05/H014 and MSM0021620838.

7 References

Ondřej Bojar and Zdeněk Žabokrtský. 2006. CzEng: Czech-English Parallel Corpus, Release version 0.5. *Prague Bulletin of Mathematical Linguistics*, 86:59–62.

Ondřej Bojar. 2002. Automatická extrakce lexikálně-syntaktických údajů z korpusu (Automatic extraction of lexico-syntactic information from corpora). Master's thesis, ÚFAL, MFF UK, Prague, Czech Republic. In Czech.

Ondřej Bojar. 2005. Budování česko-anglického slovníku pro strojový překlad. In Peter Vojtáš, editor, *ITAT 2005 Information Technologies – Applications and Theory*, pages 201–211, Košice, Slovakia, September. University of P. J. Šafařík.

Nicoletta Calzolari, Ralph Grishman, Partha Palmer, et al. 2001. The ISLE Survey of Main Approaches towards Bilingual and Monolingual Lexicons. Available at `http://www.ilc.cnr.it/EAGLES96/isle/clwg_doc/ISLE_D2.1-D3.1.zip`.

Martin Čmejrek, Jan Cuřín, Jiří Havelka, Jan Hajič, and Vladislav Kuboň. 2004. Prague Czech-English Dependecy Treebank: Syntactically Annotated Resources for Machine Translation. In *Proceedings of LREC 2004*, Lisbon, May 26–28.

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. 1987. RUSLAN: an MT system between closely related languages. In *Proceedings of the third conference on European chapter of the Association for Computational Linguistics*, pages 113–117. Association for Computational Linguistics.

Jan Hajič. 2004a. Complex Corpus Annotation: The Prague Dependency Treebank. Bratislava, Slovakia. Jazykovedný ústav Ľ. Štúra, SAV.

Jan Hajič. 2004b. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Nakladatelství Karolinum, Prague.

Jan Kocek, Marie Kopřivová, and Karel Kučera, editors. 2000. *Český národní korpus - úvod a příručka uživatele*. FF UK - ÚČNK, Praha.

Philipp Koehn. 2004. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In Robert E. Frederking and Kathryn Taylor, editors, *AMTA*, volume 3265 of *Lecture Notes in Computer Science*, pages 115–124. Springer.

M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19:313–330.

Guido Minnen, John Carroll, and Darren Pearce. 2000. Robust, applied morphological generation. In *INLG '00: Proceedings of the first international conference on Natural language generation*, pages 201–208, Morristown, NJ, USA. Association for Computational Linguistics.

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

Karel Pala and Pavel Smrž. 2004. Building Czech Wordnet. *ROMANIAN JOURNAL OF INFORMATION, SCIENCE AND TECHNOLOGY*, 7(1-2):79–88.

Adwait Ratnaparkhi. 1996. A Maximum Entropy Part-Of-Speech Tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*, University of Pennsylvania, May.

Pavel Rychlý and Pavel Smrž. 2004. Manatee, Bonito and Word Sketches for Czech. In *Proceedings of the Second International Conference on Corpus Linguisitcs*, pages 124–131.

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

2003. WinGED – Translation Dictionaries. `http://www.rewin.cz/`.

Dan Zeman, Jiří Hana, Hana Hanová, Jan Hajič, Barbora Hladká, and Emil Jeřábek. 2005. A Manual for Morphological Annotation, 2nd edition. Technical Report 27, ÚFAL MFF UK, Prague, Czech Republic.