Problems of Inducing Czech Constraint-Based Grammar

Ondřej Bojar
obo@cuni.cz

September 3, 2004

Outline

- Motivation
- Properties of Czech
- Constraints used to model surface syntax
- Experiments and results of my grammar
- Summary, identified problems

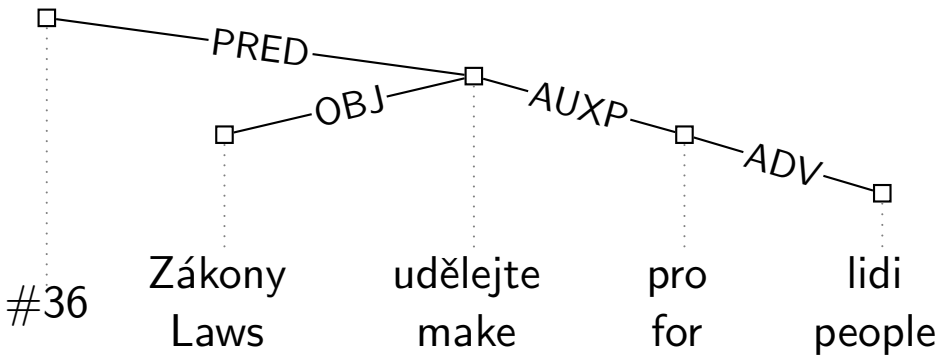
Motivation

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

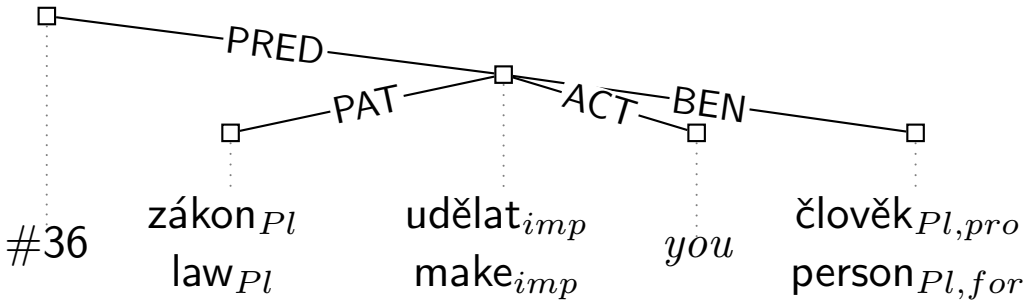
$\Rightarrow$ Goal: Acquire a large coverage grammar for Czech.

Analysis of Czech

Analytic (surface syntactic):



Tectogrammatical (deep syntactic):



Morphological:

