

# 08 – Parsing of Czech: Between Rules and Stats

## IA161 Natural Language Processing in Practice

A. Horák

NLP Centre, FI MU, Brno

November 11, 2022

# Parsing – motivation

## Example

Obehnat Šalounův pomník mistra Jana Husa na pražském Staroměstském náměstí živým plotem z hustých keřů s trny navrhuje občanské sdružení Společnost Jana Jesenia.

## Example (Human translation)

Civic association of Jan Jesenius Community proposes to surround the Solomon's monument of Master Jan Hus in Prague's Old Town Square with thick hedges with thorns.

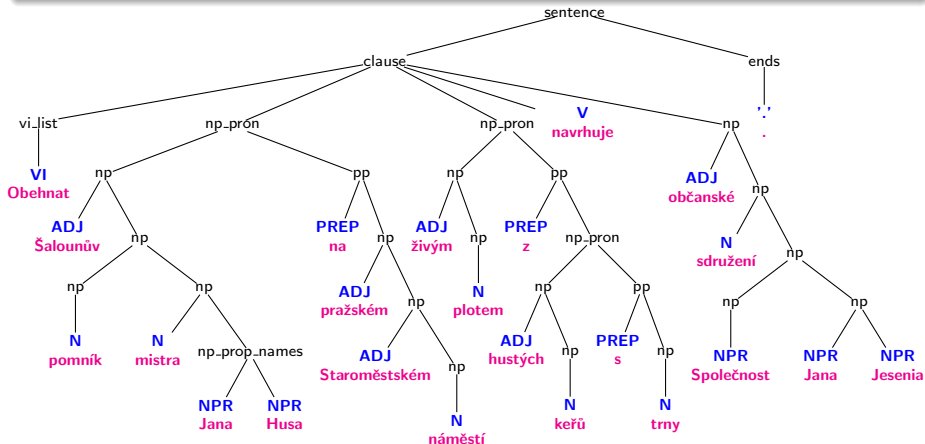
## Example (Google translate)

To surround Solomon's monument to Master Jan Hus in Prague's Old Town the square is designed by a civic association with thick hedges with thorns Company of Jan Jesenia.

# Parsing – motivation

## Example

Obehnat Šalounův pomník mistra Jana Husa na pražském Staroměstském náměstí živým plotem z hustých keřů s trny navrhuje občanské sdružení Společnost Jana Jesenia.



# Syntactic analysis – motivation

- syntactic units are carriers of **meaning**
  - ▶ “in the city”
  - ▶ meaning of “in”, “the” is unclear, complicated
  - ▶ meaning of “in the city” = **where**
- words are **not enough**
  - ▶ **red brick house** vs. **brick house red** vs. **red house brick**
  - ▶ **Honey, give me love** vs. **Love, give me honey**
- starting point for intelligent natural language **applications**:
  - ▶ extraction of facts & question answering
  - ▶ logical analysis
  - ▶ punctuation detection & grammar checking
  - ▶ natural text generation
  - ▶ authorship detection
  - ▶ machine translation

- 1 Motivation
  - Motivation
- 2 Morphology
  - Morphology
  - Guesser
  - Diacritics
  - Industrial applications
- 3 Parsing and Fact Extraction
  - Syntactic analysis
  - Syntactic trees
  - Extraction of facts
  - Grammar checking
  - Statistical parsing
  - Parsing @NLPCentre

# Word Level Analysis

“clustering” of word forms in text:

<i>států</i>					<i>stojíš</i>	
<i>státy</i>					<i>stály</i>	
<i>státech</i>	$\iff$	<i>stát<sub>noun</sub></i>		<i>stát<sub>verb</sub></i>	$\iff$	<i>stojíme</i>
<i>státu</i>		<i>a state</i>		<i>to stand</i>		<i>stůjte</i>
...						...

