A System for Syntactic Annotation of Large Czech Corpora

Tomáš Jelínek

Institute of Theoretical and Computational Linguistics

Faculty of Arts

Charles University in Prague

Corpus linguistics – 2013, Saint Petersburg State University

Outline of the talk

0. Automatic Syntactic Annotation of Corpora

1. Prague Dependency Treebank

2. Czech Dependency Parsing

3. Improving Parsing by Preprocessing of Text

4. A Rule-Based Correction System

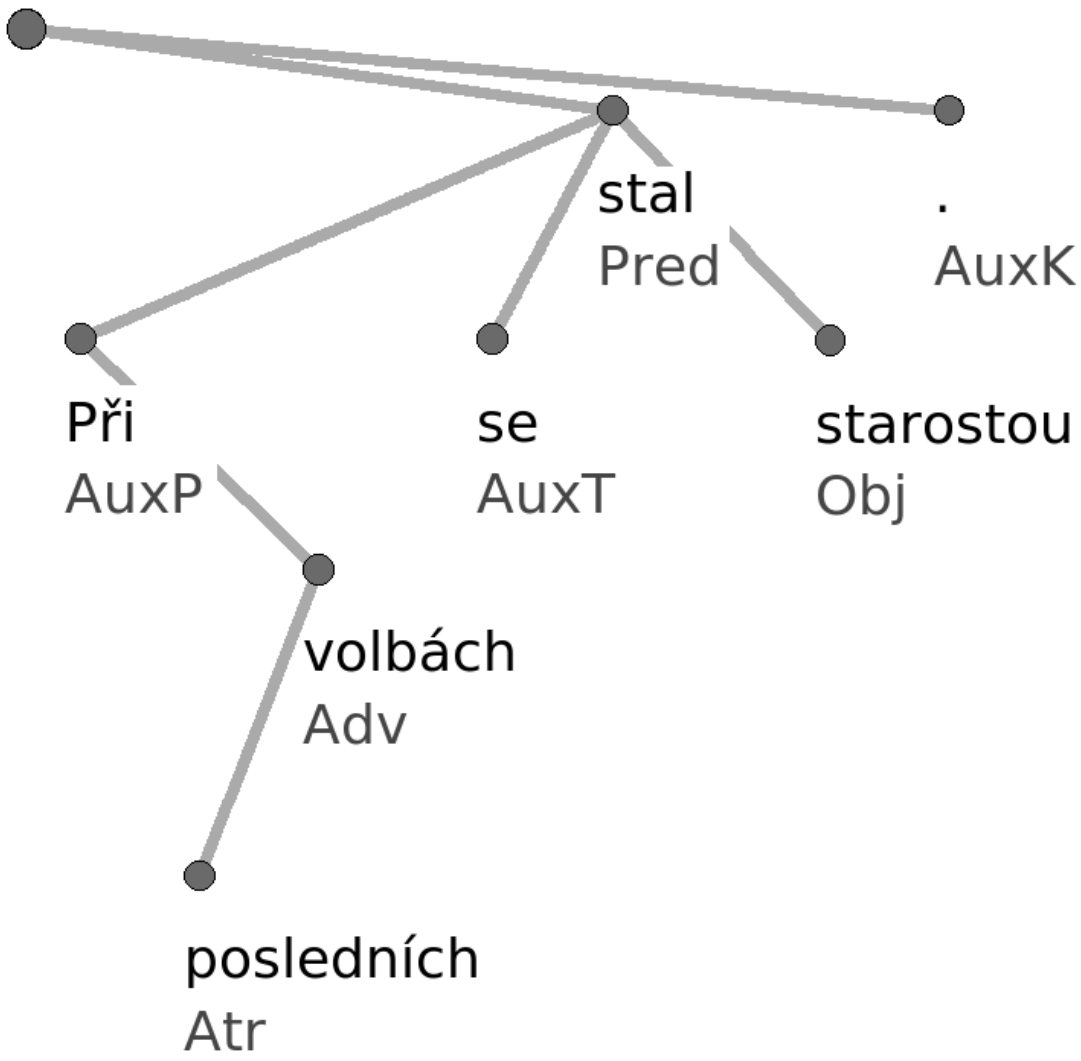
5. Conclusion

0. Automatic Syntactic Annotation of Corpora

- **Useful additional linguistic information**
- **Fully automatic annotation necessary**
- **Limited manually annotated training data**
- **Error rate too high**
- **New methods for improving the reliability of parsing are urgent!**

