07 – Parsing of Czech: Between Rules and Stats

IA161 Advanced Techniques of Natural Language Processing

A. Horák

NLP Centre, FI MU, Brno

November 19, 2020

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.

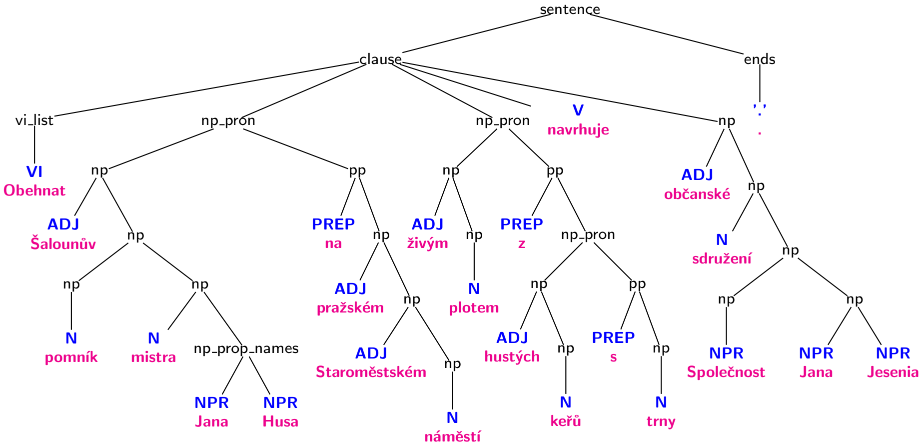
Example (Google translate)

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

Word Level Analysis

"clustering" of word forms in text:

států					*stojíš*
státy					*stály*
státech	$\Longleftrightarrow$	*stát*$_{noun}$	*stát*$_{verb}$	$\Longleftrightarrow$	*stojíme*
státu					*stůjte*
...					...

lemmatization, tagging –

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

Data for Czech Morphology

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

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 noun

a_modifier	938517	-0.8
spojený	223381	12.28
členský	137993	11.83
americký	29942	9.01
demokratický	12202	8.46

gen_2	274456	-0.7
hlava	20922	8.7
zastupování	2716	8.24
složka	5263	7.9
majetek	5793	7.85

stát verb

has_subj	942837	-3.7
zázrak	4433	7.12
nehoda	4438	6.87
socha	3587	6.72
kostel	3714	6.39

post_v	184481	-1.5
čelo	11624	9.36
pozadí	2507	7.83
fronta	2654	7.72
přepočet	1098	7.35

Processing Unknown Words

out-of-vocabulary words:

- terms: *polydaktylie*
- neologisms: *klausoviny*
- typos: *bizardního*
- colloquial words: *plaťáky*, etc.

flective languages – use word ending:

- lemma: *klausoviny* ⇒ *klausovina*
- grammatical information: *bizardního* ⇒ genitive, etc.
- derivational relations: *plaťáky* ⇔ *plaťákový*

grouping unknown word forms:

- *polydaktylie, polydaktiliích, polydaktylií, ...* ⇔ *polydaktylie*
  ⇒ data extension, precise "guessing"

