FIN-CLARIAH 2024/25

Authors:
Raine Koskimaa
Mikko Laitinen
Masoud Fatemi
Steven Coats
Ilkka Lähteenmäki
Maria Valaste
Adeline Clarke

WP4.1 ANALYTICAL SUPPORT FOR COMPUTATIONAL SSH

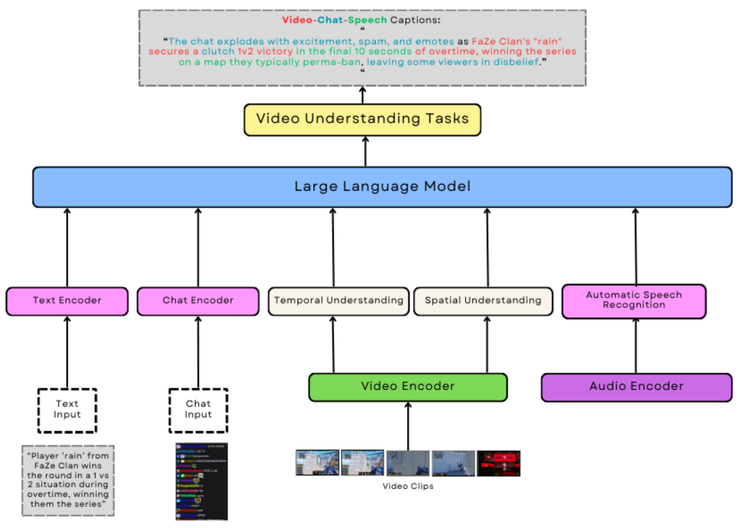
The objective of this WP is to **produce a set of efficient analytical support tools to enable researchers to utilise large born-digital or digitized multimodal data effectively**, including but not limited to game studies, computational sociolinguistics and dialectology, cultural heritage studies and computational social science.

D4.1.1 Analysis of video stream interactions with AI

Twitch streams offer a rich opportunity to develop new ML methods **for analyzing both the visual content and chat data of streams**, enabling a deeper understanding of the complex interactions between viewers and streamers.

Tools will **support comprehensive analysis using multimodal LLMs**. These models will be trained to understand the spatial and temporal aspects of livestream videos, enabling understanding of ongoing events, peak engagement moments, and recognition of scene transitions to provide the overall narrative and context of the stream (see image below)

Team lead: Raine Koskimaa, JYU. Date: Fall 2025

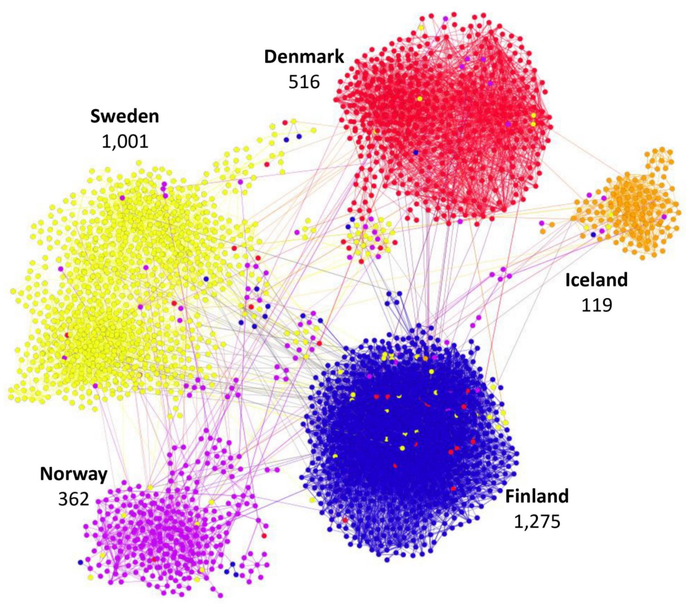


Analysis tools for

D4.1.2 Benchmark social media & D4.1.3 Interaction and language variation in social media

