

Disruptive lexical dynamics in the *Shangshu*

Persistence and change in cultural transmission

Kristoffer L Nielbo

`kln@cas.dk`

`knielbo.github.io`

`chcaa.io`

Center for Humanities Computing
Aarhus University, Denmark

- 1 Background
 - CTEXT database
- 2 Erroneous class
 - Classifier
 - Misclassification
 - Document distance
 - Erroneous features
- 3 Disruptive dynamics
 - Document representation
 - Qualitative similarities
 - Novelty & resonance
 - Results

Background

CTEXT database

Erroneous class

Classifier

Misclassification

Document distance

Erroneous features

Disruptive dynamics

Document representation

Qualitative similarities

Novelty & resonance

Results

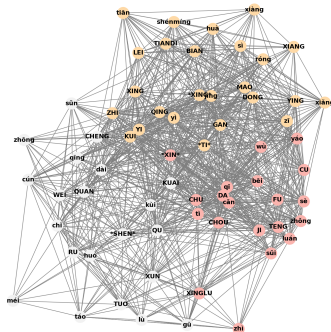


CTEXT database

Corpus of 96 Classical Chinese texts
from the CTEXT database
– ~ 1000 BCE - 200 CE
– > 6M tokens on ~ 16K types

Shangshu of 58 chapters:

Period	Chapters
Pre-Warring	10
Warring	16
Late Warring - Early Han	7
Han	25



Disruptive lexical
dynamics in the
Shangshu

Kristoffer L Nielbo
kln@cas.dk
knielbo.github.io
chcaa.io

Background

CTEXT database

Erroneous class

Classifier

Misclassification

Document distance

Erroneous features

Disruptive dynamics

Document representation

Qualitative similarities

Novelty & resonance

Results

Slingerland, E., Nichols, R., Nielbo, K., & Logan, C. (2017). The Distant Reading of Religious Texts: A Big Data Approach to Mind-Body Concepts in Early China. *Journal of the American Academy of Religion*, 85(4), 985–1016.

Nichols, R., Slingerland, E., Nielbo, K., Bergeton, U., Logan, C., & Kleinman, S. (2018). Modeling the Contested Relationship between Analects, Mencius, and Xunzi: Preliminary Evidence from a Machine-Learning Approach. *The Journal of Asian Studies*, 77(01), 19–57.

Text-age (mis-)classifier

Disruptive lexical
dynamics in the
Shangshu

Kristoffer L. Nielbo
kln@cas.dk
knielbo.github.io
chcaa.io

Train a simple and transparent learning model to explore the boundaries of age classification

- compare age-central to age-peripheral chapters of the *Shangshu*
- **error semantics** of the *Shangshu*

The probability of a document d being in class c , $P(c | d)$ is computed as:

$$P(c | d) \propto P(c) \prod_{i=1}^m P(t_i | c)$$

and the class of a document d is computed as:

$$c_{MAP} = \arg \max_{c \in \{c_1, c_2\}} P(c | d)$$

Background

CTEXT database

Erroneous class

Classifier

Misclassification

Document distance

Erroneous features

Disruptive dynamics

Document representation

Qualitative similarities

Novelty & resonance

Results



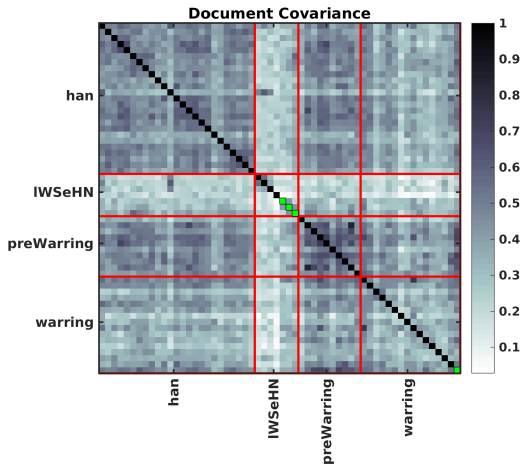


Figure 1: Documents covariance matrix for all chapters of the *Shangshu*

Three documents from the *Late Warring - Early Han* and one from *Han* are age ambiguous



