

White Paper Report

Report ID: 106128

Application Number: HD5157012

Project Director: Neil Coffee (ncoffee@buffalo.edu)

Institution: SUNY Research Foundation, University at Buffalo

Reporting Period: 4/1/2012-2/28/2014

Report Due: 5/31/2014

Date Submitted: 5/29/2014

Final White Paper

National Endowment for the Humanities Office of Digital Humanities
Start-up Phase II Grant: HD-51570-12

Tesserae: A Search Engine for Allusion

4/1/2012 – 2/28/2014

Project Directors: Neil Coffee and Jean-Pierre Koenig

University at Buffalo, SUNY

May 29, 2014

PROJECT ACTIVITIES AND ACCOMPLISHMENTS

The Tesseræ Project offers on its website (<http://tesseræ.caset.buffalo.edu/>) tools for the study of literary and linguistic influence, or intertextuality, in and among works in ancient Greek, Latin, and English. The major goals for the grant period were to improve the performance and expand the function of the site as well as further disseminate knowledge of its capabilities.

The core function of the site allows users to choose two texts to compare, adjust the parameters for comparison, and then receive a list of parallel passages in the texts that share common phrases. The default search looks for lines of poetry or sentences in each text that share two or more word roots (lemmata). So, for example, the Latin expression for “crow with a [black] throat” appearing in one text (*atro guttere corvus*, Catullus 108.5) will show up as a match for “throat of a crow” in another (*guttere corvi*, Vergil *Georgics* 1.423), despite the fact that the words for “crow” in the different texts (*corvus* and *corvi*) have different spellings following their different inflections (Figure 1).

Figure 1. Results of Tesseræ comparison search for Vergil’s *Georgics* and the poems of Catullus.

The screenshot shows the Tesseræ website interface. The browser address bar displays http://tesseræ.caset.buffalo.edu/cgi-bin/read_multi.pl. The page title is "Tesseræ" and the main heading is "INTERTEXTUAL PHASE MATCHING". Navigation links include "SEARCH HOME", "HELP", "ABOUT", and "RESEARCH". A menu bar contains "BASIC SEARCH", "ADVANCED FEATURES", "OLDER VERSIONS", "GREEK", and "EXPERIMENTAL".

Search controls include: Sort **decreasing** by **score** and format as **html**. Show **100** results at a time. 745 results in 8 pages. Go to page: 1 2 3 4 5 [next] [last]

	target phrase	source phrase	matched on	score	cross-ref (score)
1.	verg. g. 1.423 et laetae pecudes et ovantes guttere corvi .	cat. 108.5 effossos oculos voret atro guttere corvus ,	coruus, guttur	9	
2.	verg. g. 2.281 directaeque acies , ac late fluctuat omnis	cat. 63.56 cupit ipsa pupula ad te sibi derigere aciem ,	acies, dirigo	8	vergil.aeneid 7.523 (8)
3.	verg. g. 3.256 et pede prosubigit terram, fricat arbore costas	cat. 23.22 quod tu si manibus teras fricesque ,	frico, terra	8	
4.	verg. g. 4.133 nocte domum dapibus mensas onerabat inemptis.	cat. 64.304 large multiplici constructae sunt dape mensae ,	daps, mensa, mensus, metior	8	horace.ars_poetica 198 (7) ovid.fasti 5.521 (6) ovid.metamorphoses 8.572 (8) silius_italicus.punica 2.523 (7), 11.421 (8), 13.272 (8) vergil.aeneid 1.706 (8), 7.125 (7), 11.738 (6)

Users can thus find instances such as this one where one author (Catullus) seems to have influenced another (Vergil). Identifying such text reuse helps us understand the artistry of composition as well as discover allusions from one text to another, principal areas of investigation in the classical literary studies of the last several decades. Unlike with previously existing search tools, the user is not required to come up with an individual phrase to search for. The user can instead simply choose two texts for comparison to reveal all the qualifying phrases they share.

