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 of Linguistics and Language Technology
University of Malta
{sbol,manex.aguirrezabal,paggio}@hum.ku.dk

Abstract

In this work we propose a data-driven methodology for identifying temporal trends in a corpus of medieval charters. We have used perplexities derived from RNNs as a distance measure between documents and then, performed clustering on those distances. We argue that perplexities calculated by such language models are representative of temporal trends. The clusters produced using the K-Means algorithm give an insight of the differences in language in different time periods at least partly due to language change. We suggest that the temporal distribution of the individual clusters might provide a more nuanced picture of temporal trends compared to discrete bins, thus providing better results when used in a classification task.

1 Background

Several recent approaches have looked at the task of identifying temporal trends in document collections using NLP methods. An example is the diachronic text evaluation challenge (Popescu and Strapparava, 2015) in SemEval 2015, where newspaper text snippets from 1700-2010 had to be classified into time intervals of different sizes. Models for diachronic text classification are trained based on the way lexical, morphological, syntactic and stylistic features change over time (Abe and Tsumoto, 2010; Garcia-Fernandez et al., 2011; Popescu and Strapparava, 2015; Štajner and Zampieri, 2013; Szymanski and Lynch, 2015; Zampieri et al., 2016; Boldsen and Paggio, 2019).

Diachronic text classification, however, is a simplification. Firstly, no assumption is made about texts from two time spans close to each other being closer than others belonging to time spans further away. Furthermore, how the time spans should be chosen, both in terms of their size and the exact placing of the boundaries between them, seems often a rather arbitrary decision.

Important insights relevant to the issue may come from research dealing with language distance and language identification. The underlying assumption in this area is that the more difficult it is to identify differences between two languages or language varieties, the shorter is the distance between them. Perplexity has been proposed as a measure of language distance, and recently used to distinguish formal from colloquial tweets (González Bermúdez, 2015), to measure distance between languages (Gamallo et al., 2016, 2017), and, interestingly for our purposes, between historical varieties of the same language (Pichel Campos et al., 2018).

In this paper, we propose a data-driven approach to the identification of temporal trends in a corpus of medieval charters. This is a particularly interesting test-bed in that medieval manuscripts often lack explicit reference to when they were produced, and this knowledge is crucially important for their philological interpretation. We first derive perplexity measures that reflect how similar the documents are to one another, and how this similarity correlates with the time difference between them, and then we cluster the documents based on perplexity. The groups obtained through clustering are evaluated with respect to a manually determined classification into discrete 50-year time periods, a method often used to distinguish historical variants of a language, and which was applied to medieval charters in Boldsen and Paggio (2019).

We believe the idea of clustering documents based on perplexity measures as a method to discover temporal trends in a document collection is a novel and, as we argue below, promising one.

Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pages 86–91
Florence, Italy, August 2, 2019.

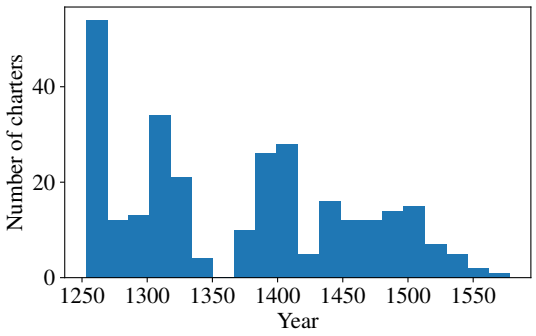

Figure 1: Plot of the distribution of the charters along the temporal line.

2 Methodology

In this section we introduce the dataset and the methods that we have employed in this work.

2.1 Dataset

The dataset in this study consists of 291 charters belonging to a larger collection of charters from St. Clara Convent in Denmark, which is part of the interdisciplinary research project *Script and Text in Time and Space*[1] studying the development of medieval Danish language and script. The charters, which are being prepared for a scholarly edition, document the property and status of the convent from when it was founded in 1256 till it was closed after the Reformation. Two different transcription levels are included in the dataset: (i) the facsimile transcription, where allographic variation is annotated, and (ii) the diplomatic transcription, where this variation is normalised, while spelling variants are kept. Most of the documents are either in Latin or in Danish, with a shift during the 15th century to documents being written in Danish. There are also two texts in Low German from earlier than 1450, and two Swedish ones from 1500-1550. The number of charters available from the various periods varies, as shown in Figure 1.

