CJBB84 Morfologie a korpus (st.12.30-14.00 G022)
[2]) každou lichou středu 12:30--14:05 (29.2., 14.3., 28.3., 11.4., 25.4., 9.5.)

Kontaktní výuka dle rozvrhu bude probíhat každou „lichou středu“, tedy v dny označené daty v harmonogramu (viz níže).
Na sudé týdny budou zadány úkoly. Ty je třeba odevzdat nejpozději den dopředu (v úterý ráno), a to jako přílohu (text word, rtf) na můj mail.

Harmonogram seminářů

29. 2. Praktické problémy lemmatizace : Když jednomu tvaru může odpovídat více lemmat.
a) Na www http://ucnk.ff.cuni.cz/ - Dohody a registrace si opatřete uživatelská přístupová práva k programu Bonito.
b) Na www http://ucnk.ff.cuni.cz/ - Manuál a instalace - Instalace si stáhněte na svůj domácí počítač program Bonito.
c) Na www http://ucnk.ff.cuni.cz/ - Dostupné korpusy si nastudujte informace o korpusech ČNK.
d) Na www http://ucnk.ff.cuni.cz/ - Manuál a instalace si nastudujte informace o tom, jak zadávat dotazy v programu Bonito.
7. 3. Úkol – Chybná disambiguace a jak se s ní vypořádat.
14. 3. Jak vyhledávat substantiva podle typu skloňování ?
21.3. Úkol – pravidla distribuce e/ě v koncovkách sb. flexe
28. 3. Vyhledávání v korpusu bez použití lemmatizace a morfologických značek (příklad tvary infinitivu).
4.4. Úkol – l-ové příčestí
11. 4. Vyhledávání víceslovných gramatických forem v korpusu.
18. 4. Úkol : Minulý čas složený.
25.4. Syntetické futurum v češtině
2. 5. Úkol: Porovnání tvarů analytických a syntetických (tento úkol zpracujete jako prezentaci). Každý student bude zpracovávat jiné sloveso/slovesný význam.
9. 5. Předtermín

Podmínky pro ukončení (5 kr. zk)

1. Odevzdání všech úkolů (4)
2. Odevzdání prezentace + ústní zkouška doplňující otázky k úkolům a prezentaci

29. 2. Praktické problémy lemmatizace : Když jednomu tvaru může odpovídat více lemmat.
a) Na www http://ucnk.ff.cuni.cz/ - Dohody a registrace si opatřete uživatelská přístupová práva k programu Bonito.

b) Na www http://ucnk.ff.cuni.cz/ - Manuál a instalace - Instalace si stáhněte na svůj domácí počítač program Bonito.
c) Na www http://ucnk.ff.cuni.cz/ - Dostupné korpusy si nastudujte informace o korpusech ČNK.
d) Na www http://ucnk.ff.cuni.cz/ - Manuál a instalace si nastudujte informace o tom, jak zadávat dotazy v programu Bonito.

Praktické problémy lemmatizace : Když jednomu tvaru může odpovídat více lemmat - disambiguace a homonymie
Vyhledáme v korpusu SYN2010 tvar *sál*
Zjistíme frekvenční distribuci lemmat
Pozorujeme vybraná data a zjišťujeme chyby
Je ruční analýza jedinou možností ?
Co musí platit, aby slovní tvar *sál* byl tvarem substantiva?
Substantivum může být součástí jmenné skupiny. Vlevo od substantiva může stát adjektivum a další jména ve stejném pádě (gramatická shoda), také předložka pojící se s příslušným pádem stojí před substantivem a nemůže stát před slovesem.
N-filtrem lze odstranit řádky, které jsou prokazatelně správně tagovány jako lemma *sál* a tag *NN..[14].**.
Co musí platit, aby slovní tvar *sál* byl slovesným tvarem ?
Česká věta obsahuje jen jeden tvar verba finita. Pokud bude v okolí tvar slovesa *být* (VB.....[12].* - *jsem, jsi, jsme, jste*, nebo Vc,* - *bych, bys, ...*), půjde o sloveso *sát* v minulém čase, nebo v kondicionálu.
Budeme prohledávat jednotlivé intervaly.
V intervalu <-1,-1> jsme našli doklady.
Jaký bude další postup ?
7. 3. Úkol:
Mohou se tvary slovesa *být* vyskytnout i na dalších pozicích ?
Jaká omezení budou platit a proč?
Popište další třídění dat směřující k odstranění chyb, které jsou výsledkem chybné desambiguace.