Úvod do korpusové lingvistiky

11

KORPUSY A KVANTITATIVNÍ DATA

Frekvenční seznam – slovních tvarů – slov – lemmat – pos – tag

- Rozdíl mezi kvantitativní a kvalitativní analýzou korpusu spočívá v tom, že kvantitativní data přirozeně čerpatelná z korpusových textů nejsou součástí lingvistických rysů, které se datům přiřazují.
- Jsou pouze bází pro analýzu, která musí pokračovat dále. V kvalitativně zaměřené analýze jsou málo frekventované jevy zkoumány se stejnou pozorností jako jevy silně frekventované.
- Cílem analýzy korpusu není konstatování obvyklých a řídkých jevů v jazyce, nýbrž detailní popis jazyka jako celku.

příklad : adjektivum *rudý*

- kvantitativní analýza odhalí pouze počet výskytů
- kvalitativní analýza se zaměří na významy užití (nejen barva, ale i např. politická příslušnost atd.)

Reprezentativnost korpusu

- Ve velmi malých korpusech je možné, že se okrajové jevy vůbec nevyskytnou a frekventované jevy nebudou zastoupeny dostatečně.
- Velký korpus zaručuje možnost dobře zkoumat frekventované jevy.
- Ncméně platí, že výskyt hapaxových jevů je stabilní (zvětšujeme-li rozsah textů, neklesá podíl – kvantitativní i kvalitativní – tzv. hapax legomena, ale i dalších hapaxových jevů).

Vzorky

- Výběr vzorku – neexistuje obecně platná metoda, jak určit reprezentativnost vzorku.
- Podobné metody výběru vzorku zaručí, že data z korpusů zpracovaných co do výběru vzorků stejnými metodami budou srovnatelná navzájem.

Zpracování dat kvantitativními metodami

- V korpusové lingvistice jde kvantitativní analýza ruku v ruce s analýzou kvalitativní.
- Běžně užívané techniky matematické statistiky, které v rámci KL následují za prostým počítáním frekvenčních výskytů jazykových jevů obsažených v korpusu.
- Díky těmto metodám se lingvisté snaží získat z korpusů nejen prostá kvantitativní data, ale dojít k interpretaci jejich závažnosti, a to pomocí exaktních matematicky ověřených postupů.

metody matematické statistiky užívané v KL

- Jsou to např. metody, při jejichž užití je možné brát zřetel na takové okolnosti, jako je typ okolí jednotky (kolokace), vzorku (žánr) atd.
- Přehled je pouze omezený (nejsem matematický statistik a úvod do mat. stat. není cílem naší přednášky).
- (více: Statistics for Corpus Linguistics v řadě edinburských učebnic empirické lingvistiky, internet).

Frekvenční analýza

- matematické sečítání počtu jednotek (tokens)
- v případě klasifikovaných jednotek typů (lemmat, tagů, pos, …)
- u anotovaných korpusů obecně můžeme počítat a) se snazší prací a b) s lepšími výsledky
- u anotovaných korpusů je třeba mít na zřeteli, že počítáme pouze výsledky anotací, nikoliv to, co skutečně v korpusu je

Proporcionalita

- Prosté počítání frekvencí je jen jako první krok další analýzy.
- Hlavní nevýhodou prostých frekvenčních výpočtů je, že výsledky, které jimi získáme, se mohou značně lišit v případě, kdy jeden a týž jev spočítáme v různých korpusech (např. psaném a mluveném).
- Jak získané výsledky porovnat?
- Výsledky ze dvou korpusů, které nejsou stejně velké: vypočítáme frekvenci jako procento z celkového počtu tokens v korpusu. Výsledek srovnání percentuálního zastoupení nám může říci něco spolehlivého.

Porovnání výskytu tokenů v korpusech různého rozsahu

- Např. zjistíme, že jev A se v korpusu psaného jazyka o 1 mil. slovních tvarů vyskytuje 500 krát a v korpusu mluveného jazyka o 100 000 slovních tvarů 50 krát.

Vypočítáme percentuální výskyt, a to takto:

- mluvený korpus (50: 100 000) x 100 =0,05%
- psaný korpus (500: 1000 000) x 100=0,05%
- V obou případech nám vyjde stejný výsledek.
- Vypočítali jsme, že s ohledem na různost proporcí vzorků je frekvence stejná.
- Vždy se vychází z poměru mezi velikostí vzorku a počtem výskytů.
- ratio=počet výskytů typů /počet výskytů tokens v celém vzorku

Testování významnosti výsledků frekvenčních analýz

- chi-square test
- MI-score
- T-score
- z- score

Kolokace

- Metody se používají pro vyhledávání statisticky významných kolokací.
- Kolokace (souvýskyt slov) jsou z lingvistického hlediska zajímavé.
- Gramatika
- Lexikon – idiomatika
- MWE