

# Sentence Level Text Analysis

Vojtěch Kovář

Natural Language Processing Centre  
Faculty of Informatics, Masaryk University  
Botanická 68a, 60200 Brno  
xkovar3@fi.muni.cz

Workshop of the Natural Language Processing Centre  
28 May 2013

# Simon spoke about sex with Britney Spears



## Zkolaboval katastr nemovitostí , lidé musejí přespávat v parcích

### Zkolaboval katastr nemovitostí

**kdo/co** | katastr nemovitostí

**přisudek** | Zkolaboval

### lidé musejí přespávat v parcích

**kde** | v parcích

**kdo/co** | lidé

**přisudek** | musejí přespávat

zdroj: [www.infobaden.cz](http://www.infobaden.cz)

# Sentence level (syntactic) analysis

- Natural language syntax

- describes relationships among words

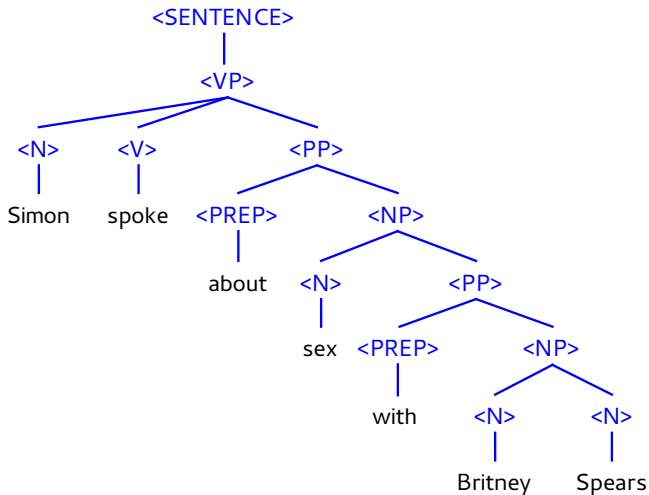
- Automatic syntactic analysis

- revealing inter-word relationships on various levels
- detection of noun (prepositional, verb, ...) phrases, clauses
- finding relationships (dependencies) among the units

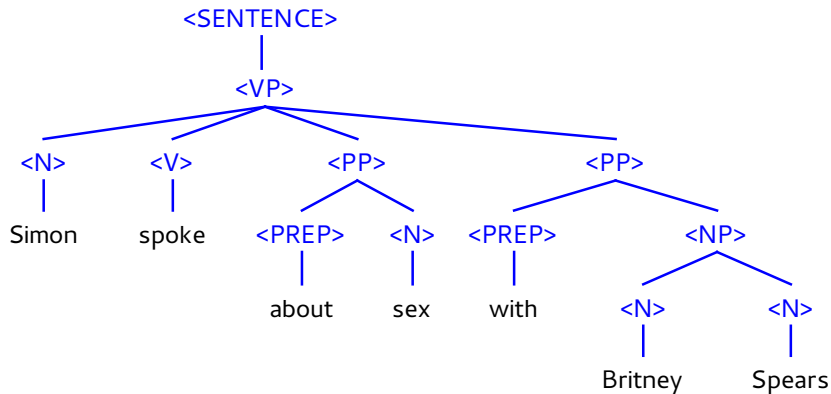
- | Simon | spoke | about sex | with Britney Spears |

