Beyond Verbal Identity

Contextual Features for Intertext Discovery

C. W. Forstall[1], L. Galli Milić[1], N. Coffee[2] and D. Nelis[1]

1. Université de Genève 2. University at Buffalo, the State University of New York

Motivation

We use the text-reuse detection tool Tesserae to locate potentially interesting allusions in Flavian epic poetry. While thematic resemblance at the scene level is often important to establishing the connection between two passages and thus the significance of an allusion, Tesserae presently focuses on localized re-use of specific phrases and may miss higher-level contextual cues.

We are testing the viability of larger-scale, "thematic" features targeted at the scene or paragraph level. Our goal is to modify the rankings of verbal correspondences identified by Tesserae according to the similarity of the respective contents of the phrases.

For example, the pair of phrases below was ranked 379th of 912 results by Tesserae, but in the context of systematic structural similarity (see right), otherwise lower-ranking text-reuse becomes more interesting.

> … etenim dat **candida** certam
> **nox** Helicen.
>
> (Val. Flac. 5.70)

> adspirant aurae in **noctem** nec **candida** cursus
> luna negat, splendet tremulo sub lumine pontus.
>
> (Verg. Aen. 7.8)

Thematic similarity

We see similar thematic elements in the openings of Aeneid 7 and Valerius Flaccus' Argonautica 5, in both cases at (what was likely) the mid-point of the narrative.

Vergil, *Aeneid* 7	*[BOOK DIVISION]*
7.1-7	Death and burial of Caieta; departure.
7.8-24	Voyage along the coast; Trojans pass Circe's land.
7.25-36	Dawn and arrival in Tiber.
7.37-106	Invocation of the Muse Erato and the situation in Latium.
7.107-147	Meal. Prayer of Aeneas; sacrifice.
7.148-285	Trojans make their way to the city and palace of Latinus.
7.286-640	Divine intervention: Juno and Allecto. War.

Valerius Flaccus, *Argonautica* 5	*[BOOK DIVISION]*
5.1-70a	Mariandyni; death and burial of Idmon and Tiphys; Erginus chosen as helmsman.
5.70b-176	Departure, voyage along southern coast of Black Sea; Argonauts pass the Chalybes, Carambis and Prometheus.
5.177-216	Evening and arrival in the Phasis. Prayer of Jason.
5.217-277	Invocation of a Muse (*dea*) and the situation in Colchis.
5.278-295	Divine intervention: Juno and Minerva. War.
5.296-328	Argonauts make their way to the city and palace of Aietes.

Methods

Corpus and text preparation

Our corpus was primarily epic, enlarged to include Ovid's *Heroides* and Seneca's *Medea*, which we felt might show affinities of style and content to our text of interest, Valerius Flaccus' *Argonautica.*

Each sample was 30 lines of text. Iinflected forms were reduced to lemmata, using methods comparable to those in Tesserae.

All preprocessing and subsequent analysis was done using R, with the help of the *cluster, mclust, tm* and *topicmodels* packages.

Unsupervised classification

We used k-means clustering to search for stable clusters of passages that shared similar language across works. Clustering was performed on two different feature sets:

1) TF-IDF weighted scores for all the words in the corpus common to two or more 30-line samples. Each sample was represented by a vector of approximately 8,000 frequencies.
2) A set of 50 topics generated using Latent Derichlet allocation (LDA). Each sample was represented by 50 values, representing its scores for each of the topics.

