GAK – CJBB84

st. 8.00-9.30

3.10. 2018

Vyhledávání v korpusu s/bez použití lemmatizace a morfologických značek

- **Základní vyhledávání v korpusu**
- **Obsah:**
  Vyhledávání tvaru slova nebo slovního spojení
  Vyhledávání podle atributu **lc** (lowercase)
  Vyhledávání podle atributu **lemma**
  Vyhledávání podle atributu **tag** (morfologická značka)
  Nastavení implicitního atributu
  Hledání v rámci jedné věty

vyhledejte v korpusu SYN2010

- slovní tvar *hnát*
- lemma *hnát*
- všechna slovesa v infinitivu
- všechna slovesa na *-át*

Otázky

- Pomocí funkce statistik zjistěte, jak je v korpusu SYN rozložen poměr tvarů slovních druhů u slovního tvaru *hnát* a u lemmatu *hnát*.
- Pokuste se pomocí P-filtru získat seznam slovesných/substantivních tvarů a zjistěte, jaká je chybovost značkování.

Možnost vyhledávat tvar bez použití morfologické značky

- Jak vyhledat tvary infinitivu bez morfologické značky ?
- Hledání pomocí formy – na co v češtině končí infinitiv ?

Zakončení

- -t
- -ti
- -ci
- -ct

Je možná nějaká přesnější specifikace ?

- .*t (? život, dost, opět, část, procent, ...)
- .*ti (? proti, ti, společnosti, děti, části, ...)
- .*ci (?práci, věci, noci, ulici, rámci, ...)
- .*ct (?patnáct, dvanáct, čtrnáct, jedenáct, ...)

Jak ?

- Existují nějaká předpověditelná tvrzení týkající se toho, co může/nemůže stát před *-t, -ti, -ci, -ct* ?
- Co nám může pomoci ?
- Gramatika
- Pozorování korpusových dat

Co předchází ?

- V
- C

V

- at, át
- et, ět, ét
- it, ít
- ot, ót
- ut, ůt
- yt, ýt

C

- .*[cčdďfghjklmnňpqrřsštťvwxzž]t
- .*[cčdďfghjklmnňpqrřsštťvwxzž]ti

Co předchází ?

- KmV
- KoV v případě otevřeného kořene (s 0 KmV)
- finála uzavřeného kořene

KmV infinitivu : gramatika nebo náhoda

- dělat, ʔzvířat, ʔstát, ʔtentokrát
- ʔlet slyšet ʔopět, ʔpět, ʔestét
- mít, ʔpocit, mluvit, ʔsít
- ʔživot, ʔkvót
- ʔminut, dosáhnout, ʔaut, ʔlhůt
- být, ʔbyt, ʔKamarýt

KoV

- III. třída vzor krýt
- atematická slovesa

finála kořene

- I. třída vzor *nést*
- I. třída vzor *péci*

dotaz

- .*[aáeěiíýy]t | .*out
- .*[aáeěiíýy]ti | .*outi

A co

- hřbet
- internet
- karet
- ret
- cigaret
- kvartet
- kulomet

?

- Existují nějaká další omezení ?
- Týkají se všech V v roli KmV/Kov ?

Dotaz

- .*[aáeěiíýt | .*out
- .*[aáeěiíýti | .*outi
- N-filtr
- .*[bpfvmdtnr]et
- .*[bpfvmdtnr]eti

Dotaz

- .*[szc]t
- .*[szc]ti
- N-filtr
- .*[cčdďfghjklmnňpqrřsštťvwxzžaeěioóúy][szc]t
- .*[cčdďfghjklmnňpqrřsštťvwxzžaeěioóúy][szc]ti

Všimněme si

- .*náct
- nárůst

Gramatika

- Uzavřené a otevřené třídy

Samostatné řešení úkolů z morfologie a tvoření slov

- Stejným způsobem popište postup, který může pomoci při vyhledávání l-ových příčestí.
- Pomocí analogických strategií lze vyhledávat také podstatná jména slovesná na *ní/tí*. Popište jak.