Prior to the grant period, existing Tesseræ search functions had significant limitations. Earlier testing showed that Tesseræ missed a number of phrase parallels that it should have caught. This led to a goal for the grant period of capturing more than 50% of the meaningful parallels from a given benchmark set. As it turned out, this goal was easily reached. With some minor adjustment of the code, our tests showed that we could capture some 70-80% of the parallels found by traditional methods, which we had previously determined to be the maximum that could be found within our benchmark set using our base method of matching a minimum of two stems (bigram lemma matching).

A more significant challenge was the fact that, for comparisons of average-sized texts, the result of our search was long lists of thousands of undifferentiated word parallels, some likely to be of minimal literary interest, as when text shared prepositions or pronouns. Some such less interesting parallels could be excluded by creating a stop list of words to exclude from search results, but this expedient solved only part of the problem. The major accomplishment of the grant period here was thus to devise and implement on the website a formula for automatically sorting matched phrases, so that those likely to be of greater artistic and interpretive significance could be featured at the top of the results list. The formula, given in Figure 2, privileges parallels where the matched words in each text are close together and where the individual words are relatively rare (within the whole corpus or the compared texts, depending upon user preferences).

Figure 2. Equation for Tesseræ Version 3 scoring system

$$\text{core} = \ln \left(\frac{\sum \frac{1}{f(t)} + \sum \frac{1}{f(s)}}{d_t + d_s} \right)$$

where

$f(t)$ is the frequency of each matching term from the target phrase in a selected corpus;

$f(s)$ is the frequency of each matching term from the source phrase in a selected corpus;

d_t is the distance between matching words in the target;

d_s is the distance between matching words in the source.

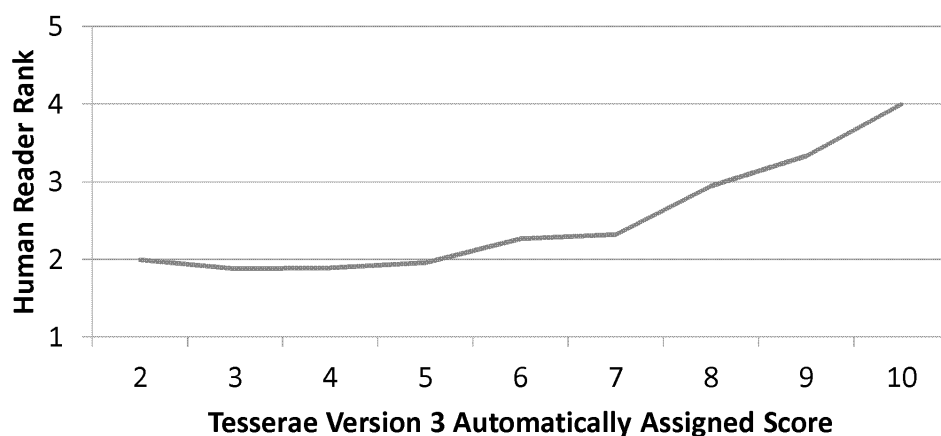
Frequency is the number of times a word occurs in its respective text divided by the total number of words in that text. The frequency of the same word may thus be different in different texts.

Where an allusion involves more than two shared words, *distance* is measured between the two lowest-frequency matching words in a phrase in order to determine scores based on words likely to be of the most literary interest. So, for example, the Vergil *Georgics* - Catullus search

illustrated in Figure 1 reveals lines in both texts containing forms of *exustus* (“used up”), *ager* (“fields”), and *cum* (“when”). Catullus 68B62 is: *cum gravis exustos aestus hiulcat agros*. *Georgics* 1.107 is: *et, cum exustus ager morientibus aestuat herbis*. The word *cum* is very frequent in Latin; the words *exustus* and *ager* are rarer. So for this part of the score calculation *cum* is excluded and the distance between *exustus* and *ager* is measured. In this case, since the words *exustus* and *ager* are adjacent in both texts, this parallel receives a higher score than it would have if the distance between *cum* and one of these words had been measured.