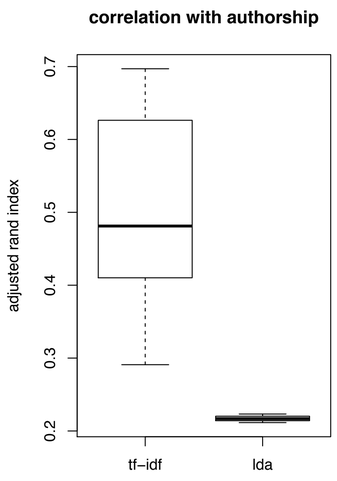
Correlation between clusterings

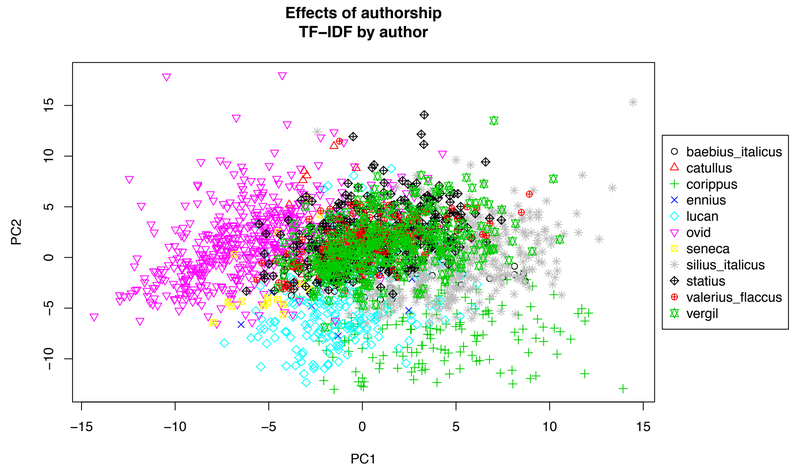
We tested correlation between the clusters generated by k-means using the adjusted rand index. This gives, for two classifications, a measure of their correlation above what is expected by chance.

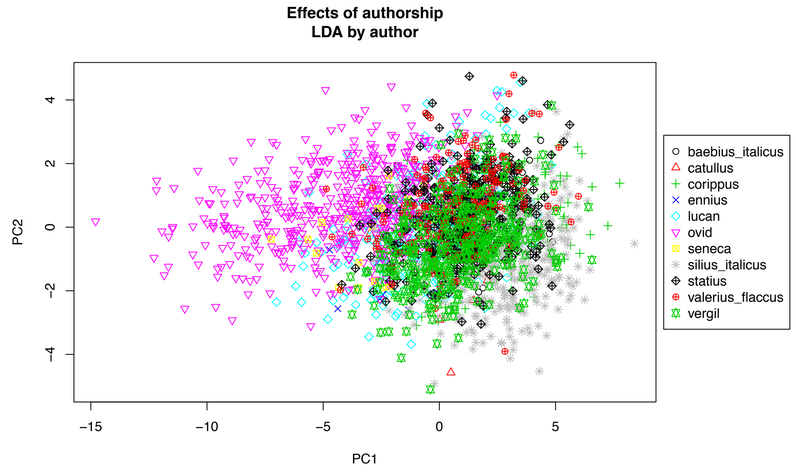
The effects of authorship

The box plot at right shows correlation between k-means clustering and true authorship, over 10 repetitions of the clustering for each treatment: tf-idf scores on the left, and LDA topic scores on the right. We chose k = 11, the number of authors in the corpus.

LDA was effective at reducing the otherwise significant impact of authorship on the classification.





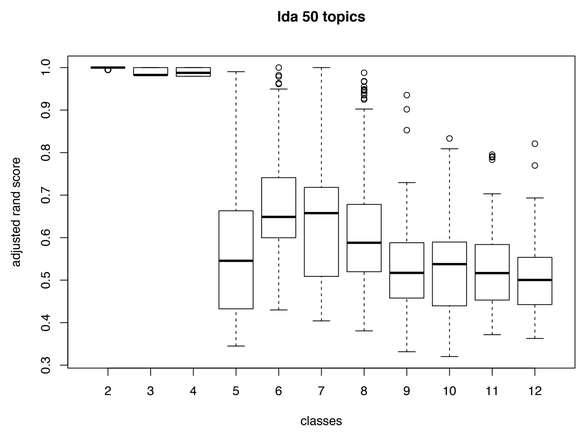


The figures above show the author effect graphically: for the TF-IDF features the distinctness of authors such as Ovid, Seneca, Lucan, Silius Italicus and Corippus from the central cloud is apparent. Under the LDA treatment, only Ovid maintained the same degree of separation.

Cluster stability

We varied k, the number of classes, from 2 to 12, and for each value of k we generated 15 clusterings. Adjusted rand indices were calculated for each of 105 possible pairs of clusters for a given value of k.

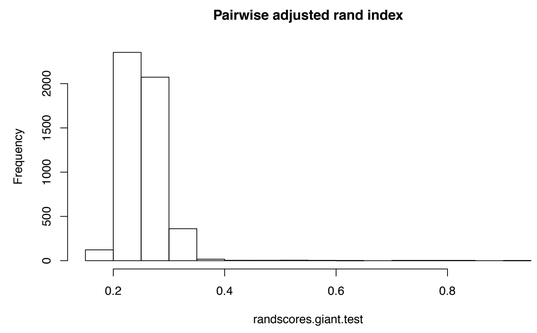
The distributions of these (right) provide an indication of the stability of each configuration of classes: small numbers of classes are highly stable; among larger values of k, divisions into 6 and 7 classes are most stable.



