Automated Compositional Change Detection in Saxo Grammaticus' *Gesta Danorum*

Kristoffer L. Nielbo[1,2], Mads L. Perner[3], Christian Larsen[3], Jonas Nielsen[4] and Ditte Laursen[5]

[1]Department of History, University of Southern Denmark, Odense, Denmark
[2]SDU eScience Center, University of Southern Denmark, Denmark
[3]Department of History, University of Copenhagen, Copenhagen, Denmark
[4]Society for Danish Language and Literature, Copenhagen, Denmark
[5]Royal Danish Library, Copenhagen, Denmark

Submitted: February 2nd, 2019

Abstract

Saxo Grammaticus' medieval source Gesta Danorum *("Deeds of the Danes") represents the beginning of the modern historical research in Denmark. The bipartite composition of* Gesta Danorum *has however been subject to much academic debate. In particular the nature and location of a transition between early Pre-Christian and late Christian content have given rise to two competing accounts. In this paper, we argue that the debate can be represented as a problem of intratextual dynamics and we combine models from Information Retrieval and Natural Language Processing with techniques for time series analysis in order to reevaluate the debate. Results indicate that the transition is gradual, starting in book eight and ending in book ten, but that a point-like interpretation is possible in book nine. We argue that the approach exemplifies scalable"automated close reading", which has multiple applications in text-based historical research.*

Keywords— Change Detection, Cultural Heritage, Medieval Literature, Text Analysis, Topic Modeling

Introduction

The medieval writer Saxo Grammaticus (c. 1160 - post 1208) represents the beginning of the modern day historian in Scandinavia. Saxo's history of the Danes *Gesta Danorum* ("Deeds of the Danes") is the single most important written source to Danish history in the 12th century. *Gesta Danorum* is tendentious, contains elements of fiction, and its composition has been an academic subject of debate for more than a century. The recent debate concerns the bipartite composition of

Gesta Danorum and centers on two related issues: 1) is the transition between the old mythical and new historical part located in book eight, nine, or ten; and 2) is this transition gradual (continuous) or sudden (point-like)? Most scholars have used qualitative observations and contextual knowledge to argue for a particular change in content and composition [1] . In this paper, we combine automated techniques from Natural Language Processing (NLP) and Information Retrieval (IR) with time series analysis in order to reassess *Gesta Danorum*'s composition debate.

Gesta Danorum (GD henceforth) was written by Saxo at the request of Archbishop Absalon (1128-1201), who was a highly influential political figure in medieval Scandinavia. GD covers the history of Denmark from a pre-Christian Norse beginning to a Christian kingdom in the end of the 12^{th} century (i.e., Saxo's present). GD is divided into sixteen books and a preamble. There are several competing accounts of GD's composition. Two accounts in particular that focus on the role of religion dominate the debate. According to the *traditional account*, whose primary objective has been to sort historical fact from fiction, books one to eight are based on Pre-Christian popular legends, books nine to twelve are a mixture of legendary tales and paraphrased written sources, and books thirteen to sixteen are eye-witness account from sources contemporary to Saxo (Skovgaard-Petersen 1969). On this account, it is claimed that books one to nine are mostly concerned with Norse mythology, while the content shifts to Christianity in books ten and onwards with the baptism of king Harold Bluetooth (Henriksen, Hansen, Maegaard, Pedersen, and Povlsen 2014). The transition to Christianity in book ten is, on this account, considered sudden and point-like.

In 1969 an *alternative account* was introduced, which grouped the books in phases of gradual consolidation of the Danish Christian Kingdom. On this account books one to four represent the time before the birth of Christ; books five to eight, the time until the arrival of Christianity in Denmark; books nine to twelve, the gradual acceptance of the new religion; and books thirteen to sixteen, the establishment of the Danish archbishopric in Lund (Skovgaard-Petersen 1969). This fourfold composition pattern matches GD's development of moral themes (Johannesson 1978). The alternative account emphasizes two issues related to the transition to Christianity in GD: 1) the process is gradual and continuous; and 2) the process starts earlier than previously thought in book eight. At the end of book eight Charlemagne finishes his efforts to convert the northernmost parts of Germany to Christianity and thus, from the beginning of book nine, all is set for the Christianity to disseminate towards the north across the Danish border. This alternative account of GD's composition has however not won over the majority of academics, who still favor the traditional account (Kværndrup 1999).

