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

K.L. Nielbo, M.L. Perner, C. Larsen, J. Nielsen and D.Laursen

`kln@cas.dk`
`knielbo.github.io`

Center for Humanities Computing
Aarhus University, Denmark

March 7, 2019

# Outline

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

K.L. Nielbo, M.L.
Perner, C. Larsen, J.
Nielsen and D.Laursen
kln@cas.dk
knielbo.github.io

Introduction
Saxo Grammatricus

Methods
Data
Vector space
Seeded LDA
Signal generation

Results
Keyword change
points
Distance matrices

Discussion

# Saxo's *Gesta Danorum*

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

K.L. Nielbo, M.L.
Perner, C. Larsen, J.
Nielsen and D.Laursen
kln@cas.dk
knielbo.github.io

**Saxo Grammaticus**

- The medieval writer (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 12$^{th}$ century.

- *Gesta Danorum* is tendentious, contains elements of fiction, and its compositions has been an academic subject of debate for more than a century.

**Composition debate**

- Debate regarding the bipartite composition *Gesta Danorum*

  1. is the transition between the old mythical and new historical part located in book eight, nine, or ten?
  2. is this transition gradual (continuous) or sudden (point-like)?

- combine NLP and IR with time series analysis in order to propose a solution

# Data

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

K.L. Nielbo, M.L. Perner, C. Larsen, J. Nielsen and D.Laursen
kln@cas.dk
knielbo.github.io

Introduction
Saxo Grammatricus
Methods
Data
Vector space
Seeded LDA
Signal generation
Results
Keyword change points
Distance matrices
Discussion

**Data set**

- all 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.

**Normalization**

- books were concatenated and sliced in non-overlapping windows at a size of 50 sentences

- unigrams were casefolded and numerals removed

- data-specific frequent words were removed
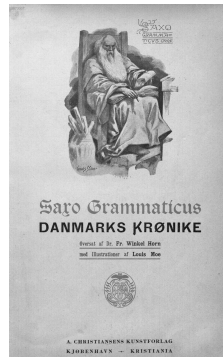
# Naive baseline model

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

K.L. Nielbo, M.L. Perner, C. Larsen, J. Nielsen and D.Laursen
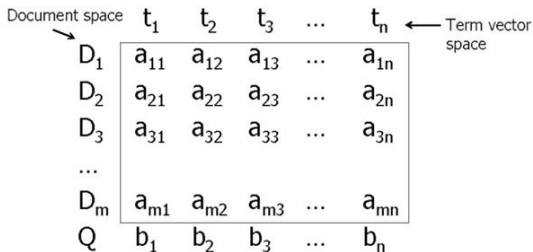kln@cas.dk
knielbo.github.io

Figure 1: Geometrical document representation, where each document is a high rank word vector over the full vocabulary.

# Alternative model

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

K.L. Nielbo, M.L.
Perner, C. Larsen, J.
Nielsen and D.Laursen
kln@cas.dk
knielbo.github.io

**Algorithm 1** Classical LDA
1: **for** *each $k = 1...T$* **do**
2:    *choose $\phi_k \sim Dir(\beta)$*
3:    **for** *each $d$ choose $\theta_d \sim Dir(\alpha)$* **do**
4:      **for** *each token $i = 1...N_d$* **do**
5:       *select a $z_i \sim Mult(\theta_d)$*
6:       *select a $w_i \sim Mult(\phi_{x_i})$*
7:      **end for**
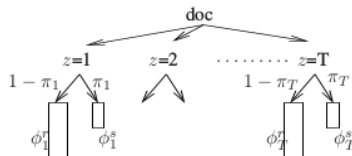8:    **end for**
9: **end for**



Figure 2: In LDA each document is represented as a low rank dense vector (i.e., a probability distribution over a small set of latent topics). Seeds improve both topic-word distributions and to improve document-topic distributions

Jagarlamudi, J., Iii, H. D., & Udupa, R. (2012). Incorporating Lexical Priors into Topic Models. Proceeding EACL 12, 204–213

# Document distance and change detection

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

K.L. Nielbo, M.L. Perner, C. Larsen, J. Nielsen and D.Laursen
kln@cas.dk
knielbo.github.io

1. Distance $D$ between every combination of two document slices $s_1$ and $s_2$ is computed for the baseline model using cosine distance $D_C$:

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

and the alternative model using relative entropy $D_{KL}$:

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

2. A semantic change signal $\Delta_D$ was estimated for each model by averaging over the distances from slice $s^j$ the preceding slices from $s^1 \dots s^{j-1}$:

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

3. Two 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 $\alpha$-level of .01.

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

K.L. Nielbo, M.L.
Perner, C. Larsen, J.
Nielsen and D.Laursen
kln@cas.dk
knielbo.github.io

Introduction
Saxo Grammatricus

Methods
Data
Vector space
Seeded LDA
Signal generation

Results

Keyword change
points
Distance matrices

Discussion

Figure 3: Keyword/entity counts with a mean-shift model. Notice that Archbishop Absalon is introduced in book 14.

