

# Morfologická analýza Desambiguace

CJBB85

Mgr. Dana Hlaváčková, Ph.D.

# Proč to potřebujeme?

- **morfologické značkování korpusů**
  - zvýšená informační hodnota korpusu
- možnost hledání v korpusu podle morfologických kategorií
- možnost samostatného použití analyzátoru jako morfologické databáze
- předpoklad pro **další stupně analýzy** jazyka
  - syntaktická, sémantická analýza
- předpoklad pro navazující aplikace
  - např. Word Sketch Engine, Morfio
- zapojení do dalších nástrojů pro práci s jazykem
  - kontrola pravopisu, překladače, slovníky, webové prohlížeče
- možnost adaptace pro jiné slovanské jazyky

# Základní pojmy

- **morfologická značka** (*tag, index*)
  - kód přiřazený k jednotlivým tvarům slov nesoucí informaci o jejich morfologických charakteristikách
- **tagset**
  - soubor používaných morfologických značek
- **značkování** (*tagování, tagging, anotace, indexování*)
  - automatické, poloautomatické, ruční
- **morfologický analyzátor** (*morphological analyzer, tagger*)
- **desambiguace** (*disambiguace, disambiguation*)
  - zjednoznačnění, výběr správné morfologické značky v závislosti na kontextu slova

# Značkovací systémy pro češtinu

- **poziční systém** (Hajič, Hlaváčová)
- **atributivní systém** (Osolsobě)
- analyzátor **MORČE** (MORfologie ČEštiny)
  - včetně desambiguace (pravděpodobnostní model)
  - Raab (ÚFAL MFF UK)
- analyzátor **AJKA** (Analyzátor JazyKA)
  - Sedláček (FI MU), **MAJKA** – Šmerk
- **MorphCon** (<http://morphcon.webnode.cz>)
  - převodník českých morfologických systémů
  - Pořízka, Schäfer, Zeman (Olomouc, Bonn, Praha)

# Morfologický analyzátor ajka

- autor – Radek Sedláček, Pavel Šmerk, Marek Veber
- formální (algoritmický) popis morfologie (Klára Osolsobě)
- systém **atribut – hodnota**
- slovo = *řetězec znaků ohraničený z obou stran mezerami*
- **segmentace slova**
  - **KMZ – IS – T**
    - *kmenový základ, intersegment, koncovka*
- koncovkové množiny
- slovník kmenů
- slovník intersegmentů
- seznam vzorů

# Průběh morfologické analýzy

- rozeznání neohebných slovních druhů
  - po rozeznání analýza skončí
- rozeznávání slova od začátku
  - záporka –ne
  - superlativní prefix –nej
- segmentace slova od konce
  - koncovka
  - intersegment
  - kmenový základ
  - přiřazení ke vzoru
- **nej-ne-oblíben-ějšími**

# Desambiguace

- Ruční, automatická (statistická), pravidlová, hybridní
- Některé tvary nelze desambiguovat – není možné jednoznačně vybrat správnou značku ani na základě kontextu

*Německá firma Tebis v Hannoveru představila kompaktní zařízení pro firemní modelárny.*

*Technické řešení těsnění nádrží a podlah...*

*Myrha je přírodní pryskyřice, aloe je vonné dřevo.*

*V osmi letech měl za sebou účinkování v mnoha televizních show...*

*Dolní listy jsou obvejčité, čepel se zužuje v ouškatý řapík.*

*Jak lze z názvu vytušit, jde o nástroje pro zprostředkování databázových transakcí a tvorbu dotazů prostřednictvím standardu SQL.*

*Jak nám řekl ředitel tohoto závodu, nebyla to jejich chyba...*

jak – k1, k6, k8, k9

# Odkazy

- <http://nlp.fi.muni.cz/projekty/wwwajka>
- <http://ufal.mff.cuni.cz/morce>
- CQL [tag=„“]
  - *Corpus Query Language*