From a methodological perspective, the composition debate is fundamentally a problem of intratextual dynamics, that is, of how the behavior of a system of words and sentences (i.e., lexical content of a GD) can be described as a function of time (i.e, its relative position in GD). The challenge is to identify a formalism that supports sufficiently simple text representations, in order to extract a content-based signal from a small collection of related documents. We can then apply

[1] For an exception see (Henriksen, Hansen, Maegaard, Pedersen, and Povlsen 2014) who uses summary statistics based entity extraction.

change detection techniques to the signal in order to identify possible content-based transitions.

Areas that research automated indexing and analysis of large text collections such as Natural Language Processing (NLP), Information Retrieval (IR), and Digital Humanities (DH), offer a range of techniques for capturing textual dynamics. Story arcs that use lexicon-based sentiment analysis represent the dynamics of a story plot as a function of its affective valence (Reagan, Mitchell, Kiley, Danforth, and Dodds 2016). A story arc is essentially a windowed sentiment analysis of a long text, the behavior of which can then be analyzed using methods from signal processing (Jockers 2017), nonlinear adaptive filtering (Gao, Jockers, Laudun, and Tangherlini 2017), and fractal scaling (Hu, Liu, Thomsen, Gao, and Nielbo in review). The accuracy of lexicon-based sentiment analysis is however dependent on appropriate dictionaries for, in our case, the archaic Danish in GD and story arcs are best suited for novels with a coherent narrative (Hu, Liu, Thomsen, Gao, and Nielbo in review).

An alternative technique is to study lexical variation over time in a more or less homogeneous collection of texts (Zhang 2016). It has been shown that Type-Token Ratio carries information about the author's mental condition (Berisha, Wang, LaCross, and Liss 2015; Garrard, Maloney, Hodges, and Patterson 2005; Lancashire and Hirst 2009; Snowdon D.A., Kemper S.J., Mortimer J.A., Greiner L.H., Wekstein D.R., and Markesbery W.R. 1996) and is linked to text creativity (Hu and Yu 2011; Zhu, Xu, and Khot 2009). Similar measures from information theory, specifically entropy, has been used to reconstruct mental change points from the collected writings of single authors (Barron, Huang, Spang, and DeDeo 2017; Murdock, Allen, and DeDeo 2015; Nielbo, Baunvig, Liu, and Gao in press). While these measures are relevant to the mental dynamics underlying Saxo's production of GD, measures of lexical variation are too far removed from the content in order to describe variation in GD's composition.

NLP and IR offer several techniques for simplifying the representation of a large text collection, which have dynamic interpretations. The classical vector space model, for instance, describes a document as a high-rank vector that represents the document as a distribution over a text collection's lexicon. The advantage of this model is that it allows for vector manipulations of documents, such that document similarity can be expressed as the (cosine to the) angle between two document vectors. Dynamically, we can then convert a text collection into a signal of document similarities based on the documents' relative position (e.g., for each sentence, how similar is its vector representation on average from the antecedent sentence vectors). While promising, the classical vector space model results in sparse high-rank vectors. As the number of documents increase in the text collection, so does the lexicon (Heaps 1978) resulting in sparse document representations and, in many cases, searching and partitioning the vector space becomes difficult, a phenomenon known as the curse of dimensionality (Bellman 1961). Partly as a solution to this problem, latent variable models were introduced in IR (Deerwester, Dumais, Furnas, Landauer, and Harshman 1990). Topic models, in particular models based on Latent Dirichlet Allocation (LDA), are probabilistic latent variable models that have become very influential in NLP and DH (Blei and Lafferty 2006). LDA describes a document as a low-rank dense vector that represents a discrete probability distribution over a small set of latent variables called topics. Each topic represents a content-related facet of the text