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

K.L. Nielbo, M.L.
Perner, C. Larsen, J.
Nielsen and D.Laursen
kln@cas.dk
knielbo.github.io

Introduction
Saxo Grammaticus

Methods
Data
Vector space
Seeded LDA
Signal generation

Results
Keyword change
points
Distance matrices

Discussion
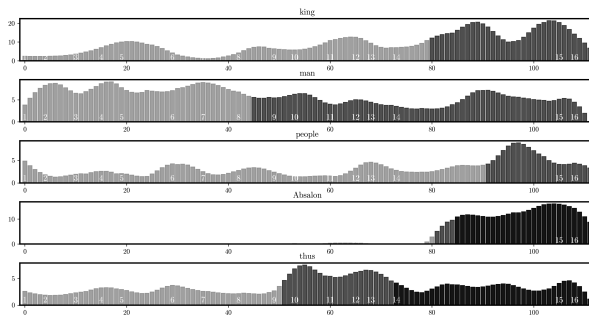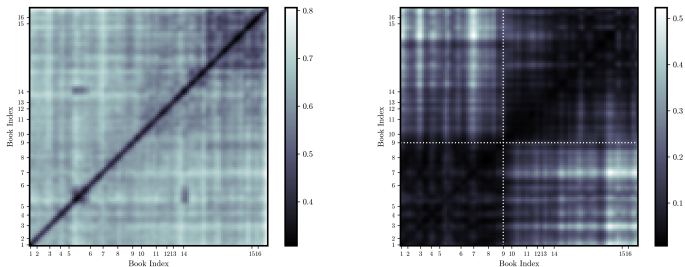
Figure 4: Distance matrices for baseline (left) and alternative (right) models. Left pattern can be explained by the burstiness of language, while the right pattern indicates a bipartite structure. Notice books rectangle from book 14-16 on the left.

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

K.L. Nielbo, M.L.
Perner, C. Larsen, J.
Nielsen and D.Laursen
kln@cas.dk
knielbo.github.io

Introduction
  Saxo Grammaticus
Methods
  Data
  Vector space
  Seeded LDA
  Signal generation
Results
  Keyword change
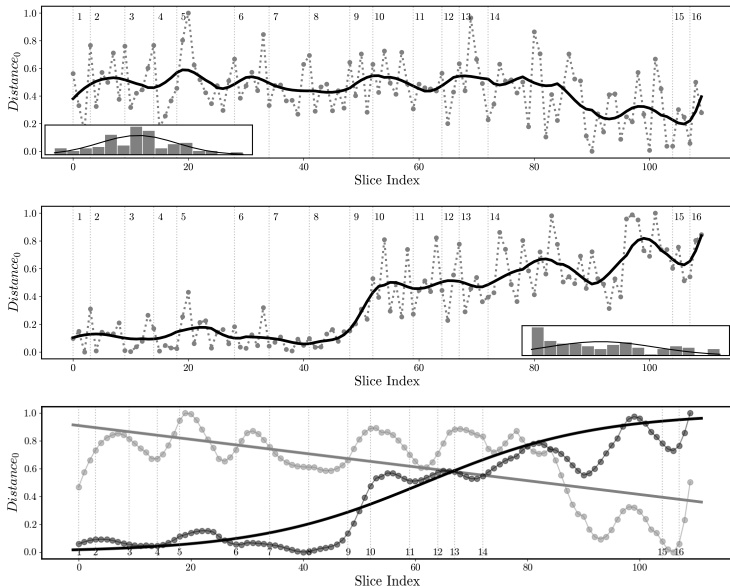  points
  Distance matrices
Discussion

Figure 5: 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 $R^2 = 0.86$

# Summary

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

K.L. Nielbo, M.L.
Perner, C. Larsen, J.
Nielsen and D.Laursen
kln@cas.dk
knielbo.github.io

**Findings**

- Baseline shows no reliable change

- Alternative show gradual change starting in book 8 (latter part) and ends in book 10

- Greatest rate of change in book 9

- Both models indicate change in book 14

**Interpretation**

- strongest support for the continuous transition claim

- although the book 14 is the second book dealing with Saxo's contemporaries, it introduces Archbishop Absalon

- baseline favors text slices that are strictly similar, while the alternative is sensitive to relational similar slices

- LDA gives us a simple technique for clustering documents on a set of hidden topics, which when combined with time series analysis offers great potential for historical research

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

K.L. Nielbo, M.L.
Perner, C. Larsen, J.
Nielsen and D.Laursen
kln@cas.dk
knielbo.github.io

Introduction
Saxo Grammatricus
Methods
Data
Vector space
Seeded LDA
Signal generation
Results
Keyword change
points
Distance matrices

**Discussion**

# A collaborative approach to research & development

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

K.L. Nielbo, M.L. Perner, C. Larsen, J. Nielsen and D.Laursen
kln@cas.dk
knielbo.github.io

**TAK**

kln@au.dk
knielbo.github.io

slides: http://knielbo.github.io/files/kln_dhn19.pdf

**& tak til**
Mads Rosendahl Thomsen, Comparative Literature, School of Communication and Culture, Aarhus University, DK
Jianbo Gao and Bin Liu, Institute of Complexity Science and Big Data, Guangxi University, CHN
Culture Analytics @ Institute of Pure and Applied Mathematics, UCLA, US