The Ancient Finnish Kings: a Computational Study of Pseudohistory, Medievalism and History Politics in Contemporary Finland and Russia

HELDIG Summit

7 November, Helsinki

Reima Välimäki, Heta Aali, Harri Hihnala, Mila Oiva, Anna Ristilä

TURUN YLIOPISTO

The Ancient Finnish Kings: a computational study of pseudohistory, medievalism and history politics in contemporary Finland and Russia

- Project funded by Emil Aaltonen Foundation 2019-21
- History culture, history politics and pseudohistory in Finland and Russia
- Narratives about Viking Age, Late Iron Age and the Middle Ages - > Medievalism
- Finnish and Russian Internet discussions: webpages, blogs, discussion forums and comment feeds, newspaper articles.

Prince Rurik and his brothers. A picture circulating in pseudohistory blogs. Original source: Ilja Glazunov: ”Grandchildren of Gostomysl: Rurik, Truvor, Sineus” (1986)

The Ancient Finnish Kings: a computational study of pseudohistory, medievalism and history politics in contemporary Finland and Russia

- Research team:
  - PI Reima Välimäki
  - Postdoctoral researchers Heta Aali and Mila Oiva
  - Junior researcher Harri Hihnala and Anna Ristilä

- Computational methods: text reuse using BLAST

- Languages analysed separately
  - Qualitative comparison of Finnish and Russian medievalist themes

- NOT only a digital phenomenon:
  - Printed pseudohistory (often self-published)

TURUN YLIOPISTO

Pseudohistory

- Part of the broader phenomenon of pseudoscience
- Pseudohistory ≠ popular or amateur history
- Narratives that claim to present facts, even research about history but do not follow the scholarly practices of history, archaology, linguistics or other disciplines involved.
- Misinformation, exaggeration and outdated interpretation
  - Arbitrary and selective use of sources
- Bypassing current scholarship and facts that contradict own interpretation
- Conspiracy theories: researchers, museums, archives or the state hide the true facts
- Demarcation problem

TURUN YLIOPISTO

Ancient Finnish Kings

The most prominent pseudohistorical narrative:

- Finnish kingdom(s) in an era before written history (iron age / Viking age)
- Strong political and military entities
- Finnish kings as ancestors of royal families of Norther Europe
- The history has been forgotten or hidden -> conspiracy theories

Why to study internet medievalism and pseudohistory?

- New forums for pseudohistory
- Some pseudohistorical blogs attract thousands of readers
- Opportunity for more aggressive engagement with scholars
- Far-right medievalism: an international phenomenon
- Is pseudohistory more popular?

Finland and Russia

- Shared medieval history
  - Shared medievalist imaginations?

- Medievalism in Finland
  - Medievalism in nation building in the 19th and early 20th century
  - Currently not a prominent part of the public historical understanding or school teaching
  - History re-enactment and Viking and medieval fairs.
  - Lively internet communities discussing the Viking and Middle Ages
  - Anti-Russian attitudes in the early 20th-century medievalism: what about know?

- Medievalism in Russia
  - “Millenium of Russia”
  - Part of state history policy
  - Problems with the legacy of Rus

Image: Rurik in the Millenium of Russia -monument (1862)

TURUN YLIOPISTO

Expected results

- Prominent narratives in the medievalist pseudohistory
- Origin and transmission of these narratives
- Circulation of identical texts
- Argumentation strategies in pseudohistory
- Textual communities of pseudohistory
- New insights into historical imagination of viking age and medieval past
    - How could scholars contribute for better and more balanced understanding?