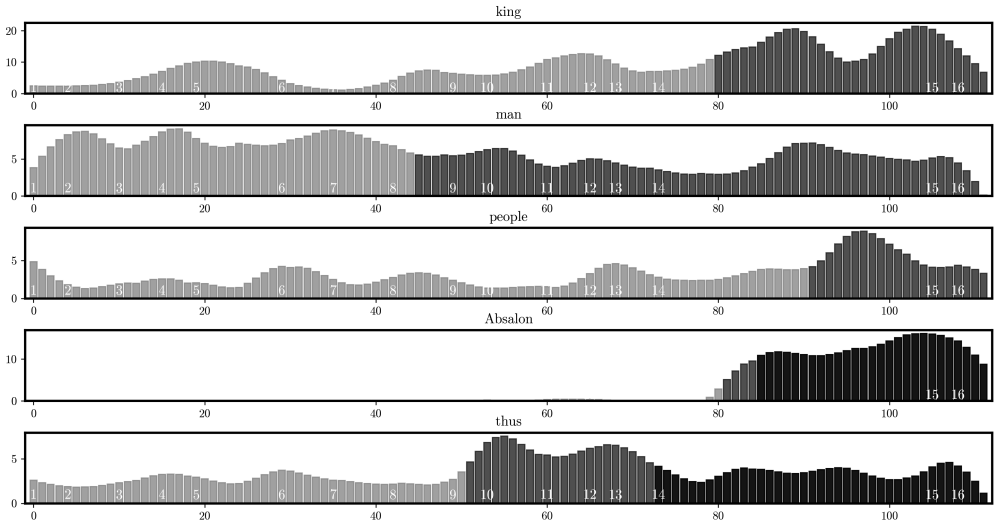


Fig. 1: Frequent keywords from GD in text slices of 50 sentences. Coloring indicates statistically significant change points.

collection as a power-law-like distribution over the full lexicon. In order to generate a content-based signal, we can compute the relative entropy between consecutive documents based on their position (e.g., for each paragraph, how much does its topic distribution differ from previous paragraphs topic distributions) (Murdock, Allen, and DeDeo 2015).

For the purpose of reassessing GD's composition, we propose to use a signal based on the cosine distance between text slices in a classical vector space model of GD as baseline (see Appendix A, equation 1 and 3). The baseline will be contrasted with a signal derived from relative entropy of text slices in a topic model of GD (see Appendix A, equation 2 and 3). Two simple change detection techniques, a mean- and variance-shift technique respectively, are then used to detect statistically reliable compositional changes in the baseline and contrast models respectively. It is important to point out that topic modeling has a full dynamic implementation in dynamic LDA (Blei and Lafferty 2006) and that embedding techniques also offer similar dynamic implementations in, for instance, dynamic Bernoulli embeddings (Rudolph and Blei 2017). Both techniques do however require substantial amounts of data and do not allow for direct comparison to our baseline. Both classical vector spaces and LDA do generate relatively simple document representations, offer a simple dynamic interpretation based on distance metrics, and their signal behaviors are comparable.

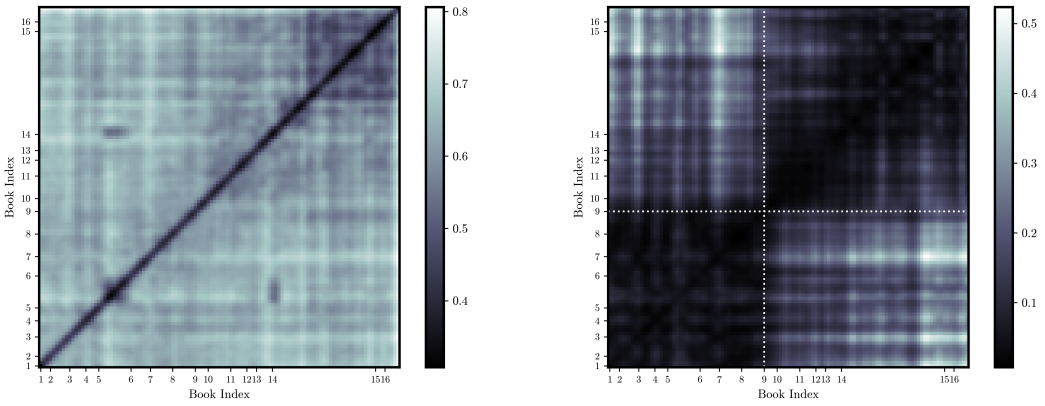


Fig. 2: Distance matrices for GD text slices based on cosine distance for the baseline model (left) and relative entropy for the contrast model (right). The intersecting white dotted lines indicates an observable transition.