1. Prague Dependency Treebank

- **A multi-level annotated corpus of written Czech**
- **1.5 million words manually annotated on a surface syntactic level (analytical layer)**
- **A dependency representation of syntactic relations: each word / token depends either on another token or on an artificial « root » token**
- **Each word is assigned a syntactic function (basic or auxiliary)**



Pred, Sb, Obj, Adv, Atr...

AuxP, AuxT, AuxV...

AuxK, AuxG...

Při posledních volbách se stal starostou.

In the last election he became mayor.

2. Czech Dependency Parsing

Parser	Accuracy*
Combination	**85,8%**
MaltParser	**85,6%**
MSTParser	**84,7%**

*UAS (**u**nlabeled **a**ttachment **s**core), e-test data

MSTParser is approx. **20 times faster** than **MaltParser**

MSTParser (Maximum Spanning Tree Parser)
by Ryan McDonald

Při posledních volbách se stal starostou.

In the last election he became mayor.

Při	*posledních*	*volbách*	*se*	*stal*	*starostou*	*.*
R6	**A6**	**N6**	**P4**	**Vp**	**N7**	**Z:**
5	**3**	**1**	**5**	**0**	**5**	**0**
AuxP	**Atr**	**Adv**	**AuxT**	**Pred**	**Obj**	**AuxK**

3. Improving Parsing by Preprocessing of Text

- **Limited, sparse training data**
  53,3% word forms occur exactly once in the training data
- **Training data cannot cover the whole dictionary**
  30.000 noun lemmas in PDT, over 200.000 noun lemmas in Czech dictionary
- **Morphological information used is not adequate**
  no distinction between syntactic nouns and adjectives, etc.
- **Solution: pre-process the text, « condense » data**

3. Improving Parsing by Preprocessing of Text

- Word-forms (used in the original settings) replaced by **lemmas**
- Narrow word-classes manually selected, replaced by one representative
- Lists of words with exactly the same syntactic properties extracted from a large corpus (1,3 bil. words) unrelated to PDT data
- Approx. 80 word-classes implemented

3. Improving Parsing by Preprocessing of Text

Examples:

Nouns

given names: Jan, Petr, Pavel, Josef... ► **Jan**

surnames: Novák, Jelínek, Petkevič... ► **Novák**

months: leden, únor, březen, duben... ► **leden**

cities (fem.sg.): Praha, Varšava, Sofia... ► **Praha**

...

Vladimír Petkevič jel do Moskvy. ►*Jan Novák jel do Prahy.*

Vladimír Petkevič went to Moscow. ►*Jan Novák went to Prague.*

3. Improving Parsing by Preprocessing of Text

Examples:

Adjectives

deverbal: smějící se, chechtající se... ► smějící se

possessive: Janův, Petrův, Pavlův... ► **Janův**

quantified: pětiletý, padesátiletý, dvaasedmdesátiletý...

... ► **xletý**

Numerals

cardinal: pět, šest, sedm, osm... ► **pět**

ordinal: pátý, šedesátý, stopadesátý... ► **pátý**

3. Improving Parsing by Preprocessing of Text

Verbs: only careful replacement of less frequent verbs with only the basic, accusative valency, having no other (even potential) valency frame

Result: reduction of variability by 17% – 20%.

train **PDT**	original data	simplified data	number decrease
forms	125.369	103.016	**17,8%**
lemmas	54.610	44.126	**19,2%**

3. Improving Parsing by Preprocessing of Text

MSTParser retrained and tested:

1. improved morphological tagging

2. reduced variability (« condensed » data)