Word form *stát* (a state/to stand, to stop) has 3 interpretations:

- lemma *stát*, noun in nominative
- lemma *stát*, noun in accusative
- lemma *stát*, verb in infinitive

# Data for Czech Morphology

**lemmatization** (*stát*), **tagging** (*noun/verb*) –

- for **indexing**, **searching**, ... and almost all NLP tools
- ambiguity resolution according to the **context**
- word form **generation**
- **spellchecking**, **diacritics** restoration

**12 M word forms** (incl. colloquial forms):

- **lemma** (canonical form, dictionary form)
- **grammatical** information: part of speech, number, case etc.

**very fast** analysis – **1 million** word forms **per second**

# Resolving Ambiguities Using Context

## Disambiguation of *stát*:

- **verb**: *Celá továrna musela hodinu stát.* (The factory had to stop for an hour.)
- **noun, nominative**: *Stát jsem já.* (I am the state.)
- **noun, accusative**: *Budujme stát pro 40 milionů.* (Let's build the state for 40 millions.)

## stát<sub>noun</sub>

<u>a_modifier</u>	<u>938517</u>	<u>-0.8</u>	<u>gen_2</u>	<u>274456</u>	<u>-0.7</u>
spojený	<u>223381</u>	12.28	hlava	<u>20922</u>	8.7
členský	<u>137993</u>	11.83	zastupování	<u>2716</u>	8.24
americký	<u>29942</u>	9.01	složka	<u>5263</u>	7.9
demokratický	<u>12202</u>	8.46	majetek	<u>5793</u>	7.85

## stát<sub>verb</sub>

<u>has_subj</u>	<u>942837</u>	<u>-3.7</u>	<u>post_v</u>	<u>184481</u>	<u>-1.5</u>
zázrak	<u>4433</u>	7.12	čelo	<u>11624</u>	9.36
nehoda	<u>4438</u>	6.87	pozadí	<u>2507</u>	7.83
socha	<u>3587</u>	6.72	fronta	<u>2654</u>	7.72
kostel	<u>3714</u>	6.39	přepoččet	<u>1098</u>	7.35



# Processing Unknown Words

out-of-vocabulary words:

- terms: *polydaktylie*
- neologisms: *klausoviny* (after V. Klaus)
- typos: *bizardního* (corr. *bizarního*)
- colloquial words: *pláťáky* (*linen trousers*), etc.

flective languages – use **word ending**:

- lemma: *klausoviny*  $\Rightarrow$  *klausovina*
- grammatical information: *bizardního*  $\Rightarrow$  genitive, etc.
- derivational relations: *pláťáky*  $\Leftrightarrow$  *pláťákový*

grouping unknown word forms:

- *polydaktylie, polydaktiliích, polydaktylí, ...*  $\Leftrightarrow$  *polydaktylie*  
 $\Rightarrow$  data extension, precise “guessing” Polydaktylie je vrožená vada, která se vyznačuje nadpočetnými prsty na ruce či noze.

# Spellchecking and Diacritics Restoration

Result of tool CZ accent

Pred domem zastekal cerny pes.

Před domem zaštěkal černý pes.

Morphology processing techniques:

- tuned for a **specific domain**
- other **languages** – Slovak, Polish, German, English, . . .

# Universality and Real-World Applications

**industrial** applications:

- **Seznam.cz, Yandex.ru, Aukro.cz, Václav Havel Library**
  - ▶ indexing and searching very big texts
- **Information System** of Masaryk University
  - ▶ MU + tens of other universities/schools (FHS UK, JAMU, VŠFS, ...)
  - ▶ affiliate projects (theses.cz, odevzdej.cz, repozitar.cz)
  - ▶ indexing, searching and plagiarism detection
- **Internet Language Reference Book** (of Czech)
  - ▶ online authoritative source on Czech orthography and grammar
  - ▶ widely used – 50,000 accesses per day

- 1 Motivation
  - Motivation
- 2 Morphology
  - Morphology
  - Guesser
  - Diacritics
  - Industrial applications
- 3 Parsing and Fact Extraction
  - Syntactic analysis
  - Syntactic trees
  - Extraction of facts
  - Grammar checking
  - Statistical parsing
  - Parsing @NLPCentre

## Simon speaks about sex with Britney Spears



?