Results

Change detection in GD's composition can be illustrated by the development of lexical items (Fig. 1). In this case we inspect change in the distribution of frequent content words (e.g., *king*, *man*, *people*) and stylistic markers (e.g., *thus*) in GD. At a more advanced level, we can assess changes in the occurrence of entities such as persons (e.g., *Absalon*) or institutions. In order to detect a change in the mean of the word frequency signal, we apply a mean-shift technique. The technique detects change points in all five words as indicated by shaded regions in Fig. 1. For *Absalon*, which represents both a word and an entity, the model correctly detects a shift in book 14, which corresponds to the location at which *Absalon* becomes an actor in GD. Notice that with the exception of *man*, all words in Fig. 1 indicate a reliable lexical change in book fourteen. While detection of change in individual lexical items is useful, it assumes an atomistic understanding of semantic dynamics, because change is tied to individual words and not their context. Saxo could have prepared the reader to change associated with *Absalon* before the Archbishop's entry point in book fourteen through the thematic structure. The atomistic assumption can be loosened by having domain experts aggregate multiple domain relevant items into a concept (e.g., $Christian\ Language = Fr(Christ) + Fr(God) + ...$) (Henriksen, Hansen, Maegaard, Pedersen, and Povlsen 2014). Reliance on domain experts that a priori select and prioritize relevant items does however run the risk of introducing a selection bias by disregarding behavior of substantial parts of the content.

The classical vector space model (baseline) and topic model (contrast) avoid theory-driven selection biases by modeling each document as a distribution over the full lexicon or the full set of latent topics. The change signals in these models are therefore estimated from the full lexical content of each document. Fig. 2 shows the distance matrices based on document slices for the baseline (left) and contrast (right) model respectively with the identity line on the antidiagonal. For the baseline, neighboring slices are similar as indicated by the small rectangular structures

around the antidiagonal. There is no clear indication of a global change pattern, because most off-diagonal points are similarly distanced from each other. The baseline matrix does show a slight negative trend, which indicates that change attenuates over time (darkness increases along the antidiagonal). Starting in book fourteen, a rectangular dark structure emerges in the upper right corner, which indicates a more similar content structure in books fourteen to sixteen of GD. This finding conforms to the analysis of lexical items in Fig. 1. A global change pattern is, on the other hand, discernible from the contrast, where two large rectangular structures at each side of book nine, one dark rectangle in the lower left corner (books one to nine) and one rectangle in the upper right corner (books nine to sixteen). This is a typical bi-modal change pattern that splits a process into two parts which are internally similar and consistent. The pattern gives a clear indication that GD's topic space is partitioned in two parts around book nine. Importantly, notice that the process is gradual as indicated by the blurred transition along the antidiagonal between the lower and upper rectangle. At closer inspection, the gradual transition starts in book eight and continues until the beginning of book ten. Within the upper rectangle (books nine to sixteen), a secondarily change is further indicated by the two embedded rectangles. This pattern is similar to the baseline's separation of the last three books, but starts earlier in book fourteen.

