UNIVERSITY OF COPENHAGEN
CENTRE FOR LANGUAGE TECHNOLOGY
DEPARTMENT OF NORDIC STUDIES AND LINGUISTICS

Identifying Temporal Trends Based on Perplexity and Clustering: Are We Looking at Language Change?

Sidsel Boldsen[1], Manex Agirrezabal[1], Patrizia Paggio[1,2]

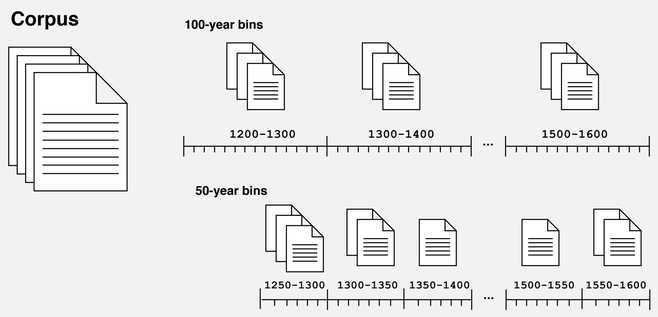
[1]Centre for Language Technology, University of Copenhagen
[2]Institute for Linguistics and Language Technology, University of Malta
{sbol,manex.aguirrezabal,paggio}@hum.ku.dk

Diachronic text classification

Recent approaches have looked at the task of identifying temporal trends in document collections using NLP methods. However:

(I) No assumption is made about texts from two time spans close to each other being closer than others belonging to time spans further away.

(II) How the time spans should be chosen, both in terms of their size and the exact placing of the boundaries between them, seems arbitrary.



We propose a data-driven approach to the identification of temporal trends in a corpus of medieval charters:

(I) We derive perplexity measures reflecting how similar documents pairs are, and how this similarity correlates with the time difference between them.

(II) We cluster the documents based on perplexity.

Perplexity as a measure of language change?

Perplexity has been proposed as a measure of language distance, and recently used to distinguish formal from colloquial tweets, to measure distance between languages, and between historical varieties of the same language.

Given a test set consisting of a sequence of characters (CH) and a character-based language model (LM), perplexity is defined by the following equation:

$$\mathrm{PP(CH,\,LM)} = \sqrt[n]{\prod_{i}^{n} \frac{1}{\mathrm{P}(\mathrm{ch}_i \mid \mathrm{ch}_i^{i-1})}}$$

A character based LSTM language model shows a moderate correlation for Latin texts (r=0.50, p<0.01) and only a weak one for the Danish texts (r=0.20, p<0.01).

Identifying temporal trends

Having trained a language model, LM_i, for each of the documents, d_i, in the collection, D, we let each of the documents in D be represented by a vector, X_i, of size $|D|$, where each value $x_{i,j}$ corresponds to the perplexity of LM_i applied to a document d_j, resulting in a distance matrix, M:

$$\mathrm{M} = \begin{bmatrix} 1.2 & 7.3 & \cdots & 3.2 \\ 5.6 & 1.8 & \cdots & 2.9 \\ \vdots & \vdots & \ddots & \vdots \\ 4 & 3.9 & \cdots & 1.1 \end{bmatrix}$$

$$\mathrm{K{-}Means}(\mathrm{X}_1,\ \mathrm{X}_2,\ \ldots,\ \mathrm{X}_N)$$

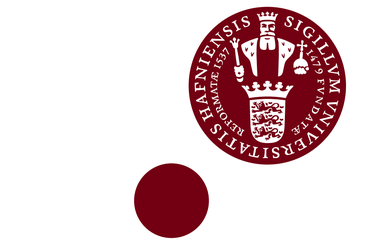
We use K-Means clustering to perform cluster analysis of the documents in the collection based on these perplexity measures. If perplexity is indicative of language change as a measure of (dis)similarity, our hypothesis is that such an analysis will give insights into how a collection of documents changes over time.

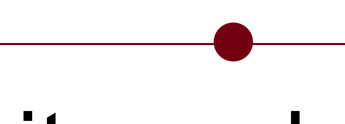
Data

291 charters belonging to a larger collection of charters from St. Clara Convent in Denmark

Most documents either in Latin or in Danish (shift to Danish during the 15th century)

Length: 351-3099 characters after subsampling.





Results

We run k-means clustering for all values k ∈ {2, ..., 10} and found that k = 7 provided a good fit in terms of intra- and inter-cluster distance.

Vectors were projected onto two components using t-SNE and documents were colour-coded according to assigned temporal bins (Figure 2).

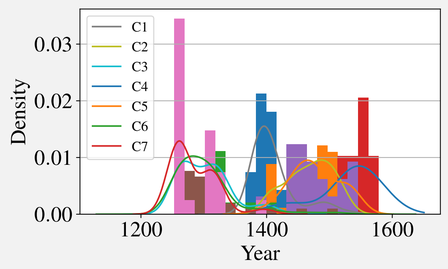


Figure 1
Year distribution for each cluster

(1) The t-SNE projection shows two main groups, corresponding to the Latin (left) and the Danish (right) documents.

(2) The group of Latin documents is sub-divided into two, with colours indicating one earlier (dark red + orange), and one later (light orange + yellow).

(3) Temporal outliers can be observed within the clusters.

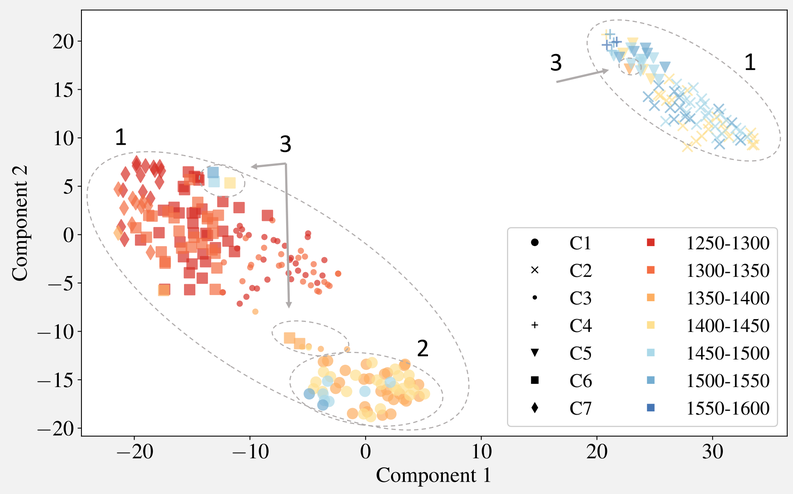


Figure 2
T-SNE projection of the documents in our dataset. For each document, the shape represents the cluster to which the document belongs based on K-Means. The colour shows the year-span to which the document belongs.

Conclusion

- We have proposed a methodology for the identification of temporal trends in a document collection as an alternative to using fixed year spans.
- The perplexities calculated by document specific language models correlate moderately with time differences.
- Performing K-Means with K=7 based on perplexity measures allowed us to discover groups reflecting language change.
- However, the distribution of the data-driven temporal bins (Figure 1) shows an overlap between the identified clusters. Further investigation is needed to get a better understanding of the clusters.
- We suggest, however, that the temporal distribution of the clusters may still give a more nuanced picture of temporal trends compared to discrete bins, and provide better results when used in a classification task.

This study was conducted within the project *Script and Text in Time and Space*, a core group project supported by the Velux Foundation.