Morfologická analýza
Desambiguace

Počítačové nástroje pro češtinu
jaro 2021

Markéta Audy Masopustová

Co je morfologická analýza?

- přiřazení morfologických informací každému slovu, tzn.:
  - základní tvar slova (*lemma*)
  - morfologické údaje ve formě značky *(tagu*), který obsahuje slovní druh, pád/osoba, číslo, vid, slohový příznak atd.
- probíhá automaticky

Proč ji 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
- guesser
  - „hádá“ interpretaci slov bez morfologického slovníku na základě pravidel/statistiky

Značkovací systémy pro češtinu

- Pražský (poziční)
  - autoři: Hajič, Hlaváčová (ÚFAL)
  - korpusy ČNK
  - 16 **povinných** pozic
  - každá pozice odpovídá víceméně nějaké kategorii z gramatiky (slovní druh, stupeň, osoba, číslo, …)

```
kočku /kočka/NNFS4-----A-----  honil /honit/VpIS----R-AA---I
kočku /kočka/NNFS4-----A-----  honit /honit/Vf--------A----I
kočka /kočka/NNFS1-----A-----  honila /honit/VpFS----R-AA---I
```

- Brněnský (atributový)
  - autorka: Osolsobě
  - většina korpusů ve Sketch Enginu (př. řada czTenTen)
  - proměnlivý počet pozic podle toho, které kategorie slovo vyjadřuje
  - úspornější, přehlednější, snadno rozšiřitelný

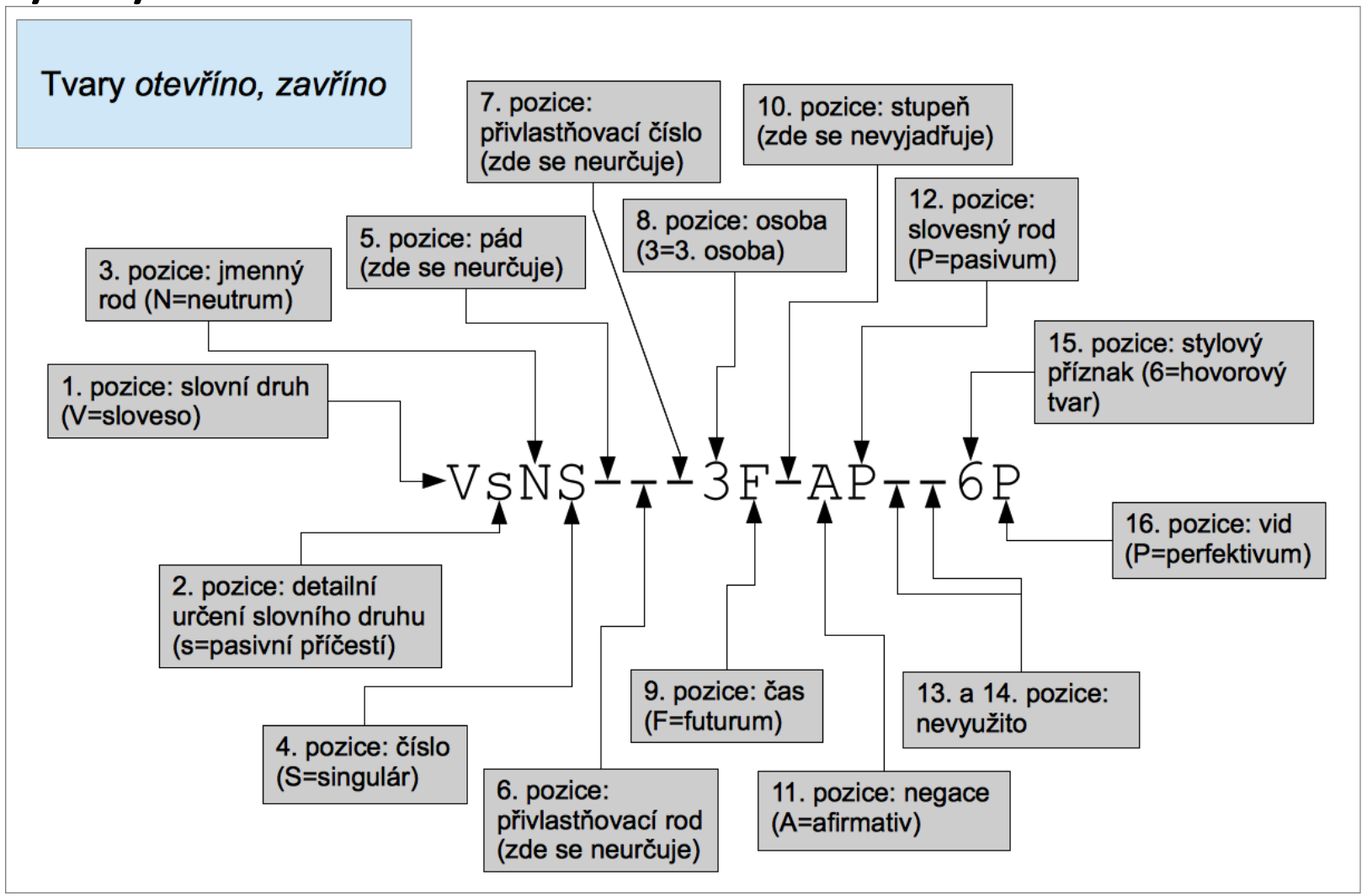
```
kočka /k1gFnSc1/kočka  honí /k5eAaImIp3nS/honit
Kočky /k1gFnPc1/kočka  nehoní /k5eNaImIp3nP/honit
kočky /k1gFnPc4/kočka  honila /k5eAaImAgFnS/honit
```

