Úvod do korpusové lingvistiky 2

KORPUS V MODERNÍM SLOVA SMYSLU A BUDOVÁNÍ KORPUSŮ

Korpus

- V širším slova smyslu soubor textů
- Sbírka textů
- Korpus v moderním slova smyslu

Definice korpusu v moderním slova smyslu

- vzorky (sampling) a reprezentativnost
- konečná velikost (omezený a vymezený rozsah)
- strojově čitelná forma (MRF)
- standardní reference

Reprezentativnost korpusu

- Texty mají reprezentovat jazyk, a to buď obecně v jeho různých podobách (psané/mluvené), nebo speciálně (např. žánrově vymezené korpusy, autorské korpusy, žákovské korpusy).
- Vzorky – z textů, z nichž se skládá korpus, se vybírá vzorek (reprezentativní část textu), nebo je text zařazen do korpusu jako celek.

Velikost korpusu

- Vymezený obsah i rozsah
- Rozsah psaných a mluvených korpusů s ohledem na žánr
- Rozsah a obsah autorských korpusů
- Rozsah a obsah specializovaných korpusů

Strojově čitelná a přístupná podoba

- Konverze textů existujících ve strojově čitelné podobě do jednotného formátu
- Převedení textů, které neexistují ve strojově čitelné podobě
- OCR metody
- Ruční přepis
- Budování pravidel pro ruční přepis jako metodologie

Standardní reference

- Vnětextové značkování
- Vnitrotextové značkování
- Tokenizace
- Tagging
- Tree bank
- Sémantické anotace
- Fonetický přepis

Budování korpusu

- Určit typ a účel
- Sběr dat
- Zajištění právní ochrany poskytnutých dat
- Zajištění automatických nástrojů pro budování korpusu
- Zajištění kvalifikovaných anotátorů
- Zajištění nástrojů pro přístup ke korpusům

Hlavní zásady anotační praxe

- Anotační schéma by mělo vycházet z teoretických východisek, která by měla být jasně formulovaná a přístupná každému konečnému uživateli korpusu. Mnohé korpusy byly anotovány ručně (existence subjektivních interpretací zaviněných osobou anotátora ve sporných případech). Značkování by pak mělo být doplněno komentáři, z nichž by byl důvod příslušné volby patrný.

Co má uživatel korpusu vědět o anotaci, chce-li ji použít

- Mělo by být jasné JAK a KDO anotaci provedl (JAK – ručně x automaticky x poloautomaticky, s postkorekcí x bez korekce) (KDO – počítačový program, anotátor - člověk)
- Uživatel korpusu by si měl být vědom toho, že anotace nejsou nějakou nedotknutelnou neomylnou instancí. Anotace je pouze více či méně užitečným nástrojem. INTERPRETACE.
- Anotační schéma by mělo být založeno na široce schvalovaných a teoreticky nezatížených principech. Není na škodu i zjednodušující přístup.
- Žádné anotační schéma nemá právo být pokládáno za standardní. Je-li nějaké řešení uznávanější, děje se tak pouze z praktických důvodů.