The performance of the scoring system was then tested by comparing the automatic Tesseract identification of significant parallels with the identification by traditional scholarly methods. The object of study for the test was the intertextual relationship between two epic poems in Latin: Vergil’s 1st century BCE epic *Aeneid* and Book 1 of Lucan’s *Civil War*, an epic composed about a century later. Scholars have demonstrated that Lucan often draws phrasing from the *Aeneid* for artistic effect. The question was how well Tesseract could replicate these scholarly findings by identifying the most meaningful parallels from among all the instances where Lucan used two or more word stems (lemmata) within the same sentence that Vergil had used similarly. The test compared how Tesseract scored the parallels it found with how human commentators scored an overlapping set of parallels. Members of the project team hand-ranked a sample of some 3,000 parallels drawn from existing scholarly commentaries and previous Tesseract searches (the latter to include low scoring results), using a scale of 1 (least interest) to 5 (greatest interest) scale. The Tesseract automatic scoring system then ranked the same parallels from 1 (of least interest) to 11 (of highest interest). As illustrated in Figure 3, the test demonstrated a significant correlation between human ranks and those of the automatic scoring system. That is, on average, the scoring system ranked as most significant the same results that human interpreters did. With this improvement in place on the website, the core Tesseract lemma search instantly became much more useful, since it brought the most potentially interesting results to the top.

Figure 3. Average Reader Ranking Per Automatic Score Level for Phrase Parallels in Lucan *Civil War* 1 and Vergil *Aeneid*.



Another major goal was then driven by user requests: the expansion of the corpus of texts that could be searched. In order to deliver rapid results, texts must be added to the system by the Tesseract team for

pre-processing. The text addition procedure is computationally simple, but requires scholarly understanding of how to segment the texts properly. At the start of the grant period, Tesseract had a relatively small corpus of Latin texts. By the end of the grant period, the site had incorporated all of the Latin texts on the Perseus Project website and all of its ancient Greek texts. The project team also worked in collaboration with graduate students in the University at Buffalo Departments of English and Linguistics to mount a sample set of English texts. These accomplishments required considerable efforts on the part of a graduate RA on the project and a group of undergraduate volunteers, but ultimately resulted in the development of a semi-automated workflow that will improve the efficiency of future text addition.

In addition to improving search of two texts by similarity of lemma, the team also created other new features on the site. Multi-text search (<http://tesseract.caset.buffalo.edu/multi-text.php>) is an extension of lemma matching that takes into account not just two texts for comparison, but others in the corpus as well. This feature is illustrated in Figure 1 above, where the final column, entitled “cross-ref,” shows additional locations in the corpus where a phrase found in the two compared texts appears. The first result in Figure 1 shows that the phrase consisting of the words *guttur* (“throat”) and *corvus* (“crow”) in the same line of verse appears nowhere else in the Perseus Latin corpus. This phrase thus forms a unique and distinctive link between the two texts, making it more likely that it constitutes an allusion. Conversely, in the fourth result shown, the co-occurrence of the words *daps* (“feast”) and *mensa* (“table”) is found not only in the compared works of Catullus and Vergil, but also in five later works, some with multiple occurrences. Users can click on the underlined numerical location identifiers to produce a pop-up window showing the text in which the matched words occur. In this case, the researcher might conclude that, since the phrase is still relatively rare (common phrases have hundreds of instances), this is not an instance of ordinary language reuse, but constitutes artistic and conceptual dialogue between Catullus and later poets, all writing on a similar theme of feasting with similar language and imagery.

During the grant period, Tesseract also developed and implemented on the website several types of intertextual search that are not based in similarity of lemma. The sound-matching feature, available as an option on the main search pages for each language, matches passages between two texts based on the frequency of shared three-letter sequences (character trigrams). Matching of sound similarity can reveal shared sound play among poets as well as overall trends in literary and linguistic sound patterns. The standalone topic modeling feature (<http://tesseract.caset.buffalo.edu/cgi-bin/lisa.pl>) employs an algorithm to match passages with overall similar meaning. The word-level semantic matching feature involves the matching of individual words by meaning, rather than by lemma identity. A version of this search live on the site allows users to compare ancient Greek and Latin texts for passages that share similar content, in the form of synonyms or related words (<http://tesseract.caset.buffalo.edu/cross.php>). Figure 4 gives the top results from a comparison of Homer’s Greek *Iliad*, written down around the end of the 8th century BCE, with Vergil’s Latin *Aeneid*, composed in the 1st century BCE.