In addition to the language variation, the charters also vary in length. Therefore, the dataset was resampled by normalising the length of the individual documents. This was done by finding the outliers in the distribution (documents longer than approximately 3000 characters) and randomly subsampling text from them to get as close as possible to the average length of the rest of the collection. This process produced a more balanced dataset of 291 documents of length between 351 and 3099 characters.

[1] https://humanities.ku.dk/research/digital-humanities/projects/writing-and-texts-in-time-and-space

2.2 Perplexity and language modelling

Perplexity is a metric that expresses how well a language model fits a test sample. It is based on the computation of the probability of each sentence in the test set as predicted by the language model. A low perplexity corresponds to a high probability of the sentences in the test sample.

Given a test set consisting of a sequence of characters (CH) and a character-based language model (LM) with n-gram probabilities $P(ch_i|ch_1^{i-1})$, perplexity (PP) is defined by the following equation (Pichel Campos et al., 2018, 148):

$$PP(CH, LM) = \sqrt[n]{\prod_i^n \frac{1}{P(ch_i|ch_1^{i-1})}} \quad (1)$$

We train document-specific character-based language models and test each model on the remaining documents in the collection. A perplexity measure is then computed for each pair of language model and test document. The measure is used as an estimator of the distance between each document pair. Since the charters represent different stages of language development during a time period of about 350 years, we expect the perplexity related to pairs of language models and test manuscripts to increase with the temporal distance between the text from which a language model is derived and the text to which the same model is applied.

2.3 Language models

As a baseline we used character trigrams to estimate character language models for each of the documents in our corpus, and then calculated the perplexity of each document, given each language model (Stolcke, 2002). We estimated the probabilities by Maximum Likelihood and dealt with zero-counts using Witten-Bell smoothing.

To get more representative language models we then trained Recurrent Neural Network Language Models (Elman, 1990) with LSTM (Hochreiter and Schmidhuber, 1997). The main advantage of RNNs is that the Markov assumption from the trigram language model is relaxed, and thus, the

quality of the language model is expected to be better. Our RNN also makes use of an embedding layer that projects each character to a numeric representation. This numeric representation is given as input to the LSTM cell, which, together with the previous layers content, generates a probability distribution of the possible next characters. Then, calculating the perplexity of a language model in a test corpus is relatively simple, if we consider the probability of the whole test sequence.

2.4 Clustering

Having trained a language model, LM, for each of the documents, d, 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 a language model, LM_i, trained on document d_i, and applied to a document, d_j.

We use k-means clustering to perform cluster analysis of the documents in the collection. In k-means the objective is to find the best k clusters which minimise the distance between cluster centroids and the data points within the clusters (Bishop, 2006). Thus, when applying k-means to the collection of documents, we find clusters of documents which are similar in terms of perplexity. If perplexity is indicative of language change as a measure of (dis)similarity, our hypothesis is that such an analysis will give insights to how a collection of documents changes over time.

3 Results and discussion

In this section first we discuss the usefulness of the perplexity measures as predictors of distance between documents on the temporal line, and then we give an account of the clustering results.

3.1 Perplexity as a predictor of language change

To evaluate whether perplexity was a good basis on which to cluster the charters, in other words whether the perplexity measures modelling similarity between documents are actually related to temporal change, we run a correlation between those measures on the one hand, and differences in years between each document pair on the other. The expectation was that the higher the perplexity between a model and a text is, the greater the temporal distance between them.

The correlation is moderate when using the perplexity calculated by the baseline (Pearson'r = 0.49, p-value < 0.01), and even higher when using the values provided by the RNN model (Pearson'r = 0.65, p-value < 0.01). It thus shows that the neural language model does a better job than the baseline.

However, language change from Latin to Danish during the 15^{th} century might be the main factor behind the correlation strength. To test this, we partitioned the perplexity data from the RNN model into two groups based on the language, and run correlation tests for each partition separately. Although we still found a moderate correlation for the Latin texts (Pearson'r = 0.50, p-value < 0.01), only a weak one was observed for the Danish ones (Pearson'r = 0.20, p-value < 0.01). Nevertheless, for the majority of the charters in the dataset, perplexity still appears potentially useful for the task of modelling temporal change, and was indeed used to drive the clustering.

3.2 Results of clustering

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

To visualise the results, the document vectors were projected onto two components using t-SNE (Maaten and Hinton, 2008). The resulting projections can be seen in figure 2, in which three groups of documents are clearly distinguished: two groups to the left - one at the top and one at the bottom - and one in the top right corner. The clusters from the k-means clustering are indicated through shapes, revealing that clusters 3, 6 and 7 are gathered in the top left group, whereas the group in the middle mostly consists of instances from cluster 1, and the top right group is mostly made up of texts from clusters 2 and 5. Temporal outliers can be observed in all three groups.