TURUN YLIOPISTO

Data & Methods of the Project

Mila Oiva & Anna Ristilä

Turun yliopisto
University of Turku

CC BY

Scraping the Data

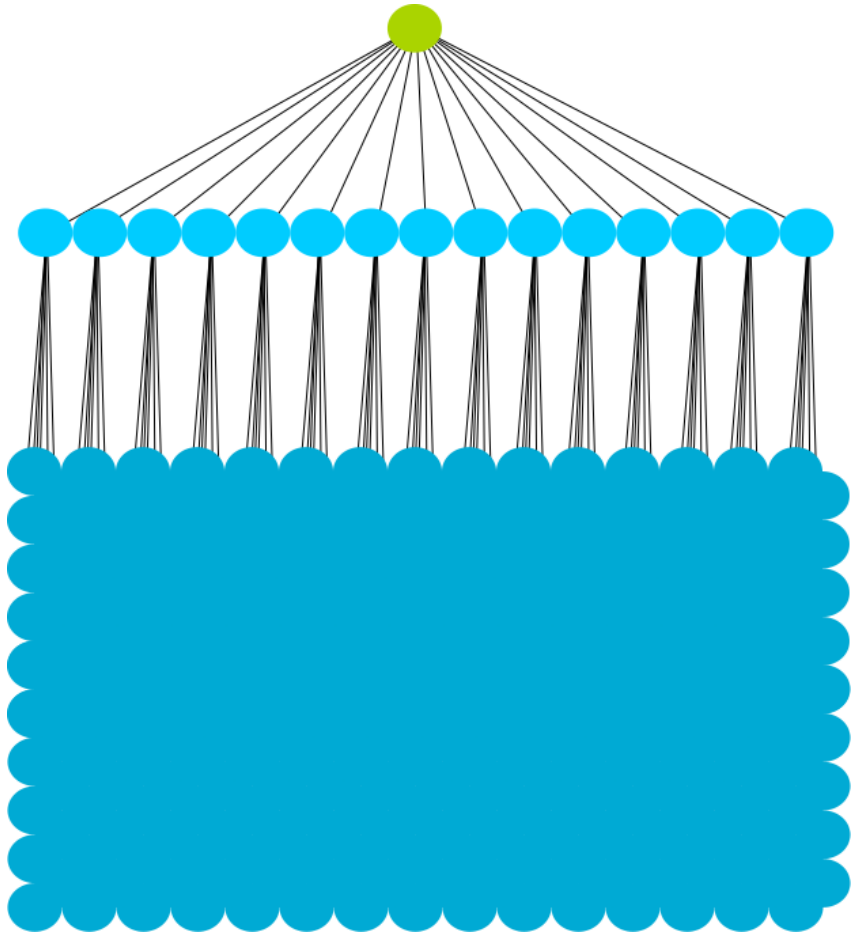
- manually searched seed links
- scraping to the third step

Scraping the data

- The textual data was scraped with a programmed web scraper (named Phist). The scraper was coded with Python 3.5 and it collected data in three cycles.
- First, the seed links were fully scraped. In the following two cycles all unique outgoing links from the previous cycle were scraped. This way the scraper was able to collect material "two steps down" from the seed links.
- No YouTube or Instagram links were scraped. Facebook, Twitter, VKontakte and LinkedIn links were collected but not scraped.

Exponentiality of the Internet

- Seed links: 378
- 2nd level links: 37’701
- 3rd level links: 1’888’084



Data

- scrape date
- page url
- full text (no menus or equivalent)
- all outgoing links
- publish dates from 4 select domains
- authors from 1 select domain
- embedded comments and their authors from 1 select domain

The scraper recorded the data in .json format.

Amount of data (Russian internet)

- Links in total: ~2 million
- Total computing time: ~80 days
- Wall-clock computing time: ~2 days
- Data amount: ~250 GB

Methods

- Which texts spread widely in the internet? What kinds of texts they are? Are they somehow connected? Why these texts spread?
- > text reuse detection with BLAST
- > clusters of reused texts
- > networks of reused texts

BLAST

Applying BLAST to Text Reuse Detection in Finnish Newspapers and Journals, 1771–1910

Aleksi Vesanto
Turku NLP Group
Department of FT
University of Turku
aleksi.vesanto@utu.fi

Asko Nivala
Cultural History and
Turku Institute for Advanced Studies
University of Turku
asko.nivala@utu.fi

Heli Rantala
Cultural History
University of Turku
heli.rantala@utu.fi

Tapio Salakoski
Turku NLP Group
Department of FT
University of Turku
tapio.salakoski@utu.fi