Turning to the change signal as estimated by cosine distance and relative entropy, respectively, the observations from our distance matrices are corroborated. While the baseline shows little indication of change (Fig. 3.a), except for a dip in book fourteen, the contrast exhibits a tilted S-like shape as a rapid change initiates in book eight, centers in book nine, and stabilizes in book ten (Fig. 3.b). A simple way to analytically confirm a potential change point, is to compute the derivative of the signals. For the baseline the maximal change in the signal is located in book five, but its magnitude is rivaled by multiple other locations, which makes it hard to decide on the relevant point. The derivative of the contrast model supports the argument that maximal signal change is located in book nine. Derivatives will identify signal change irrespective of the robustness or statistical reliability of the change point. In order to confirm that the change is reliable, we need to use a statistical technique that estimates the significance of a change in the signal's central tendency or variance. For this purpose we apply a mean and variance shift technique to the baseline and contrast. There is no indication of change in the overall mean of the signal for the baseline model, while the LDA-based contrast shows a robust change point in book nine. This change is also reflected in the contrast's variance with an addition of a change in book fourteen. The variance shift model for the baseline is complicated indicating multiple changes in books seven, nine, eleven and fourteen. This resembles the pattern indicated by the derivative, where multiple points are competing due to an almost constant level of change combined with small oscillations. At a higher level of abstraction, we model each signal with a no-change/constant change linear model and compare it with a sigmoid model that captures a single dominant change point (Fig. 3.c). For the baseline model a no-change model explains more variance $R^2 = 0.52$ than the sigmoid model $R^2 = 0.02$. The pattern is reversed for the contrast model, where the sigmoid model explains more variance $R^2 = 0.93$ in comparison with the linear model $R^2 = 0.86$.

To summarize the findings, we find unequivocal support for a) a change point in the LDA-based

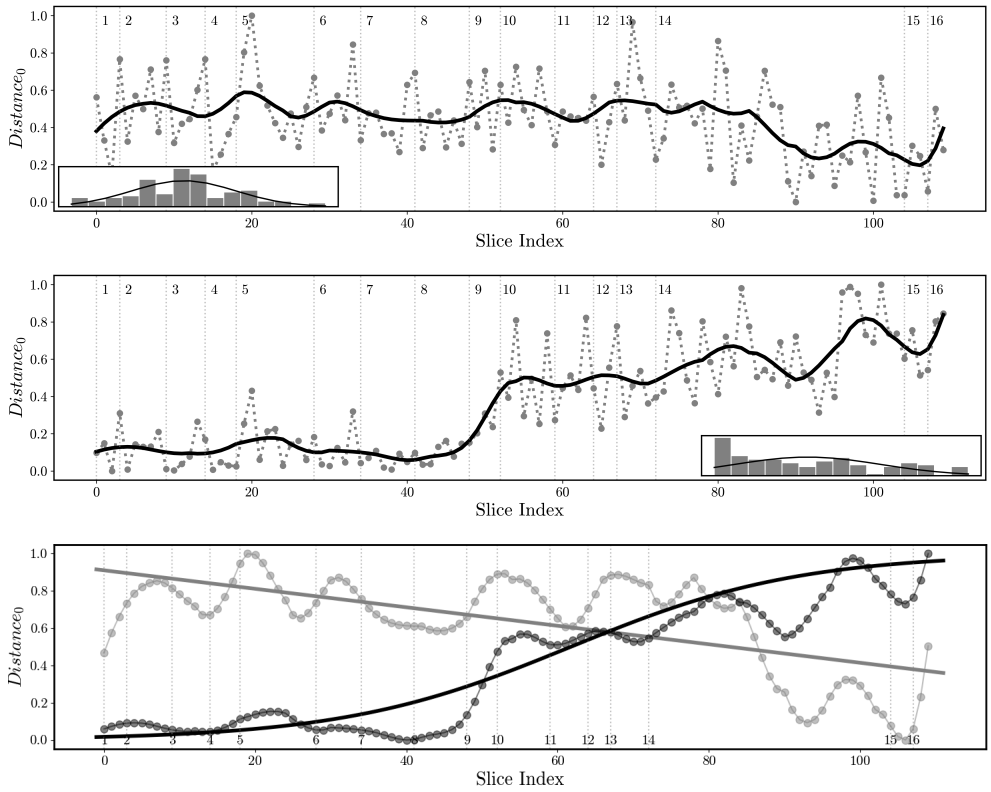


Fig. 3: Signals for the baseline (upper) and contrast (middle) model based on the slice distance from the beginning ($Distance_0$), see Appendix A. Smoothing (black lines) are generated with an adaptive nonlinear filter. Lower figure compares the fits to the two signals (gray for baseline and black for contrast) that explain the most variance.

contrast model in comparison to the vector space baseline model, b) a change point that gradually evolves from the end of book eight to the beginning of book ten, but with the point estimate located in book nine, and c) there are some indications of a secondary change point located in book fourteen.