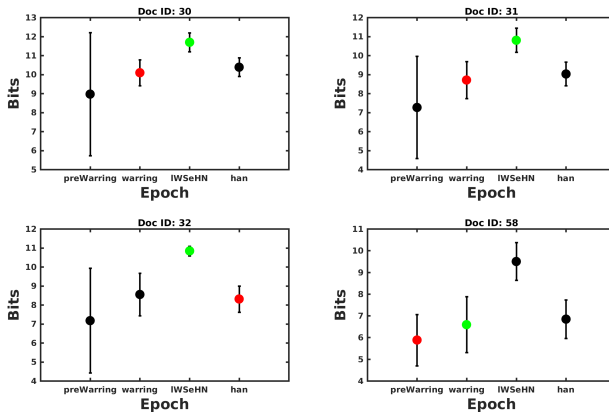


Figure 2: Average distance* to all classes show that the **error class** is closer to the document than the **correct class**

The distance between documents $s^{(1)}$ and $s^{(2)}$:

$$D_{KL}(s^{(1)} | s^{(2)}) = \sum_{i=1}^K s_i^{(1)} \times \log_2 \frac{s_i^{(1)}}{s_i^{(2)}}$$



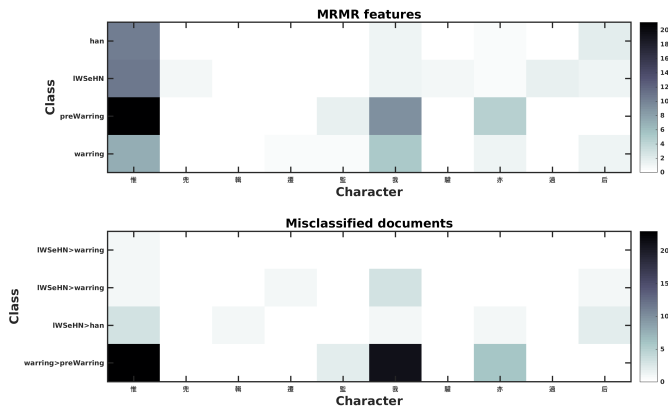


Figure 3: Features that collectively are most central for classification. Signal from feature one is sufficient to explain the error.

Disruptive age effect

Disruptive lexical
dynamics in the
Shangshu

Kristoffer L Nielbo
kln@cas.dk
knielbo.github.io
chcaa.io

Background

CTEXT database

Erroneous class

Classifier

Misclassification

Document distance

Erroneous features

Disruptive dynamics

Document representation

Qualitative similarities

Novelty & resonance

Results

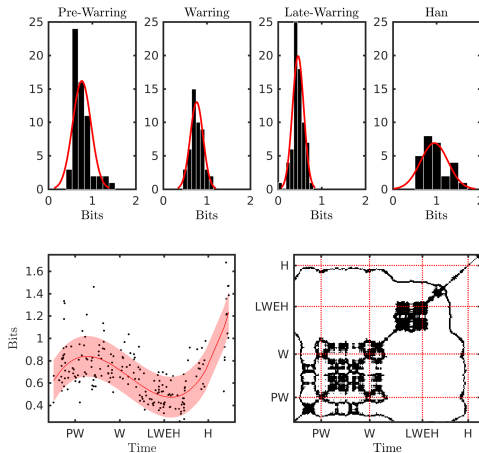


Figure 4: Length normalized lexical density over time for the Shangshu, notice the change points around and laminar region during *Late Warring - Early Han*



Dense document representation

- model semantic disruption as “variation on a theme” \Rightarrow use a simple Bayesian model to capture lexical semantics
- model each document as a distribution on lexical topics, e.g., $s = [0.09 \ 0.78 \ 0.11 \ 0.2]$, where each ‘topic’ is a distribution on words, and compare document similarity as the distance between any two documents with *chapter – index j* and *k*:

$$D_{KL}(s^{(j)} | s^{(k)}) = \sum_{i=1}^K s_i^{(j)} \times \log_2 \frac{s_i^{(j)}}{s_i^{(k)}}$$