Hannu Salmi
Cultural History
University of Turku
hannu.salmi@utu.fi

Filip Ginter
Turku NLP Group
Department of FT
University of Turku
filip.ginter@utu.fi

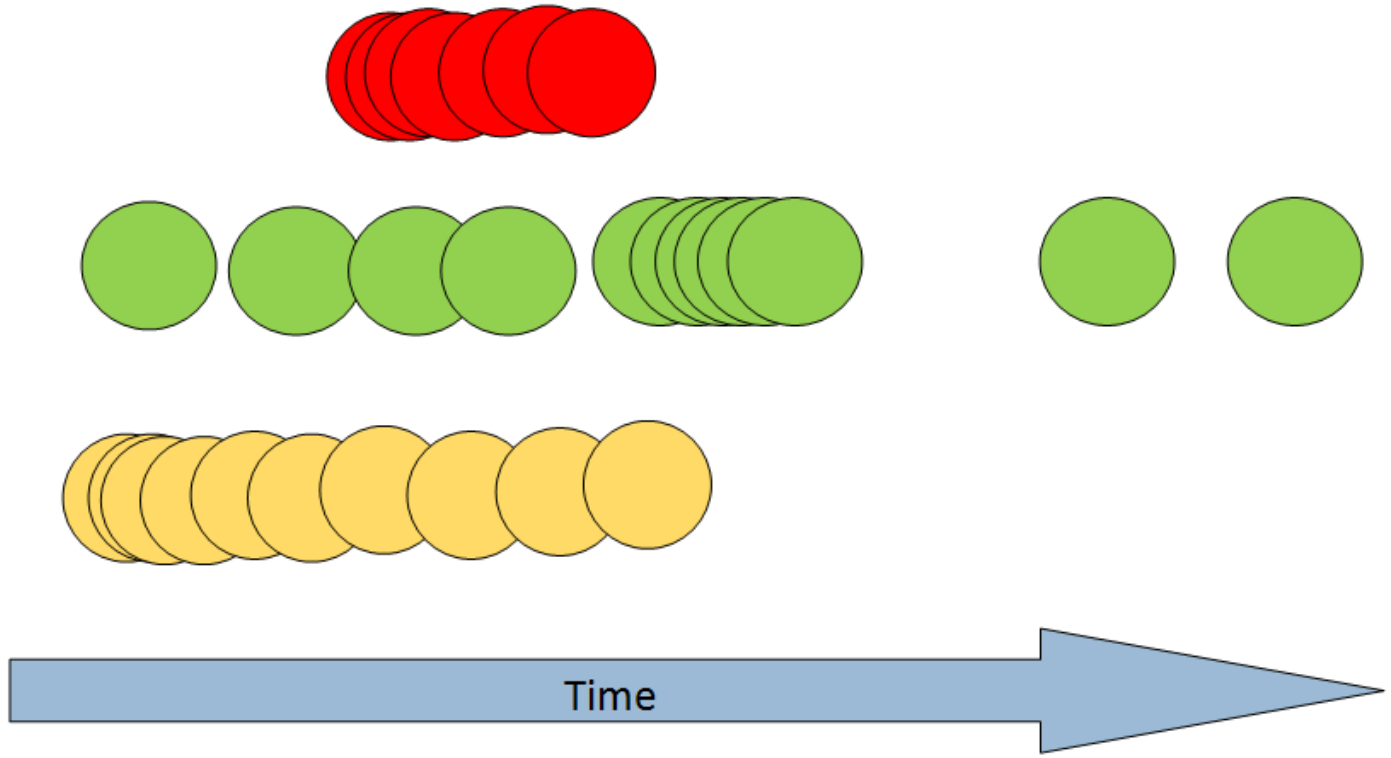
Abstract

We present the results of text reuse detection, based on the corpus of scanned and OCR-recognized Finnish newspapers and journals from 1771 to 1910. Our study draws on BLAST, a software created for comparing and aligning biological sequences. We show different types of text reuse in this corpus, and also present a comparison to the software Passim, developed at the Northeastern University in

of nineteenth-century US newspapers by Ryan Cordell, David A. Smith and their research group (Cordell, 2015; Smith et al., 2015). However, in contrast to the US press, the nineteenth- and early twentieth-century Finnish newspapers were typically printed in the *Fraktur* typeface, which (together with other possible sources of noise) poses unusual difficulties for Optical Character Recognition (Kettunen, 2016). To solve this problem, we have developed a novel text reuse detection solution based on BLAST (Vesanto et al., 2017) that is

