Chapter 3: Working Notes on Exploring Parsing Resources from Bilingual Texts

Douglas Jones, Cynthia Kuo

Abstract

We conducted a preliminary survey of the prospects of using bilingual corpus to generate a grammar for monolingual parsing. The purpose of the survey was primarily educational. We wanted to learn about the current state of the art in parsing as applied to large-scale corpora. We examined the Czech Readers Digest corpus, which consists of around 2 million words of aligned Czech and English sentences and has been preprocessed in Prague. Since we had parsers available for both languages, we parsed both sides of the corpus and surveyed the parse structures. Although we did not have time over the summer for a large-scale project to automatically generate one grammar from the other, we felt that there were enough systematic structural correspondences to warrant further work in this direction. The purpose of this report is to describe some details about this very valuable bilingual corpus and to comment on the tools we used for our survey.

1. **Motivation**

   We conducted a preliminary survey of how we might use a bilingual corpus to generate a grammar for monolingual parsing. In particular, we examined the Czech *Readers Digest* corpus, which consists of around 2 million words of aligned Czech and English sentences. These sentences were taken from the Czech *Readers Digest* articles paired with the original English *Readers Digest* articles from which they were translated. The essential idea is that correspondences in the bitext imply structure for the two monolingual halves of the text. In the ideal case, we would like to infer the structure itself, based on the knowledge that the aligned sentences are translations of each other, or perhaps to parse the English side and use those structures to infer the Czech parses. In these cases, we would have a way to build up a new treebank, based in part on an analysis of the English side of the bilingual corpus. Furthermore, since we do have the Czech treebank, we could compare any new results with it as a reference point. Our somewhat more modest goal was to see to what extent the English parses and Czech parses correspond, given the corpus, parsers, and other resources that are available at the workshop. Since very large Czech and English treebanks are now available, we are able to train parsers to parse both sides of the corpus.

   Our original motivation for trying this experiment at the workshop is to provide a very different source of grammatical information for Eric's "Super Parser". Our expectation is that the Super Parser will make the best improvement over the individual parsers when the individual parsers behave very differently. Being able to infer the Czech parses from the English side of course would have met the criterion of being a very different source of grammatical information. Moreover, to the extent that such an exercise is possible, we would have a means of building trainable parsers for languages for which treebanks are not available, but for which bilingual texts are available.

2. **The Czech *Readers Digest* Corpus**

We started with the larger set of sentences: about 23 thousand sentences that were perfectly aligned and processed in Prague. We begin looking for isomorphic or nearly isomorphic parses. We also looked at small set of around 50 sentences by hand.

Some sample sentences are shown in Figure 1.

to je zvlá¹tní , pomyslel si .	that ’s strange , he thought .
sáhl na sklo a zjistil , ¾e chvìní mù¾e sice zmírnit , ale ¾e ho nezastaví úplnì .	feeling the glass , he discovered he could dampen the vibration but could n’t make it stop .
" vylep¹uju ho , " odpovìdìl les .	‘‘ i’m making it better , ’’ les responded

Figure 1. Sample of Aligned Sentences

The sentences in the aligned corpus were perfectly matched. Consequently, some of the text from the original *Readers Digest* text that did not match was left out. Also, the formatting had been removed and the text was in all lower case. This presented a problem because the parsers and taggers were trained on mixed-case text. However, since we also had access to the full original text, we were able to reconstruct the formatting information from the original. We felt this was easier than re-aligning the original text to get the full formatting.

As you can see in (2), the sentences corresponded quite closely in length. The mean sentence length for the English sentences is 17. Five words whereas the median Czech sentences is a little over 16 words. The English sentences were a little bit longer. The reason for that appeared to be that various English function words (such as particles, preposition, determiners, and so on) corresponded to morphological inflections in Czech.

	English length	Czech length	Absolute Difference
Mean	17.5	16.1	7.6%
Median	16	15	6.7%
Stdev	8.8	8.2	6.5%

Figure 2. Comparison of Sentence Lengths

3. Exploration of Data Space

We divided the data space into two main categories: correspondences, those that can be transformed and those that cannot. Of those that can be transformed, some small number will actually be isomorphic structures. The remainder is the transformable structures. Of those that cannot be transformed in a straightforward way, some will be because of the inherent freedom of

translation. Others will be because of parse errors or alternative analyses that are incompatible.

