Miten tutkijat kuvailevat aineistojaan: historiallinen korpuslingvistiikka

Samuli Kaislaniemi, Akatemian tutkijatohtori
(Englannin filologia)
samuli.kaislaniemi@uef.fi
Twitter: @samklai

Historiallinen korpuslingvistiikka

- Historiallinen kielitiede tutkii kielen muutosta
- Kielen historialliset variantit: yleensä kirjoitettua kieltä

- Korpuslingvistiikka = tietokoneavusteista kielentutkimusta tekstikorpusten l. digitoitujen tekstikokoelmien avulla
- Juuret 1960-luvulla, historiallisten korpusten osalta 1980-luvulla
- *Helsinki Corpus of English Texts* (1991)
  - https://varieng.helsinki.fi/CoRD/corpora/HelsinkiCorpus/

Historiallinen korpuslingvistiikka: aineistot

- korpus = tekstikokoelma tietokoneluettavassa muodossa
  - järjestelmällisesti kerätty
  - luonnollista kieltä, kirjoitettua tai puhetta (litteroituna)
- tekstiä on usein käsitelty jotenkin tutkimusta helpottamaan
  - rakenteita ja tekstin jäsentelyä muutettu (esim. kappalejako, otsikot)
  - kirjoitusasua modernisoitu; käytetään nykyaakkostoa (Ei tihrusteta fraktuuraa)
  - kielen piirteitä voi olla annotoitu (esim. sanaluokka)

Historialliset korpukset: metadataa 1

Yleisesti:

