CJBB75 Základy využití korpusu pro praxi

st. 9.10-10.50 G13

Různé korpusy a rozdíly v anotačních schématech

- Tokenizace, automatická anotace, d[ei]sambiguace
- Anotace velkých synchronních korpusů ČNK
- Anotace mluvených korpusů
- Anotace KSK
- Pražský a brněnský systém anotací
- Specifika anotací SYN2005
- Co se skrývá za označením slovní druh X.*

Tokenizace, automatická anotace, d[ei]sambiguace

- Tokenizace – rozdělení textu na jednotky, s nimiž se dále pracuje při strojovém zpracování PJ.
- Automatická anotace – automatická morfologická analýza – slovník (word/lemma/tag) – je obecně víceznačná.

Tokenizace

- grafické slovo
- slova se spojovníkem
- spřežky
- zkratky

Morfologická homonymie – víceznačnost formy

- Zdraví je velký dar.

zdraví

- ✖ *zdraví/zdraví/NNNS1.**
- ✖ *zdraví/zdraví/NNNS2.**
- ✖ *zdraví/zdraví/NNNS3.**
- ✖ *zdraví/zdraví/NNNS4.**
- ✖ *zdraví/zdraví/NNNS5.**
- ✖ *zdraví/zdraví/NNNS6.**
- ✖ *zdraví/zdraví/NNNP1.**
- ✖ *zdraví/zdraví/NNNP2.**
- ✖ *zdraví/zdraví/NNNP3.**
- ✖ *zdraví/zdraví/NNNP4.**
- ✖ *zdraví/zdraví/NNNP5.**

zdraví

- *zdraví/zdravý/AAMP1.**
- *zdraví/zdravý/AAMP5.**
- *zdraví/zdravit/VB-S---3P.**
- *zdraví/zdravit/VB-P---3P.**
- *zdraví/zdravět/VB-S---3P.**

je

- *je/být/VB-S---3P.**
- *je/on/PPXP4—3.**
- *je/on/PPNS4—3.**

velký

- *velký/velký/AAIS1.**
- *velký/velký/AAIS4.**
- *velký/velký/AAIS5.**
- *velký/velký/AAMS1.**
- *velký/velký/AAMS5.**

dar

- *dar/dar/NNIS1.**
- *dar/dar/NNIS4.**

d[ei]sambiguace

- Zjednoznačnění – volba kontextově správné varianty.
- stochastické metody
- pravidlové metody
- hybridní metody

Zdraví je velký dar.

- *zdraví/zdraví/NNNS1.**
- *je/být/VB-S---3P.**
- *velký/velký/AAIS1.**
- *dar/dar/NNIS1.**

Anotace velkých synchronních korpusů ČNK

- Tzv. pražský systém založený na morfologické analýze (slovníku) J. Hajiče
- Stochastické metody disambiguace
- Pravidlové metody disambiguace
- Guessery/hadače

Struktura značky

http://ucnk.ff.cuni.cz/bonito/znacky.php

- Každá značka je řetězcem 16 znaků (16. pozice chybí pouze v korpusech SYN2000 a ORWELL).
- Značka je konstruována tak, aby každá pozice odpovídala jedné morfologické kategorii podle víceméně tradičního lingvistického pojetí.
- Každé hodnotě v dané kategorii odpovídá jeden znak, převážně písmeno velké abecedy (např. 'P' pro plurál, neboli množné číslo), výjimečně i jiný znak (např. 'f' pro infinitiv, nebo ',' pro podřadicí spojky).
- Hodnota, která nedává smysl (např. pád u sloves), je reprezentována znakem '-' (pomlčka).

Anotace mluvených korpusů

- Ruční
- Není široce přístupná

Anotace KSK

- Upravená verze morfologického slovníku (Osolsobě 1996) a morfologického analyzátoru *ajka* (Sedláček 2004).
- Ruční disambiguace.

Pražský a brněnský systém anotací

- Projekt nové národní morfologie

Specifika anotací SYN2005

- Testování guesserů

Co se skrývá za označením slovní druh X.*

- Slova, kterým nelze na základě morfologického slovníku přiřadit žádnou interpretaci.
- Méně obvyklá slova.
- Méně obvyklé tvary.
- Překlepy.

Doporučená četba pro zájemce o probíranou problematiku:

- Jelínek, T.: Nové značkování v Českém národním korpusu. *Naše řeč 91*, 2008, s. 13–20.
- Jelínek, T., Petkevič, V.: Systém jazykového značkování korpusů současné psané češtiny. In Petkevič, V. – Rosen, A. (eds.) *3. Gramatika a značkování korpusů*, Praha : Nakladatelství Lidové noviny/Ústav Českého národního korpusu, 2011, s. 154–170.

Doporučená četba pro zájemce o probíranou problematiku:

- Osolsobě, K.: Popis gramatických významů (hodnot) jednoduchých slovesných tvarů v anotacích českých (slovenských) korpusů. *SPFFBU* A 55, Brno : FF MU, 2007, s. 201–218.
- Petkevič, V.: Reliable Morphological Desambiguation of Czech: Rule-Based Approach is Necessary. In: Šimková, M. (ed.), *Insight into the Slovak and Czech Corpus Linguistics*, Bratislava : Veda, 2006, s. 26–44.

Doporučená četba pro zájemce o probíranou problematiku:

- Petkevič, V.: Využití vidu ke zkvalitnění automatického značkování češtiny. In Bičan, A. – Klaška, J. – Macurová, P. – Zmrzlíková, J. (eds.), *Karlík a továrna na lingvistiku. Prof. Petru Karlíkovi k životnímu jubileu*, Host : Brno, 2010,s. 368–387.
- OSOLSOBĚ, Klára. *Česká morfologie a korpusy*. 1. vyd. Praha: Karolinum, 2014.

Úkol na 11. 3. 2015

- pražský‘ tagset (http://ucnk.ff.cuni.cz/bonito/znacky.php)
- brněnský‘ tagset (http://nlp.fi.muni.cz/projekty/ajka/tags.pdf)
- OSOLSOBĚ, Klára. Popis gramatických významů (hodnot) jednoduchých slovesných tvarů v anotacích českých (slovenských) korpusů. *Linguistica Brunensia*, Brno: Masarykova Univerzita, 2007, A 55, č. 1, s. 201-218.

Kvíz

- ***spal* – l-ové příčestí/imperativ**
- **Pokuste se najít co nejvíce analogických homonym**
- **Pokuste se formulovat nějaké lingvisticky založené hypotézy, jak takové doklady lze hledat**