Recall that the Prague Dependency Treebank encodes dependency relations. The Penn Treebank, on the other hand, encodes constituent structures. We were able to produce both dependency structures and constituent structures for both the English and the Czech sides of our corpora, using the conversions developed at the workshop. For our initial survey, we looked at the dependency structures.

We will now step through examples of each type of data. Figure 3 illustrates what we mean by an isomorphic parse. What is required is that the topology of the parse be the same (of course) and that the nodes at each point correspond. The second half is not entirely trivial since the parts of speech do not correspond perfectly in Czech and in English. Figure 3 shows an example of an isomorphic parse for both sides of the corpus.

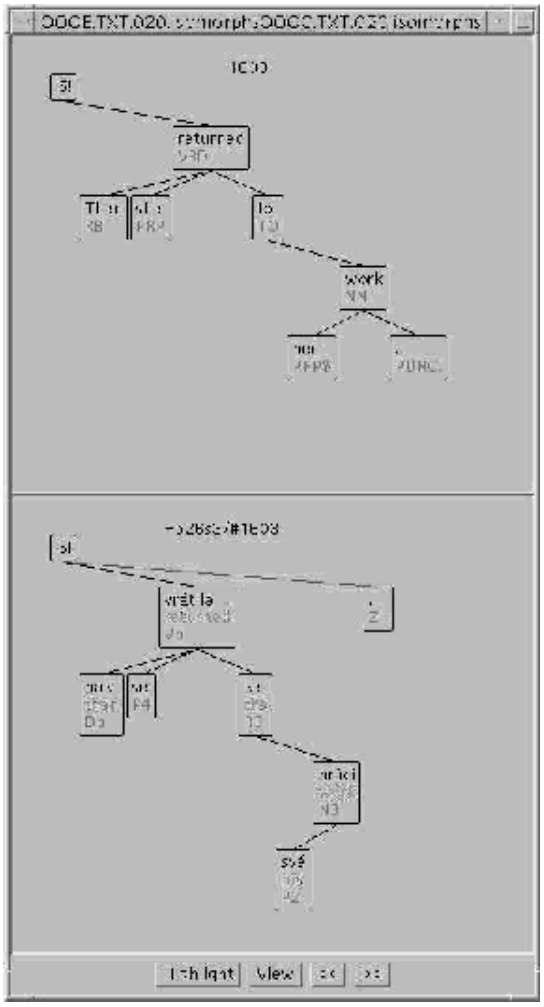

Figure 3. Isomorphic Parse

Transformable Parses

Naturally, isomorphic parses were very rare. We focussed most of our attention on parses that

appeared easily transformable. In Figure 4, the only problem is that in the Czech parse, the punctuation is attached high in the tree, whereas in the English parse, it is attached low. This does not reflect a linguistic difference, but rather, a difference in encoding schemes in the two treebanks. Regardless, this is the kind of parse that would be easy to transform.

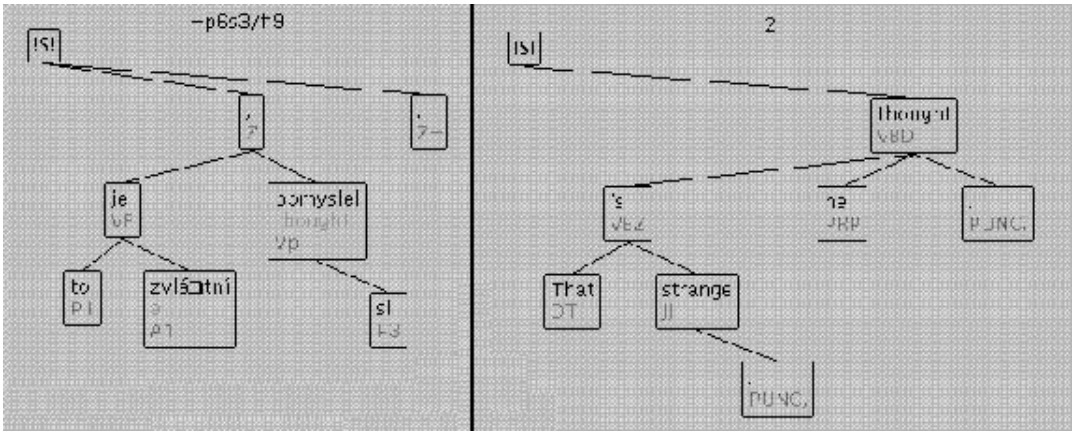

Figure 4 contains a sketch of some of the types of transformations that could be applied to create isomorphic correspondences.

