Machine-Assisted Dating of Medieval Texts with Cluster Analysis

Oksana Dereza

National Research University “Higher School of Economics”,
Lomonosov Moscow State University

oksana.dereza@gmail.com

Dating Medieval Texts

Problems

- Manuscript date ≠ language period
    - Older texts copied throughout centuries
    - High literary style = archaic forms in later texts
- Old editions
    - More linguistic knowledge now
    - Some texts are overedited
- Lack of data = lack of confidence
    - 550 – 1140
    - 1400 – 1600
    - before XII century

Dating Medieval Texts

Traditional methods

- Linguistic
  - Scribal errors
  - Occasional late grammar forms
- Extralinguistic
  - Records about scribes, patrons etc. (available only for late manuscripts)

What if there are some orthographic patterns on character level unseen by human eye?

Research Questions

- Do the manuscripts written in the same language (according to the editor) cluster together?
- Do the manuscripts dated to the same period (according to the editor) cluster together?
- Which texts cluster together and can it help us to date them more accurately?

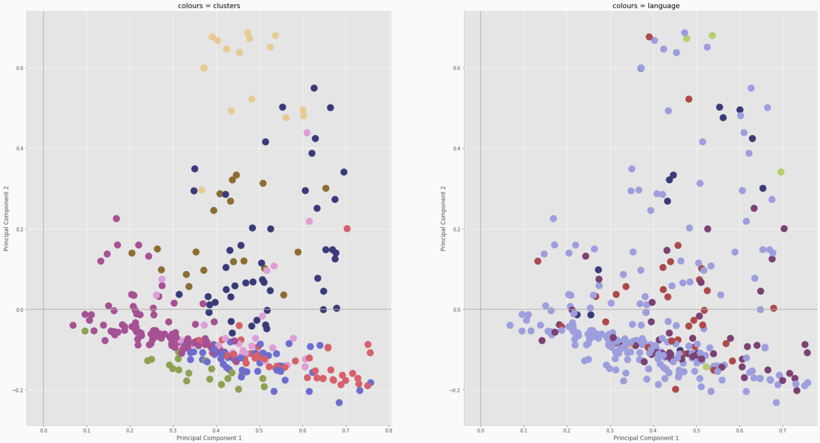
Data

- Digital editions from CELT
- 329 texts, both prose and verse
- 1,513,785 tokens; 447,262 unique
- 7 different language labels, 141 date labels, 46 editors
- Languages
    - Old Irish – 32 texts
    - Old Irish; Middle Irish – 54 texts
    - Middle Irish – 223 texts
    - Middle Irish; Early Modern Irish – 4 texts
    - Early Modern Irish – 14 texts
    - Old Irish; Middle Irish; Early Modern Irish – 1 text

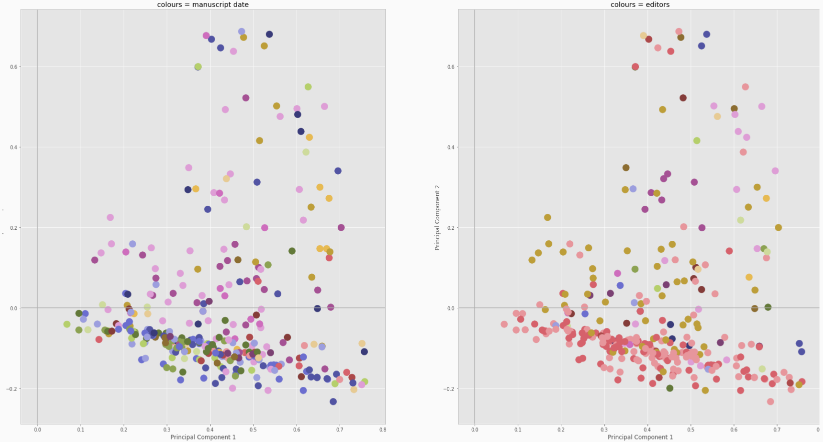
Pipeline

- Vectorising texts with tf-idf
- Character-level n-grams, from bigrams to 5-grams
- Spectral clustering algorithm
- LSA and t-SNE algorithms for dimension reduction
- Comparing with editors’ judgement