Concluding Remarks

This study has combined techniques from NLP and IR with time series analysis in order to revisit a longstanding academic debate regarding the bipartite composition of an influential historical source. Two competing positions, the traditional and alternative account respectively, were identified in the debate. The traditional account argues for a point-like transition located in book nine, while the alternative account argues for a gradual transition that starts in book eight. In order to decide between the two accounts, we introduced a baseline model based on a classical IR technique and a contrast model based on a more recent NLP technique. The baseline model was not able to reliably detect any change point in GD, although there were some indications of a change in book fourteen. Analysis of the contrast model's behavior, on the other hand, shows clear indications of a gradual transition that starts in the latter part of book eight and ends in book ten. However, if we assume that the transition is point-like in accordance with the traditional account, the contrast model exhibits the greatest rate of change in book nine. As such, the contrast models behavior can explain central claims from the traditional as well as the alternative account. We do however favor the alternative account of GD's bipartite composition, because it is more sensitive to the texts' detailed dynamic behavior. Regarding the change point in book fourteen, which both models detect, it is important to notice that while it is the second book that deals with Saxo's contemporaries, it introduces Archbishop Absalon who requested GD.

On the methodological side, a comment on why the probabilistic formalism underlying LDA was better suited for detecting a bipartite structure than the geometric formalism underlying the vector space, is necessary. The vector space only takes the lexical representation of each document into account disregarding relational information between words that appear in similar contexts. The baseline distance measure therefore favors text slices that are strictly similar, while relative entropy in the contrast models is sensitive to text slices that are relational similar. This is reflected in the distance matrix visualization in figure 2, where only neighboring slices share similarity in the vector space. This type of similarity is to be expected according to the burstiness of words (i.e., the probability of a word re-occurring after it has appeared once decays in a power-law like fashion). LDA, and several other kinds of topic models, offers a simple technique for clustering documents on a set of hidden topics that encode the co-occurrence structure of the entire data set. We believe that the technique has great potential for solving problems in historical research that require comparison and grouping of source texts.[2]. We are therefore more inclined to value a topic model's capacity to

[2]Topic models are not necessarily confined to lexical features, but can also be utilized for grouping objects on, for instance, visual features or auditory features.

downsize sparse high-rank vector representations in a probabilistic framework, than to use it for exploring the topicality structure that runs through a collection of documents.

Related to the value of topic modeling in historical research, it is worth mentioning the approach to automated text analysis that is indirectly advanced in this study. While DH in particular has become the proponent of so-called distant reading, that is, large-scale computer-assisted text analysis, this study has tried to develop an "automated close reading" which uses a small collection of more or less homogeneous texts written by one author. This approach is more akin to qualitative text analysis as it is practiced in the humanities and social sciences, but it utilizes the transparency and formal rigor of mathematics and computation. The approach also scales, that is, with the easy access to fast computing facilities and digitized text collections, it can easily be turned into a large-scale project. The opposite is however not the case, distant reading cannot necessarily contribute substantial information to close reading.

Finally it is important to point out that the reconstruction of system dynamics from a combination of topic models and distance metrics is applicable beyond a single author. Similar methods have already been successfully applied to study the topical evolution of more temporally dispersed document from a single author (Murdock, Allen, and DeDeo 2015) as well as larger cultural systems (Barron, Huang, Spang, and DeDeo 2017). With some minor modifications such methods are ideal for novelty and trend detection in newspapers and social media (Wevers, Nielbo, and Gao in review). We are therefore optimistic in terms of these future applications of these methods in various humanities and social science domains that study the behavior of text collections.

Acknowledgements

The authors want to thank *Arkiv for Dansk Litteratur* (*Archive for Danish Literature*) for access to data and HUMlab's Datasprint initiative at Copenhagen University Library for the opportunity to collaborate.