Transformation	Applicable language	Reason
Paraphrasing / changing word order	Czech & English	Linguistic differences. In particular, Czech is a free word order language
Eliminating determiners	English	Czech does not use determiners, such as "the" or "a." Some determiners are, however, translated as pronouns in Czech.
Combining modal verbs and infinitives with main verb	English	Because of Czech's rich morphology, modal verbs and the infinitive "to" do not exist in Czech; the main verb is inflected.
Eliminating punctuation	Czech & English	Design differences between Penn Treebank and Prague Dependency Treebank. The Czech translations also contain added punctuation.
Skip prepositions	English	Where English uses a preposition, Czech may use a case marker on the noun (object of the preposition).
Dropping subjects	English	Czech sometimes drops the subject of a sentence, when the inflection on the main verb makes the subject clear.

Figure 4. Sketch of Transformation Types

Free Translations

Because of the inherent freedom in translation, some of the sentences correspond only very loosely. Figure N shows an example of such a "free" translation.

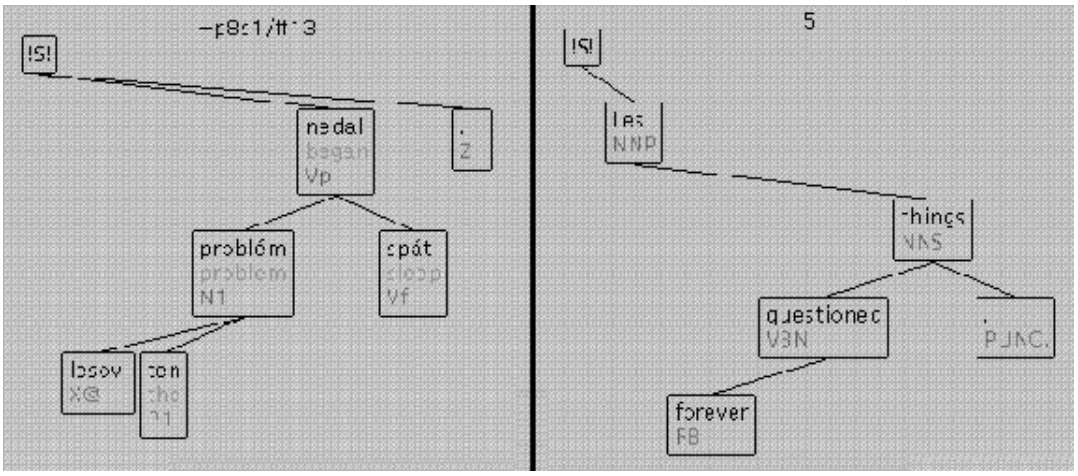

Figure 5. Free Translations not easily Transformable

Incompatibilities from Alternative Analyses

In some cases, the two treebanks simply encode relations differently. In the Penn Treebank, the category of a coordinate structure matches the elements coordinated. For example, the category of *Noun and Noun* is itself *Noun*. When the constituents are converted to dependency structures, the result is that the head of the phrase is one of the coordinates. In the Prague Dependency Treebank, on the other hand, the head of the coordinated structure is the coordinator. So the head of *Noun and Noun* is *and*, not *Noun*. Naturally, these structures do not match. We do expect them to be transformable.