Figure 4. Cross-language intertextual search via semantic matching between Greek (Homer *Iliad*) and Latin (Vergil *Aeneid*). Matching words are highlighted in red onscreen.

The screenshot shows the Tesseractae website interface. At the top, there is a logo and the name 'TESSERAЕ'. Below this, there are navigation links for 'LATIN', 'GREEK', 'ENGLISH', 'OTHER TOOLS', and 'SOURCES'. A search bar is present with a search icon. Below the search bar, there are filters for 'Sort' (decreasing), 'by' (score), and 'and format as' (html). A 'Change Display' button is also visible. Below the filters, it says 'Show 100 results at a time.' and '6125 results in 62 pages. Go to page: 1 2 3 4 5 [next] [last]'. The main content is a table with the following columns: 'target phrase', 'source phrase', 'matched on', and 'score'.

	target phrase	source phrase	matched on	score
1.	verg. aen. 4.652 accipite hanc animam, meque his exsolvite curis .	hom. il. 23.62 εὔτε τὸν ὕπνος ἔμαρπτε λύων μελεδήματα θυμοῦ	cura, exsoluo	11
2.	verg. aen. 4.200 centum aras posuit , vigilemque sacraverat ignem ,	hom. il. 10.46 Ἐκτορείοις ἄρα μάλλον ἐπὶ φρένα θῆχ' ἰσροΐσιν :	sacer, pono	11
3.	verg. aen. 4.589 terque quaterque manu pectus percussa decorum,	hom. il. 13.282 ἐν δέ τέ οἱ κραδίη μεγάλα στέρνοισι πατάσσει	percutio, pectus	11
4.	verg. aen. 4.83 incubat, illum absens absentem auditque videtque ;	hom. il. 18.53 εἶδετ' ἀκούουσαι ὄς ἐμῶ ἐνὶ κήδεα θυμῶ.	audio, uideo	11
5.	verg. aen. 4.410 prospiceres arce ex summa , totumque videres	hom. il. 14.228 ἀκροτάτας κορυφάς : οὐδέ χθόνα μάρπτε ποδοῖν:	superus, arx	10

In the first result shown, in the right-hand source phrase Homer describes how sleep released the hero Achilles from his cares (λύων μελεδήματα). In the Latin target phrase from Vergil on the left, the heroine Dido, faced with disastrous defeat and humiliation, calls on the gods to release her from her cares (*exsolvite curis*). Semantic matching thus automatically reveals a similar moment of crisis and release for two major characters of classical epic poetry. Such semantic matching relies on a dictionary of related words. In this case, the dictionary indicates that the Greek word μελεδήματα (“cares”) and the Latin word *curis* (“cares”) are functionally equivalent. In order to create this feature, since no digital Greek-Latin translation dictionary existed, the Tesseractae team pursued two approaches to create one. The first, “pivot” method involved identifying related words by the similarity of their English dictionary definitions in online Greek and Latin dictionaries. The second, “alignment” method involved deducing related words from parallel digital editions of the New Testament in Greek and Latin.

DISSEMINATION

Beyond offering updated tools on its website, the Tesseractae team has pursued a number of avenues within the grant period to disseminate its work. A blog site (<http://tesseractae.caset.buffalo.edu/blog/>) offers guidance on the use of the site and reports on project activities. The team has published or had accepted for publication several articles that describe the project methodology and use of the site, offer new interpretations of literary works using Tesseractae tools, and address theoretical consequences for the study of intertextuality.

Scheirer, W., C. Forstall, and N. Coffee. (under review). "The Sense of a Connection: Automatic Tracing of Intertextuality by Meaning."