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

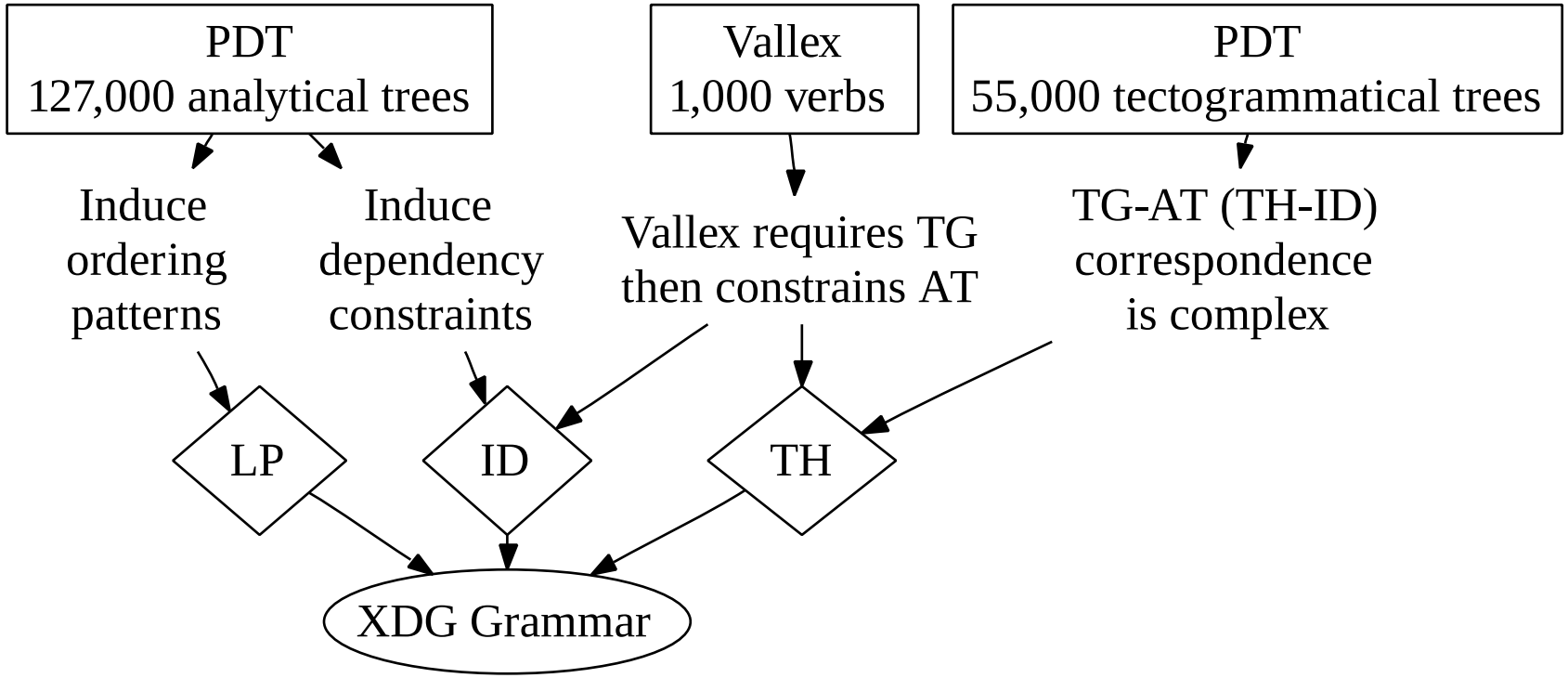
	Czech	English
Rich morphology	$\geq$ 4,000 tags possible, $\geq$ 1,400 seen	50 used
Word order	free	rigid

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

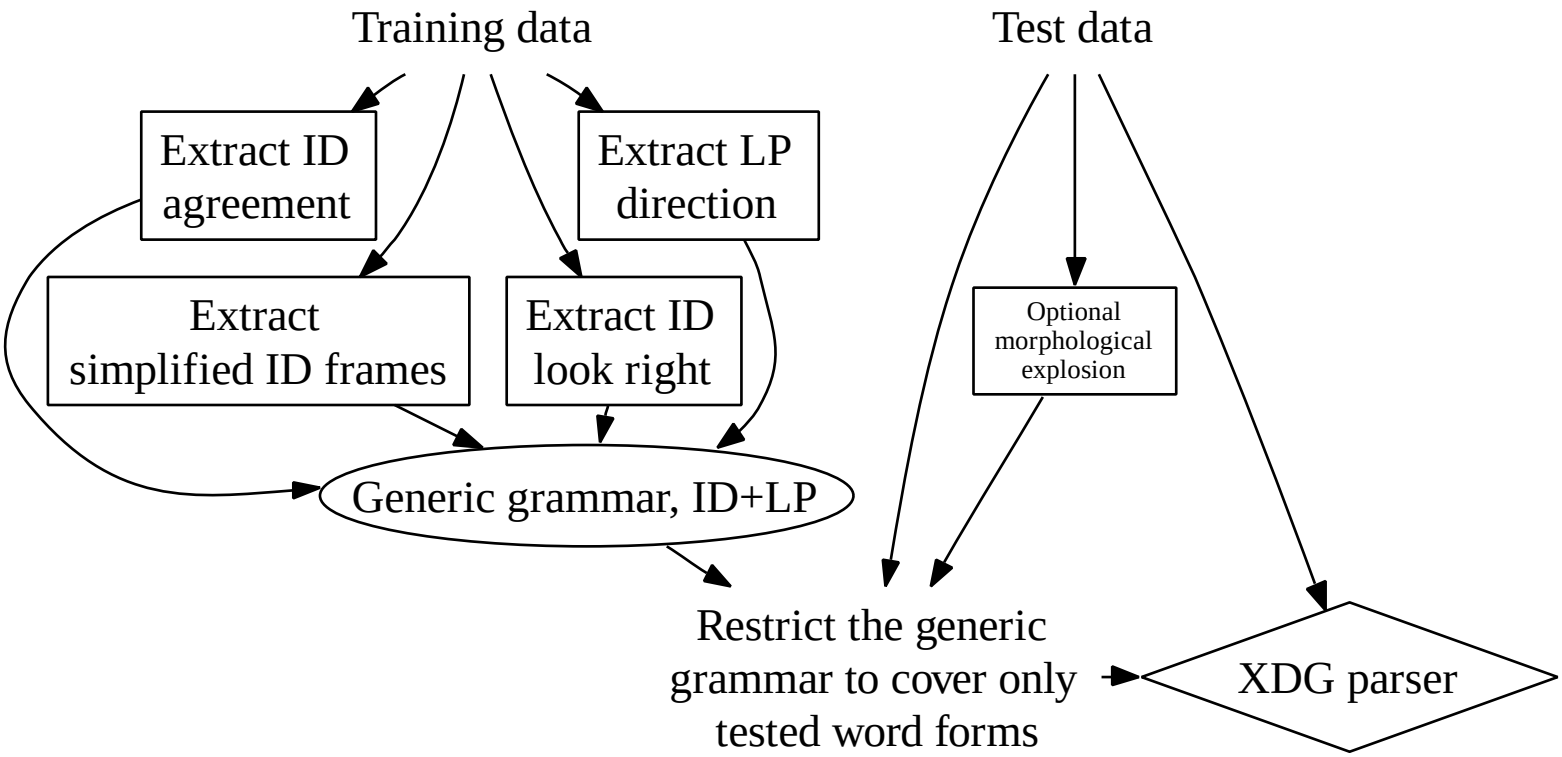
Nonprojective sentences	16,920	23.3%
Nonprojective edges	23,691	1.9%