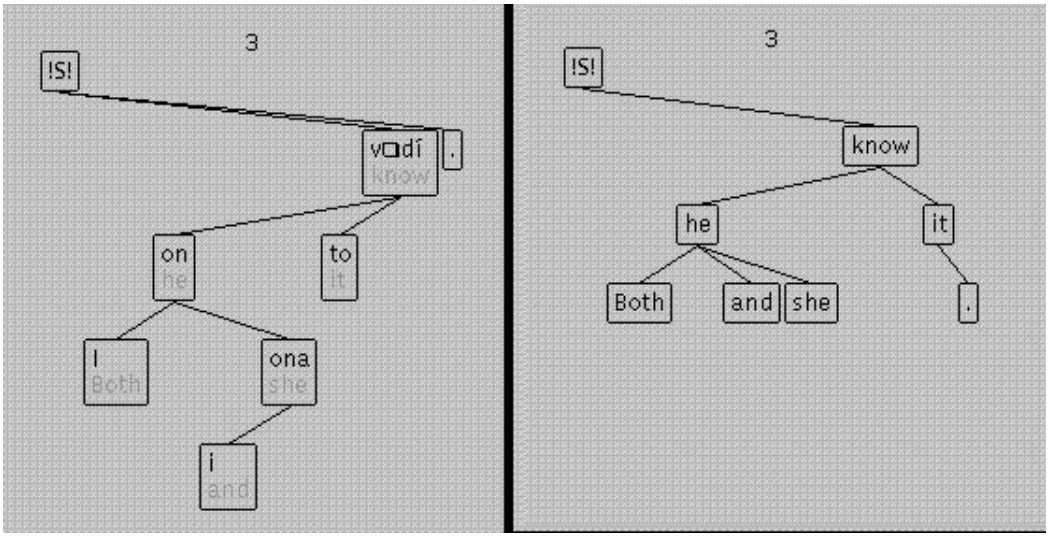

Figure 6. Alternative Analyses

Parse Errors

Since neither parser is perfect, we expect some of the mismatches in structure to be because of parse errors. The accuracy rate for Michael Collins's parser was 88% on the English data and 79% on the Czech data. The English sentence in Figure 7 is parsed incorrectly, possibly because of a tokenization error. *I* is tagged as a noun, when it should be a pronoun. Also, *it* and *better* should modify the verb *make*, not *respond*

