

Emitonet: en minimal tolkningsenhet for humanistisk tekstanalyse med språkmodeller

Lars Gunnarsønn Bagøien Johnsen (Nasjonalbiblioteket)

Forskere møter ofte et tilbakevendende problem med digitale metoder: hvordan forene fortolkende dybde med storskala beregning. Statistiske teknikker som kollokasjonsanalyse eller topic modeling er kraftige verktøy, men kan fremstå som «svarte bokser» for dem uten teknisk bakgrunn.

Språkmodeller (LLM) tilbyr mer intuitive tekstoperasjoner – oppsummering, parafrasering, klassifisering – men hva finnes mellom fulltekstopsummeringen og de enkelte datapunktene?

Vi introduserer emitonet som en mikroenhet for fortolkning. Emitonet oppstår når en språkmodell knytter et tekstfragment til en konseptuell tolkning, som at *Ola* → *intellektuell* fra fragmentet *Ola leser bøker*, eller *spise* → *verb* fra *De forsøkte å spise kake*. Et emiton består av (1) et tekstfragment, (2) et fokuselement (navn, entitet, nøkkelord) og (3) en tolkning.

Emitonet skiller seg fra oppsummeringer, embedding-representasjoner og klassifikasjoner: det er lite, tolkningsbart og kan telles, noe som gjør det egnet for semantisk annotasjon, konseptkartlegging og kombinasjon av kvalitative og kvantitative analyser. Vi viser også hvordan et lite, representativt utvalg av konkordanser – ofte et par tusen linjer selv ved millioner av treff – kan gi et solid grunnlag for analyse. Kombinert med LLM gir dette rask, etterprøvbar innsikt som i mange tilfeller kan supplere eller erstatte fullskala kollokasjonsanalyser.

Vi viser hvordan emitonet kan brukes til å kartlegge meningsfelt i historisk litteratur, og hvordan det kan fungere som et lett, generativt format for semantisk kartlegging i digital humaniora.

Emitons: a Minimal Unit of Interpretation for Humanistic Text Analysis with Language Models

Researchers often face a recurring challenge in digital methods: how to reconcile interpretive depth with large-scale computation. Statistical techniques such as collocation analysis or topic modeling are powerful tools, but they can appear as “black boxes” to those without technical backgrounds. Language models (LLMs) offer more intuitive text operations – summarization, paraphrasing, classification – but what exists between full-text summaries and individual data points?

We introduce the emiton as a micro-unit of interpretation. An emiton arises when a language model links a text fragment to a conceptual interpretation, such as *Ola* → *intellectual* from the fragment *Ola reads books*, or *eat* → *verb* from *They tried to eat cake*. An emiton consists of (1) a text fragment, (2) a focus element (name, entity, keyword), and (3) an interpretation.

The emiton differs from summaries, embedding representations, and classifications: it is small, interpretable, and countable, making it suitable for semantic annotation, concept mapping, and the combination of qualitative and quantitative analyses. We also demonstrate how a small, representative sample of concordances – often just a few thousand lines even when millions of hits are available – can provide a solid foundation for analysis. Combined with LLMs, this yields rapid, verifiable insights that in many cases can complement or even replace full-scale collocation analyses.

We show how the emiton can be used to map semantic fields in historical literature, and how it can function as a lightweight, generative format for semantic mapping in the digital humanities.