MSTParser	accuracy*
original results	**84,7%**
condensed data	**86,1%**

* UAS, e-test data set of PDT

4. A Rule-Based Correction System

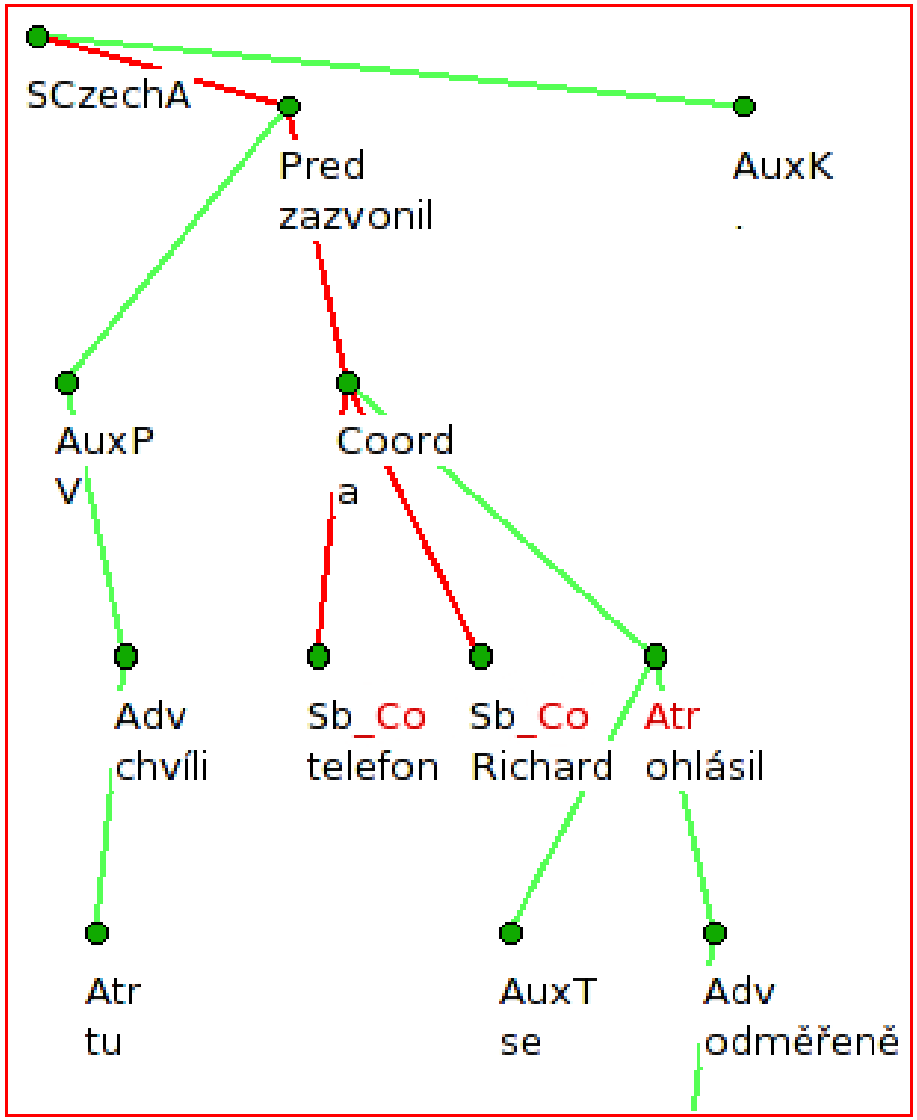
- Independent work prior to the data simplification process
- Post-processing of syntactically annotated text
- Correction of most frequent errors of syntactic annotation
- Typical errors found in a large corpus (100 mil. words) annotated with original settings of MST-P

4. A Rule-Based Correction System

- 1. Error identification
- 2. Choice of correction algorithm
- 3. Correction (if possible)