Spectral clustering + LSA



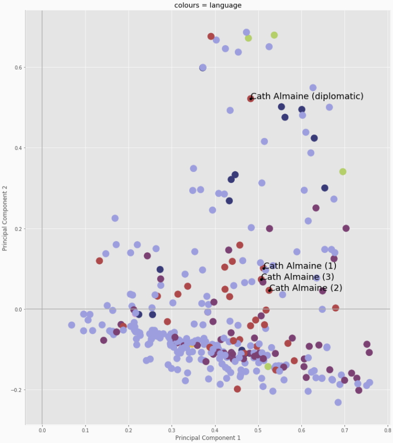
Spectral clustering + LSA

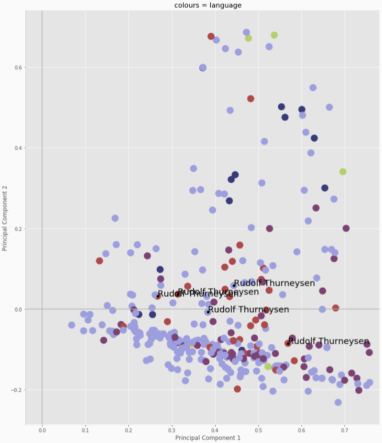


colours = manuscript date

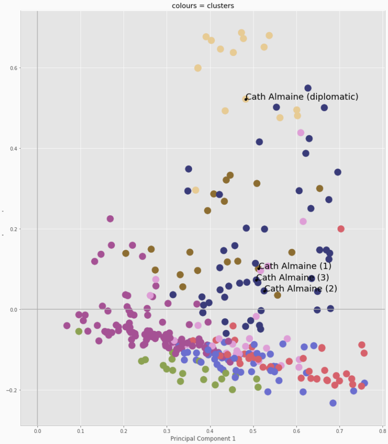
colours = editors

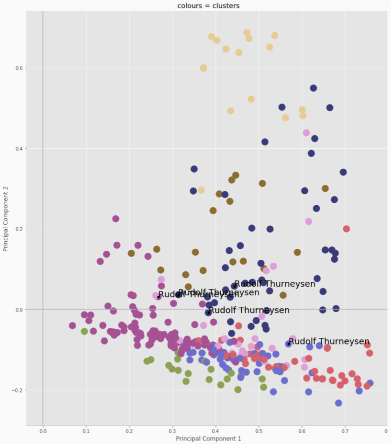
Cath Almaine / Rudolf Thurneysen



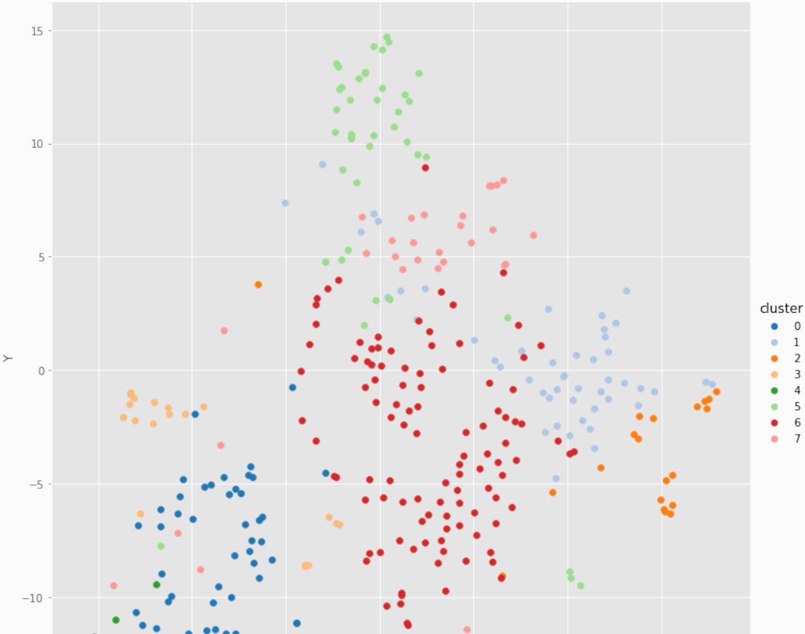


Cath Almaine / Rudolf Thurneysen

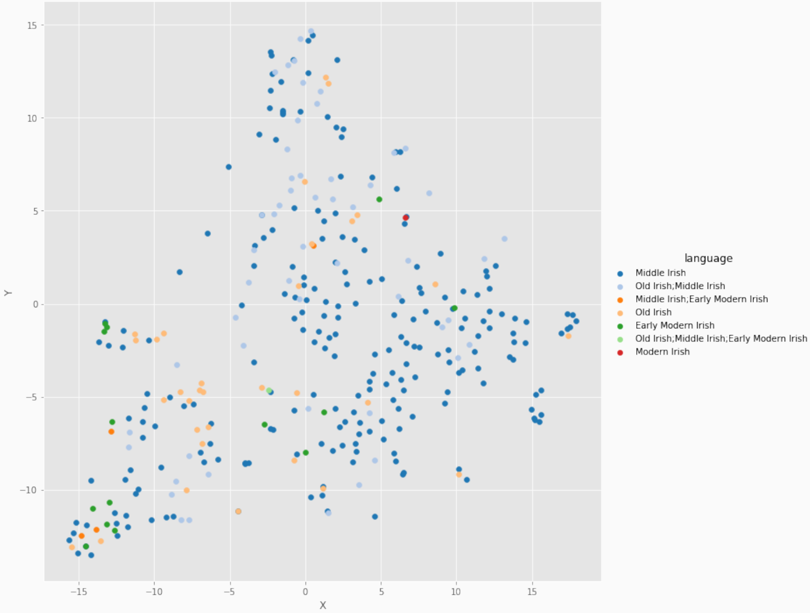




LSA + t-SNE

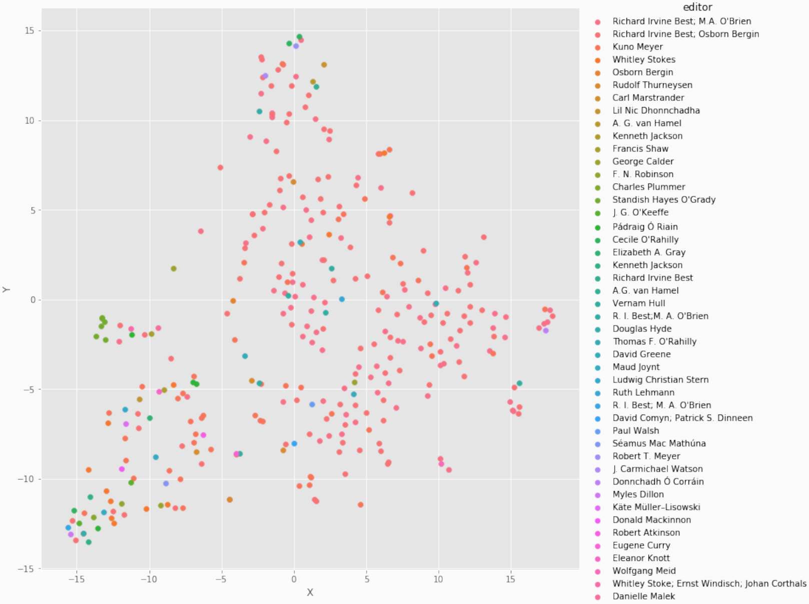


LSA + t-SNE



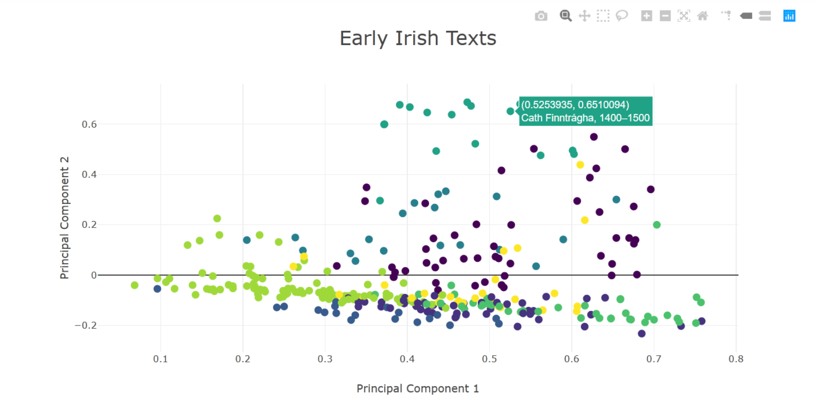
15
10
5
0
−5
−10
Y
−15
−10
−5
0
5
10
15
X
language
● Middle Irish
● Old Irish;Middle Irish
● Middle Irish;Early Modern Irish
● Old Irish
● Early Modern Irish
● Old Irish;Middle Irish;Early Modern Irish
● Modern Irish

LSA + t-SNE


editor
Richard Irvine Best; M.A. O'Brien
Richard Irvine Best; Osborn Bergin
Kuno Meyer
Whitley Stokes
Osborn Bergin
Rudolf Thurneysen
Carl Marstrander