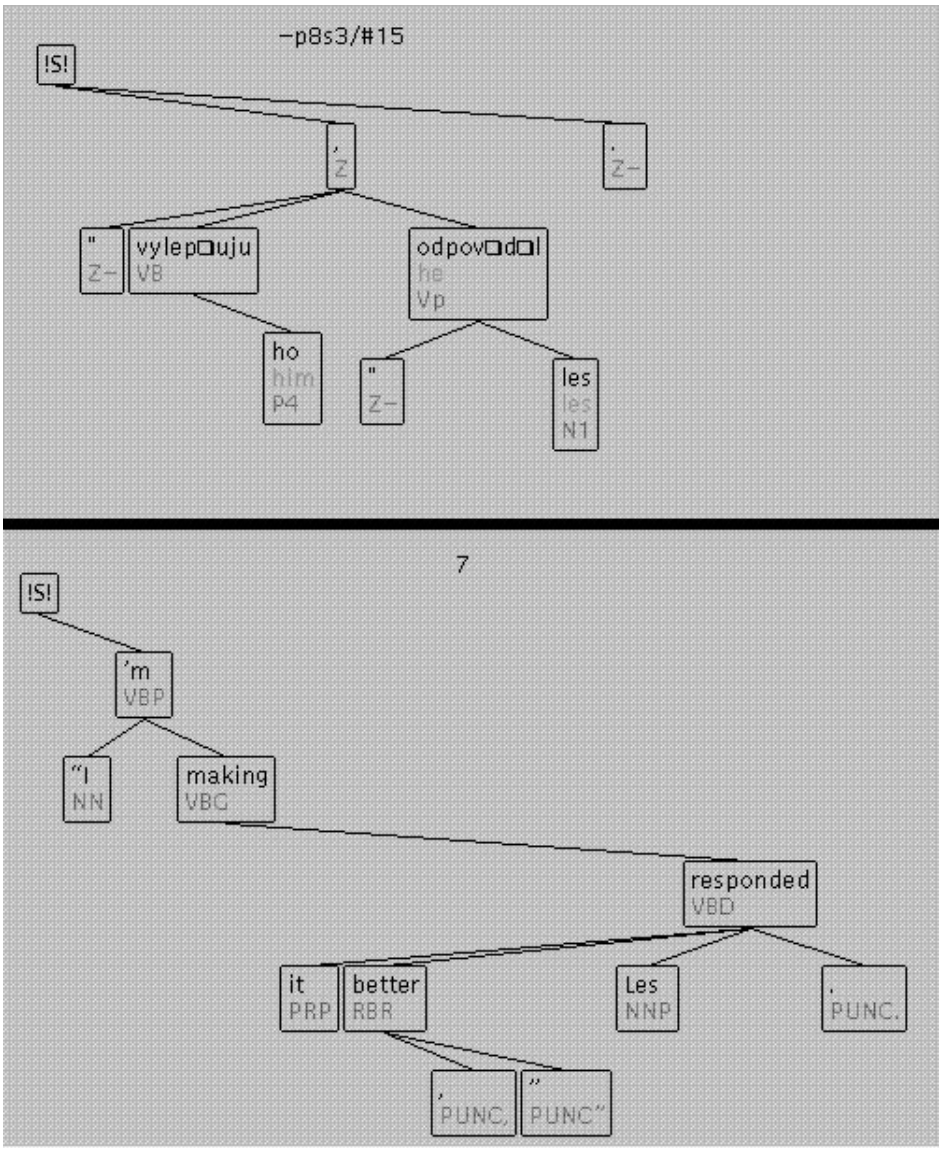

Figure 7. Parse Errors

4. Tools and Data Format

Preparing to conduct these experiments required some data formatting work to be done. Although substantial preprocessing for the text was already done in Prague, we still needed to get the English side into the proper format for parsing. Tagging and tokenizing presented some obstacles, since the aligned sentences had been pre-tokenized, but in a way that was

not completely compatible with what the Collins parser needed for input.

To process a portion of the corpus, we usually created a Unix Makefile (see bin\Makefile) to apply each process to the original. The Makefile would take care of the following processing:

Alignment

For the alignment, we used the two files E000.TXT and C000.TXT, which contained one pair of perfectly aligned sentences on each line. These sentences were in lower case.

Tokenization

The problem with the lower case aligned sentences was that the Collins parser had been trained on the Penn Treebank, which was in upper and lower case. Our solution was to restore the original formatting of the Readers Digest articles. We felt this was easier than re-aligning the text. After restoring the formatting, we re-tokenized and tagged the text with standard tools.

- Czech: bin\restore-cz-sgml.prl
- English: bin\restore-en-sentences.prl, bin\tokenize-en

Tagging

For the English side, we used Eric Brill's tagger to tag the text, and adjusted the output to match the nonterminal symbols from the Collins parser. For the Czech side, we used the tagger from the workshop.

- Czech: bin\tag-cz
- English: bin\tag-en

Parsing

We were then able to parse the English Readers Digest text with Model 2 of the Collins parser. We were able to produce both constituent trees and dependency trees for the parses.

- Czech: bin\parse-cz, bin\just-parse.prl
- English: bin\parse-en, bin\just-parse.prl

Viewing the Parses

We then used Oren Schwartz's viewer to look at the parses. (The viewer is available at http://www.clsp.jhu.edu/ws98/projects/nlp/doc/9807/PV.html)

