Me, in the Past Few Years
Properties of Czech
Czech-English Word Alignment and MT

Ondřej Bojar
obo@cuni.cz

October 27, 2005

Outline

- Background: Computer Science at Charles University Prague
- My master's (Diplomarbeit): Picking nice examples
- Properties of Czech, analysis of Czech
- Relevant Czech data
- My experiments here in Aachen

Background: Computer Science

Master Study at Charles University culminates with two (separate) tasks:

- Software Project
  Joint work of 3–6 students. Should take 1 year, never takes less than 1.5 or 2. The goal: experience team work on a large scale project, submit a usable piece of software.
- Master Thesis (Diplomarbeit)

Our Project: The Ents (2000–2002)

The Goal: A simulation of human-like environment (a family house) with user- and computer-controlled inhabitants (ents).

The Result:

- 6 students, 2 years (not so intensive work as performed here)
- a distributed (client-server) unix application
- $>$ 100,000 lines of code in C, C++, Pascal, Mercury, Perl
- 5000 lines of code in a new scripting language E
- 500 pages of documentation in Czech

My contribution: E scripts + NLP module implemented in Mercury:

- understanding definite descriptions of objects in the environment
- concretization – a process of further communication to identify an object uniquely

$\Rightarrow$ ents respond to commands in Czech

My Master's: Picking Nice Examples (2002/3)

Motivation:

- Accuracy of parsing Czech is limited, especially around the verbs.
- Valency of verbs is (supposedly) crucial for many NLP tasks.

$\Rightarrow$ Goal: Automatically extract nice examples, i.e. sentences easy to parse.

The result:

- a scripting language for partial parsing and filtering sentences.
- a script of 15 filters and 21 rules for Czech:
  - selects 10–15% of sentences
  - improves parsing accuracy by 5–10% absolute (correct dependencies) or 10–15% absolute (correct verb modifications)

Other Recent Experiments (2003–2005)

Constraint-based parsing of Czech didn't work out (Bojar, 2004):

- Local constraints on tree structure induced from a treebank were too weak
  $\Rightarrow$ exponentially many analyses remained possible (though not correct).

Inter-annotator agreement of verb-frame disam. (Lopatková et al., 2005):

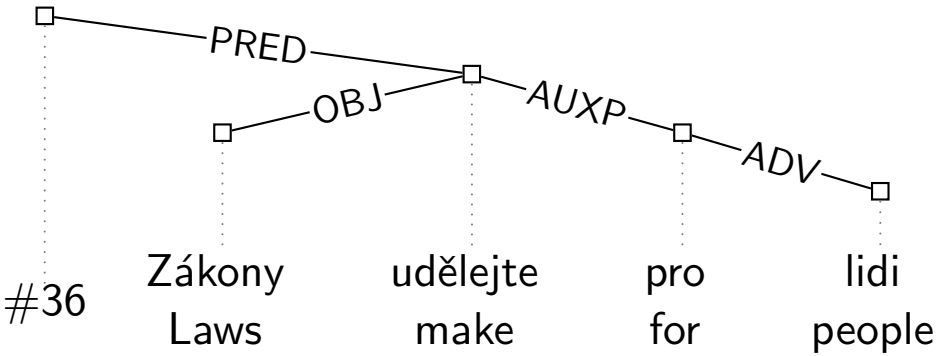
- Allowed to check quality of a Czech (deep) valency lexicon of verbs.
- Results comparable with others (PropBank etc.), best for Czech so far.