Topic Stability

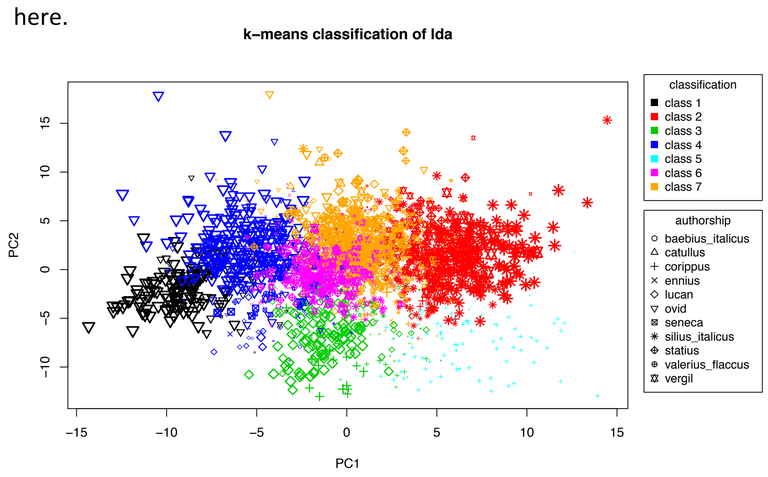
To test the stability of LDA, we generated 100 different LDA models of 50 topics, performing k-means clustering on each one with k = 7.

The figure at right shows the distribution of adjusted rand index values for 4950 pairwise comparisons between the 100 classifications produced. Correlation is consistent but low, at around 0.25, with one or two outlier cases having high agreement.

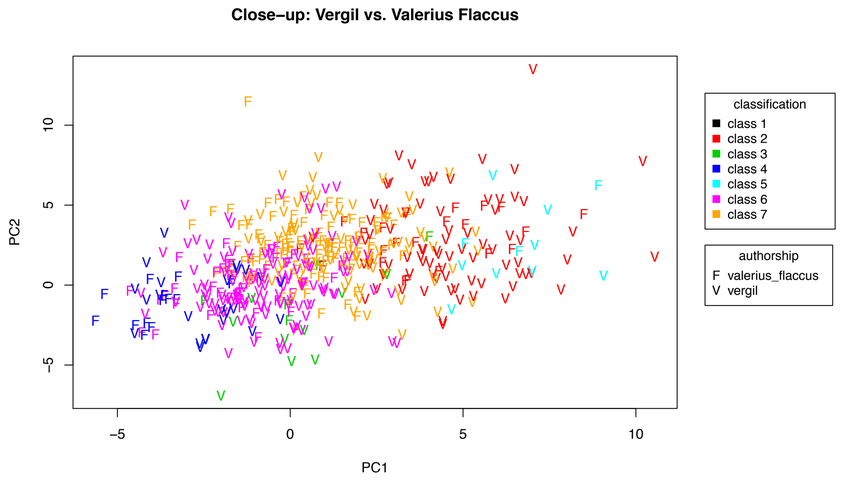


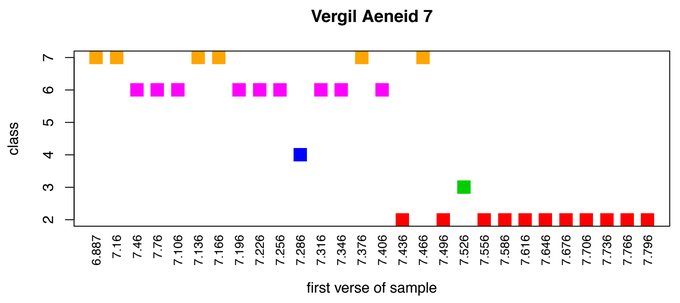
Sample results

Below we show one example of k-means clustering into 7 classes, taken from the topic stability experiments described above. Point size shows how often, in 100 different tests, each sample fell into the class shown here.



Below: a close-up showing only Vergil's *Aeneid* and Valerius Flaccus' *Argonautica*. This is the type of result that we are looking for: samples fall into multiple classes and are not segregated by author.





Above: book 7 of the *Aeneid*. The first half of the book, which features more peaceful content, alternates between classes 6 and 7, the most general of the epic classes. The preparations for war in the book's second half group with class 2. Two passages affiliate with more author-specific groups: Juno's speech at 286 falls in the group dominated by Ovid's *Metamorphoses*, while the single brief battle scene groups with Lucan's *Civil War* in class 3.

TESSERAE

FNSNF Swiss National Science Foundation

University at Buffalo The State University of New York

Université de Genève

This work was funded by FNS project number **100015_146976**, "Intertextuality in Flavian Epic Poetry."