Od Novega mesta do Njujorka

(vaje)

I. Kako zapisujemo tujejezične elemente?

1. Se pogosteje uporablja citatni zapis ali poslovenjene oblike?
    a. Odpremo JANES v. 0.4.: http://nl.ijs.si/noske/janes04.cgi/first_form?corpname=janes.04 (UI: **janes**, geslo: **9neztandart6**)
    b. V iskalno okence za enostavno iskanje vpišemo naslednje besede v citatnih oblikah in poiščemo njihovo pogostnost.
        - like

          _____
        - good

          _____
        - please

          _____

2. Poleg citatnih oblik se, kot smo videli, uporabljajo tudi različne poslovenjene oblike.
    a. Katere so po vašem mnenju poslovenjene oblike zgornjih besed?
    b. Napišite te oblike v okence za enostavno iskanje in poiščite njihovo frekvenco. Prepišite najdene vrednosti za citatne oblike (vse skupaj) in poslovenjene oblike (vse skupaj).
        - Oblike za *please* in njihova frekvenca:

          _____

          _____

          _____

          _____

        - Oblike za *good* in njihova frekvenca:

          _____

_____

_____

_____

- Oblike za *like* in njihova frekvenca:

_____

_____

_____

_____

3. Katere oblike torej prevladujejo pri izbranih besedah, citatne ali poslovenjene?

_____

II. Koliko je tujejezičnih elementov v JANES-u in iz katerih jezikov so?

1. Najprej si poglejmo, koliko je tujejezičnih elementov v korpusu.
   a. V korpusu JANES so besede, ki jih je avtomatski označevalnik prepoznal kot tuje, vendar program ne ločuje med različnimi jeziki. Besede so preprosto označene s posebno, vedno enako kodo. Ko iščemo s CQL, lahko uporabimo to kodo in vse take besede izluščimo naenkrat.
      - V polje CQL vpišite iskalni ukaz [tag="Nj"] (Nj je oznaka za tujejezične elemente)
      - Koliko je vseh takih konkordanc ? _____
   b. Za primerjavo odprimo tokrat govorni korpus GOS in preverimo, koliko je tujejezičnih elementov v tem korpusu.
      - Vtipkajmo enak iskalni ukaz v polje CQL. Rezultat je _____.
      - Zakaj prihaja do take razlike?

   c. Ali se prisotnost tujejezičnih elementov spreminja glede na besedilni tip? Poiščite skupno število tujejezičnih elementov v vsakem izmed štirih podkorpusov JANES-a.
      - Na vrhu izberite podkorpus, nato kot vrsto iskanja izberite CQL in v ukazno polje vpišite [tag="Nj"].

- Rezultati:

  Blogi (JANES v0.4 Blog): _____

  Komentarji (JANES v0.4 News): _____

  Forumi (JANES v0.4 Forum): _____

  Tviti (JANES v0.4 Tweet): _____

2. Videli smo, da se poleg angleščine, ki nedvomno prevladuje, v spletni slovenščini pojavljajo tudi besede iz drugih jezikov. Kako bi ugotovili, kateri jeziki se poleg angleščine še pojavljajo?
   a. Izberite si enega izmed podkorpusov, poiščite vse tujejezične elemente v tem podkorpusu. Napravite frekvenčni seznam besed. Nato na prvih desetih straneh frekvenčnega seznama »ročno« preverite, kateri jeziki se pojavljajo (poleg angleščine). Zapišite te jezike (in primere rabe, ki ste jih našli) spodaj:

      _____

      _____

      _____

3. Kako točni so naši rezultati?
   a. Ker avtomatsko označevanje ni 100 %, se pojavljajo napake: med besedami, označenimi kot neslovenskimi, je tudi nekaj takih, ki to zagotovo niso. Poiščite jih 5.

      _____

      _____

      _____

      _____

      _____

   b. Zakaj prihaja do takih napak?

III. Kdaj in kako se uporablja tujejezične elemente?

1. Kako se pojavljajo tujejezični elementi? Kakšne vsebine bodo izražene v tujih jezikih?
   a. Razmislite, na kakšen način bi se lahko pojavljali tujejezični elementi v korpusu JANES (naštejte, kar se spomnite s predavanja).

      _____

_____

_____

b. Ustvarite seznam konkordanc vseh tujih besed v korpusu JANES.
   - Izberite korpus JANES v0.4.
   - Pri vrstah iskanj izberite CQL in v iskalno polje vpišite [tag="Nj"].

c. Preglejte dobljene konkordance in ocenite, ali se primeri, ki ste jih našli, ujemajo z vašimi predvidevanji (in spominom). Navedite vsaj 5 različnih načinov in po en primer za vsak način.

_____

_____

_____

_____

_____

d. Kateri način vam je najbližji oz. kaj sami uporabljate v podobnih okoliščinah (socialna omrežja, forumi, blogi ipd.)?
   Napišite tri besede/besedne zveze, ki jih najpogosteje uporabljate, in primerjajte njihovo pogostnost v vseh štirih podkorpusih.
   - Besedo/besedno zvezo preprosto vpišite v polje za enostavno iskanje, pri čemer najprej izberite vsak podkorpus posebej. Prepišite, kolikokrat na milijon se pojavlja v posameznem podkorpusu.

Besed/Bes. zveza	Tweet	Forum	Blog	News

_____

_____

_____

e. Kdaj se po vašem mnenju avtorji odločajo za **preklapljanje** – preskok iz slovenščine v drugi jezik in uporabo tujejezičnih elementov (če odmislimo naslove, lastna imena ipd.)?
   - Poiščite vse tujejezične besede v podkorpusu Tweet.
   - Napišite, kateri so razlogi za preklapljanje v opazovanih primerih:

     _____

     _____

     _____

_____

_____

- Ali avtorji izrazito preklapljajo na kakem posebnem mestu v svojih tvitih?