- aineiston nimi

- kieli
- tekstilaji
- tekstien lukumäärä
- sanamäärä
- aikakausi

- annotaatio
- formaatti/formaatit
- lisätiedostot
- saatavuus
- lisenssi

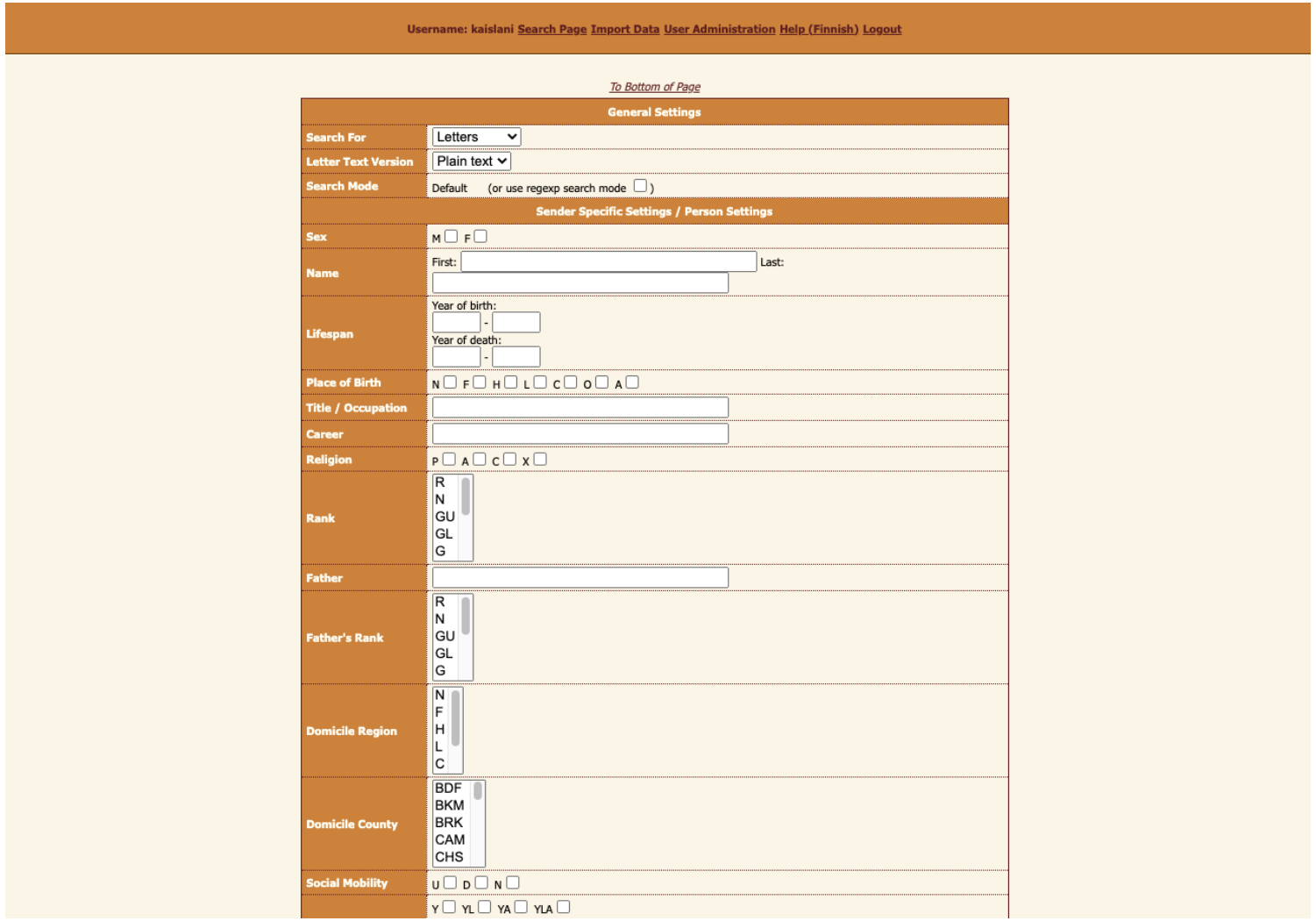
Historialliset korpukset: metadataa 2

- korpuksen kokoaja(t)
- kokoamispaikka
- kokoamisaika/julkaisuvuosi
- kokoamisprojektin rahoittaja(t)
- projektin nettisivut
- korpuksen versiot
- korpuksen manuaali
- korpusohjelmisto(t)
- miten viitataan
- jne

Esimerkki: *Corpora of Early English Correspondence* (CEEC)

- englanninkielisiä kirjeitä
- 1400–1800
- n.12 000 kirjettä, 5+ miljoonaa sanaa
- n.1 200 kirjoittajaa
- kerätty painetuista editioista
- sosiolingvistiikkaa varten: paljon metadataa
  - taustatiedot kirjoittajista koottu tietokantaan (sukupuoli, sääty, koulutus, ym jne)

Korpuksen metadatahakukone: CEECer



- (Ei julkinen)
- Metadataa 60 kenttää, mm:
  - nimi
  - sukupuoli
  - sääty
  - isän sääty
  - elinalue
  - suhde vastaanottajaan
  - sanamäärä
  - ikä kirjoittaessa