- Coffee, N. and C. Forstall. (forthcoming). "Claudian's Engagement with Lucan in his Historical and Mythological Hexameters" in *Lucaïn et Claudien face à face: une poésie politique entre épopée, histoire et panégyrique*, eds. V. Berlincourt, L. Galli-Milic, D. Nelis. Winter Verlag.
- Forstall, C., N. Coffee, T. Buck, K. Roache, and S. Jacobson. (2014). "Modeling the Scholars: Detecting Intertextuality through Enhanced Word-Level N-Gram Matching." *Literary and Linguistic Computing (LLC)*. doi: 10.1093/lc/fqu014 [[Abstract](#)] [[Preprint](#)] [**See Appendix 4**]
- Coffee, N., J. Gawley, Christopher Forstall, Walter Scheirer, Jason Corso, David Johnson, and Brian Parks. (2014). "Modeling the Interpretation of Literary Allusion with Machine Learning Techniques." *Journal of Digital Humanities*3.1. [[Poster](#)]
- Coffee, N., J.-P. Koenig, S. Poornima, C. Forstall, R. Ossewaarde, and S. Jacobson. (2012). "The Tesserae Project: Intertextual Analysis of Latin Poetry" *Literary and Linguistic Computing* 28.1: 221-8. doi: 10.1093/lc/fqs033 [[Abstract](#)] [[Preprint](#)]
- Coffee, N., J.-P. Koenig, S. Poornima, C. Forstall, R. Ossewaarde, and S. Jacobson. (2012). "Intertextuality in the Digital Age." *Transactions of the American Philological Association* 142.2: 383-422. [[Abstract](#)] [[Preprint](#)]

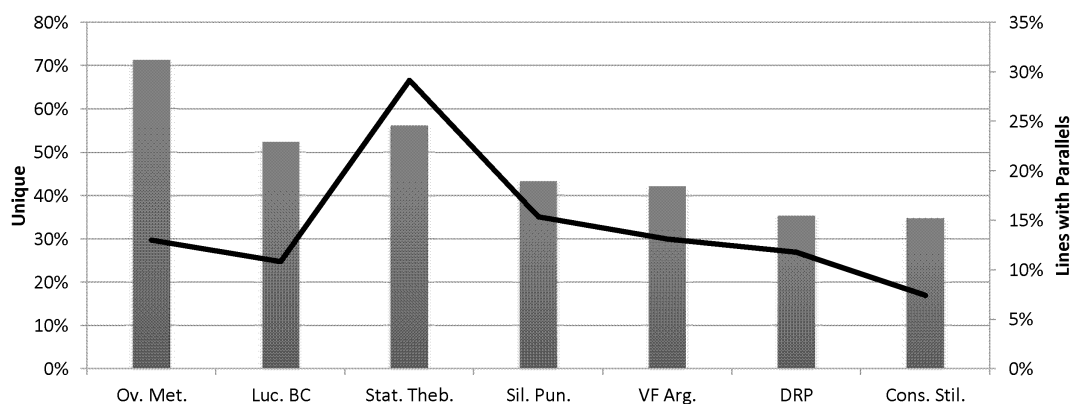
The Tesserae team and its collaborators have likewise given numerous peer-reviewed conference papers and invited lectures during the grant period.

- Gawley, J. and Forstall, C. (2014). "Automating the Search for Cross-Language Text Reuse." Short paper, *Digital Humanities 2014*. July 11. Lausanne, Switzerland.
- Scheirer, W. and Forstall, C. (2014). "Euterpe's Hidden Song: Patterns in Elegy." Poster, *Digital Humanities 2014*. July 10. Lausanne, Switzerland.
- Coffee, N. (2014). Participant in panel "Rethinking Text Reuse as Digital Classicists." *Digital Humanities 2014*. July 10. Lausanne, Switzerland.
- Coffee, N. (2014). "Modeling the Scholars: Detecting Intertextuality through Enhanced Word-Level N-Gram Matching" *International Workshop on Computer Aided Processing of Intertextuality in Ancient Languages*. June.
- Bernstein, N., Gervais, K., Lin, W. (2014). "Comparative Rates of Text Reuse in Latin Epic." *American Philological Association / American Institute of Archaeology Annual Meetings*. January 3. [[Screencast](#)]
- Gervais, K. (2014). "Flavian Intertextuality: A Digital Approach," 35th ASCS Conference, January 2014, Auckland, NZ [[Paper](#)]
- Coffee, N. (2013). "Roses and Lilies: Digital Adventures in Intertextuality." Invited lecture, Yale University. December 5.
- Coffee, N., J. Gawley, C. Forstall, W. Scheirer, D. Johnson, J. Corso and B. Parks. (2013). "Modeling the Interpretation of Literary Allusion with Machine Learning Techniques." Electronic poster presented at *Digital Humanities 2013*, University of Nebraska – Lincoln, July 18. [[Abstract](#)] [[Slides](#)]
- Coffee, N., C. Forstall, and J. Gawley. (2013). "What is Allusion? A Digital Approach." Poster presented at the *Digital Classics Association* conference, University at Buffalo, April 5.