# Syntactic analysis

## Natural language syntax

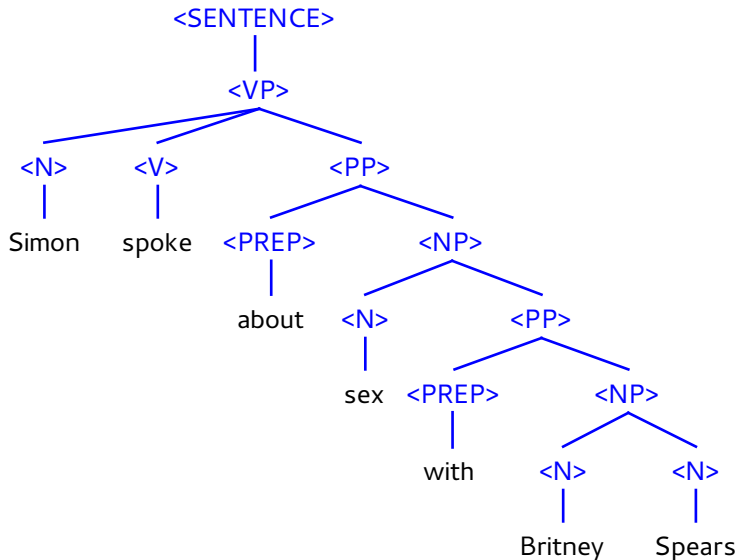
- describes relationships among words

## Automatic syntactic analysis

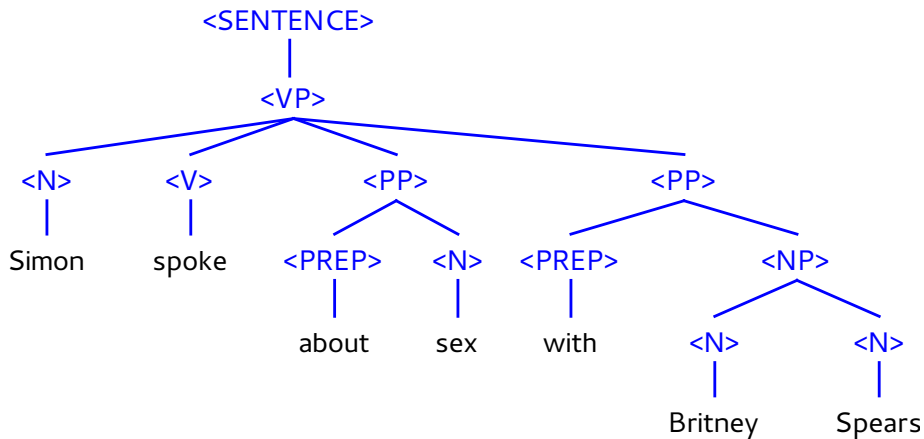
- revealing inter-word relationships on various levels
- detection of noun (prepositional, verb, ...) phrases, clauses

— Simon — speaks — about sex — with Britney Spears —  
— Simon — speaks — about sex with Britney Spears —