Esimerkki: CEEC ja korpusta kuvaileva metadata

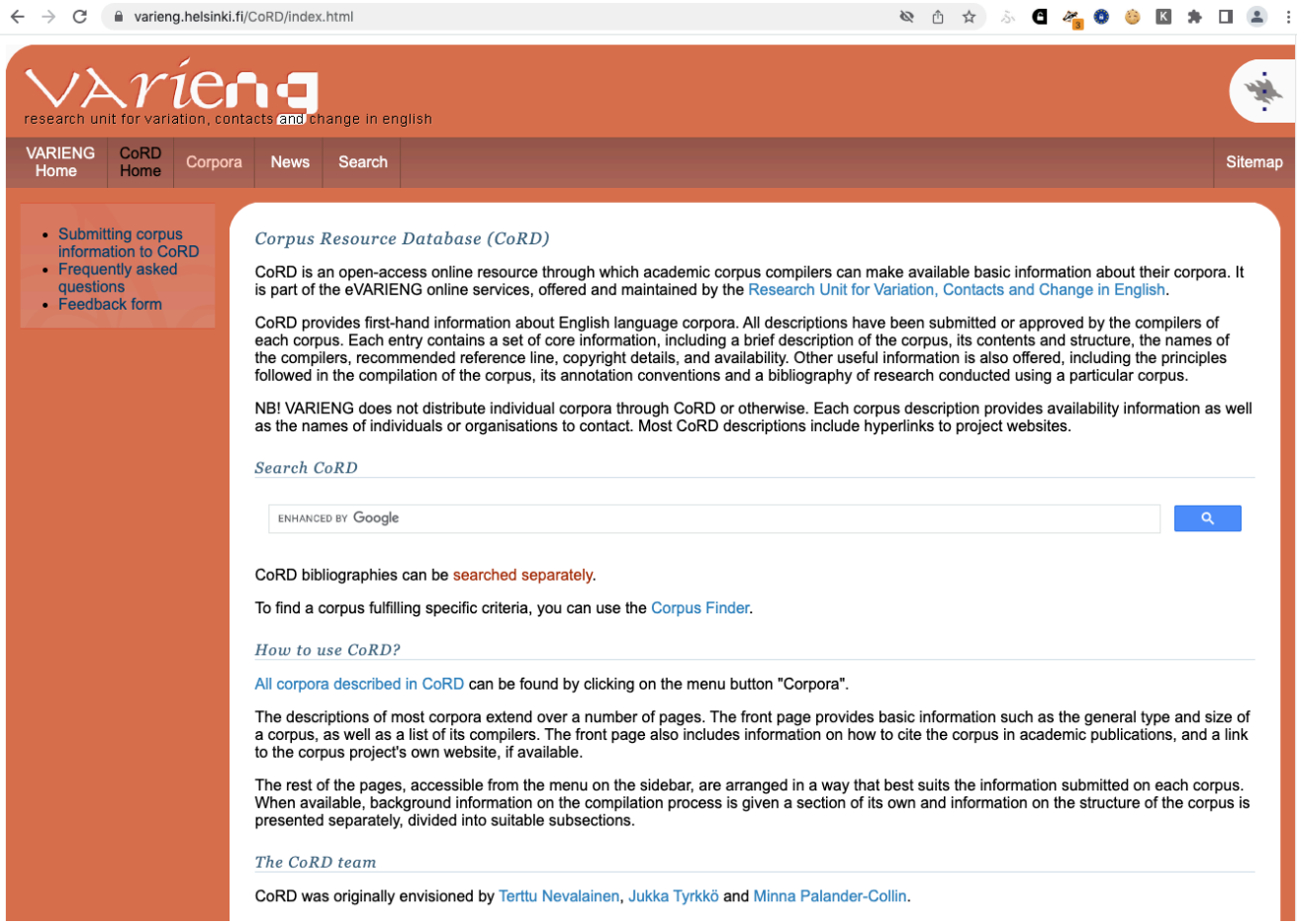
- *metadatan tuottaminen projektin aikana (tieteenalan sanastot, standardit, ontologiat)*
- *onko työkaluja/alustoja metadatan kirjaamiseen*

Historialliset korpukset – metadatan tallentaminen & julkaiseminen

- korpuskuvailuja löytyy:
  - julkaistun korpuksen yhteydestä (ETED, Lampeter; OTA)
  - julkaisuista (artikkelit ja kirjat)
  - manuaaleista (CEECS, PCEEC, CEECES)
  - metadata-alustoista (Etsin, CoRD)

Corpus Resource Database (CoRD)

- https://varieng.helsinki.fi/CoRD/index.html
- Tarve metadatatietokannalle – tehtiin itse!
- Ts. olemassaolevia kanavia ja työkaluja ei tunneta, eikä niitä juuri käytetä



vaikeuksia ja ongelmia

- korpus ei ole ‘julkaisu’
- ei laajasti tutkijatasolla tunnettuja standardeja tai periaatteita, sanastoa tai työkaluja
- dokumentoinnista ei saa pisteitä (priorisointi, aikataulujen venyminen)
- vrt DMP

Omia kokemuksia

- Zenodo
- DMP