UNIVERSITETET I OSLO

Emnemodellering med R: Noe for meg?

Luigi Maglanoc, Jarle Ebeling og Heidi Karlsen

The Fish Fork (bruk av korpusmetoder i litteraturforskning)

“Fish [Stanley, 1973] famously critises stylistics in general for being ‘circular’ and ‘arbitrary’, due to its selective attention to data. Either we select a few linguistic features that we know how to describe, and ignore the rest; or we select features which we know are important, describe them, and then claim they are important. Since stylisticians can neither describe everything, nor attach definitive meanings to specific formal features, they are apparently caught in a logical fork.” (Michael Stubbs, 2015)

Ut av the Fish Fork

1. Rejection. Stylistic analysis adds nothing to close reading. Worse, it is circular, since it merely describes a few things which we already knew were important. Furthermore, the frequency of a linguistic feature does not determine its literary importance.
2. A weak defence. Perhaps stylistics analysis tells us nothing new, but it provides precise linguistic description. Computers can count (some) things accurately across large text collections, and there is, after all, some relation between frequency and salience.
3. A stronger defence. Quantitative analysis can discover linguistic features which even expert scholars have not noticed and/or which cannot be discovered by manual analysis: for example, relative frequencies of patterns in a text and in a large reference corpus.
4. The strongest(?) defence. A systematic stylistic analysis not only describes new things, but also helps to explain readers' literary reactions, which are based on unconscious linguistic knowledge of norms of language use and stylistics variation.

(Michael Stubbs, 2015)

Emnemodellering: Hva er det og hva bør man tenke gjennom før man går i gang?

Aller viktigst: Forskningsspørsmål / Hva vil jeg finne ut? Kjenn dine tekster!

- Grovkalibret, frekvensbasert tilnærming til ‘Hva leste du nå?’
- Basert på frekvens og sammenfall av ‘enheter/tegn’ (ord) i en “tekst”
- *Ting å tenke gjennom som vil påvirke resultatet:*
  - ➢ Hva skal utgjøre mine “tekster”?
  - ➢ Stoppordliste (fjerne ord): Kun funksjonsord? Egennavn?
  - ➢ Fjerne tall?
  - ➢ Ordlengde og ordfrekvens – Hva med forkortelser?
  - ➢ Antall emner vil sterkt påvirke resultatet
  - ➢ Enkeltord versus n-grams/ fraser
  - ➢ Lemmatisering
  - ➢ Ordklassetagging
  - ➢ Modellerer datasettet flere ganger (jf. «seed»)

Forskningsspørsmål:

Hvor mye sammenfall/forskjell kan vi identifisere mellom emnene i 4 engelske krimromaner publisert på 1920-tallet?

Teksene (to kvinnelige og to mannlige forfattere):

- AgaChr1
- DorSay1
- EdgWal1
- RonKno1

NB! Få tekster og (for) lite material for oversiktens skyld i dette kurset.