

# **GAK – CJBB84**

čt. 12.30-14.00

6.10.2011



# 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

- `.*[cčdd'fghjklmnňpqrřsštt'vwxyzž]t`
- `.*[cčdd'fghjklmnňpqrřsštt'vwxyzž]ti`

# Co předchází ?

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

# KmV infinitivu : gramatika nebo náhoda

- dělat, ?zvířat, ?stát, ?tentokrát
- ?let slyšet ?opět, ?pět, ?estét
- mít, ?pocit, mluvit, ?sít
- ?život, ?kvót
- ?minut, dosáhnout, ?aut, ?lhůt
- být, ?byt, ?Kamarýt

# KoV

- III. třída vzor krýt
- atematická slovesa

# finála kořene

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

# dotaz

- `.*[aáeěííý]t|.*out`
- `.*[aáeěííý]ti|.*outi`



# A co

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

?

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

# Dotaz

- `.*[aáeěiíý]t|.*out`
- `.*[aáeěiíý]ti|.*outi`
- N-filtr
- `.*[bpfvmdtnr]et`
- `.*[bpfvmdtnr]eti`

# Dotaz

- `.*[szc]t`
- `.*[szc]ti`
- N-filtr
- `.*[cčdd'fghjklmnňpqrřsštt'vwxyzžaeěioóúy][szc]  
t`
- `.*[cčdd'fghjklmnňpqrřsštt'vwxyzžaeěioóúy][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.