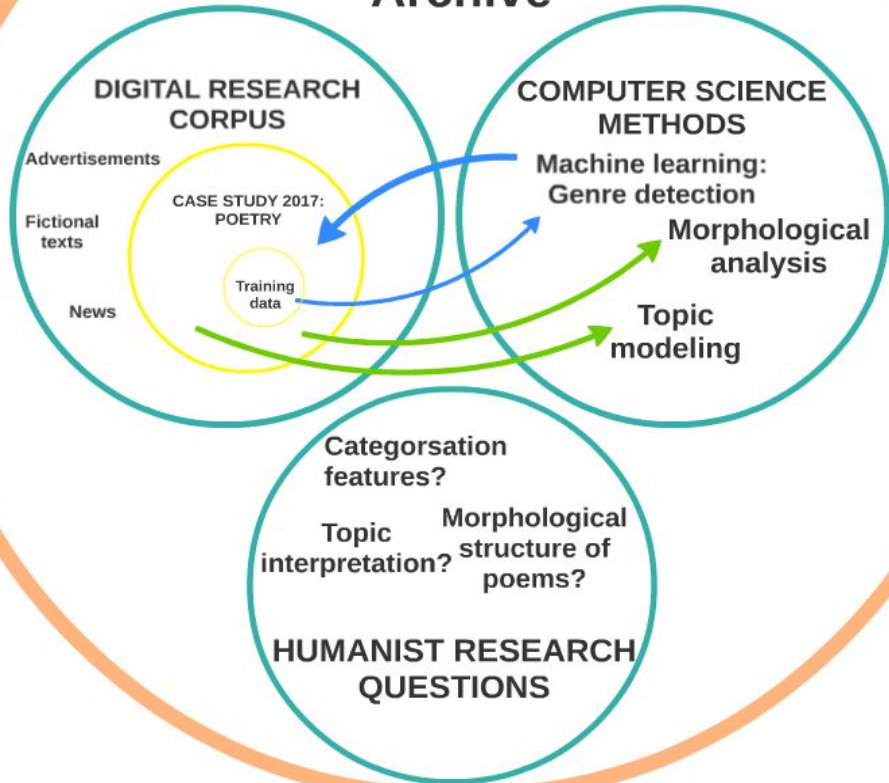


Big data approach to 19th-century Finnish newspaper literature

Mikko Koho, Agata Dominowska, Elsi
Hyttinen, Péter Ivanics, Elizabeth Oakes, Ilona
Pikkanen, Leena Tulkki, Risto J. Turunen

A New Literary History of the Digital Newspaper Archive



DH Hackathon 2017

Goals:

Complementing previous research

Revealing a larger picture of 19th-century Finnish newspaper literature

Distant reading: going beyond the literary canon

Case study: poetry

National Library's Digital Collections

- 5.6 million OCR'd pages of historical newspapers
- <http://digi.kansalliskirjasto.fi>
- Available as data

M a t i l l e !

*Oi vanha Matti kuitenkin
Kuin ootkaan nyt jo heikko;
Pois ruso on sun poskiltas:
Niin oot kuin vanha peikko!*

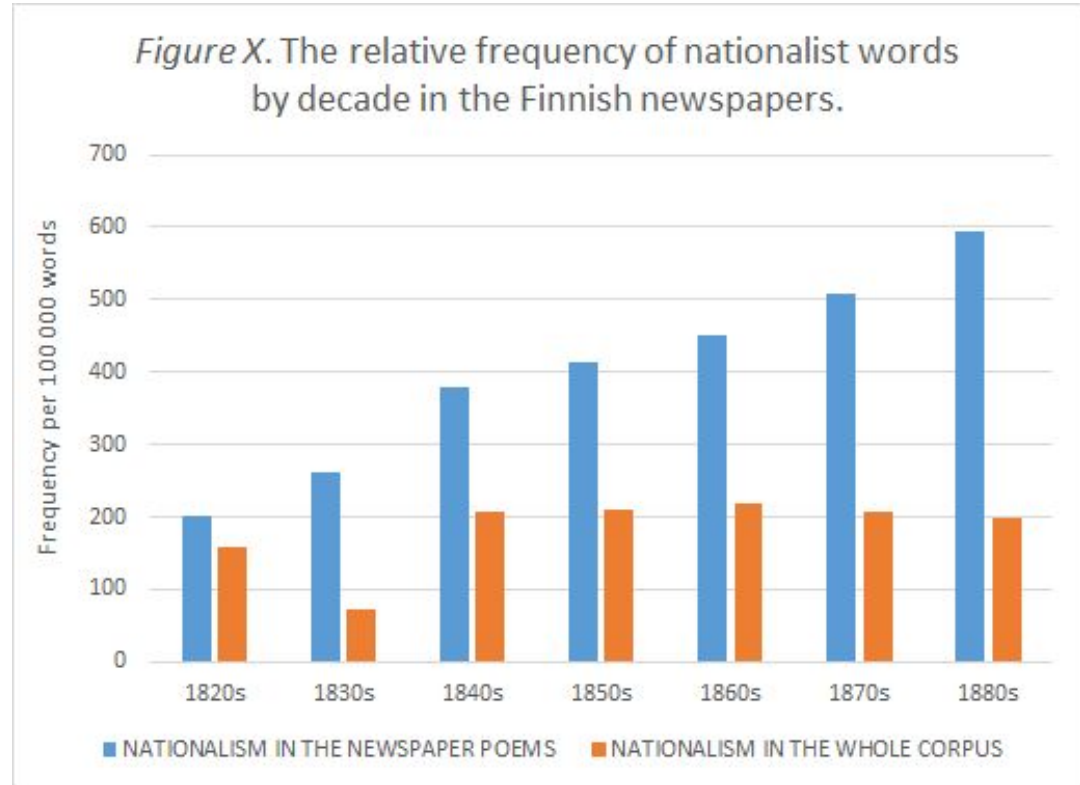
A. J. P.

Method for extracting poems

- Support Vector Machine classifier
 - 2 classes: poems & non-poems
 - Features
 - Word frequencies
 - 5 additional ad-hoc features
- Manually collected training data

Results & Analysis

- Found 18,591 poems
- Precision 87%
- Recall 31%
 - Varied highly between years



Thank you!