- bracket concrete semantics (\sim reduce interpretive load) and only compare relative entropy between documents on topics (“variation on a theme”)

Disruptive lexical
dynamics in the
Shangshu

Kristoffer L. Nielbo
kln@cas.dk
knielbo.github.io
chcaa.io

Background

CTEXT database

Erroneous class

Classifier

Misclassification

Document distance

Erroneous features

Disruptive dynamics

Document representation

Qualitative similarities

Novelty & resonance

Results



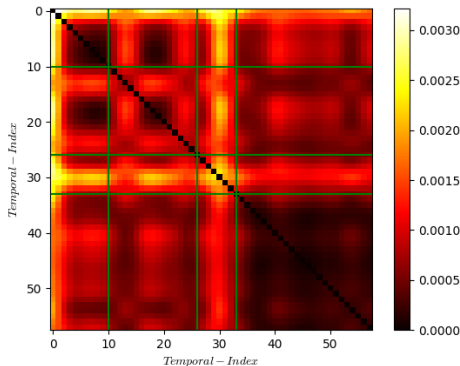


Figure 5: Distance matrix indicate some similarities with the sparse model (Fig. 1), notice the disruptive effect of *Late Warring - Early Han* centered on $i = 30$.

Background

CTEXT database

Erroneous class

Classifier

Misclassification

Document distance

Erroneous features

Disruptive dynamics

Document representation

Qualitative similarities

Novelty & resonance

Results

Compute disruption as a combination of resonance on novelty: Novelty over window w :

$$\mathbb{N}_w(j) = \frac{1}{w} \sum_{d=1}^w D_{KL}(s^{(j)} \mid s^{(j-d)})$$

with Transience:

$$\mathbb{T}_w(j) = \frac{1}{w} \sum_{d=1}^w D_{KL}(s^{(j)} \mid s^{(j+d)})$$

for Resonance

$$\mathbb{R}_w(j) = \mathbb{N}_w(j) - \mathbb{T}_w(j)$$

S

Background

CTEXT database

Erroneous class

Classifier

Misclassification

Document distance

Erroneous features

Disruptive dynamics

Document representation

Qualitative similarities

Novelty & resonance

Results

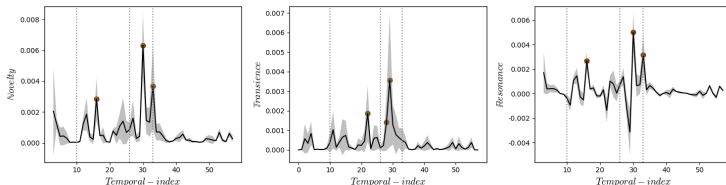


Figure 6: *Shangshu* Chapters' Novelty, Transience, and Resonance for $w = 3$

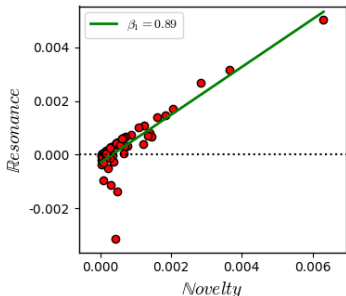


Figure 7: *Shangshu* Chapters' Resonance on Novelty for $w = 3$

in conclusion

- Late Warring - Early Han display age class atypical behavior
- lexical density shows global minimum and laminar behavior during Late Warring - Early Han
- two disruptive maxima are located in Late Warring - Early Han
- saturation followed by innovation
- class-dependent findings (study 1) confirmed by class-independent model (study 2)

Disruptive lexical dynamics in the *Shangshu*

Kristoffer L. Nielbo
kln@cas.dk
knielbo.github.io
chcaa.io

Background

CTEXT database

Erroneous class

Classifier

Misclassification

Document distance

Erroneous features

Disruptive dynamics

Document representation

Qualitative similarities

Novelty & resonance

Results



Thank you for your attention

kln@au.dk
knielbo.github.io
chcaa.io

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

Background

CTEXT database

Erroneous class

Classifier

Misclassification

Document distance

Erroneous features

Disruptive dynamics

Document representation

Qualitative similarities

Novelty & resonance

Results