- | Simon | spoke | about sex with Britney Spears |

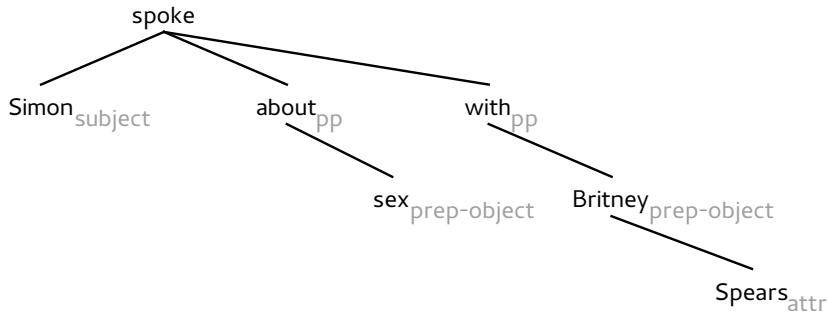
# Syntactic trees



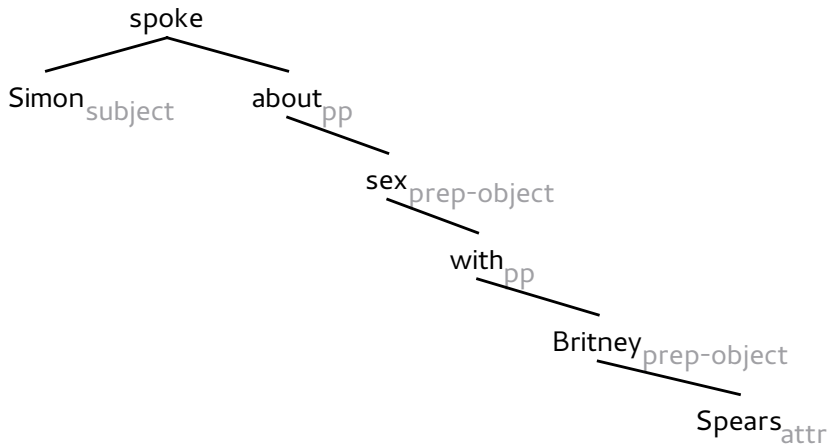
# Syntactic trees



# Syntactic trees



# Syntactic trees





# Why are we doing this?

- Syntactic units are carriers of meaning
  - “in the city”
  - meaning of “in”, “the” is unclear, complicated
  - meaning of “in the city” is simply **where**
- Words are sometimes 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

# Example: Extraction of facts

**Zkolaboval katastr nemovitostí , lidé musejí přespávat v parcích**

**Zkolaboval katastr nemovitostí**

**kdo/co** katastr nemovitostí

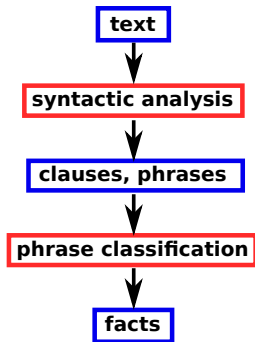
**přisudek** Zkolaboval

**lidé musejí přespávat v parcích**

**kde** v parcích

**kdo/co** lidé

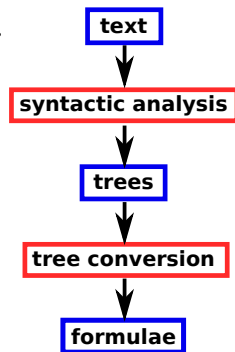
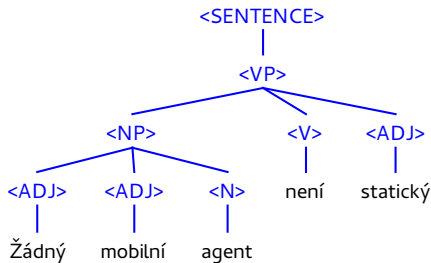
**přisudek** musejí přespávat



zdroj: [www.infobaden.cz](http://www.infobaden.cz)

## Example: Logical analysis

Žádný mobilní agent není statický .



$\neg \exists x (mobilni(x) \wedge agent(x) \wedge staticky(x))$

$\lambda w_1 \lambda t_2 [ \mathbf{Not}, [ \mathbf{True}_{w_1 t_2}, \lambda w_3 \lambda t_4 (\exists i_5) ([ \mathbf{statický}_{w_3 t_4} i_5 ]$   
 $\wedge [ [ \mathbf{mobilní, agent}_{w_3 t_4}, i_5 ] ] ) ] ] ] \dots \Pi$

## Example: 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 :)
- Similarly with other grammar phenomena
  - “This is worth try” → “This is worth try~~ing~~”

# How to analyze natural language syntax?

## ■ Prerequisites

- **word level analysis** (part of speech, gender, number)
- named entity recognition
- lexical semantic 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 analyze 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

# Syntactic analysers in the NLP Centre

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