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

The big picture: Data sources for Czech



The small picture: What has been implemented



Collecting data

Throughout the grammar:

- information about words is collected *nonlexicalized*
- based on simplified morphological tag (POS, SUBPOS, CASE, NUMBER and GENDER).

Reason:

After having seen	20,000	75,000	sentences
a new lemma (i.e. word) 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

ID Agreement

- Every mother of an ID edge allows only daughters of specific morphological properties.

- Every daughter of an ID edge accepts incoming edges only from mother nodes with specific properties.

Simplified ID Frames

XDG valency principle: Every mother allows only specific combination and cardinalities of outgoing ID edges.

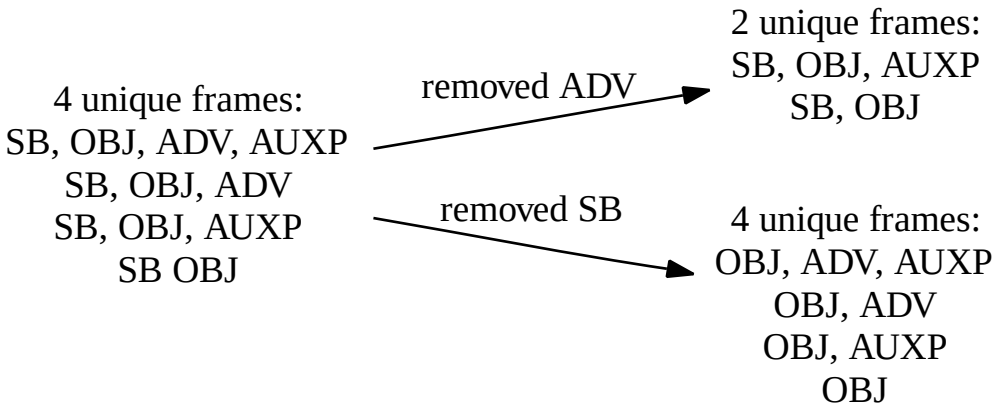
- Current approaches[1] aim at distinguishing COMPLEMENTS vs. ADJUNCTS.
- Current formalisation of XDG makes no use of this distinction if the DELETABILITY (elipticity) of modifications is taken into account.
- What counts in XDG are restrictive implications.
  Currently expressed simply by enumerating the allowed combinations.
- No approach aims at discovering the interdependencies of particular modifications, so far.

[1] See (Sarkar and Zeman, 2000) for comparison and references.