Spellchecking and Diacritics Restoration

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

Simon speaks about sex with Britney Spears

Syntactic analysis

Natural language syntax

- describes relationships among words

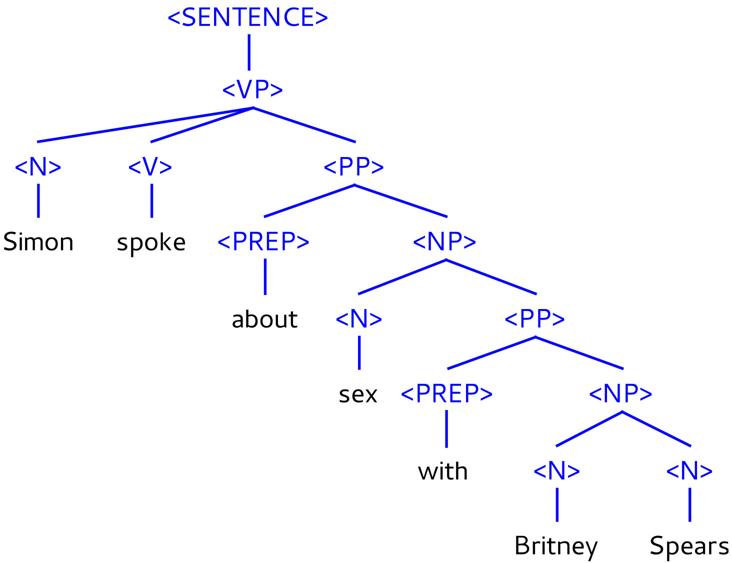
Automatic syntactic analysis

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

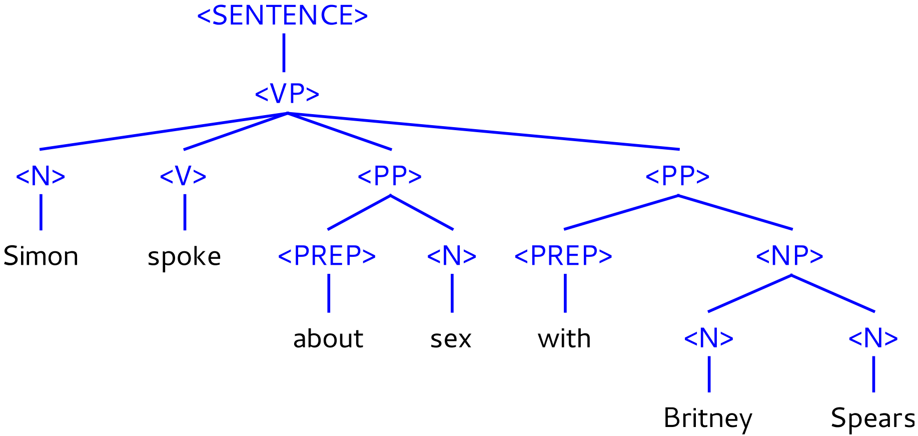
— Simon — speaks — about sex — with Britney Spears —

— Simon — speaks — about sex with Britney Spears —

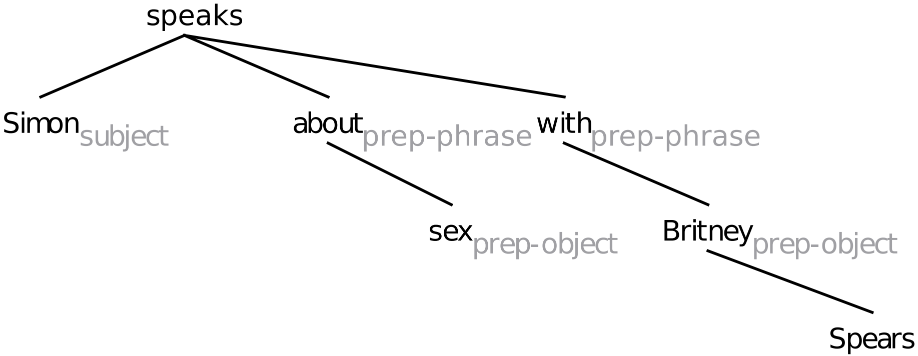
Syntactic trees



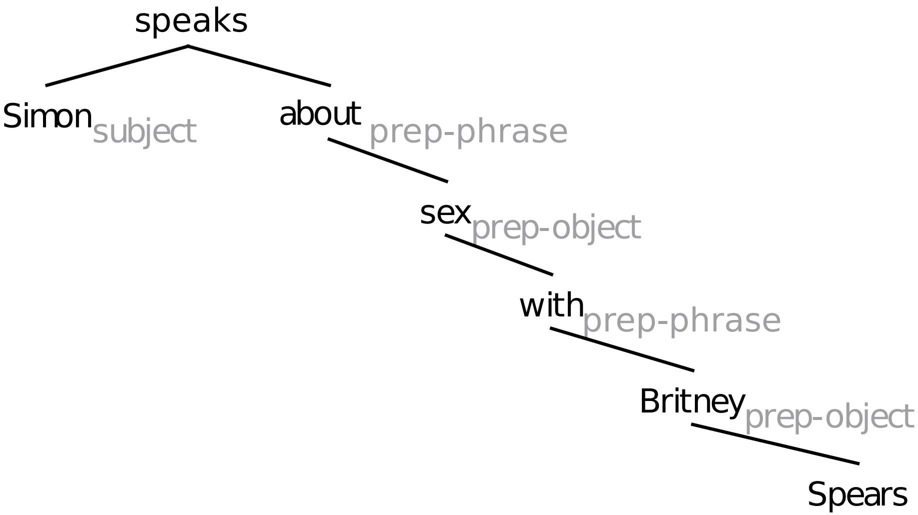
Syntactic trees



Syntactic trees



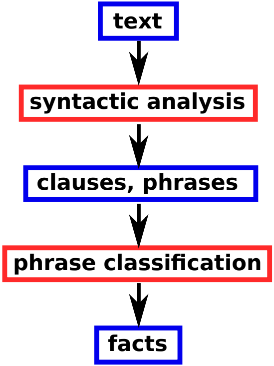
Syntactic trees



Extraction of structured information (facts)

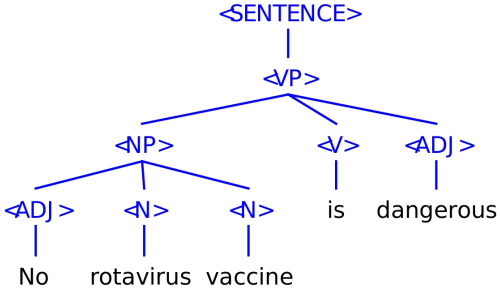
Patient died because he was given wrong meds

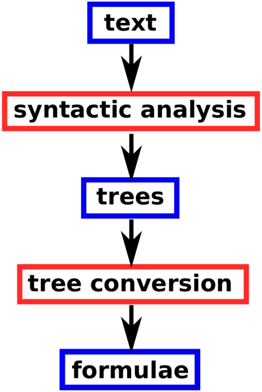
- **Patient died**
  - **subject** patient
  - **action** died
- **because he was given wrong meds**
  - **subject** he
  - **action** was given
  - **object** wrong meds



Example: Logical analysis

No rotavirus vaccine is dangerous.