In order to evaluate what the clusters can tell us about the temporal development of documents in the collection, we colour-coded the documents according to their manually assigned temporal bins. Figure 2 shows the distribution of earlier (warmer colours) and later (cooler colours) documents.

First of all, had we coloured the documents to highlight the different languages, we would see that the left groups correspond to the Latin documents and the right one to all the rest, i.e. Danish, Swedish and Low German. This result is highly

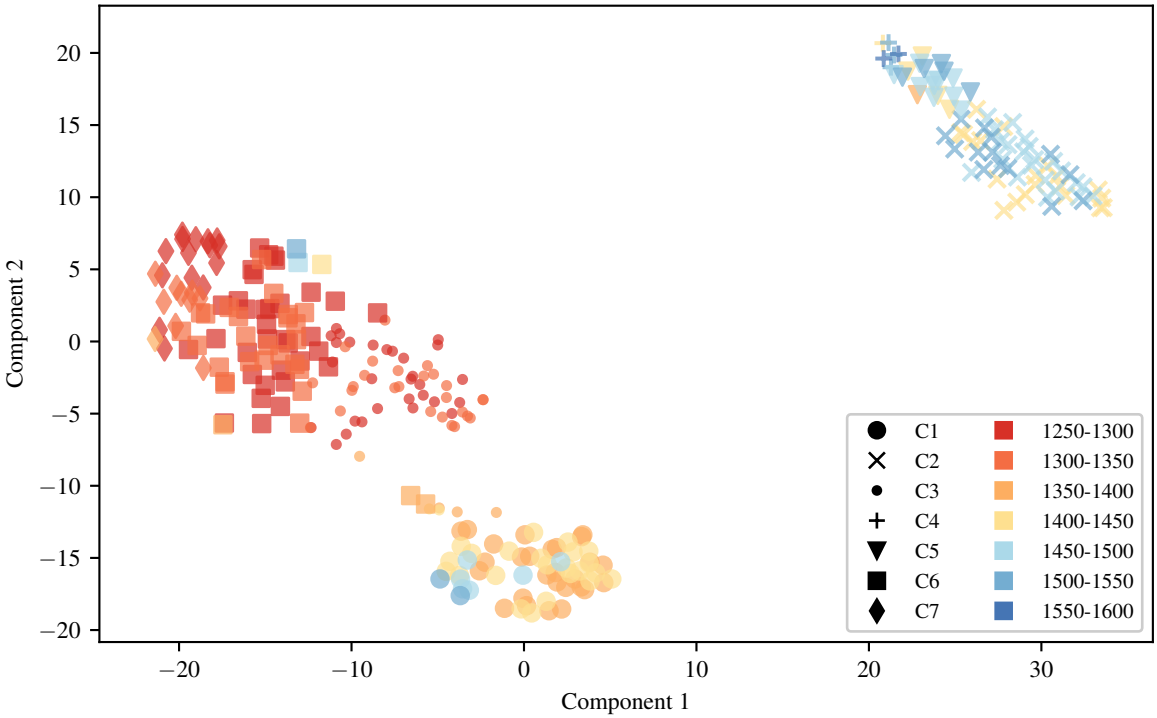

Figure 2: T-SNE projection of the documents in our dataset. Each document is represented as a vector of perplexities. For each document, the shape represents the cluster to which the document belongs based on K-Means. The colour shows the year-span to which the document belongs.

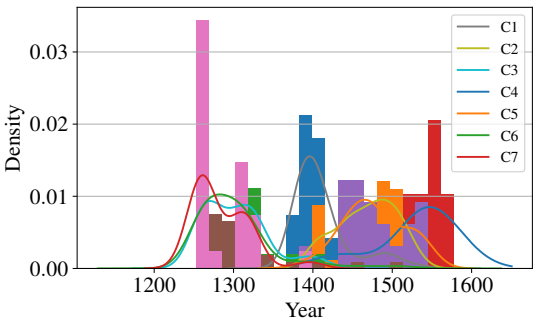

Figure 3: Year distribution for each cluster.

expected given what we know about the language distribution.

Secondly, the top left clusters seem to represent earlier documents (red, dark orange) relative to the internal temporal distribution of the Latin documents, while the lower cluster represents the later ones (light orange, yellow). It remains to be seen if this partition corresponds to time-related language change or some other difference (different scribe, different register, etc.) or whether it is due to a gap in the data just before 1350 (see figure 1).