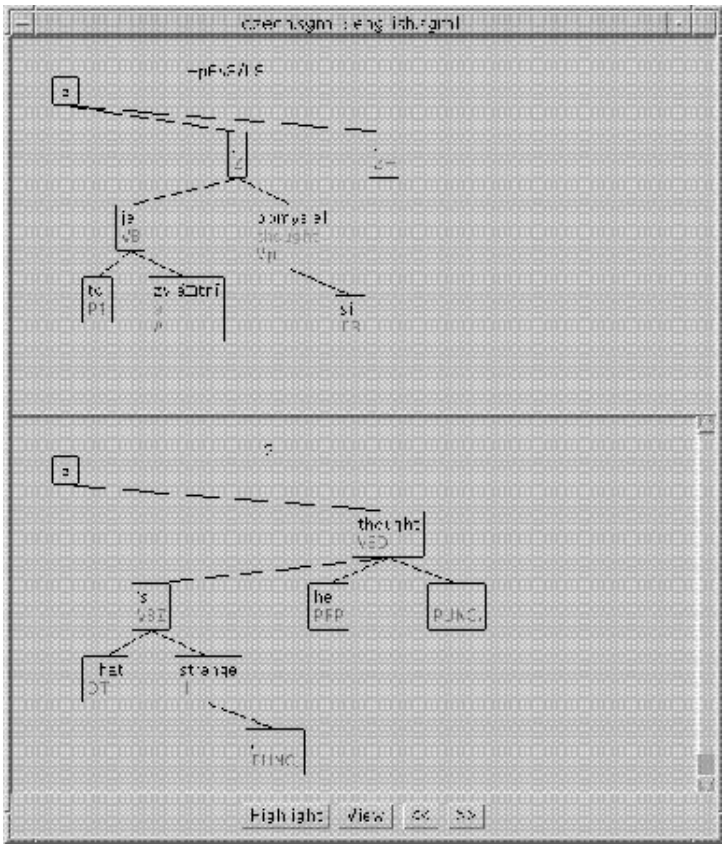

Figure 8. Oren Schwartz's Tree Viewer

We also used Radu Florian's tool for looking at the constituent structures: (shows partially matching trees).

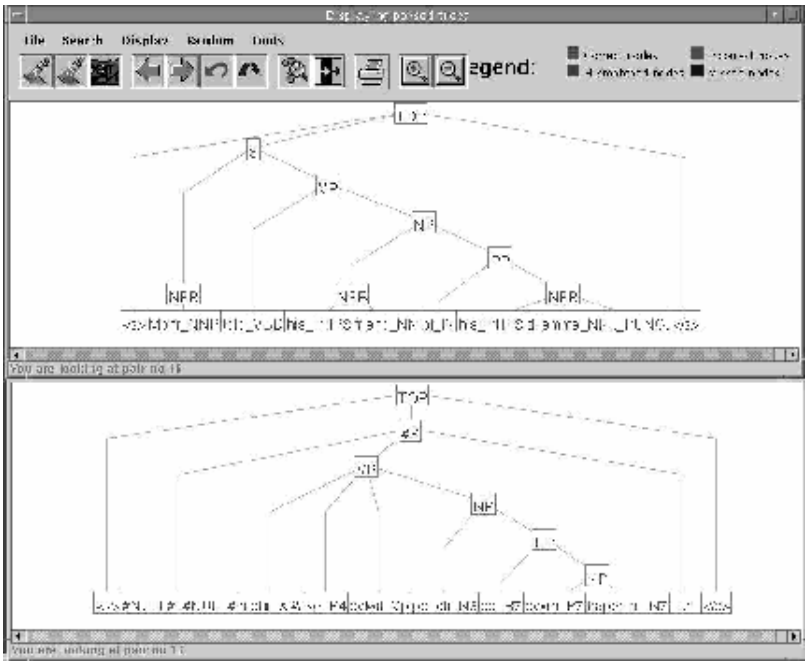

Figure 9. Radu Florian's Tree Viewer

Conversion between Dependency Structures and Constituent Structures

Once we had parses for the text, we used Lance Ramshaw's tools to convert these to sgml (see bin\tosgml.prl)

The English parsers had richer constituent structure than the Czech parses. The Collins parser was trained on constituents derived automatically from the Czech dependency trees.

Morphological Anchors

We then began looking for systematic correspondences, using the part of speech tags, for which there were relatively close affinities as well as a probabilistic bilingual word lexicon built in Prague from the Readers Digest Corpus, and began looking for ways to infer the Czech structure. The Czech data was tagged using the conventions set for the Prague Dependency Treebank. The English parsers were developed using the tags in the Penn Treebank. The following table truncates the tags on both sides and matches the basic parts of speech. We used these tag affinities to anchor nodes in the corresponding dependency trees.