Experiments towards machine translation:

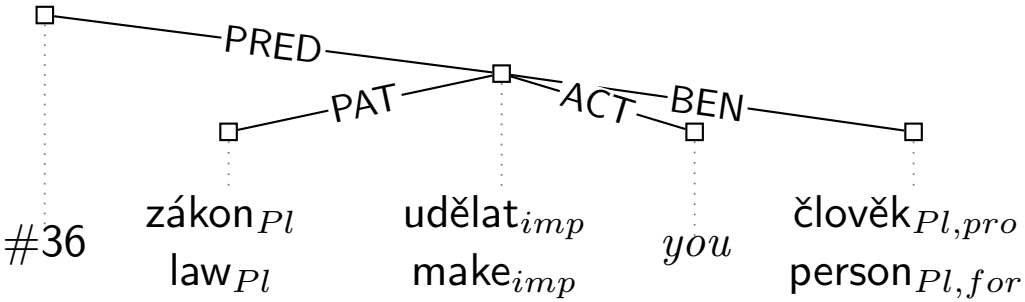
- Augmenting machine-readable dicts. with syntactic information (Bojar, 2005)
- (Rather unsuccessful) attempts at reusing an old rule-based MT system (Bojar, Homola, and Kuboň, 2005)
- First experiments with extracting parallel verb frames (Bojar and Hajič, 2005)

Analysis of Czech

Analytic (surface syntactic):



Tectogrammatical (deep syntactic):



Morphological (ambig.):

Form	Lemma	Morphological tag
zákony	zákon	NNIP1-----A----
zákony	zákon	NNIP4-----A----
zákony	zákon	NNIP5-----A----
zákony	zákon	NNIP7-----A----
udělejte	udělat	Vi-P---2--A----
udělejte	udělat	Vi-P---3--A---4
pro	pro-1	RR--4----------
lidi	člověk	NNMP1-----A----
lidi	člověk	NNMP4-----A----
lidi	člověk	NNMP5-----A----

Properties of Czech language

	Czech	English
Rich morphology	$\geq$ 4,000 tags possible, $\geq$ 2,300 seen	50 used
Word order	free	rigid

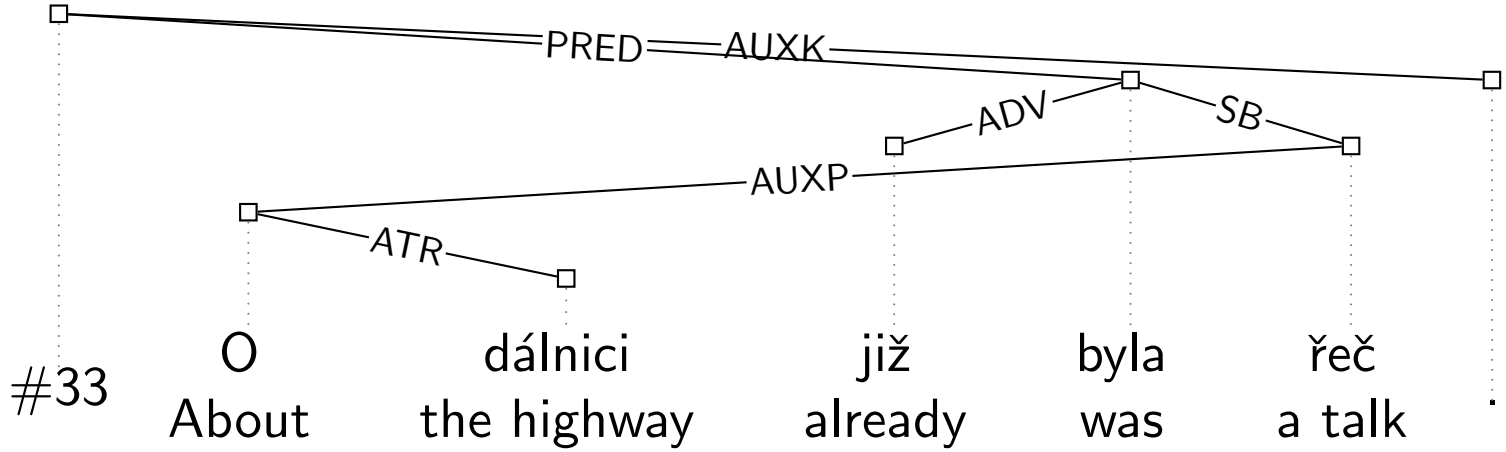
- rigid global word order phenomena: clitics
- rigid local word order phenomena: coordination, clitics mutual order

Nonprojective sentences	16,920	23.3%
Nonprojective edges	23,691	1.9%

Known parsing results	Czech	English
Edge accuracy	69.2–82.5%	91%
Sentence correctness	15.0–30.9%	43%

Data by (Collins et al., 1999), (Holan, 2003), Zeman (http://ckl.mff.cuni.cz/~zeman/ /projekty/neproj/index.html) and (Bojar, 2003). Consult (Kruijff, 2003) for measuring word order freeness.

