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

Outline

Disruptive lexical dynamics in the Shangshu

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

Background CTEXT database

Classifier

Mindreifine

Misclassification

ocument distance

ruptive dynamic

ocument representation

ualitative similaritie

Novelty & resonance

Results

1 Background CTEXT database

2 Erroneous class
Classifier
Misclassification
Document distance
Erroneous features

3 Disruptive dynamics
Document representation
Qualitative similarities
Novelty & resonance
Results



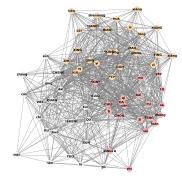
CTEXT database

Corpus of 96 Classical Chinese texts from the CTEXT database

- \sim 1000 BCE 200 CE
- ->6M tokens on ~16 K types

Shangshu of 58 chapters:

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



Disruptive lexical dynamics in the Shaneshu

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

Background CTEXT database

CTEXT databas

Litolieot

Misclassification

Cocument dis

Disruptive dynamics

Oocument 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

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, $P(c \mid d)$ is computed as:

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

and the class of a document d is computed as:

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

Disruptive lexical dynamics in the Shangshu

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

Background
CTEXT databa

Erroneous clas

Classifier

Misslassificatio

Document distance

Document representation

Qualitative similarities

Novelty & resonance



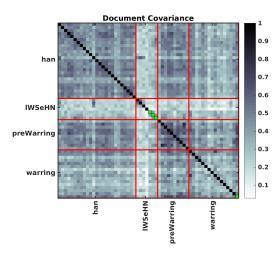


Figure 1: Documents covariance matrix for all chapters of the Shangshu

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

Disruptive lexical dynamics in the Shangshu

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

CTEXT database

Classifier

Misclassification

Erroneous features

Disruptive dynam

Qualitative similarities

Novelty & resonance

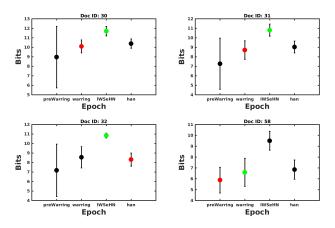


Figure 2: Average distance* to all classes show that the error class is closer to the document than the correct class

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

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



Disruptive lexical dynamics in the Shangshu

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

Background CTEXT database

Classifier

Misclassification

Document distance

Disruptive dynar

Document representation

Qualitative similarities

Novelty & resonance

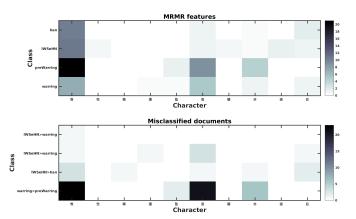


Figure 3: Features that collectively are most central for classification. Signal from feature one is sufficient to explain the error.



Disruptive lexical dynamics in the Shangshu

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

Background

Classifier

Document distance

Erroneous features

visitaptive dynami

Dualitative similarities

Novelty & resonance

Disruptive age effect

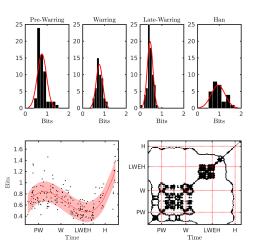


Figure 4: Length normalized lexical density over time for the Shangshu, notice the change points around and laminar region during Late Warring - Early Han

Disruptive lexical dynamics in the Shangshu

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

Background

Erroneo

Misslassificatio

Document di

Disruptive dynamics

D------

Qualitative similarities

Novelty & resonance

Posulto

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 = [0.09.78.11.2], where each 'topic' is a distribution on words, and compare document similarity as the distance between any two documents with chapter - index i and k:

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

– 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

Classifier

Classifier

Misclassification

Erroneous features

Document representation

Qualitative similarities
Novelty & resonance



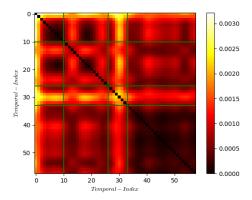


Figure 5: Distance matrix indicate some similarities with the sparse model (Fig. 1), notice the disruptive effect of Late Warring - Early Han centered on i = 30.

Disruptive lexical dynamics in the Shangshu

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

Background

Erroneous cia

Classifier

Document distance

Erroneous features

isruptive dynam

cument representation

Qualitative similarities

Novelty & resonance



Disruptive dynamics

Disruptive lexical dynamics in the Shangshu

Kristoffer I Nielbo kln@cas dk knielbo.github.io cheaa io

Novelty & resonance

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

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

with Transience.

window w:

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

for Resonance

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

S

Murdock, J., C. Allen, S. DeDeo, 2015, Exploration and Exploitation of Victorian Science in Darwin's Reading Notebooks, arXiv:1509.07175 .

Nielbo, K.L., M.L. Perner, C Larsen, J Nielsen, Laursen D. 2019. Automated Compositional Change Detection in Saxo Grammaticus' Gesta Danorum, hal-02084682

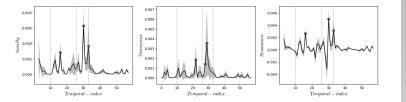


Figure 6: Shangshu Chapters' Novelty, Transience, and Resonance for w=3

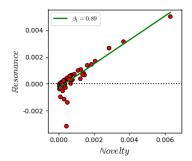


Figure 7: Shangshu Chapters' Resonance on Rovelty for w=3

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 Shaneshu

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

Background CTEXT database

Erroneous

Misclassification

rroneous features

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

Disruptive lexical dynamics in the Shangshu

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

Novelty & resonance