Appendix A: Methods

Data

The data set consisted of the full sixteen books of *Saxos Danmarkshistorie* translated from Latin by Peter Zeeberg and published by Det Danske Sprog- og Litteraturselskab and G.E.C.Gads Forlag in 2000. The books were concatenated into one string and sliced in non-overlapping windows at a size of 50 sentences. The window size was selected in order to have sufficient words for topic modeling while retaining sensitivity for minor topical structure. Before model training, the document slices were lowercased, numerals were removed, and the sentences tokenized at the word-level. A set of data-specific frequent words that were not directly content relevant (e.g., Saxo and Danmarkshistorie) were removed in order to increase the models' discriminatory power.

Document Representations

The classical vector space model (i.e., the baseline model) consisted of a document-term co-occurrence matrix that represents each document as a sparse vector of word features. The model was generated at the level of unigrams only and therefore relies on a bag-of-words assumption, which ignores any syntactical information that cannot be identified at the level of individual words. This assumption was retained in the contrast model, which was a simple topic model trained over unigrams using Latent Dirichlet Allocation (LDA). LDA, and topic models more generally, provide a handle in documents' lexical semantics based on their n-gram co-occurrence structure. LDA is a three-level hierarchical Bayesian model that is typically used to identify a topicality structure that runs through a large collection of documents (Blei, Ng, and Jordan 2003). LDA can also be viewed as a probabilistic approach to downsizing the sparse high dimensional vectors of a document-term co-occurrence matrix to dense low-rank vectors.

Document Distance and Change Detection

In order to test for content change in the sequential structure of GD, we utilized a three step procedure on the models. First, the distance D between every combination of two document slices s_1 and s_2 was computed for the baseline model using cosine distance D_C (Manning, Raghavan, and Schütze 2008):

$$D_C(s_1, s_2) = \frac{s_1 \cdot s_2}{\| s_1 \| \| s_2 \|} \tag{1}$$

where the numerator is the dot product of slice vectors s_1 and s_2 and the denominator is product of their lengths. For the contrast model, relative entropy D_{KL} substituted the distance measure[3]

[3]Relative entropy, or Kullback-Leibler divergence, is asymmetrical such that $D_{KL}(s_1 \mid s_2) \neq D_{KL}(s_2 \mid s_1)$ and therefore not strictly speaking a measure of distance, but difference between two probability distributions.

(Murdock, Allen, and DeDeo 2015):

$$D_{KL}(s_1 \mid s_2) = \sum_{i=1}^{n} s_{i1} \times \log_2 \frac{s_{i1}}{s_{i2}} \tag{2}$$

A slice representation in the topic model is a discrete probability distribution over a set of topics and relative entropy measures the difference between slices. Second, a semantic change signal Δ_D was estimated for each model by averaging over the distances from slice s^j the preceding slices from $s^1 \ldots s^{j-1}$:

$$\Delta_D(s_j) = \frac{1}{N} \sum_{i=1}^{j-1} D(s_j, s_i) \tag{3}$$

where D is determined by the model. A more general information theoretical version of this change measure is referred to as novelty $\mathcal{N}$ in the literature (Barron, Huang, Spang, and DeDeo 2017). Third, two simple change detection techniques, a mean- and variance-shift technique, were applied to each signal in order to identify statistically reliable change points in their respective mean and variance at an α-level of .01 (Kulkarni, Al-Rfou, Perozzi, and Skiena 2015; Tyler 2000).

References

Barron, Alexander T J, Jenny Huang, Rebecca L Spang, and Simon DeDeo. 2017. "Individuals, Institutions, and Innovation in the Debates of the French Revolution." *arXiv:1710.06867* p. 8.

Bellman, Richard E. 1961. *Adaptive Control Processes: A Guided Tour*. Princeton University Press.

Berisha, Visar, Shuai Wang, Amy LaCross, and Julie Liss. 2015. "Tracking Discourse Complexity Preceding Alzheimer's Disease Diagnosis: A Case Study Comparing the Press Conferences of Presidents Ronald Reagan and George Herbert Walker Bush." *Journal of Alzheimer's Disease* 45:959–963.