According to recent estimates, around 90 % of the world population use various social media applications. Yet, representative datasets from social media are scarce. **The objective of D4.1.2 is to build representative benchmark collections ("Suomi24's" of various social media)**. One such dataset is the Nordic Tweet Stream (NTS) that contains nearly 75 million messages from nearly 900k Twitter/X users from the Nordic region between January 2013-May 2023. The data are equipped with an intuitive user interface and back-end operations that make use of data enrichment tools developed by UEF group. These search tools make it easy for researchers to subset their own material and to carry out analysis of very large populations. **D4.1.3 produces tools for improved social network analysis in sociolinguistics**. Using NTS and "sister datasets" from the UK, the US and Australia. This data enrichment enables zoom into networks with differing properties, ranging from close-knit communities to loose-knit ones. We plan on continuing this data enrichment with an ML solution that aims at predicting users' age range, but more on this in 2025.

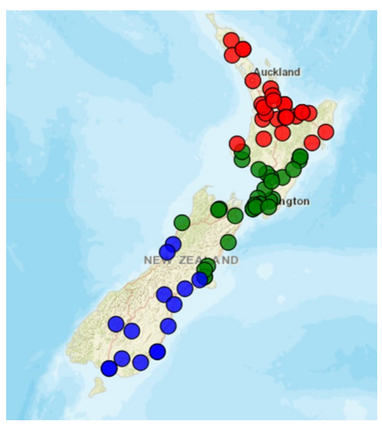
Team lead: Mikko Laitinen, UEF. Dates: D4.1.2 Fall 2024 / D4.1.3 Winter 2025



D4.1.4 Analysis of multimodal properties of naturalistic speech

The goal is to develop **easy-to-use notebook environments to analyse regional language variation, with a focus on naturalistic speech**. for the automatic transcription, alignment, segment extraction, and phonetic analysis of speech signals, allowing consideration of regional variation in vowel quality and quantity, prosody, or other features. A further focus may be the incorporation of models for the automatic analysis of the video content from which audio and transcripts have been extracted. This approach can be used, for example, to identify dialect regions based on spatial autocorrelation derived from F1 and F2 formant values.

Team lead: Steven Coats, OULU. Date: Fall/Winter 2025



This image depicts New Zealand English vowel regions, based on monophthongs from 17,854 YouTube videos uploaded to channels of New Zealand councils.

D4.1.5 Analysis of multimodal cultural heritage

This task surveys the **needs for analytical and conceptual tools of researchers utilizing visual cultural heritage data, and to identify key differences between current visual cultural heritage research practices and data intensive computational visual cultural heritage.** To achieve this objective, together with D3.2.1, we will map research interests and reach out to active visual communities such as the VIHI network (Visuaalisen historian verkosto). One option is to take part in the Helsinki Digital Humanities Hackathon, another is to organize scholarly workshops to map out needs and current bottlenecks of visual cultural heritage research.

Team lead: Ilkka Lähteenmäki, OULU. Date: Winter 2025

D4.1.6 Enrich survey data with register data and unstructured text

In Finland, it is typical that the frame population of a sample is formed by using register data or that this data is included in the actual survey. The combination of the two sources allows for a rich data set, which often contains textual data that are rarely fully utilized. This provides an opportunity to explore how it could be done while simultaneously taking into account the sampling design which allows the results obtained to be generalised.

This D4.1.6 will extend the *finnsurveytext* R package produced in 2022/23 to help social science and humanities researchers analyse and understand responses to open-ended survey questions in Finnish. In this update an additional functionality will be included to integrate with the popular *survey* package through the *svydesign* object furthering our package's useability in survey analysis and enabling analysis of open-ended questions to be better integrated with analysis of closed questions. Additionally, the new version will upgrade comparison functions and visualisations and enable inclusion of weights, clustering, and stratification within analysis.

Team: Maria Valaste & Adeline Clark, UHEL (CSDS). Date: Spring/Summer 2025

Download *finnsurveytext* package in R

UNIVERSITY OF EASTERN FINLAND
JYVÄSKYLÄN YLIOPISTO UNIVERSITY OF JYVÄSKYLÄ
HELSINGIN YLIOPISTO HELSINGFORS UNIVERSITET UNIVERSITY OF HELSINKI
UNIVERSITY OF OULU
DARIAH-FI
KIELIPANKKI The Language Bank of Finland
FIN-CLARIN