XDG vs. Czech

Ondřej Bojar
obo@cuni.cz

March 17, 2004
Outline

- Motivation
- Short characterization of Czech
- Experiments and results of my grammar
- Summary, identified problems
- Further research
Motivation

- Severe limitations of parsers available for Czech.
  Statistical: Collins, Zeman, Horák; Handcrafted: Žabokrtský
  Single solution only, restricted to analytical (ID) trees.

- No constraint-based large-coverage grammar tested on Czech.

- XDG has nice theoretical properties wrt. to constraint parsing.

- No large grammar has yet been implemented in XDG.

- There is plenty (though not enough) annotated data available for Czech.

⇒ Goal: Acquire a large coverage grammar for Czech.
### Properties of Czech language

<table>
<thead>
<tr>
<th>Rich morphology</th>
<th>Czech</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥ 4,000 tags possible, ≥ 1,400 seen</td>
<td>free</td>
<td>rigid</td>
</tr>
<tr>
<td>Word order</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Rigid global word order phenomena:** clitics

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

<table>
<thead>
<tr>
<th>Known results</th>
<th>Czech</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge accuracy</td>
<td>69.2–82.5%</td>
<td>91%</td>
</tr>
<tr>
<td>Sentence correctness</td>
<td>15.0–30.9%</td>
<td>??</td>
</tr>
</tbody>
</table>

Data by (?, ?), Zeman (http://ckl.mff.cuni.cz/~zeman/projekty/neproj/index.html) and (?). Consult (?) for measuring word order freeness.
The big picture: Data sources for Czech

- PDT: 127,000 analytical trees
  - Korthals partial model
  - PHMM, GJ not available yet

- Vallex: 1,000 verbs
  - Vallex requires TG then constrains AT

- PDT: 55,000 tectogrammatical trees
  - TG-AT (TH-ID) correspondence is complex

- XDG Grammar
  - LP
  - ID
  - TH

My simple grammar

Ondřej Bojar

XDG vs. Czech

March 17, 2004
The small picture: What has been implemented

Training data

- Extract ID agreement
- Extract ID simplified frames
- Extract ID allowed edges over
- Extract ID look right

Generic grammar, ID+LP

Test data

- Optional morphological analysis
- Restrict the generic grammar to cover only tested word forms

XDG parser