Blei, David M. and John D. Lafferty. 2006. "Dynamic topic models." In *Proceedings of the 23rd international conference on Machine learning*, pp. 113–120. ACM.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent dirichlet allocation." *the Journal of machine Learning research* 3:993–1022.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. "Indexing by latent semantic analysis." *Journal of the American society for information science* 41:391.

Gao, Jianbo, Matthew Jockers, John Laudun, and Timothy Tangherlini. 2017. "A multiscale theory for the dynamical evolution of sentiments in novels."

Garrard, Peter, Lisa M. Maloney, John R. Hodges, and Karalyn Patterson. 2005. "The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author." *Brain* 128:250–260.

Heaps, Harold S. 1978. *Information Retrieval, Computational and Theoretical Aspects*. New York: Academic Pr.

Henriksen, L., D.H. Hansen, B. Maegaard, B.S. Pedersen, and P. Povlsen. 2014. "Encompassing a spectrum of LT users in the CLARIN-DK Infrastructure." In *LREC Proceedings*, pp. 2176–2181.

Hu, Qiyue, Bin Liu, Mads Rosendahl Thomsen, Jianbo Gao, and Kristoffer L Nielbo. in review. "Dynamic evolution of sentiments in Never Let Me Go: Insights from multifractal theory and its implications for literary analysis."

Hu, Xiao and Bei Yu. 2011. "Exploring The Relationship Between Mood and Creativity in Rock Lyrics." In *ISMIR*, pp. 789–794.

Jockers, Matthew. 2017. "Package 'syuzhet'." .

Johannesson, Kurt. 1978. *Kurt Johannesson: Saxo Grammaticus. Komposition och världsbild i Gesta Danorum*. Number 31 in Lychnos bibliotek. Stockholm: Almquist & Wiksell International.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. "Statistically Significant Detection of Linguistic Change." pp. 625–635. ACM Press.

Kværndrup, S. 1999. *Tolv principper hos Saxo. En tolkning af Danernes Bedrifter.* Denmark: Multivers Aps Forlag,.

Lancashire, Ian and Graeme Hirst. 2009. "Vocabulary Changes in Agatha Christie's Mysteries as an Indication of Dementia: A Case Study." Toronto.

Manning, Christopher, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. New York: Cambridge University Press, 1 edition edition.

Murdock, Jaimie, Colin Allen, and Simon DeDeo. 2015. "Exploration and Exploitation of Victorian Science in Darwin's Reading Notebooks." *arXiv preprint arXiv:1509.07175* .

Nielbo, Kristoffer L., Katrine F. Baunvig, Bin Liu, and Jianbo Gao. in press. "A Curious Case of Entropic Decay: Persistent Complexity in Textual Cultural Heritage." *Digital Scholarship in the Humanities* pp. 1–17.

Reagan, Andrew J., Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds. 2016. "The emotional arcs of stories are dominated by six basic shapes." *arXiv preprint arXiv:1606.07772* .

Rudolph, Maja and David Blei. 2017. "Dynamic Bernoulli Embeddings for Language Evolution." pp. 1–11.

Skovgaard-Petersen, I. 1969. *Saxo, Historian of the Patria*, volume Medieval Scandinavia. Denmark.

Snowdon D.A., Kemper S.J., Mortimer J.A., Greiner L.H., Wekstein D.R., and Markesbery W.R. 1996. "Linguistic ability in early life and cognitive function and alzheimer's disease in late life: Findings from the nun study." *JAMA* 275:528–532.

Tyler, W. A. 2000. "Change-point analysis: A powerful new tool for detecting changes."

Wevers, Melvin, Kristoffer L. Nielbo, and Jianbo Gao. in review. "Tracking the Consumption Junction: Temporal Dependencies in Dutch Newspaper Articles and Advertisements."

Zhang, Yanhui. 2016. "Entropic evolution of lexical richness of homogeneous texts over time: A dynamic complexity perspective." *Journal of Language Modelling* 3:569–599.

Zhu, Xiaojin, Zhiting Xu, and Tushar Khot. 2009. "How creative is your writing? a linguistic creativity measure from computer science and cognitive psychology perspectives." In *Proceedings of the workshop on computational approaches to linguistic creativity*, pp. 87–93. Association for Computational Linguistics.