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.

$\Rightarrow$ Goal: Acquire a large coverage grammar for Czech.

Properties of Czech language

	Czech	English
Rich morphology	$\geq$ 4,000 tags possible, $\geq$ 1,400 seen	50 used
Word order	free	rigid

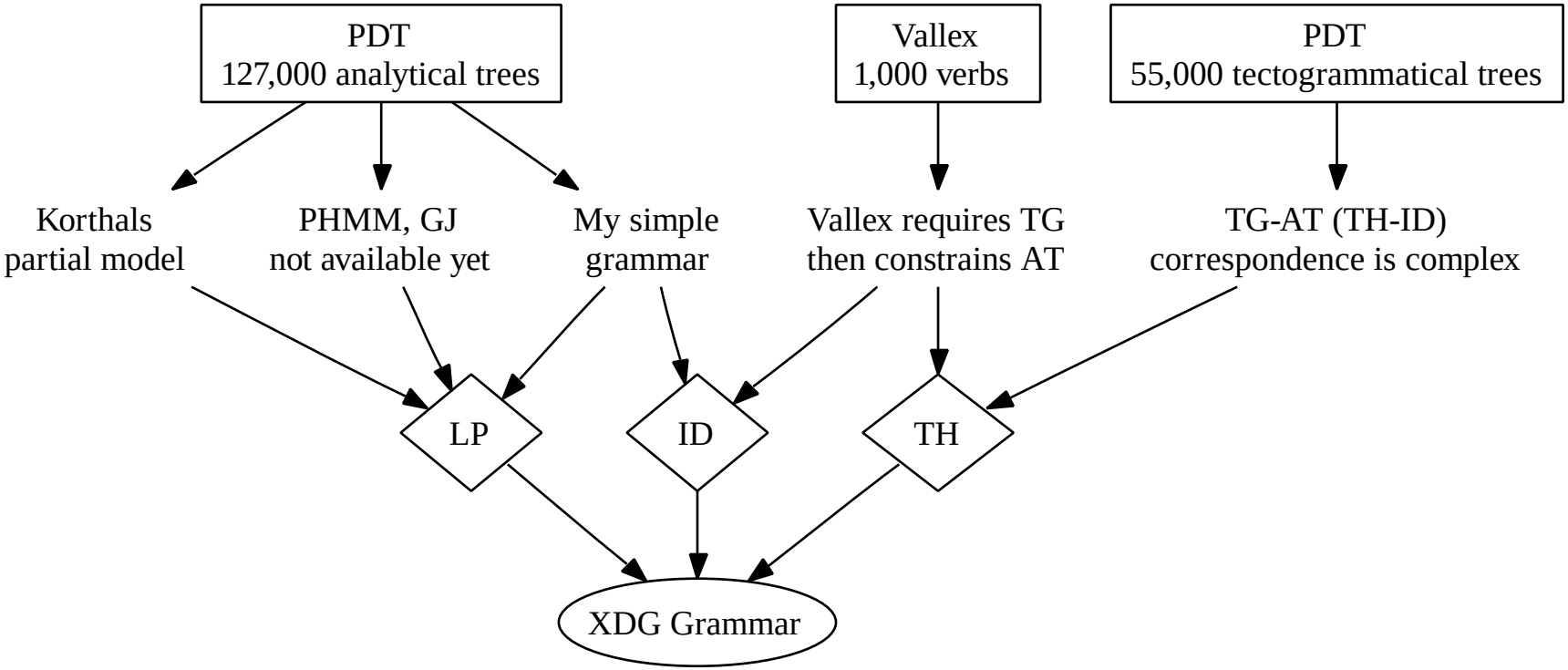
- rigid global word order phenomena: clitics

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

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

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



The small picture: What has been implemented