# Syntactic trees

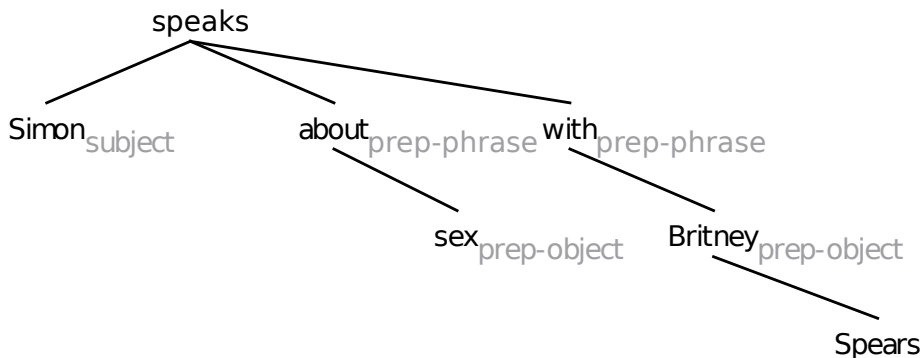


# Syntactic trees

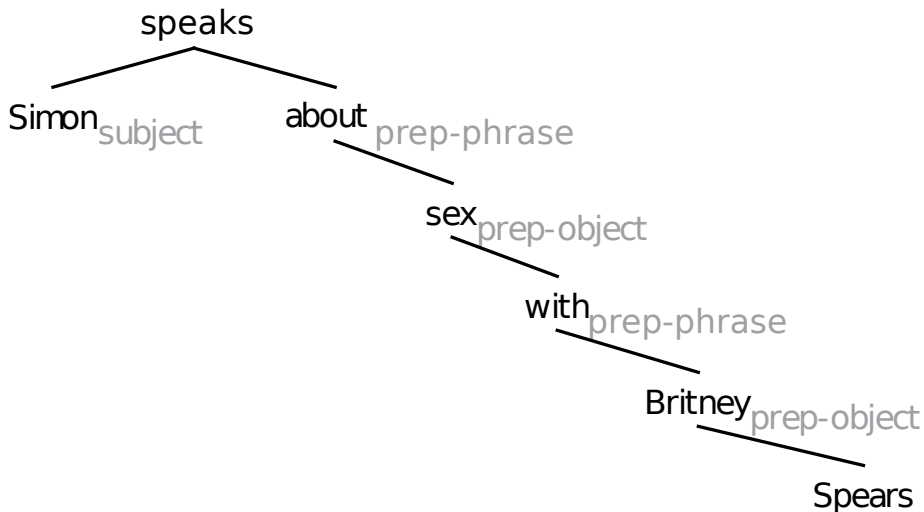




# Syntactic trees

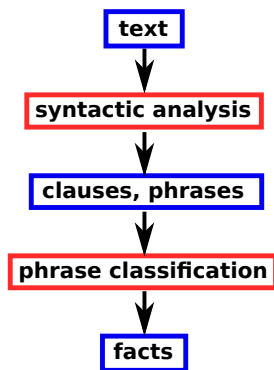
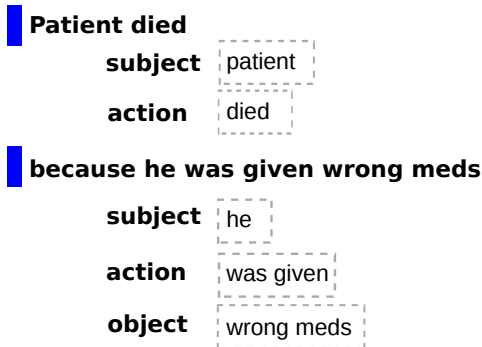


# Syntactic trees



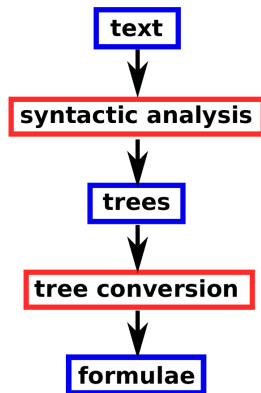
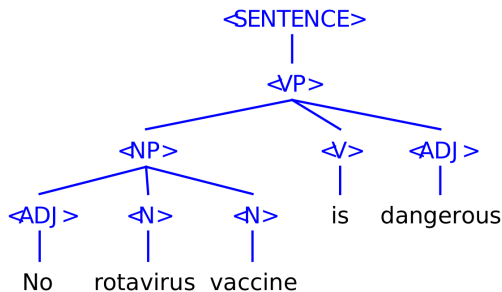
# Extraction of structured information (facts)

## Patient died because he was given wrong meds



## Example: Logical analysis

No rotavirus vaccine is dangerous.