Part of speech	Czech tag	English tag
adjective	A-	J-
adverb	D-	R-
conjunction	J-	CC, (IN)
determiner	--, (P-)	D-
existential (English "there is," "there was")	--	EX
interjection	I	INTJ
modal	(V-)	MD
noun	N-	N-
number	C-	CD
particle	T-	PRT
possessive	P-	POS, -$
preposition	R-, (J-)	IN
pronoun	P-	PRP
punctuation	Z-	PUNC
to (English word)	--, R-	TO
wh- word (who, how, etc)	P-	W-
verb	V-	V-
unknown, misc.	X-	F-, L-, S-, U-, etc.

Figure 10. Tag Affinities

A short sample of unigram frequencies for the part of speech tags is shown in Figure 11. These numbers were compiled from a set of aligned Reader's Digest sentences. The set contained sentences with a combined length (English sentence length + Czech sentence length) of twenty or less. Notice that there is a similar distribution of these tags.

POS	Czech	English
Conjunction	433	231
Determiner	--	892
Modal	--	173
Particle	74	4
Preposition	658	684
Pronoun	1297	1262
Unknown	531	22
Total number of	9317	11214

words

Figure 11. Unigram Frequencies of Tags for Function Words.

The unigram frequencies for all of the tags in the data set are shown in Figure 12. Notice the close correspondences for the major parts of speech - adjective, adverb, noun, preposition, pronoun, verb. The English sentences tend to be slightly longer than the Czech sentences. This is probably due to Czech's rich morphology; some words necessary in English are unnecessary in Czech. These results indicate a close match, word for word, between the Czech and English translations.

Part of speech	Czech tag	Number of occurrences	English tag	Number of occurrences
adjective	A-	655	J-	643
adverb	D-	883	R-	920
conjunction	J-	433	CC	231
determiner	--, (P-)	--	D-	892
existential (English "there is," "there was")	--	--	EX	33
interjection	I	3	INTJ	0
modal	(V-)	--	MD	173
noun	N-	2112	N-	2868
number	C-	197	CD	203
particle	T-	74	PRT	4
possessive	P-	[see pronoun]	POS, -$	75
preposition	R-	658	IN	684
pronoun	P-	1297	PRP	1262
punctuation	Z-	[removed]	PUNC	[removed]
to (English word)	--, R-	--	TO	211
wh- word (who, how, etc)	P-	[see pronoun]	W-	106
verb	V-	2420	V-	2442
unknown, misc.	X-	531	F-, L-, S-, U-, etc.	22
total number of words	--	**9317**	--	**11214**

Figure 12. Unigram Frequencies of All Tags.

Lexical Anchors

Although we found that the part of speech tags were quite reliable for finding matching structures, even when we did not have information about lexical correspondences, translations, etc, we did explore using

lexical anchors to identify correspondences. The English translations shown beneath the Czech words in the parse were inserted automatically using the probabilistic lexicon built by Cmejrek and Curin in Prague. We simply chose the word that was listed as most probable.

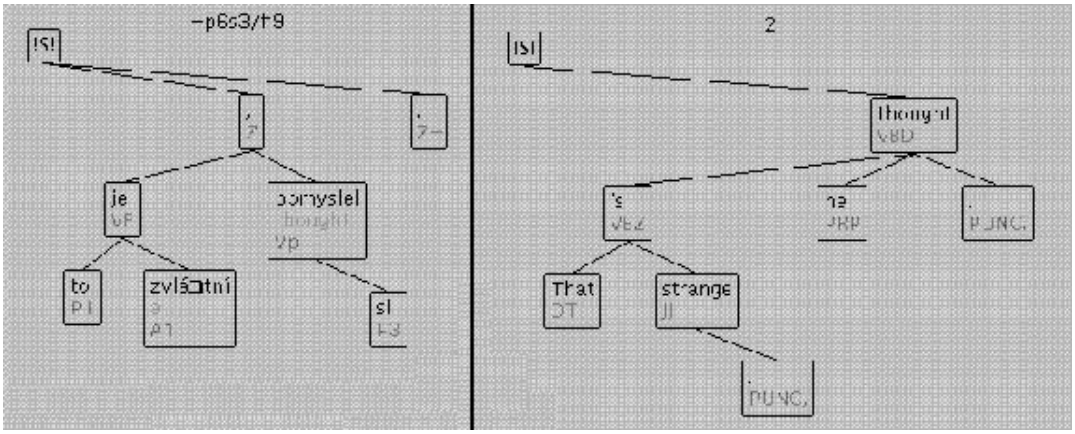

Figure 13. Lexical Anchors

The work by Cmejrek (1998) and Curin (1998) produced about 12,000 words with possible translations ranked in order of probability. Figure 14 shows the possible translations for *absurdní*, namely, *absurd*, *possible*, and *preposterous*. In this case, we are happy with the top choice for the lexical anchor.