At 2000 sentences (2% of PDT), there were seen unique	
14394	lemma and list of sons' lemmas
11596	lemma and list of sons' simplified tags
11514	lemma and bag of sons' simplified tags
11349	lemma and set of sons' simplified tags
6420	simplified tag and list of sons' simplified tags
6263	simplified tag and bag of sons' simplified tags ←
6009	simplified tag and set of sons' simplified tags
4742	lemmas
4154	list of sons' simplified tags
3851	bag of sons' simplified tags
3488	set of sons' simplified tags
468	simplified tags

⇒ I introduced a rather simplistic approach that simplifies what will not hurt and stores the complexity of interdependencies by enumerating.

Example: Observed under a verb:



$\Rightarrow$ ADV is more optional than SB.

Simplification used iteratively, until the number of unique frames is acceptable. Torn-off modifications are added to all the frames as optional.

Simplifying modifications of verbs: 4$\rightarrow$(AUXP)$\rightarrow$2$\rightarrow$(ADV)$\rightarrow$1

The resulting frame:

SB, OBJ, AUXP*, ADV*

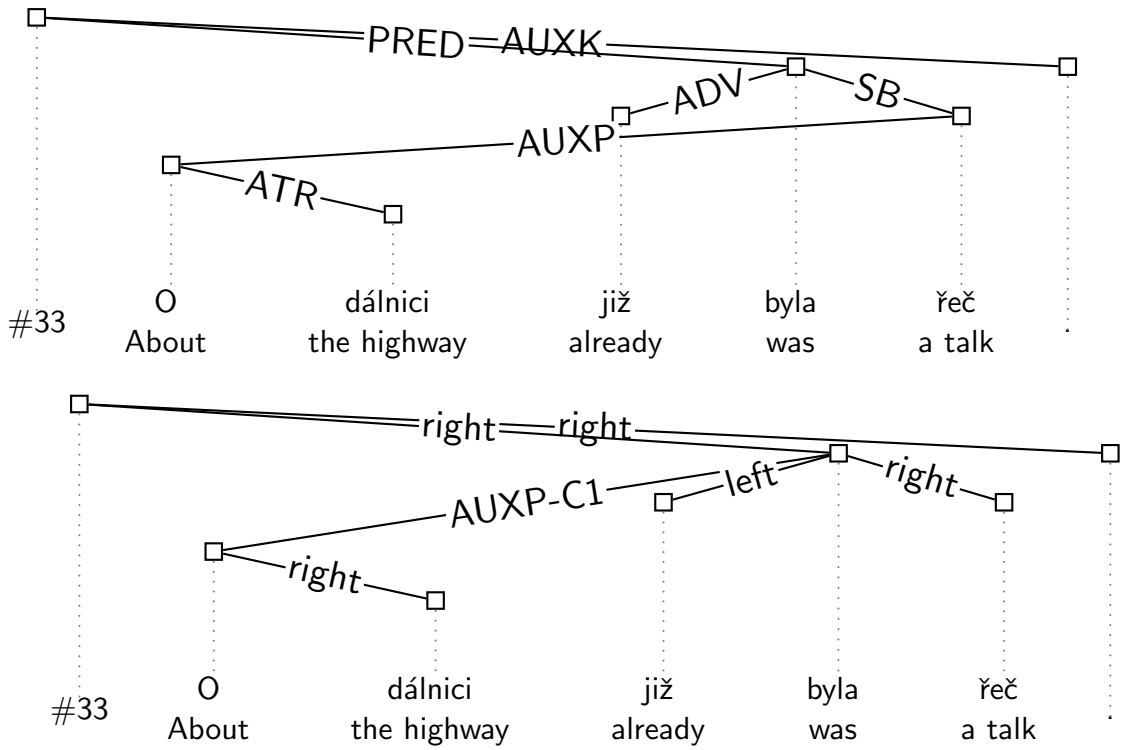
LP Direction

- No theory of topological fields for Czech ⇒ model edge direction only.
- Every LP mother offers 3 fields: left*, head, right*.
- Every LP mother allows only observed combinations of ID label and LP field for outgoing edges. ("A preposition offers the noun only to the right.")

⊖ Doesn't capture non-projectivity.

⊖ Doesn't capture global phenomena like clitics.