$$\lambda w_1 \lambda t_2 [\mathbf{Not}, [\mathbf{True}_{w_1 t_2}, \lambda w_3 \lambda t_4 (\exists i_5) ([\mathbf{dangerous}_{w_3 t_4} i_5] \wedge [[\mathbf{rotavirus}, \mathbf{vaccine}]_{w_3 t_4}, i_5])]]] \dots \Pi$$
$$\neg \exists x (\mathit{dangerous}(x) \wedge \mathit{rotavirus\_vaccine}(x))$$

# Grammar checking

- Let's eat grandma!
  - ▶ syntactic analysis
  - ▶ detection of non-probable constructions
  - ▶ → grandma is not a usual object of eating
  - ▶ → correction suggestion
- Let's eat, grandma!
  - ▶ life saved :)
- other grammar phenomena
  - ▶ "This is worth try" → "This is worth trying"



# How to analyse natural language syntax?

## Prerequisites

- word level analysis (part of speech, gender, number)
- named entity recognition
- common sense information (e.g. “pregnant” goes with women only)

## Named entity recognition

- determine that e.g. “prof. Václav Šplíchal” is a person
- can be viewed as a sub-task of syntactic analysis

# How to analyse natural language syntax?

## Statistical methods

- people annotate corpus
- statistic methods learn rules from the corpus
- universal across languages (to some extent)
- annotation is expensive
- hard to customize for different applications
- data are usually not big enough

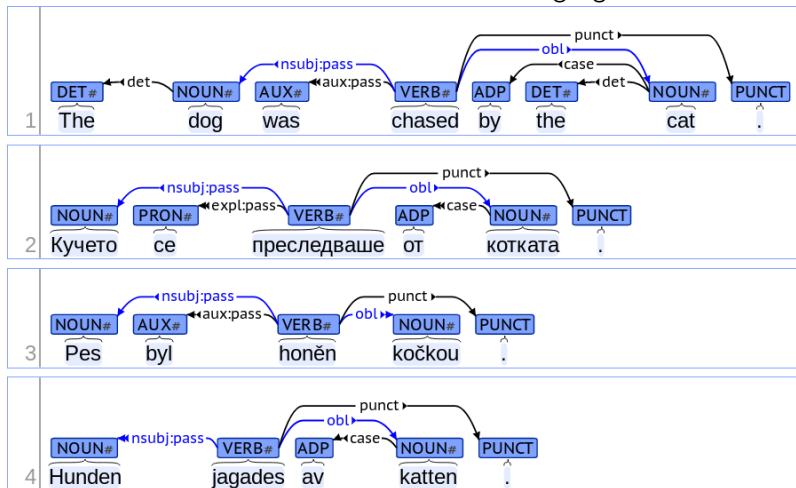
## Rule-based methods

- specialists develop a set of rules (“grammar”)
- not universal, depends on specialists
- grammar can become uneasy to maintain
- easy to customize for different applications

## Hybrids

# Statistical parsing

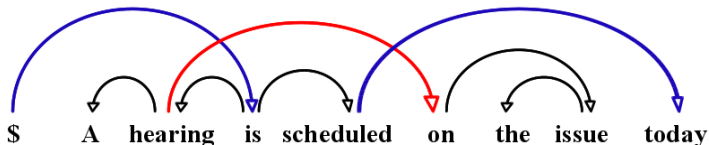
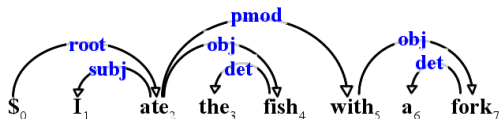
- mostly dependency parsing
- [www.universaldependencies.org](http://www.universaldependencies.org), UD
  - ▶ unified dependency annotation for different languages
  - ▶ more than 200 treebanks in more than 100 languages





