FIN-CLARIAH 2024/25

W.P 1.1. TEXT PROCESSING AND ANNOTATION ENVIRONMENTS

Author:
Jussi Piitulainen
University of Helsinki

DELIVERABLES

- Q3 (Sep 2024) D:1.1.1 Named entity annotation
- Q8 (Dec 2025) D:1.1.2 Ingesting new unstructured resources

D1.1.1: MORE ANNOTATIONS

- Explicit annotations facilitate search, filtering, counting.
- Add **name annotations** (with FiNER) to Suomi24 2001-202 (data from City Digital).
- Add **language annotations** (with HeLI-OTS) to Suomi24 2001-2020 sentences, with summary counts in paragraphs and texts.
- Also considering counts and frequency bands that can be easily computed from the data but would facilitate use if available as annotations.

```
<paragraph sum_lang="|fin:3|" type="body">
<sentence lang="fin" polarity="neut">
"         "         |
<ne name="Brad J. Gemmell" type="EnamexPrsHum">
Brad      Brad      |EnamexPrsHum-B-0|
J.        J.        |
Gemmell   Gemmell   |EnamexPrsHum-E-0|
<ne>
on        olla      |
<ne name="Teksasin yliopiston" type="EnamexOrgEdu">
Teksasin          Teksas    |EnamexOrgEdu-B-0|EnamexLocPpl-F-1|
yliopiston        yliopisto         |EnamexOrgEdu-E-0|
</ne>
merentutkimuslaitoksen  merentutkimuslaitos       |
tutkija           tutkija |
,         ,         |
ja        ja        |
```

Illustrating WP 1.1 with some new annotations (FiNER, HeLI-OTS) to Suomi24 2018.

D1.1.2: MORE CORPORA

Nordic Tweet Stream 2013-2023 (nts-src)

- http://urn.fi/urn:nbn:fi:lb-2024032221
- 10 million geolocated tweets collected through Twitter's Academic API (Mikko Laitinen, UEF)

Finnish presidential elections 2024 in social media (somepressa24-src)

- http://urn.fi/urn:nbn:fi:lb-2024030501
- social media data about presidential elections (Salla-Maaria Laaksonen ja Essi Pöyry, UHEL)
- UHEL Legal Services are investigating how these can be made available

Extended National Library historical newspaper and periodical corpus

The Finnish part klk-fi-v2 soon (already in Korp), the Swedish part klk-sv-v2 to follow; these may also be further extended by Q8.

Suomi24 to be extended with 2021-2023 data as soon as possible (we have the data).

And whatever else researchers make available.

OUTCOME / ANNOTATIONS

- The new annotations to Suomi24 will be available both in Korp (PUB) and in the downloadable versions (ACA).
- Other data and annotations will be similarly available (as PUB as we can).
- https://www.kielipankki.fi/corpora/
- https://www.kielipankki.fi/korp/
- https://www.kielipankki.fi/download/

COLLABORATORS

- Kielipankki corpus team (Tommi, Mietta, Wilhelmina, Jyrki, Jussi, Erik, Ute, Jack, Krister)
- Språkbanken (Korp software, Swedish annotations)
- CSC (storage, distribution, computational resources)

UNIVERSITY OF HELSINKI
DARIAH-FI
KIELIPANKKI The Language Bank of Finland
FIN-CLARIN