LP: Adding non-projectivity



- Extend the LP valency of specific mothers to accept climbed edges of specific type.
- No restriction on the order of climbed edges wrt. to left/right/head.

ID Look right (look ahead)

- Generally accepted: Head-daughter dependencies model syntactic analysis best. (Confirmed by several parsers of English.)

- Doubted by (Dubey and Keller, 2003): For German sister-sister dependencies (lexicalized case) are more informative.

- My experiment: Difficulty of predicting ID edge label based on morphological properties of the node and a close context.

Context		Neighbours		Sisters	
used	Head	Left	Right	Left	Right
Entropy	0.65	1.20	1.08	1.14	1.15

$\Rightarrow$ New principle: An incoming ID edge to a word must be allowed by the word class of the right neighbour.

Combining constraints

$$word\ form \;\&\; \left(\bigvee_{morphological\ variants} \left(\begin{array}{c} morphological\ information \\ \& \\ \bigvee(allowed\ incoming\ ID\ edges) \\ \& \\ \bigvee(allowed\ outgoing\ ID\ edges) \\ \& \\ \bigvee_{simplified\ frames}(IDvalency) \\ \& \\ \bigvee(allowed\ incoming\ LP\ edges) \\ \& \\ \bigvee(allowed\ outgoing\ LP\ edges) \end{array} \right) \right)$$

Results at the first choicepoint

Training sentences	2500	5000	2500	5000
	Unsolved sentences		Avg. ambiguity/node	
Without Look Right	21.1	11.9	8.09	8.91
With Look Right	25.6	15.4	8.17	9.05
	Assigned structural edges		Correct structural edges	
Without Look Right	4.4	3.3	82.3	82.5
With Look Right	4.7	3.5	81.9	81.0
	Assigned labelled edges		Correctly labelled edges	
Without Look Right	3.4	2.3	85.9	85.9
With Look Right	3.6	2.5	85.0	83.5

Sources of ambiguity

Avg. possible heads	Max. possible heads	Number of observations	Word class
22.0	59	2459	Z : (punctuation)
20.7	59	1045	J ^ (conjunctions, coordinating)
19.2	49	500	C = (cardinals)
18.9	60	1035	D b (adverbs)
18.2	57	543	J , (conjunctions, subordinating)
17.3	59	2611	R R (prepositions)
17.2	44	448	D g (adverbs)
15.6	38	130	T T (particles)
13.9	47	1568	V B (verbs)
12.8	36	519	V f (verbs, infinite)
		. . .	
6.1	49	3310	A A (adjectives)

Summary (1)

Czech vs. Ondřej Bojar + XDG: 1:0

- (Czech is a challenge!)
- Tested XDG on large data.
- Attempted at a constraint-based grammar for Czech.
  - ⊕ Support for non-projectivity.
  - ⊕ Open for adding more levels of language description.
  - ⊖ Not competitive with existing parsers.
  - ⊖ Doubts about scalability.

Summary (2): Identified problems

General:

- Syntactic ambiguity of Czech is too high to be captured by the designed local constraints (agreement, valency, look right, edge direction).

Particular:

- Multi-word expressions (dates, names) not handled separately.
- No reasonable theory of coordination (punctuation, conjunctions).
- No theory of segmentation, clausal structure.

Further Research / Open Hopes

- TH (deep syntactic) level to restrict the number of solutions:
  only if a lexicon is used to strictly restrict the TH

- Searching for situation-specific constraints:
  Approach unambiguity without over-fitting.
  General constraints are too specific (over-fitted) for some situations while too underspecified for another ones. Goal: induce constraints varying in specificity.

- Make use of frequency information to guide the search.

References

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.

Dubey, Amit and Frank Keller. 2003. Probabilistic Parsing for German using Sister-Head Dependencies. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 96–103, Sapporo.

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

Sarkar, Anoop and Daniel Zeman. 2000. Automatic Extraction of Subcategorization Frames for Czech. In *Proceedings of the 18th International Conference on Computational Linguistics (Coling 2000)*, Saarbrücken, Germany. Universität des Saarlandes.

Detailed numbers

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

[2]

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

[3]

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

[4]

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

[3] Data by (Holan, 2003).

[4] Data by (Holan, 2003).

Analytic vs. Tectogrammatical (2)