Some Computational Experiments with Czech

Ondřej Bojar
obo@cuni.cz

December 7, 2006

Outline

- Background: Computer Science at Charles University in Prague
  - Student software project: Simulated family house
  - My master's: Picking nice examples
- Properties of Czech, analysis of Czech, available data
- Some of my previous experiments
- PhD research (ongoing): Constructing verb valency frames
- Experiments towards MT
  - This year's JHU summer workshop: Moses
- My task here: tree-based machine translation
- Summary of keywords

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: Picking nice examples of linguistic phenomena

Our Project: The Ents (2000–2002)

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

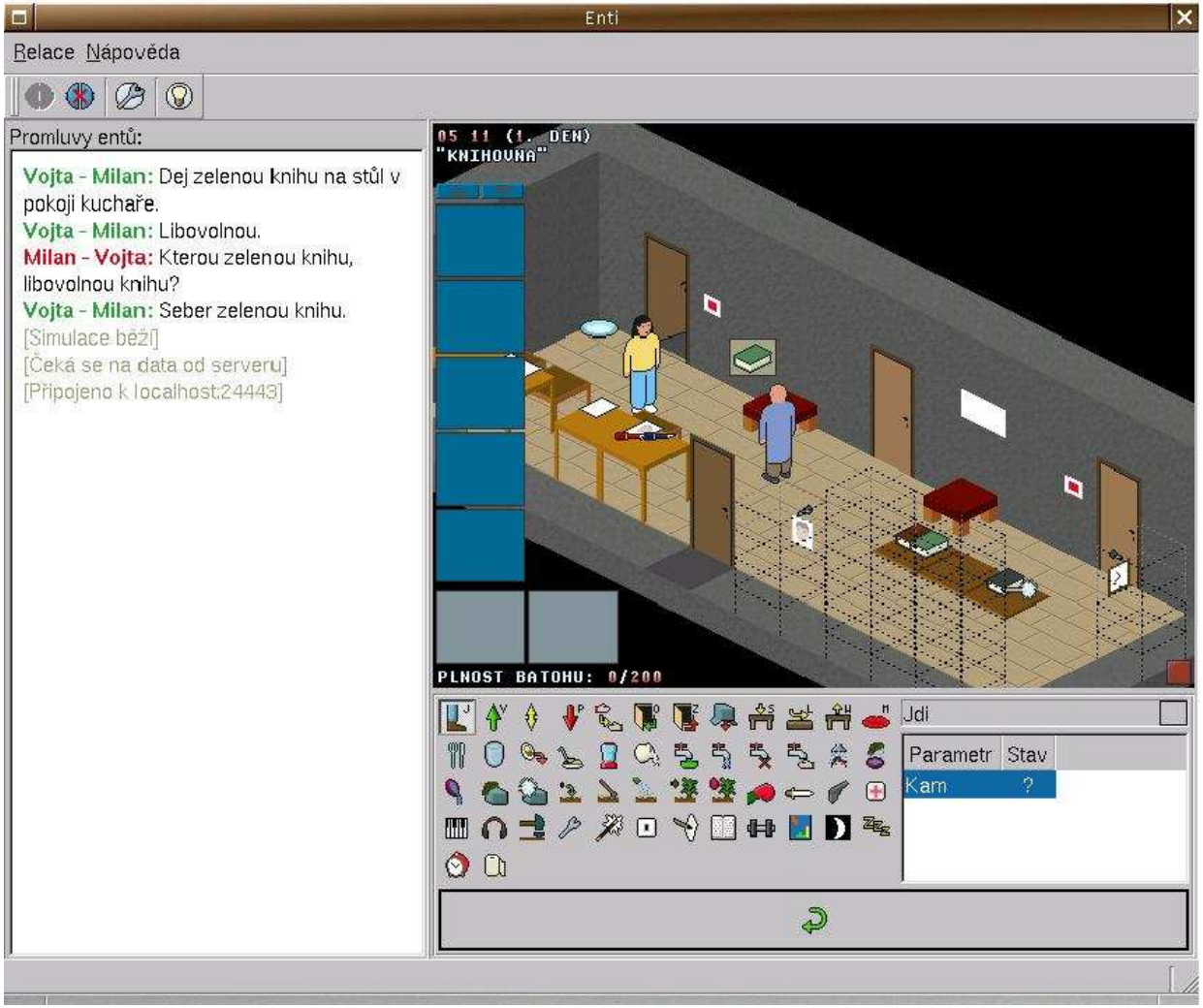
The Result:

- 6 students, 2 years (student style of intensive work)
- 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


Enti
Relace Nápověda
Promluvy entů:
Vojta - Milan: Dej zelenou knihu na stůl v pokoji kuchaře.
Vojta - Milan: Libovolnou.
Milan - Vojta: Kterou zelenou knihu, libovolnou knihu?
Vojta - Milan: Seber zelenou knihu.
[Simulace běží]
[Čeká se na data od serveru]
[Připojeno k localhost:24443]
05 11 (1. DEN)
"KNIHOVNA"
PLNOST BATOHU: 0/200
Jdi
Parametr
Stav
Kam
?