Nonprojectivity



Non-projectivity:

- does not seem to cause delays in reading experiments (Bojar et al., 2004)
- disappears at the deep syntactic level (Veselá, Havelka, and Hajičová, 2004)
- parsing ($O(n^2)$) solved only very recently (McDonald et al., 2005)

Czech Data Relevant for Me/MT

Name and version	Sents.	Tokens	Vocab.	Lemmas	Notes
Czech National Corpus (SYN2000d)	6.8M	114M	1.7M	775k	
Prague Dep Tbk (PDT 1.0)	82k	1.3M	130k	55k	

Parallel Czech-English

Name and version	Sents.	Tokens	Vocab.	Lemmas	Notes
Prague Cz-En Dep Tbk (PCEDT 1.0)	22k/49k	0.5M/1.2M	57k/30k	28k/25k	
Reader's Digest (PCEDT 1.0)	44k/44k	658k/755k	84k/36k	?	stories
Kačenka	128k/105k	1.5M/1.5M	102k/47k	39k/22k	stories
OPUS EU Constitution	11k/10k	127k/164k	?	?	bad tok.
Kolovratník	107k/107k	1.3M/1.5M	190k/92k	?	not tok.!

BEAST: a compilation of web dictionaries (400k pairs, 235k cs, 225k en entries; if rejecting multi-word expressions: 138k pairs, 58k cs, 53k en)

My Experiments in Aachen

Word Alignment (GIZA++ against hand-aligned data from PCEDT):

- for 38% tokens where GIZA misaligned, two humans had a disagreement, too
- AER improved when using lemmas (27.4%→15.0%)
  and singletons substituted with POS →14.6%
  and refined method instead or intersection →13.5%
  or symmetric alignment (Matusov, Zens, and Ney, 2004) →11.8%

Machine translation (pbt, Czech→English, 20k sents., 5M tokens, 30–60k voc):

	BLEU [%]	NIST [%]
Baseline	21	5.7
Baseline but unknown words not spoiled	23	6.0
Best alignment and unknowns unspoiled	28	7.3
+ bigger LM used	31	7.8

Anyone interested in learning the details? (And thus helping me. :-)

Options for November

- Circumventing translation of numbers and names.
  Named-entity recognizer/trivial guesser needed.
- Adding more data (corpora, BEAST) . . . boring but useful.
  Part of my task here, in fact. . .
- Detailed analysis of MT upper bounds.
  Given this little corpus and this test data, what is achiveable?
- Detailed analysis of failures.
  Which n-grams spoil my BLEU? Do they have something in common?
- Extracting parallel syntactic information.
  Homework for the dissertation.

Summary of Keywords

Keywords describing my research:

- Czech, Czech-English
- syntactic analysis
- extraction of (parallel) syntactic information about words; dictionaries

Keywords important for Prague (as far as I know):

- deep syntax, tectogrammatical layer
- valency
- information structure (topic-focus articulation, coreference)
- PDT, PCEDT, PADT (Arabic!), TrEd

References

Bojar, Ondřej. 2003. Towards Automatic Extraction of Verb Frames. *Prague Bulletin of Mathematical Linguistics*, 79–80:101–120.

Bojar, Ondřej. 2004. Problems of Inducing Large Coverage Constraint-Based Dependency Grammar for Czech. In *Constraint Solving and Language Processing, CSLP 2004*, volume LNAI 3438, pages 90–103, Roskilde University, September. Springer.

Bojar, Ondřej. 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.