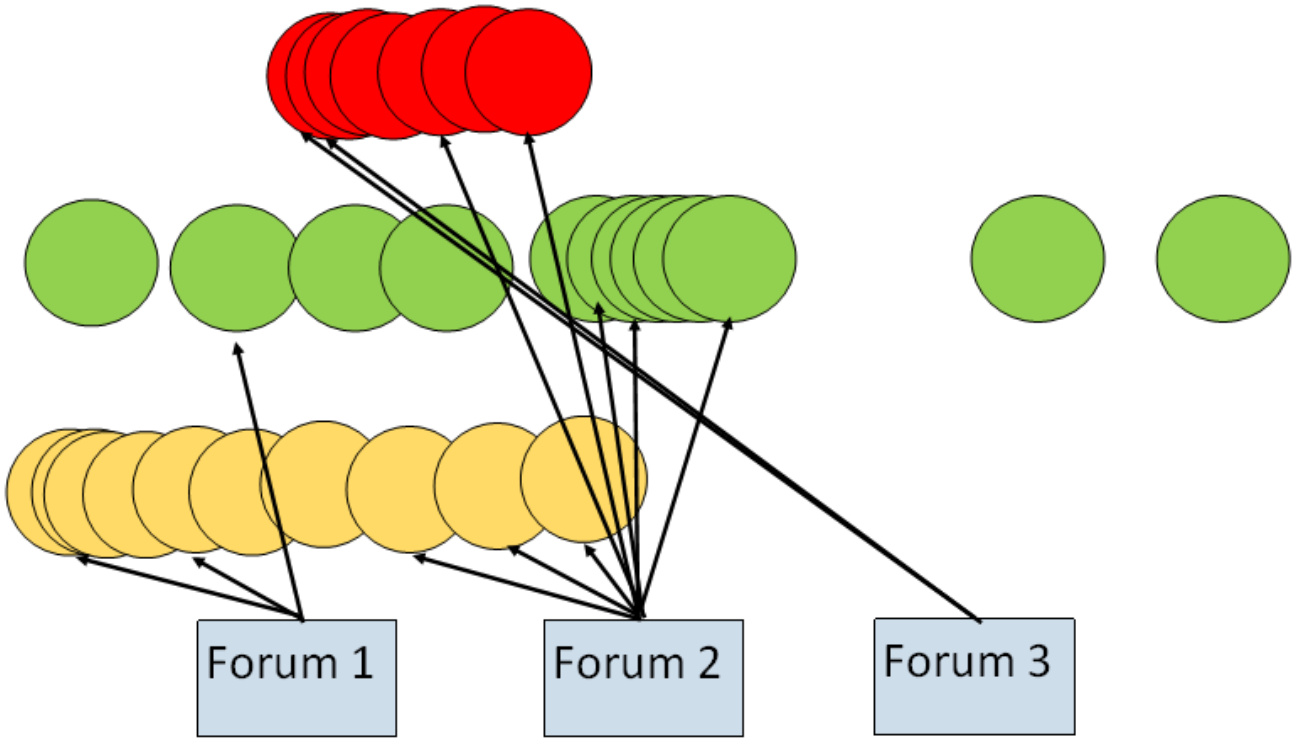
- Originally created for comparing and alingning biological sequences
- But works also with texts!
- Resistant to OCR errors

Clusters of reused texts



CC BY

Clusters and forums



Networks?

- Networks of urls sharing texts
- Networks of urls linking to each other

TURUN
YLIOPISTO

https://sites.utu.fi/pseudohistoria/