_____

_____

IV. Kdo uporablja tujejezične elemente?

1. Korpus JANES je označen tudi s podatki o tem, kdo so avtorji besedil, ki jih vključuje. Zlasti podkorpus TWEET ponuja veliko informacij v tem smislu. Poglejmo si, pri katerih uporabnikih se tujejezični elementi pojavljajo pogosteje.
    a. V podkorpusu TWEET so zasebni uporabniki označeni kot 'private', podjetja, organizacije in drugi javni subjekti pa s 'corporate'. Kateri tip uporabnikov uporablja več tujejezičnih elementov?
        - Vključimo lastnosti besedil pri iskanju.
        - Izberemo možnost 'corporate' pri izbiri TEXT.SOURCE.
        - Z iskanjem CQL poiščemo vse tujejezične elemente v izbranih tvitih.
          Rezultat za javna besedila: _____
        - Nato gremo nazaj na iskanje in pri TEXT.SOURCE izberemo možnost 'private'.
        - Z iskanjem CQL poiščemo vse tujejezične elemente v teh tvitih.
          Rezultat za zasebna besedila: _____

2. Zabeležen je tudi spol uporabnikov.
    a. Ali tujejezične elemente več uporabljajo ženske ali moški?
        - Pri lastnostih besedil imamo možnost izbire spola avtorja.
          Rezultat za ženske: _____
          Rezultat za moške: _____

3. Kdo (kateri posameznik) je v svojih tvitih uporabil absolutno največ tujejezičnih elementov?
        - Izdelamo konkordance vseh tujejezičnih elementov v podkorpusu TWEET.

- Izberemo Frekvence in v spodnjem kvadratu (Frekvenčna razporeditev po lastnostnih besedil) Izberemo prvo možnost (text.author).
- Avtor/-ica z največjim absolutnim številom tujih besed v svojih tvitih je: _____

b. Kaj pa če upoštevamo relativno frekvenco, kdo najpogosteje uporablja tujejezične elemente?

- Na zgoraj izdelanem frekvenčnem seznamu kliknemo na napis »Rel. frekvenca (%)«
- Največji odstotek tujejezičnih elementov je v svojih tvitih uporabil/a: _____.

c. Katere tujejezične elemente pa ta oseba uporablja?

- Kliknemo na »p« na začetku vrstice ob imenu avtorja, da odpremo konkordance vseh tvitov s tujejezičnimi elementi tega avtorja.
- Izdelamo frekvenčni seznam tujejezičnih elementov za te konkordance.
- Izberite še kakega drugega avtorja in preverite, ali so besede pri vrhu podobne.

V. Stopnja (ne)standardnosti in tujejezični elementi

1. Podkorpus Tweet je označen tudi glede na stopnjo nestandardnosti besedila (oznake L1, L2 in L3). Oglejmo si, ali višja stopnja nestandardnosti pomeni tudi pogostejšo uporabo tujejezičnih elementov. Kakšna so vaša pričakovanja?

a. Največ tujejezičnih elementov bo verjetno v besedilih z oznako _____.

b. Preverimo, koliko je tujejezičnih elementov v besedilih z najnižjo stopnjo nestandardnosti (torej v besedilih, ki najbolj sledijo pravilom o standardni rabi slovenščine).

- Izberemo podkorpus TWEET.
- Pri Lastnostih besedil izberemo L1.
- V iskalno polje CQL napišemo iskalni ukaz za tujejezične elemente: [tag="Nj"].

- Število tujejezičnih elementov v besedilih, označenih z L1, je _____.

c. Enako naredimo za besedila, označena z L2 in L3.
   - Število tujejezičnih elementov v besedilih, označenih z L2, je _____.
   - Število tujejezičnih elementov v besedilih, označenih z L3, je _____.

d. Razmislimo o razlogih za take rezultate. Pobrskajte po konkordancah za posamezne oznake nestandardnosti in preverite, v kakšnem sobesedilu se kje pojavljajo tujejezični elementi.

2. Preverimo lahko tudi, kdo uporablja najbolj nestandarden jezik *in* največ tujejezičnih elementov.
   a. Preverimo, kdo uporablja več tujejezičnih elementov v najbolj nestandardnih tvitih, moški ali ženske.
      - Izberemo podkorpus TWEET.
      - Pri Lastnostih besedil izberemo L3.
      - Pri Lastnostih besedil izberemo spol avtorjev (enkrat ženski, enkrat moški).
      - V iskalno polje CQL napišemo iskalni ukaz za tujejezične elemente: [tag="Nj"].
      - Skupno število tujejezičnih elementov pri ženskah v najbolj nestandardnih tvitih je: _____

        Skupno število tujejezičnih elementov pri moških v najbolj nestandardnih tvitih je: _____

   b. Poiščite tri avtorje moškega in tri ženskega spola, ki uporabljajo najbolj nestandarden jezik in hkrati največ tujejezičnih elementov.
      - Ko po zgoraj navedenem postopku dobimo konkordance, zgradimo frekvenčni seznam, pri čemer v oknu *Frekvenčna razporeditev po lastnostih besedil* izberemo »text.author«.
      - V najbolj nestandardnih tvitih največ tujejezičnih elementov uporabljajo:

        moški: _____

        _____

        _____

        ženske: _____

_____

_____