Bojar, Ondřej and Jan Hajič. 2005. Extracting Translation Verb Frames. In Walther von Hahn, John Hutchins, and Christina Vertan, editors, *Proceedings of Modern Approaches in Translation Technologies, workshop in conjunction with Recent Advances in Natural Language Processing (RANLP 2005)*, pages 2–6. Bulgarian Academy of Sciencies, September.

Bojar, Ondřej, Petr Homola, and Vladislav Kuboň. 2005. Problems Of Reusing An Existing

MT System. In *IJCNLP 2005 - Companion Volume to the Proceedings of Conference including Posters/Demos and Tutorial Abstracts*, pages 181–186, October.

Bojar, Ondřej, Jiří Semecký, Shravan Vasishth, and Ivana Kruijff-Korbayová. 2004. Processing noncanonical word order in Czech. In *Proceedings of Architectures and Mechanisms for Language Processing, AMLaP 2004*, pages 91–91, Université de Provence, September 16-18.

Collins, Michael. 1996. A New Statistical Parser Based on Bigram Lexical Dependencies. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 184–191.

Collins, Michael, 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.

Holan, Tomáš. 2003. K syntaktické analýze českých(!) vět. In *MIS 2003*. MATFYZPRESS, January 18–25, 2003.

Kruijff, Geert-Jan M. 2003. 3-Phase Grammar Learning. In *Proceedings of the Workshop on Ideas and Strategies for Multilingual Grammar Development*.

Lopatková, Markéta, Ondřej Bojar, Jiří Semecký, Václava Benešová, and Zdeněk Žabokrtský. 2005. Valency Lexicon of Czech Verbs VALLEX: Recent Experiments with Frame Disambiguation.

In Václav Matoušek, Pavel Mautner, and Tomáš Pavelka, editors, *Text, Speech and Dialogue: 8th International Conference, TSD 2005, Karlovy Vary, Czech Republic, September 12-15, 2005. Proceedings*, volume LNAI 3658, pages 99–106. Springer Verlag, September.

Matusov, E., R. Zens, and H. Ney. 2004. Symmetric Word Alignments for Statistical Machine Translation. In *Proceedings of COLING 2004*, pages 219–225, Geneva, Switzerland, August 23–27.

McDonald, Ryan, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-Projective Dependency Parsing using Spanning Tree Algorithms. In *Proceedings of HLT/EMNLP 2005*, October.

Veselá, Kateřina, Jiří Havelka, and Eva Hajičová. 2004. Condition of Projectivity in the Underlying Dependency Structures. In *Proceedings of Coling 2004*, pages 289–295, Geneva, Switzerland, August. COLING.

Detailed Numbers on Non-Projectivity

Edge length	1	≤ 2	≤ 5
English [%]	74.2	86.3	95.6
Czech [%]	51.8	72.1	90.2

[1]

Number of gaps	0	1	2
Sentences [%]	76.9	22.7	0.42

[2]

Climbing steps	1	2	3	4	5
Nodes [%]	90.3	8.0	1.3	0.3	0.1

[3]

[1] Data for English by (Collins, 1996). Data for Czech by (Holan, 2003).

[2] Data by (Holan, 2003).

[3] Data by (Holan, 2003).

Data Sparseness

After having seen	20,000	75,000	sentences
a new lemma comes every	1.6	1.8	test sentences
a new full morphological tag comes every	110	290	test sentences
a new simplified tag comes every	280	870	test sentences

Simplified morphological tag = POS, SUBPOS, CASE, NUMBER and GENDER.

Where GIZA Fails, Humans Have Troubles, Too

Percentage of running words where the alignment matches (Ok) or mismatches (With Problems):

- Humans against each other
- GIZA++ againts golden set derived by joining the human annotations

		Baseline		Improved	
Humans	GIZA++	en	cs	en	cs
With Problems	With Problems	14.3	15.5	14.3	15.5
With Problems	OK	0.1	0.1	0.2	0.1
OK	With Problems	38.6	35.7	25.2	25.0
OK	OK	46.9	48.7	60.4	59.4

Analytic vs. Tectogrammatical