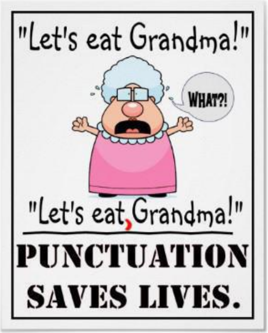


$$\lambda w_1 \lambda t_2 \Big[\mathbf{Not}, \Big[\mathbf{True}_{w_1 t_2}, \lambda w_3 \lambda t_4 (\exists i_5) \Big([\mathbf{dangerous}_{w_3 t_4}\ i_5] \wedge [[\mathbf{rotavirus}, \mathbf{vaccine}]_{w_3 t_4}, i_5] \Big) \Big] \Big] \ldots \pi$$

$$\neg \exists x (dangerous(x) \wedge rotavirus\_vaccine(x))$$

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

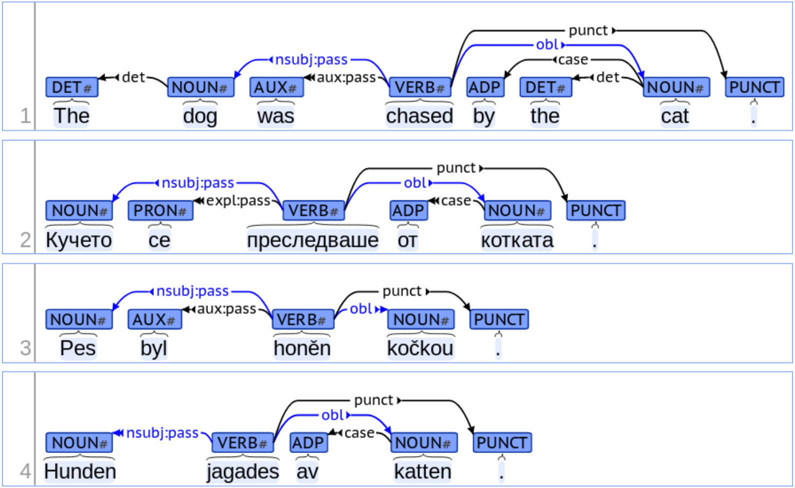
Rule-based methods

- 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

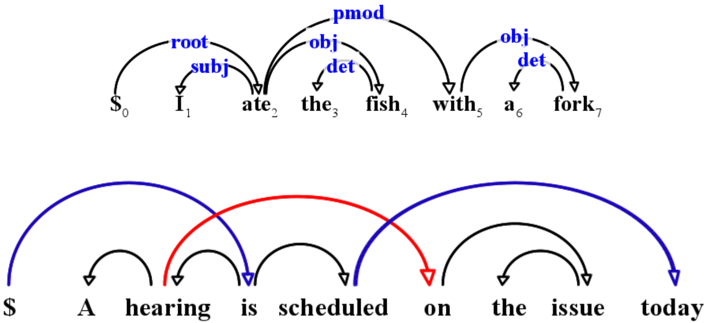
Statistical parsing

- mostly dependency parsing
- www.universaldependencies.org, UD
  - unified dependency annotation for different languages
  - more than 100 treebanks in more than 70 languages



Statistical parsing

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

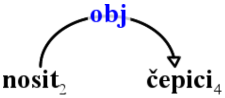


Example from "Dependency Parsing" by Kübler, Nivre, and McDonald, 2009

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:

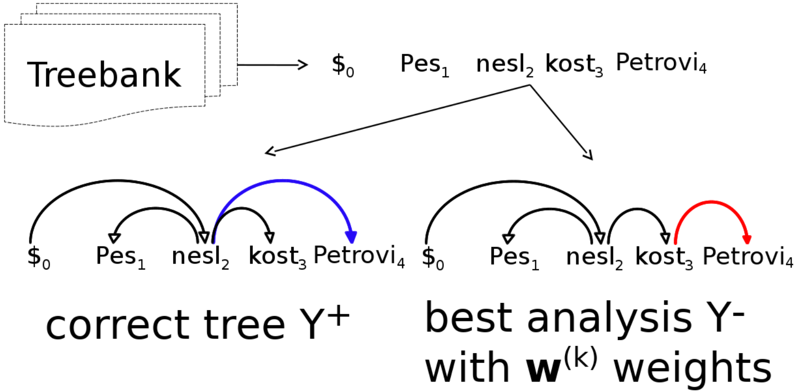
- graph-based – tree is created from the list of edges
- transition-based – sequence of actions assigning the dependency edges

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 **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.