Looking at the distribution of the temporal bins more closely, however, there is no really clear pattern to how these are distributed between the individual clusters. If we focus on the top left group corresponding to the earliest temporal bins, for instance, it is difficult to interpret the way the three clusters - 3, 6 and 7 - are distributed within the period. This is confirmed in Figure 3 where the clusters are plotted as yearly distributions using Gaussian kernel density estimation (Bishop, 2006). The plot makes it evident that the distributions of the three clusters overlap. This suggests that there may be other factors than language change as such influencing the models. For example, we know that a group of papal letters belong to the early stages of the collection. The special register that these letters use could possibly explain the creation of several clusters within a similar time period. More in-depth analysis is needed, possibly in cooperation with philologists, to understand the exact nature of the differences the clusters are capturing, particularly whether they reflect other textual characteristics than the existence of language variants due to temporal change.

4 Conclusion

In this work we have proposed a methodology for the identification of temporal trends in a document collection. To this end, we relied on perplexities derived from recurrent neural network language models and K-Means clustering.

The perplexities calculated by document-specific language models correlate moderately with time differences. Performing K-Means with `K=7` based on perplexity measures proved to be a good method for grouping documents based on intrinsic evaluation (inter- and intra-cluster distance). The method allowed us to discover groups that seem at least partially to reflect differences due to language change not only in the sense of radical change in language (from Latin to Danish), but also changes within the same language (Latin).

The remaining question is whether the clusters found can be more deeply characterised. They seem to be somewhat temporally distributed which, however, could partly be explained by the nature of the dataset. Thus, future work involves investigating how other factors could represent temporal trends in the data. This could be done by evaluating how congruent the clusters are with documented trends within the dataset, for example trends that could be caused by the existence of specific types of text such as the group of papal letters.

Another interesting problem is to see how such clusters can be used in relation to the task of temporal document classification (extrinsic evaluation). Using the temporal distribution of the individual clusters might provide a more nuanced picture of temporal trends compared to discrete bins, thus providing better results when used in a classification task.

References

Hidenao Abe and Shusaku Tsumoto. 2010. Text categorization with considering temporal patterns of term usages. In *Proceedings of the 2010 IEEE International Conference on Data Mining Workshops*, ICDMW '10, pages 800–807, Washington, DC, USA. IEEE Computer Society.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Sidsel Boldsen and Patrizia Paggio. 2019. Automatic dating of medieval charters from Denmark. In *Proceedings of the 4th Digital Humanities in the Nordic Countries Conference*.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Pablo Gamallo, Inaki Alegria, José Ramom Pichel, and Manex Agirrezabal. 2016. Comparing two basic methods for discriminating between similar languages and varieties. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 170–177.

Pablo Gamallo, Jose Ramom Pichel, and Inaki Alegria. 2017. A perplexity-based method for similar languages discrimination. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 109–114.

Anne Garcia-Fernandez, Anne-Laure Ligozat, Marco Dinarelli, and Delphine Bernhard. 2011. When was it written? Automatically determining publication dates. In *String Processing and Information Retrieval*, pages 221–236, Berlin, Heidelberg. Springer Berlin Heidelberg.

Meritxell González Bermúdez. 2015. An analysis of twitter corpora and the differences between formal and colloquial tweets. In *Proceedings of the Tweet Translation Workshop 2015*, pages 1–7. CEUR-WS. org.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605.

José Ramom Pichel Campos, Pablo Gamallo, and Iñaki Alegria. 2018. Measuring language distance among historical varieties using perplexity. Application to European Portuguese. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 145–155. Association for Computational Linguistics.

Octavian Popescu and Carlo Strapparava. 2015. Semeval 2015, task 7: Diachronic text evaluation. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 870–878.

Sanja Štajner and Marcos Zampieri. 2013. Stylistic changes for temporal text classification. In *Text, Speech, and Dialogue*, pages 519–526, Berlin, Heidelberg. Springer Berlin Heidelberg.

Andreas Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

Terrence Szymanski and Gerard Lynch. 2015. Ucd: Diachronic text classification with character, word,

and syntactic n-grams. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 879–883.

Marcos Zampieri, Shervin Malmasi, and Mark Dras. 2016. Modeling language change in historical corpora: The case of Portuguese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4098–4104, Paris, France. European Language Resources Association (ELRA).