# Statistical parsing

- one edge for each word
- difficult for non-projective trees

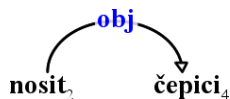


Example from "Dependency Parsing" by Kübler, Nivre, and McDonald, 2009

# Evaluation

information:

- **head** – the governing word
- **dependent** – the modifier word
- **type** – edge label



metrics (percentage):

- **Unlabeled attachment score (UAS)** – words with correct head
- **Labeled attachment score (LAS)** – words with correct head and type
- **Root Accuracy (RA)** – analysis with correct root
- **Complete Match rate (CM)** – fully correct analyses

# Statistical dependency parsing

basic approaches:

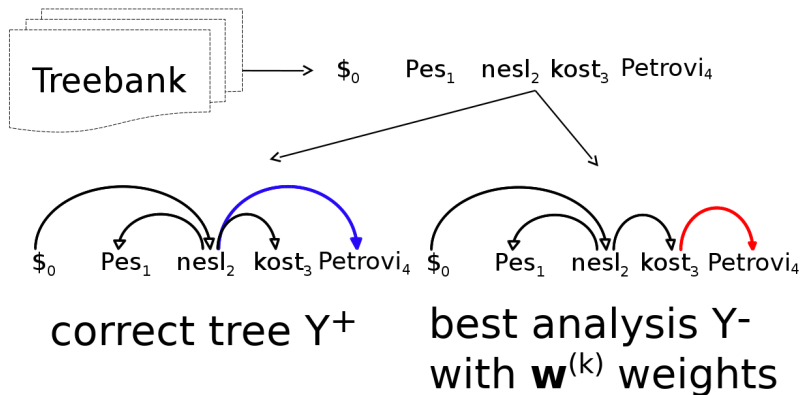
- **graph-based** – tree is created from the **list of edges**
- **transition-based** – sequence of **actions** assigning the dependency **edges**

2 tasks:

- **determine the tree** (search problem)
  - ▶ we know **edge scores**, how to find the **best tree**
  - ▶ e.g. *Maximum Spanning Tree* (McDonald et al, 2005)
- **learning problem**
  - ▶ we have the **treebank**, how to determine the **edge scores**
  - ▶ using **edge features** and **online learning**

# Online learning of dependency edge score

learning the **feature weights  $\mathbf{w}$**



$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \mathbf{f}(X, Y^+) - \mathbf{f}(X, Y^-)$$

# Syntactic analysers in the NLP Centre

- **Synt**
  - ▶ C++, **fast** (0.07 s/sentence)
  - ▶ based on an expressive **meta-grammar**
- **SET**
  - ▶ Python, slower but easily **adaptable**
  - ▶ based on a set of phrase **patterns**
- **Synt+SET**
  - ▶ **rule-based** backbone with **statistical** extensions
  - ▶ **grammars** for Czech, English and Slovak
  - ▶ accuracy **85–90 %** on newspaper texts
- **Word Sketches**
  - ▶ very fast **shallow syntax** for large corpora
  - ▶ **35 languages**

# Conclusions

## Sentence level analysis

- detection of phrases and inter-word relationships
- their further processing

## Applications

- grammar checking
- information analysis of text
- text generation

# References I



Baisa, V. and Kovář, V. (2014).

Information extraction for Czech based on syntactic analysis.

In Vetulani, Z. and Mariani, J., editors, *Human Language Technology Challenges for Computer Science and Linguistics*, pages 155–165, Cham. Springer International Publishing.



Qi, P., Dozat, T., Zhang, Y., and Manning, C. D. (2019).

Universal dependency parsing from scratch.

*arXiv preprint arXiv:1901.10457*.



Straka, M., Straková, J., and Hajič, J. (2019).

Czech text processing with contextual embeddings: Pos tagging, lemmatization, parsing and ner.

In *International Conference on Text, Speech, and Dialogue*, pages 137–150. Springer.

## References II



Zhang, Y., Zhou, H., and Li, Z. (2020).

Fast and accurate neural crf constituency parsing.

*arXiv preprint arXiv:2008.03736.*