



Miten tutkijat kuvailevat aineistojaan: historiallinen korpuslingvistiikka

Samuli Kaislaniemi, Akatemian tutkijatohtori

(Englannin filologia)

samuli.kaislaniemi@uef.fi

Twitter: [@samklai](https://twitter.com/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 tietokone luettavassa 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 (*Œ tihruŒta fracturaa*)
 - kielen piirteitä voi olla annotoitu (esim. sanaluokka)



Historialliset korpuukset: metadataa 1

Yleisesti:

- kieli
- tekstilaji
- tekstien lukumäärä
- sanamäärä
- aikakausi
- aineiston nimi
- annotaatio
- formaatti/formaatit
- lisätiedostot
- saatavuus
- lisenssi



Historialliset korpuksset: 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
- sociolinguistiikkaa 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

Username: kaislani [Search Page](#) [Import Data](#) [User Administration](#) [Help \(Finnish\)](#) [Logout](#)

[To Bottom of Page](#)

General Settings	
Search For	Letters
Letter Text Version	Plain text
Search Mode	Default (or use regexp search mode <input type="checkbox"/>)
Sender Specific Settings / Person Settings	
Sex	M <input type="checkbox"/> F <input type="checkbox"/>
Name	First: <input type="text"/> Last: <input type="text"/>
Lifespan	Year of birth: <input type="text"/> - <input type="text"/> Year of death: <input type="text"/> - <input type="text"/>
Place of Birth	N <input type="checkbox"/> F <input type="checkbox"/> H <input type="checkbox"/> L <input type="checkbox"/> c <input type="checkbox"/> o <input type="checkbox"/> A <input type="checkbox"/>
Title / Occupation	<input type="text"/>
Career	<input type="text"/>
Religion	P <input type="checkbox"/> A <input type="checkbox"/> c <input type="checkbox"/> x <input type="checkbox"/>
Rank	R N GU GL G
Father	<input type="text"/>
Father's Rank	R N GU GL G
Domicile Region	N F H L C
Domicile County	BDF BKM BRK CAM CHS
Social Mobility	U <input type="checkbox"/> D <input type="checkbox"/> N <input type="checkbox"/> Y <input type="checkbox"/> YL <input type="checkbox"/> YA <input type="checkbox"/> YLA <input type="checkbox"/>



Esimerkki: CEEC ja korpusta kuvaileva metadata

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



Historialliset korpuukset – metadatan tallentaminen & julkaiseminen

- korpuskuvailuja löytyy:
 - julkaistun korpuksen yhteydestä ([ETED](#), Lampeter; [OTA](#))
 - julkaisuista (artikkelit ja kirjat)
 - manuaaleista ([CEECS](#), PCEEC, [CEECS](#))
 - 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

varieng.helsinki.fi/CoRD/index.html

VARIENG
research unit for variation, contacts and change in english

VARIENG Home CoRD Home Corpora News Search Sitemap

- Submitting corpus information to CoRD
- Frequently asked questions
- Feedback form

Corpus Resource Database (CoRD)

CoRD is an open-access online resource through which academic corpus compilers can make available basic information about their corpora. It is part of the eVARIENG online services, offered and maintained by the [Research Unit for Variation, Contacts and Change in English](#).

CoRD provides first-hand information about English language corpora. All descriptions have been submitted or approved by the compilers of each corpus. Each entry contains a set of core information, including a brief description of the corpus, its contents and structure, the names of the compilers, recommended reference line, copyright details, and availability. Other useful information is also offered, including the principles followed in the compilation of the corpus, its annotation conventions and a bibliography of research conducted using a particular corpus.

NB! VARIENG does not distribute individual corpora through CoRD or otherwise. Each corpus description provides availability information as well as the names of individuals or organisations to contact. Most CoRD descriptions include hyperlinks to project websites.

Search CoRD

ENHANCED BY Google

CoRD bibliographies can be [searched separately](#).

To find a corpus fulfilling specific criteria, you can use the [Corpus Finder](#).

How to use CoRD?

All corpora described in CoRD can be found by clicking on the menu button "Corpora".

The descriptions of most corpora extend over a number of pages. The front page provides basic information such as the general type and size of a corpus, as well as a list of its compilers. The front page also includes information on how to cite the corpus in academic publications, and a link to the corpus project's own website, if available.

The rest of the pages, accessible from the menu on the sidebar, are arranged in a way that best suits the information submitted on each corpus. When available, background information on the compilation process is given a section of its own and information on the structure of the corpus is presented separately, divided into suitable subsections.

The CoRD team

CoRD was originally envisioned by [Terttu Nevalainen](#), [Jukka Tyrkkö](#) and [Minna Palander-Collin](#).



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