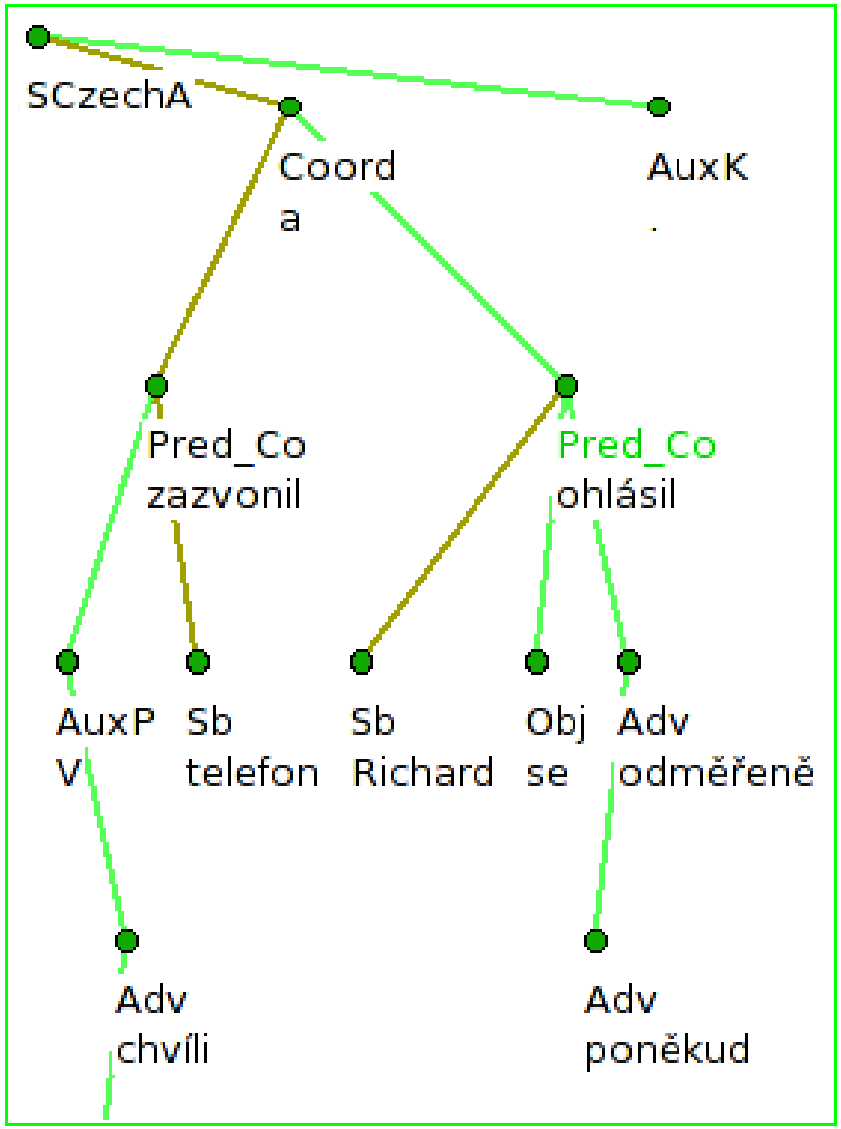
- Correction of dependency relations, syntactic functions or morphological tags (case)

- 26 rules implemented



*V tu chvíli zazvonil telefon **a** Richard se ohlásil poněkud odměřeně.*

*At that moment the phone rang **and** he answered it, a little curtly.*



*V tu chvíli **zazvonil** telefon **a** Richard se **ohlásil** poněkud odměřeně.*

*At that moment the phone rang **and** he answered it, a little curtly.*

4. A Rule-Based Correction System: Results

MST Parser	UAS	LAS
original settings	84,7%	77,6%
+ correction	85,1%	78,9%
improvement	0,4%	1,3%

BUT ! ...

4. A Rule-Based Correction System: Results

MST Parser	UAS	LAS
original settings	84,7%	77,6%
condensed data	86,1%	79,0%
+ correction	86,2%	79,4%
improvement	0,1%	0,4%

5. Conclusion

- **Data condensing procedure:**
  - **promising**
  - **less labor intensive**
  - **can be used with any parser**

- **Rule-based correction system:**
  - **only complementary**
  - **labor intensive**
  - **adaptation to parser and its settings necessary**

Спасибо за внимание!