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.

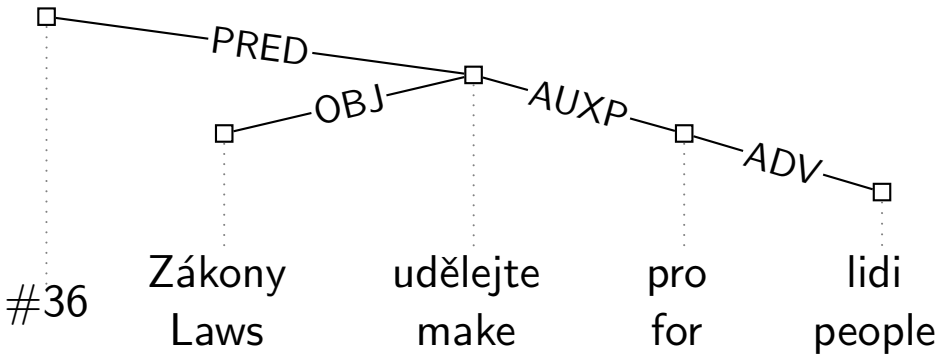
$\Rightarrow$ Goal: Automatically extract nice examples, i.e. sentences easy to parse.

The result:

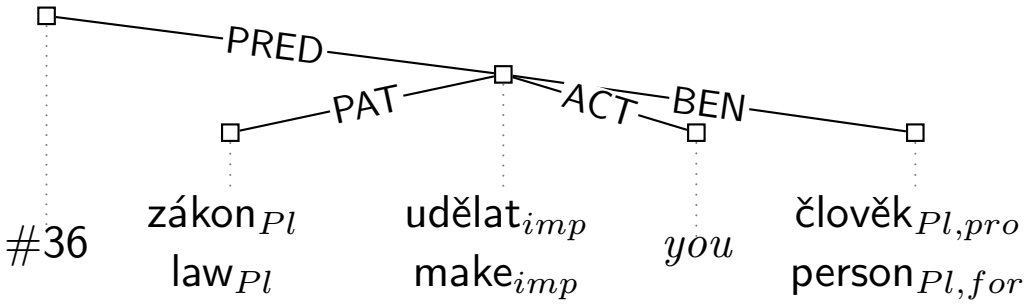
- a scripting language for partial parsing and filtering sentences
  Engine in Mercury, regular expressions over untyped feature structures.
- a script of 15 filters and 21 rules for Czech:
  - selects 10–15% of sentences
  - improves parsing accuracy by 5–10% absolute (correct dependencies) or 10–15% absolute (correct verb modifications)

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

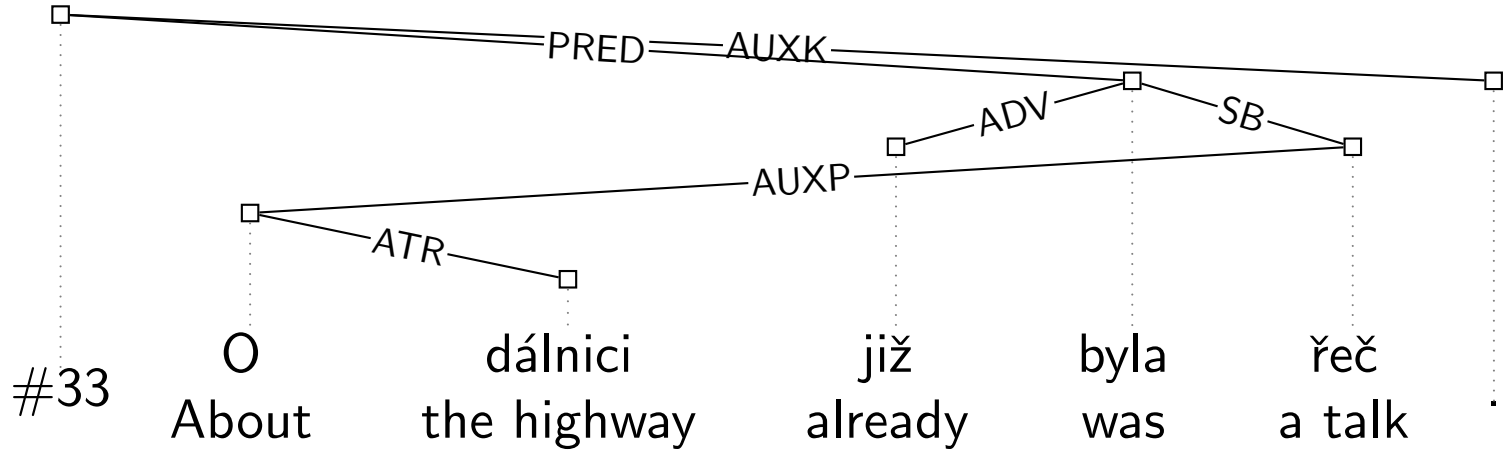
- rigid global word order phenomena: clitics
- rigid local word order phenomena: coordination, clitics mutual order

