CJBB105 – 6

Mgr. Dana Hlaváčková, Ph.D.

CJBB105

Korpusové manažery

- zpracování textů do korpusové podoby
- prohlížení korpusových dat a práce s nimi
- budování korpusů
- navazující aplikace
- webová rozhraní, desktopové aplikace, webová stránka
- často omezený přístup, nutná registrace
- KonText – ÚČNK, Praha
- Sketch Engine – MU (CZPJ FI MU + Lexical Computing, Ltd.), Brno

Možnosti zobrazení

- vybraný korpus, počet nalezených výskytů
  - počet výskytů na milion pozic, ARF (průměrná redukovaná frekvence)
- zobrazení ve formě konkordance (KWIC) nebo věta
- atributy – word, lemma, tag, lc, část tagu
- strukturní značky – hranice vět, dokumentů ad.
- reference – metainformace o textech
- šířka kontextu, počet konkordancí na stránku
- popis dotazu (konkordance)

Možnosti hledání

- konkrétní tvar slova (*slovo, slovní tvar*)
- lemma – nalezeny všechny tvary vyskytující se v korpusu
- fráze – spojení dvou a více slov
  - specifikace kontextu
- tag
  - konstrukce značky (KT)
- znak (SKE)
- podřetězec (KT)
- CQL (Corpus Query Language)
- regulární výrazy

Třídění výsledků

- náhodný vzorek, promíchání výsledků
- třídění kontextu a KWIC (podle abecedy)
  - podle atributů
  - víceúrovňové a retrográdní
- filtrování konkordancí
  - pozitivní a negativní filtry
  - pouze 1. výskyt v dokumentu

Frekvenční distribuce

- frekvenční údaje – číselné i grafické znázornění
  - KWIC (lemmata, slovní tvary)
  - tagy
  - typy dokumentů
  - víceúrovňové
- vizualizace frekvenčního rozložení přes celý korpus (SKE)

Kolokace

- výpočet kandidátů na kolokace (ustálená slovní spojení)
  - frekvence spojení (dvou a více jednotek) – vysoká
  - frekvence spojení s ostatními jednotkami – nízká
    - vztaženo k velikosti korpusu
  - kolokační paradigma, monokolokabilita (*stroužek česneku, tratoliště krve*)
  - asociační míry
- MI-score
  - pravděpodobnost současného výskytu dvou slov (mutual information)
- T-score
  - zapojeno rozložení spojení přes celý korpus, nenáhodný jev
- Dice, Log-Dice
  - nepočítají s velikostí korpusu

Další funkce

- vytvoření subkorpusu
  - podle metainformací o textech (KT)
  - z aktuálních konkordancí (SKE)
- seznam slov
  - podle frekvence
  - uživatel definuje kritéria

KonText – externí funkce

- SyD
  - korpusový průzkum variant slov
  - synchronní i diachronní korpusy
  - psaný i mluvený jazyk
- KWords
  - generování klíčových slov
  - porovnání výskytů s referenčním korpusem
- Morfio
  - vyhledání seznamů slov (až n-tic) na základě slovotvorných charakteristik

Sketch Engine

- Tezaurus – podobná slova, míra podobnosti na základě kontextů, vizualizace
  - hra Uhádni to slovo (*https://nlp.fi.muni.cz/projekty/uhadni_to_slovo/*)
- Word Sketch – slovní profily, tagging
  - tabulky zachycují okolí zadaného lemmatu podle určitých kategorií
- Sketch Diff – porovnání slovních profilů dvou lemmat
- tvorba korpusů a subkorpusů