- Forstall, C. and W. J. Scheirer. (2012). "Revealing hidden patterns in the meter of Homer's *Iliad*." Poster presented at the Chicago Colloquium on Digital Humanities and Computer Science, University of Chicago, Chicago, IL. November 17-19. [[Abstract](#)] [[Poster](#)]
- Gawley, J., C. Forstall, and N. Coffee. (2012). "Evaluating the literary significance of text re-use in Latin poetry." Poster presented at Chicago Colloquium on Digital Humanities and Computer Science, University of Chicago, Chicago, IL. November 17-19. [[Abstract](#)] [[Poster](#)]
- Coffee, N. (2012). "Large- and Small-Scale Intertextuality in Claudian's Historical and Mythical Hexameters." Paper presented at "Lucain et Claudien face à face: une poésie politique entre épopée, histoire et panégyrique," Fondation Hardt, Vandœuvres-Geneva, Switzerland, November 8-10.
- Forstall, C. (2012). "Revealing Intertextuality with Tesserae." Workshop presented at "Lucain et Claudien face à face: une poésie politique entre épopée, histoire et panégyrique," Fondation Hardt, Vandœuvres-Geneva, Switzerland, November 8-10.

One part of this work has investigated large-scale flows of language and ideas in ways made possible by comprehensive automatic identification of parallel phrases. Figure 5, from Coffee and Forstall Forthcoming, illustrates the use of phrases from the seminal Latin epic *Aeneid* by later epic poets writing in Latin.

Figure 5. Use by later Latin epic works of most significant (score 8 or above) phrases from Vergil's *Aeneid*. (From left to right, later epic works are: Ovid *Metamorphoses*, Lucan *Civil War*, Statius *Thebaid*, Silius Italicus *Punica*, Valerius Flaccus *Argonautica*, Claudian *On the Rape of Persephone*, and Claudian *On the Consulship of Stilicho*).



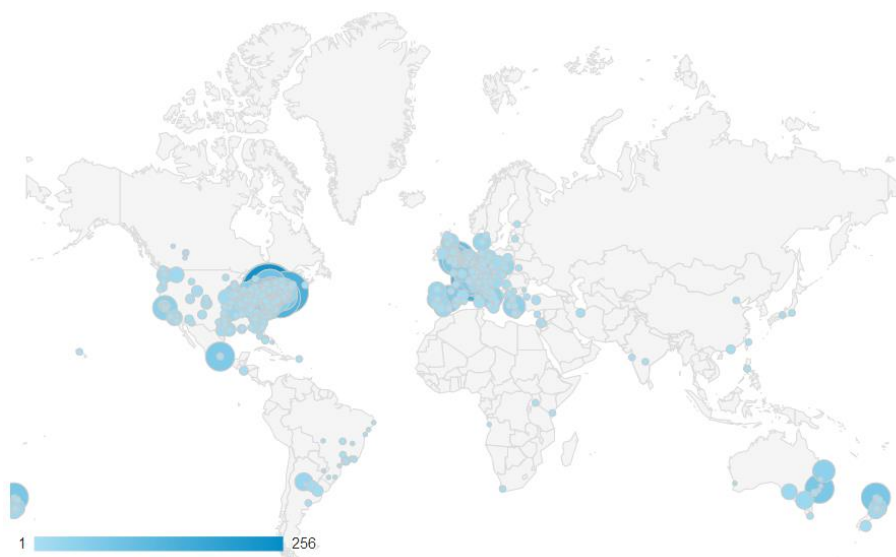
The line, indexed on the right of the chart, shows the percentage of verse lines in later epic poems that share high-scoring (8 and above) phrases with the *Aeneid*. The line rises to the peak of some 28% in the case of the *Thebaid* of Statius. This indicates that more than a quarter of all of the lines in Statius's *Thebaid* employ phrasing from Vergil's *Aeneid*. The columns, indexed to the left, show what percentage of these shared phrases occur uniquely between a given later epic and the *Aeneid*. The shrinking of columns over time, from left to right on the chart, illustrates the fact that it becomes more difficult (or less likely) for later poets to use a phrase from the *Aeneid* that has not already been used by earlier

authors. Hence we find a declining trend in the percentage of phrases shared uniquely with the *Aeneid*. Here too, Statius is remarkable, since, as his higher column indicates, he is the only poet who manages to forge more unique links with the *Aeneid* than a predecessor (Lucan).

