# Morfologická analýza pro češtinu

morfologický analyzátor

K. Pala, P. Šmerk

podzim 2019

# Úvod 1

- algoritmický popis české morfologie
- Východiska a technická řešení
- Slovní druhy, gramatické kategorie a značky (tagsety), lemmata
- Morfologická analýza a syntéza – segmentace
- Flexe – ohýbání – deklinace a konjugace, vzory, neohebné sl. druhy (prep., konj., part.)
- Notace, formalizace, vhodné pro SP

# Úvod 2

- derivační morfologie – tvoření slov, detaily?
- automaticky lze odvozovat:

- posesivní (přivlastňovací) adj. od substantiv
- deadjektivní adverbia, *dobrý, dobře*
- stupňování adjektiv (výjimky *dobrý*, *lepší* ?)
- deverbativní substantiva a adj. (*boj–ovat-ový*)
- přechylování (příjmení a rodinná jména)
- deminutiva (*dům, domek, domeček)*
- činitelská jména a další (*učit-el)*

# Ajka, Majka, morfologická analýza

Nástroje pro práci s morfologickou databází

- Morfologická databáze a slovník kmenů (cca 400 tis. položek)

- Pokud možno úplné seznamy vzorů, podvzorů, odchylek, variant, výjimky
- příklad: seznam vzorů s 1-5 kmeny
- Přiřazení mezi kmeny a vzory – seznamy dává databáze

# Segmentace slova

- Morfologická analýza (syntéza) vychází z definic vzorů podle slovních druhů (klasicky 10, H&J). Slovoforma – slovní tvar – je dán kombinací 5 složek
- Prefix (fakultativní)
- Kmenový základ (KMZ, stabilní)
- Intersegment (IS, nestabilní)
- Koncovka (T= koncovka flektivní)
- Postfix (fakultativní)

# Vzory a konc. množiny

- 5 komponentové členění umožňuje operativně zpracovat hláskové alternace kmene v závislosti na koncovce (koncovkách, *vlk - vlci*).

Vzor je definován jako kombinace

PREFIX+KMZ+IS (lexikální část)

KONCOVKOVÉ MNOŽINY

(koncovka opatřená morfologickou značkou – potencionální gramatické významy příslušné kombinace nezávislé na kontextu).

# Příklad vzoru

*poslat/pošlu*

PREFIX: po-

KMZ: -0-

IS –sla-/-slá-/-šl-/-šle-

T (koncovky): (-t,-l,-la,..)/(-n,-na,..)/(-u,-i,…)/(-š,-0,-me…)

Postfix: -0

V korpusu SYN2000 se najde 10754 tvarů

# Příklad vzoru II

- po
- +0 {maže}
- <sla> W3A,W5A
- <slav> W7
- <slá> W4C
- <slan> PRT1,PRT2
- <slán> V13,PRT3
- <slavš> PRMP
- <šl> W1B,W2B,W6A
- <šle> W1A
- <šlouc> PRMI

# Vlastnosti uvedeného popisu

- Používá se stále, pravidelně se doplňuje
- Využití: značkování korpusů, modul pro syntaktickou analýzu, součást korektorů
- Pokrytí: testováno na SYN2000, cca 96 % (4 % - číselné výrazy, zkratky, jiné jazyky)
- komplementárně je k dispozici seznam kolokací – cca 110 000 pol. – propojení
- Lze jej rozšiřovat (spisovná norma – nespisovné útvary, změny v kodifikaci)

# Nová zjištění u vzorů

- Vzor hrad/les (MČ2 – tab. Str. 306)
- Vzor žena (hláskové alternace ve kmeni kromě gen. pl.)
- Vzor píseň/kost (MČ2 str. 330-1)
- Vzor muž/obyvatel (jen v PMČ, v SYN2000 11028 příkladů, všechny uvedeny jako nom. sg. , ve skutečnosti přinejmenším ¾ z uveného počtu jsou tvary gen. pl.!!!)

# Zpřesnění vzorů

- branou – bránou (679:307)
- branám – bránám (43:1)
- branách – bránách (18:1)
- branami – bránami (307:8),
- silou – sílou (3207:53)
- silám – sílám (585:1)
- silách - sílách (1046:1)
- silami – sílami (1516:2)
- kravou – krávou (3:53)
- kravám – krávám (26:0)
- kravách – krávách (13:0)
- kravami – krávami (32:6).

# Doplnění vzorů II

- Adjektiva
- - stupňování
- - derivace adverbií

# Doplnění vzorů III

- Slovesa
- - brát – bral – brán nebrat – nebral – nebrán

(alternace kmenotvorné přípony v závislosti na počtu slabik)

- IV. třída slovesná (kolísání mezi vzory prosit/trpět-sázet)
- II. třída vzor tisknout – tvary minulého (l-ového) participia (-0-/-nu-) – kolísání – úplné seznamy

# Silná slovesa v češtině?

- Členění sloves
- I. třída
- II. třída vzor začít
- III. třída vzor krýt

(počet do 200 neprefigovaných sloves)

# Derivační vzory zájmen a číslovek

- dev-ět-0
- dev-ít-i
- dev-át-ý
- dev-ater-o/ý
- dev-ítk-a

# Morfologické značky (tagy)

- systém atribut – hodnota, např.
- Podle slovních druhů (10)
- [word=slonům & lemma="slon" & tag="k1gMnPc3"]
- [word="dobří" & lemma="dobrý" & tag="k2eAgMnPc1d1"]
- [word="brát" & lemma="brát" & tag="k5mFaI"]
- [word="beru" & lemma="brát" & tag="k5mIp1nSaI"]
- [word="bral" & lemma="brát" & tag="k5mAgMnSaI"]

# Poziční vs. atributový systém

- [tag="NNMS(1|2|3|4|5|6|7).*"] > [tag="k1gMnSc(1|2|3|4|5|6|7).*"]
- [tag="NNIS(1|2|3|4|5|6|7).*"] > [tag="k1gNnSc(1|2|3|4|5|6|7).*"]
- [tag="NNFS(1|2|3|4|5|6|7).*"] > [tag="k1gFnSc(1|2|3|4|5|6|7).*"]
- [tag="NNNS(1|2|3|4|5|6|7).*"] > [tag="k1gInSc(1|2|3|4|5|6|7).*"]

# Soubory značek?

- Brno vs. Praha, atributy vs. pozice
- Nový soubor značek (tagset)
- návrh společného – UK, MU, ÚJČ, další?
- příprava na podzim

# Co je potřeba k propojení?

- Nástroje – každý má svoje
- seznamy, slovníky, data
- výměna výsledků – oboustranně
- ostatní pracoviště?