absurdní	9	
3.636364e-01	4	absurd
1.818182e-01	54	possible
1.818182e-01	2	preposterous
9.090909e-02	23044	<EMPTY_WORD
9.090909e-02	7039	And

******** total 1.000000e+00 **********

Figure 14. Probabilistic Lexicon

Improving Lexical Anchors

However, the top-ranked choice was not always the best one. When we looked at these closely, we thought of two ways to improve the translation choice: the first was to use the evidence from Minimum Edit Distance (Levinshtein Distance) to identify cognates. Figure 15 shows how we could pick *auction* as the translation for *auk\350n\355*, skipping past the more probable (but incorrect) *at*. (Thanks to Christoph Tillman for helping us with the code for measuring the Levinshtein Distance).

Minimum Edit Distance

aukèní	at	2.50	
	auction	0.71	*****
	sale	1.5	
	sotheby's	1.0	

Figure 15. Minimum Edit Distance to Improve Guess

The other idea was to use semantic evidence to identify related translation possibilities, using WordNet. In Figure 16, we could use this evidence to accept *absurd* and *preposterous* as translations for *absurdni* because both *absurd* and *preposterous* are in the same synset in WordNet v1.6. We could then exclude the incorrect translation possible. Of course in this instance, the top choice given was correct. Nevertheless, we thought using WordNet might be a fruitful exercise.

		Same synset (Wn 1.6)
absurdní	absurd	*****
	possible	
	preposterous	*****

Figure 16. Semantic Evidence from WordNet to Improve Guess

1. Conclusion and Future Work

As we mentioned in the abstract and introduction, this survey was preliminary and its primary purpose was educational. The Czech Readers Digest corpus is clearly a very valuable resource for research in a variety of areas, including parsing and the automatic construction of resources for machine translation.

2. Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. (#IIS-9732388), and was carried out at the 1998 Workshop on Language Engineering, Center for Language and Speech Processing, Johns Hopkins University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation or The Johns Hopkins University.

3. References

1. [Charniak 1997]
   Eugene Charniak: *Statistical Techniques for Natural Language Parsing*. In: AI Magazine, Volume 18, No. 4. American Association for Artificial Intelligence, 1997.
2. [Cmejrek 1998]
   Martin Cmejrek: *Automatická extrakce dvojjazyèného pravdìpodobnostního slovníku z paralelních textù* (Master's Thesis on automatic extraction. Univerzita Karlova, Praha 1998.

3. [Collins 1996]
   Michael Collins: *A New Statistical Parser Based on Bigram Lexical Dependencies*. In: Proceedings of the 34th Annual Meeting of the ACL, Santa Cruz 1996.
4. [Collins 1997]
   Michael Collins: *Three Generative, Lexicalised Models for Statistical Parsing*. In: Proceedings of the 35th Annual Meeting of the ACL, Madrid 1997.
5. [Curín 1998]
   Jan Curín: Master's Thesis on automatic extraction of bilingual terminology. Univerzita Karlova, Praha 1998.
6. [Zeman 1998 a]
   Daniel Zeman: *A Statistical Approach to Parsing of Czech*. In: Prague Bulletin of Mathematical Linguistics 69. Univerzita Karlova, Praha 1998.
7. [Hajic 1998]
   Jan Hajic: *Building a Syntactically Annotated Corpus: The Prague Dependency Treebank*. In: Issues of Valency and Meaning, pp. 106-132 Karolinum, Charles University Press, Prague, 1998.
8. [Hajic and Ribarov 1997]
   Jan Hajic, Kiril Ribarov: *Rule-Based Dependencies*. In: proceedings of MLnet Workshop on Empirical Learning of NLP Tasks. Pages 125-135.