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

1 Background
   CTEXT database

2 Erroneous class
   Classifier
   Misclassification
   Document distance
   Erroneous features

3 Disruptive dynamics
   Document representation
   Qualitative similarities
   Novelty & resonance
   Results
Corpus of 96 Classical Chinese texts from the CTEXT database
- ~ 1000 BCE - 200 CE
- ~ 6M tokens on ~ 16K types

Shangshu of 58 chapters:

<table>
<thead>
<tr>
<th>Period</th>
<th>Chapters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Warring</td>
<td>10</td>
</tr>
<tr>
<td>Warring</td>
<td>16</td>
</tr>
<tr>
<td>Late Warring - Early Han</td>
<td>7</td>
</tr>
<tr>
<td>Han</td>
<td>25</td>
</tr>
</tbody>
</table>


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:

$$c_{MAP} = \arg \max_{c \in \{c_1, c_2\}} P(c \mid d)$$
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 \( s^{(2)} \):

\[
D_{KL}(s^{(1)} \mid s^{(2)}) = \sum_{i=1}^{K} s^{(1)}_i \times \log_2 \frac{s^{(1)}_i}{s^{(2)}_i}
\]
Figure 3: Features that collectively are most central for classification. Signal from feature one is sufficient to explain the error.
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
Dense document representation

– model semantic disruption as "variation on a theme" ⇒ 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 $k$:

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

– bracket concrete semantics (≈ reduce interpretive load) and only compare relative entropy between documents on topics ("variation on a theme")
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 dynamics

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

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

with Transience:

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

for Resonance

$$R_w(j) = N_w(j) - T_w(j)$$

---


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 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)
Thank you for your attention

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