

The Tesserae Project

Detecting Intertextuality of Meaning and Sound

Neil Coffee

(State University of New York, Buffalo, USA)

Christopher Forstall

(State University of New York, Buffalo, USA)

James Gawley

(State University of New York, Buffalo, USA)

Abstract The Tesserae Project offers a free online intertextual search tool for ancient Greek, Latin, and English. Tesserae has in the past allowed for a pairwise searching of literary texts in these languages for exact word or lemma similarities. This paper describes two new types of search now offered by Tesserae, by meaning (semantic search) and by sound.

Summary 1 Semantic matching. – 2 Sound matching. – 3 Similarity of Meaning and Sound in Intertextual Research.

Keywords Intertextuality. Text reuse. Allusion. Greek literature. Latin literature. English literature. Digital humanities.


The Tesserae Project provides a free online tool that allows users to compare two texts in ancient Greek, Latin, or English and automatically discover instances of similar language, and so identify textual phenomena including text reuse and literary allusion (<http://tesserae.caset.buffalo.edu>). The basic Tesserae search finds sentences or poetic verse lines in two different texts that share two or more similar lemmata. The lemma search is designed to capture parallels where inflectional forms differ, so that 'animo... magno' in one text would match 'animus... magnus' in another. This extends to irregular and suppletive systems so that 'gladium... fert' would match 'gladium... tulit'. The lemma-based search, documented in the project's previous publications, has been shown to capture some 60% of significant parallels found by commentators.

The project's current work involves the development of new search features, with the aim of creating a blended search that better replicates and extends human recognition of textual similarity. These forms of search remain experimental, in that they have not yet been tested and tuned. Nevertheless, they provide new capacities that users can currently employ

Antichistica 14

DOI 10.14277/6969-182-9/ANT-14-14 | Submission 2017-08-14

ISBN [ebook] 978-88-6969-182-9 | ISBN [print] 978-88-6969-183-6

© 2017 |  Creative Commons Attribution 4.0 International Public License

and explore. This article will briefly describe and give examples of two new forms of search, so that users can experiment with them in an informed way. Those who wish to explore the code directly can do so at <https://github.com/tesseract>.

1 Semantic matching

One of the new Tesseract capabilities is the ability to find passages sharing words with similar meanings, regardless of the lemma. This 'semantic search' is available for both the Latin and Greek corpora and adds the ability to compare Latin works directly to Greek. Users can conduct a semantic search from the search page by clicking 'show advanced', then choosing 'semantic' from the feature drop-down list.

Semantic search captures instances of lemma identity, but also parallels that cannot be found through lemma matching. For example, if we compare the whole of Vergil's *Aeneid* with the whole of Ovid's *Metamorphoses*, among the top twenty results is the match of *Aeneid* 3.632, "saniem eructans", with *Metamorphoses* 4.494, "saniemque vomunt". Though Vergil writes of 'belching up gore' and Ovid writes 'they throw up gore', the passage is detected as a potentially significant parallel because 'belch' and 'throw up' are near synonyms. Semantic search works similarly in matching Latin to Greek. In a comparison of all of Homer's *Odyssey* with all of Vergil's *Aeneid*, the phrase κύματ' ἔταμνεν at *Odyssey* 13.88 is identified with "alta secans fluctuque" at *Aeneid* 10.687, where both phrases refer to ships cutting through waves.

As in the lemma-based search, semantic search compares the two selected texts word-by-word. Sentences (or verse lines) sharing two or more semantically-related words are returned as a match. Words are considered related if they appear in the Tesseract related-word dictionary for the language pair being searched - Latin to Latin, Greek to Greek and Latin to Greek. These dictionaries were created by automatically associating the English definitions in lexica retrieved from the Perseus Digital Library: Lewis and Short for Latin and Liddell-Scott-Jones for Greek. Dictionary headwords were compared in a three-step process. First, the most semantically significant words within the definition for each dictionary headword were identified by using TF-IDF scores. Next, these TF-IDF scores were fed into a latent semantic indexing model in order to give each definition a numeric vector representative of its meaning. Finally, the resulting vector-representations of each definition were compared according to their cosine similarity. Pairs of headwords whose entries exceeded a certain threshold were considered 'semantically similar'. Empirical testing of the optimal threshold is ongoing. This method does not distinguish between forms of relatedness: related words could be synonyms, antonyms, or metonyms,

and some are false positives. Preliminary results for the Latin-Greek dictionary suggest, however, that approximately 80% of the correspondences it yields consist of synonyms or closely related words, as in the examples above.

2 Sound matching

Tesserae also has a sound matching capability, using orthography as a proxy for phonology. It is designed to capture instances where one author echoes not the words or meaning of another, but a pattern of sounds. Users can conduct a sound search from the search page for any language, by clicking 'show advanced', then choosing 'sound' from the feature drop-down list.

In a search for sound similarities between Vergil's *Aeneid* and Ovid's *Amores*, among the top results was the line "ipsis praecipuos ductoribus addit honores" at *Aeneid* 5.249 matching the line "haec est praecipuo victoria digna triumpho" at *Amores* 2.12.5. These lines would not have matched by bigram lemma search, since they share only one common word, "praecipuo(s)". They were returned by the sound search because of the additional similarity of the sounds in "ductoribus" and "victoria". The lines deal with related themes. In the *Aeneid*, Aeneas is giving honours to the competitors in the naval race. In the *Amores*, Ovid is celebrating his conquest of Corinna in terms of a military triumph. Both are scenes of military-style competition, victory and celebration. The absence of more than one common word between the two passages makes it difficult to see Ovid's words as an allusion to the *Aeneid*'s line but he seems to be appropriating Vergil's language, sound, and metrical rhythm as ready-made materials to lend a Vergilian feeling to a version of the lover-as-soldier trope.

There are many ways that a sound search could be constructed. Tesserae's sound detection searches for identical three-letter sequences, or character trigrams. The more trigrams shared by two sentences or lines, the higher that pair ranks. Forms from the same lemma are given no special weight, but appear in the results to the extent that they have similar character trigrams.

3 Similarity of Meaning and Sound in Intertextual Research

Traditionally, most instances of Greek and Latin intertextuality documented by scholars have consisted of either quotation or close verbal echo. Computational methods can detect these forms of intertextuality, but have the potential to identify more subtle phenomena as well. Scholars have also long considered that similarities of meaning and sound can contribute

to making a meaningful intertextual parallel, as when they are combined with the reuse of a single lemma, or even constitute such a parallel in the absence of any shared lemmata. The computational capacity to detect semantic and sound similarity across large corpora holds the potential to expand our understanding of intertextuality, allowing us to give fuller consideration to instances where the parallelism between texts is more subtle, but nevertheless significant. We hope that the meaning and sound matching capabilities of the Tesseract website will contribute to advancing this research.

Bibliography

- Coffee, Neil et al. (2012). "Intertextuality in the Digital Age". *TAPA*, 142(2), 381-419.
- Buck, Thomas et al. (2014). "Modeling the Scholars. Detecting Intertextuality through Enhanced Word-Level N-Gram Matching" [online]. *Literary and Linguistic Computing*. URL <https://academic.oup.com/dsh/article-lookup/doi/10.1093/lc/fqu014> (2017-11-13).
- Forstall, Christopher; Scheirer, Walter (2010). "Features from Frequency. Authorship and Stylistic Analysis Using Repetitive Sound". *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*, 1(2), 1-23.