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

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

⇒ 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.

⇒ 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.):

<table>
<thead>
<tr>
<th>Form</th>
<th>Lemma</th>
<th>Morphological tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>zákony</td>
<td>zákony</td>
<td>NNIP1----A----</td>
</tr>
<tr>
<td>zákony</td>
<td>zákony</td>
<td>NNIP4----A----</td>
</tr>
<tr>
<td>zákony</td>
<td>zákony</td>
<td>NNIP5----A----</td>
</tr>
<tr>
<td>udělejte</td>
<td>udělat</td>
<td>Vi-P---2--A----</td>
</tr>
<tr>
<td>udělejte</td>
<td>udělat</td>
<td>Vi-P---3--A---4</td>
</tr>
<tr>
<td>pro</td>
<td>pro-1</td>
<td>RR--4------------</td>
</tr>
<tr>
<td>lidi</td>
<td>člověk</td>
<td>NNMP1----A----</td>
</tr>
<tr>
<td>lidi</td>
<td>člověk</td>
<td>NNMP4----A----</td>
</tr>
<tr>
<td>lidi</td>
<td>člověk</td>
<td>NNMP5----A----</td>
</tr>
</tbody>
</table>

Ondřej Bojar Some Computational Experiments with Czech December 7, 2006
Properties of Czech language

<table>
<thead>
<tr>
<th>Czech</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rich morphology</td>
<td>≥ 4,000 tags possible, ≥ 2,300 seen</td>
</tr>
<tr>
<td>Word order</td>
<td>free</td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th>Edge accuracy</th>
<th>Czech</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>69.2–82.5–86%</td>
<td>91%</td>
<td></td>
</tr>
</tbody>
</table>

| 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

To by se mělo změnit.

It cond. part. refl./passiv. part. should change full stop

#45

• 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