Lil Nic Dhonnchadha
A. G. van Hamel
Kenneth Jackson
Francis Shaw
George Calder
F. N. Robinson
Charles Plummer
Standish Hayes O'Grady
J. G. O'Keeffe
Pádraig Ó Riain
Cecile O'Rahilly
Elizabeth A. Gray
Kenneth Jackson
Richard Irvine Best
A.G. van Hamel
Vernam Hull
R. I. Best;M. A. O'Brien
Douglas Hyde
Thomas F. O'Rahilly
David Greene
Maud Joynt
Ludwig Christian Stern
Ruth Lehmann
R. I. Best; M. A. O'Brien
David Comyn; Patrick S. Dinneen
Paul Walsh
Séamus Mac Mathúna
Robert T. Meyer
J. Carmichael Watson
Donnchadh Ó Corráin
Myles Dillon
Käte Müller-Lisowski
Donald Mackinnon
Robert Atkinson
Eugene Curry
Eleanor Knott
Wolfgang Meid
Whitley Stoke; Ernst Windisch; Johan Corthals
Danielle Malek
Y
15
10
5
0
−5
−10
−15
X
−15
−10
−5
0
5
10
15

Interactive visualisation

https://plot.ly/~ancatmara/86/early-irish-texts/



https://github.com/ancatmara/medieval_texts_clustering/blob/master/medieval_texts_clustering.ipynb/