Disappearing discourses: Avoiding anachronisms and teleology with data-driven methods in studying digital newspaper collections

Elaine Zosa, Lidia Pivovarova, Simon Hengchen, Jani Marjanen & Mikko Tolonen
Historian’s dilemma:
How to be true to the past and relevant for the present?
“… an irretrievable image of the past threatens to disappear if any present does not recognize it as meaningful to itself.”

“… es ist ein unwiederbringliches Bild der Vergangenheit, das mit jeder Gegenwart zu verschwinden droht, die sich nicht als in ihm gemeint erkannte.”

– Walter Benjamin, Gesammelte Schriften, I:2, p. 695. (Translation from David Armitage)
“The veto-right of the sources”

“Seeing things their way”
• Data-driven methods cannot solve the theoretical problem
• BUT, on some level, something that has been a question of the historian’s craft and perspective, now is an issue of method.
• Data-driven text-mining methods allow – more than ever before – a historian to disregard the present perspective

Which discourses in fact did disappear? How can we categorize them?
Preliminary results
Dataset and training

- Dataset: *Uusi Suometar* 1869-1917 from the National Library of Finland
- Pre-processing: stopword removal, lemmatization
- Trained a Dynamic Topic Model with 30 topics
- Dynamic Topic Model: models topic evolution
The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

Topic shares
Topic 20 (in light pink) becomes less prominent in the latter years of our corpus.
The NewsEye project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 770299.

Topic 20 (in light pink) becomes less prominent in the latter years of our corpus.
The NewsEye project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 770299.
The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

Sessions of the Diet of Finland (Suomen maapäivät)
The NewsEye project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 770299.
The NewsEye project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 770299.

Topic - Diet of Finland

1874
sääty
esitys
hyväksyä
protokolla
seurakunta
tuomari
valiokunta
päätös
ehdoitella
suomi

1875
seurakunta
esitys
edusmies
valiokunta
pappi
sääty
hyväksyä
ehoitus
kysymys
pitää

1876
edusmies
seurakunta
valiokunta
esitys
sääty
hyväksyä
puhuja
ehoitus
muuttaa
ehdoitus

1877
sääty
mietintö
hyväksyä
valiokunta
esitys
tahtoa
päättös
puhuja
edusmies
muuttaa

The NewsEye project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 770299.
The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.
The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.
The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.
Next steps

• Train on longer time periods and include other newspapers
• Use a topic model that explicitly models time distribution of static topics (Topics over Time, Wang & McCallum, 2006)
More About NewsEye & Comhis

https://www.newseye.eu/
Twitter: @NewsEyeEU
https://www.helsinki.fi/en/researchgroups/computational-history