Pražský systém značek

1. pozice: slovní druh
2. pozice: detailní určení slovního druhu
3. pozice: jmenný rod
4. pozice: číslo
5. pozice: pád
6. pozice: přivlastňovací rod
7. pozice: přivlastňovací číslo
8. pozice: osoba
9. pozice: čas
10. pozice: stupeň
11. pozice: negace
12. pozice: aktivum/pasivum
13. pozice: nepoužito
14. pozice: nepoužito
15. pozice: varianta, stylový příznak
16. pozice: vid

Detailní popis: https://wiki.korpus.cz/doku.php/seznamy:tagy

Pražský systém značek II



zdroj: https://wiki.korpus.cz/lib/exe/detail.php/seznamy:tagy.png?id=seznamy%3Atagy

Brněnský systém značek

- k1.*–k0 slovní druhy
- kA zkratky
- kY podmiňovací způsob (dříve)
- kI interpunkce

Detailní popis: https://www.sketchengine.eu/wp-content/uploads/Czech_Morphological_Tagset_Revisited_2011.pdf

Morfologické analyzátory pro češtinu

- analyzátor MORČE (MORfologie ČEštiny)
  - včetně desambiguace (pravděpodobnostní model)
  - autor Raab (ÚFAL MFF UK)
  - morfologický slovník MorfFlex, tagger MorphoDiTa
- analyzátor AJKA (Analyzátor JazyKA)
  - autor Sedláček (FI MU), navazuje na něj MAJKA (autor Šmerk)

Morfologický analyzátor Ajka

- autor Radek Sedláček
- založena na formálním (algoritmickém) popisu české morfologie (Klára Osolsobě) a strojovém slovníku češtiny, který je reprezentován datovou strukturou *trie* (nevýhodou jsou vysoké nároky na paměť)
- veškerá data pro morfologický analyzátor jsou uložena ve slovníku kmenů a souboru koncovkových množin a vzorů
- systém atribut – hodnota
- slovo = *řetězec znaků ohraničený z obou stran mezerami*
- segmentace slova

  KmZ – IS – T

  kmenový základ, intersegment, koncovka
- příliš složitá a nerozšiřitelná

Morfologický analyzátor Majka

- autor Pavel Šmerk
- navazuje na předchozí analyzátor Ajka
- nový formát slovníku a souboru vzorů
- pravidelné jevy uloženy v souboru vzorů, nepravidelné ve slovníku
- kompletně založen na konečných automatech
- jednodušší, rychlejší
- důkaz, že pro češtinu není třeba specializovaných datových struktur nebo algoritmů
- díky Majce vznikl podobný analyzátor pro slovenštinu
- rozšířena o slovenštinu, polštinu, angličtinu
- doplnění diakritiky CzAccent
- Majku využívá např. Seznam.cz

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-ějš-ími

Desambiguace

- zjednoznačnění, odstranění homonymie
- manuální, statistická (94 %), pravidlová, hybridní
- některé tvary nelze desambiguovat – ani na základě kontextu nelze jednoznačně přiřadit správnou značku

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

CQL

- Corpus Query Language
- systém závorek, klíčových slov, hledaných slov a regulárních výrazů
- umožňuje vyhledávat v korpusu jednoduché i složité a velice specifické informace
- možnost vyhledávat podle konkrétního tvaru slova (*word*), základního tvaru slova (*lemma*), nebo morfologické značky (*tag*)
  - formát [atribut = „hodnota“]
    - atribut: word, lemma, tag
    - hodnota: samotný výraz nebo výraz specifikovaný regulárním výrazem

[word ="kočkou"]

[tag="C.*"]

[lemma="nosit"][lemma="dříví"][]{1,3}[lemma="les"] within <s/>

Odkazy

- Ajka: http://nlp.fi.muni.cz/projekty/wwwajka
- atributivní systém: https://www.sketchengine.eu/tagset-referencefor-czech/
- poziční systém: https://wiki.korpus.cz/doku.php/seznamy:tagy
- MorphoDiTa: http://lindat.mff.cuni.cz/services/morphodita/
- CQL: https://www.sketchengine.eu/documentation/corpus-querying/

Úkoly v korpusu SYN2020

- najděte všechny výskyty slova *obec*
- najděte všechny výskyty slovního tvaru *pěšími*
- najděte všechna neutra v instrumentálu plurálu
- najděte všechna životná maskulina končící na *-a*

Úkoly v korpusu csTenTen17

- najděte všechny výskyty slova *obec*
- najděte všechny výskyty slovního tvaru *pěšími*
- najděte všechna neutra v instrumentálu plurálu
- najděte všechna životná maskulina končící na *-a*

Děkuji za pozornost.