Vaje III

Raba korpusov
WordSmith 4.0

Ponovitev

- Pojavnice/različnice
- Frekvenčni seznami
- Ključnost
- Kolokacije
- Načini iskanja
  - Enostavno iskanje
  - Regularni izrazi
  - Iskanje po oznakah

WordSmith

- **Wordlist**
- **1.** Izdelajte besedni seznam za slovenski korpus. Na zavihku **Statistics** si oglejte podatke o korpusu in jih interpretirajte. Koliko ima vsako besedilo pojavnic/različnic[1] (tokens/types)? Kako dolžina besedila vpliva na razmerje med številom pojavnic in različnic (TTR)?
- 2. Odprite zavihek **Frequency**. Oglejte si prvih 30 besed v pogostostnem seznamu. Katere besedne vrste se najpogosteje pojavljajo? Na katerem mestu se pojavi prva polnopomenska beseda?
- 3. Odprite zavihek **Alphabetical**. Kako različne oblike besed v slovenščini (npr. skloni) vplivajo na pogostost besednih oblik? Poiščite samostalnik *alkohol* in z miško povlecite vse njegove oblike na osnovno imenovalniško obliko. Nato ponovno uredite besedni seznam (**Edit – Resort**). V pogostostnem seznamu si oglejte, kako se sprememba odraža na vrstnem redu najpogostejših besed.
- 4. V zavihku **Alphabetical** raziščite druge načine urejanja besednega seznama (po dolžini besed, odzadnje itd.).
- 5. V nastavitvah aktivirajte seznam praznih besed (**Settings – Stop-, Lemma& matchlists**). Izdelajte nov pogostostni besedni seznam. Ocenite uporabnost takega seznama za terminografske namene.
- 6. Izdelajte indeks vašega korpusa, tako da ponovno izdelate besedni seznam, vendar zdaj namesto **Make a wordlist now** izberete možnost **Add to index**. Nato indeks odprete (**File – Open...**). Zdaj lahko s pomočjo funkcije **Clusters** izdelate seznam dvo- ali večbesednih enot. Zaenkrat izdelajte dvobesednega in med prvimi stotimi vnosi poiščite nekaj terminoloških kandidatov ter terminoloških kolokacij.

Vaja z Wordsmithom

- **Concord**
- 1. Iz pogostostnega seznama izberite besedo, ki se vam zdi terminološko zanimiva. Izdelajte konkordanco za to besedo (**Compute – Concordance**). Uredite jo po levem okolju (**Resort** - prva leva, druga leva) in izluščite večbesedne terminološke kandidate. Nato konkordanco preuredite po desnem okolju in znova izluščite večbesedne terminološke kandidate.
- 2. Preizkusite še delovanje funkcij **Plot**, **Clusters** in **Collocates**. Kaj vam povedo o frazeološkem obnašanju izbrane besede?
- **Keywords**
- 1. S pomočjo programa Wordlist izdelajte besedni seznam za vaš korpus in za primerljivi, referenčni korpus. S programom Keywords primerjajte oba seznama in izluščite ključne besede.
- [1] token (pojavnica) – Osnovni korpusni element, npr. beseda, ločilo ali številka. Velikost korpusa tipično izražamo s številom pojavnic; če rečemo, da ima korpus 100 milijonov besed, s tem v resnici mislimo na pojavnice.
- type (različnica) – Če vsako pojavnico štejemo le enkrat, dobimo seznam različnic. Če bi v stavku "Moj pes ne mara mačk, niti ne mara drugih psov." prešteli pojavnice, bi jih našteli 12, različnic pa 10.

Uporabno:

Več o uporabi Wordmitha:

- Spela Vintar (2008). Terminologija. Terinološka vedain računalniško podprta terminografija

Samodejno luščenje terminologije:

- http://lojze.lugos.si/cgitest/extract.cgi

Dodatne vaje iz uporabe korpusa Fidaplus:

- Razložite, kaj bi vam v korpusu našel iskalni pogoj "#1hiter&~#2r*".
- S pomočjo besedilnih vzorcev "znan kot" in "ali tudi" poiščite po pet primerov sopomen.
- A je "tajkun" beseda, ki je postala popularna čez noč? V kakšnih besedilih se največ uporablja?
- Poiščite pet glavnih stavkov, ki vsebujejo tri ali več vprašalnic (vprašalnih zaimkov).