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

Known parsing results	Czech	English
Edge accuracy	69.2–82.5–86%	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.

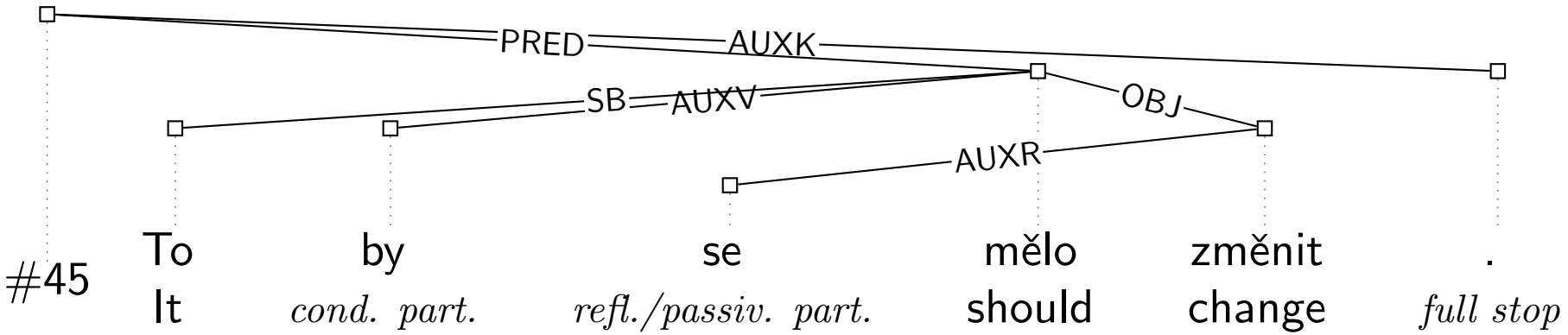
Nonprojectivity

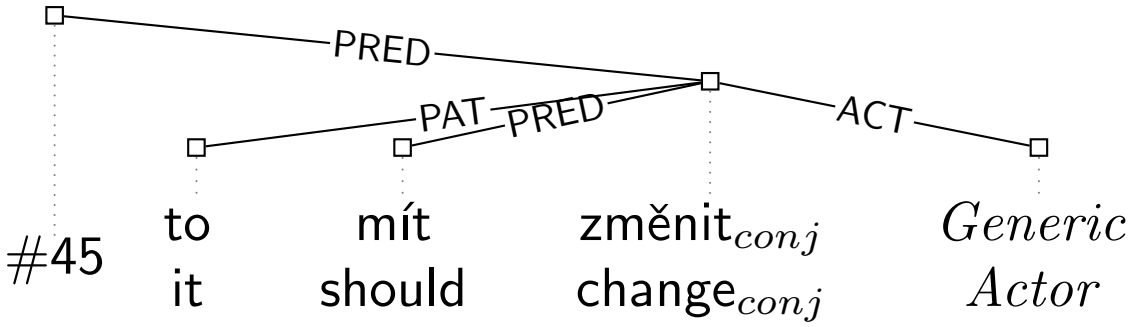


Non-projectivity:

- 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 recently (McDonald et al., 2005)

Analytic vs. Tectogrammatical





- hide auxiliary words, add nodes for "deleted" participants
- resolve e.g. active/passive voice, analytical verbs etc.
- "full" tecto resolves much more, e.g. topic-focus articulation or anaphora