AUDIENCE

On April 6, 2013, Tesseract installed Google Analytics on its website, which has allowed tracking of site usage. As of May 23, 2014, the site had been visited by 3,335 unique users (IP addresses). Figure 6 illustrates the number of visits per city over this time period, indicating that users are located principally at college and university centers across North America, Europe, and Australasia. Excluding locations of team members, the ten cities with the greatest number of visits in this period are: New York, Basel, Cambridge UK, Cincinnati, New Haven, Oxford, London, Milan, Auckland, and Ann Arbor.

Figure 6. Geographical location by city of Tesseract site users April 6, 2013 – May 23, 2014.



The enthusiasm for Tesseract evident in these numbers has also been conveyed in messages sent to the team. In one representative response, a distinguished Dutch scholar, Vincent Hunink, wrote on December 20, 2013:

“Tesseract is one of the best new digital tools that have been made available to researchers of classical poetry. I find it truly amazing: Tesseract produces result for which, in the old days, scholars would have had to search indexes, consult their own memory, ask friends, and study for many years. Now just a quick and easy online search is enough. Moreover, Tesseract is not only very fast, but also produces results which one would never have thought of oneself, and so gives ‘new food for thought.’ I am thrilled that my research text is now part of the corpus, and I am sure Tesseract will help me to analyze it in the best ways possible, Thanks very much for doing this great job for the academic community!”

EVALUATION

In the course of this work, the Tesseræ team has received evaluative feedback from a variety of sources. These have included direct user responses to the website like that of Professor Hunink. Referee reports from submitted articles and conference abstracts, as well as audience members at conferences, have given the team valuable advice. Tesseræ also receives ongoing responses from its Advisory Board and other collaborators. Based on its NEH-funded work, the team won a grant from the Swiss government for a collaborative project with the University of Geneva from 2014 to 2016 for the study of intertextuality in Latin epic poetry of the Flavian era (late 1st century CE). This collaboration has provided the team with a rich source of critique. The initial stage of this project involved the first-ever workshop on intertextuality and digital methods, held at the Fondation Hardt near Geneva in February 2014, which gathered digital and traditional scholars from several countries to discuss further development of digital methodologies for intertextual study (<http://tesseræ.caset.buffalo.edu/blog/category/workshop/>). Tesseræ responses to these evaluations have included an expansion of its text base, addition the multi-text searching feature, development of semantic matching, and the conduct of collaborative discussions with other European partners.

LONG TERM IMPACT

User responses suggest that the project will continue to have a major impact the study and understanding of classical texts. By vastly speeding up the investigation of phrase parallels and other types of repetition and variation, Tesseræ for the first time enables comprehensive study of literary and linguistic influence beyond the level of individual words and exact quotation. It also allows scholars to look at textual relationships in new ways that supplement traditional approaches. Although it is not possible to say exactly what intellectual consequences will follow, it seems likely that Tesseræ methods will facilitate advancement in several areas. The comprehensive nature of Tesseræ search will contribute to fuller and more subtle appreciations of influence in passages of literary texts. The precise nature of Tesseræ search and results has already helped spur a new round of discussion of what exactly constitutes an allusion that recalls another text, as opposed to ordinary language reuse. Tesseræ capacities should also aid ongoing research into large-scale literary and intellectual trends, known to classicists as studies in the classical tradition or classical reception. This will be particularly true if Tesseræ search methods can be applied to a larger corpus of texts in late antique period, and extended to medieval, early modern, and modern literatures. Tesseræ's English search capacity shows that its approach can be extended to languages beyond Greek and Latin. Here and in other respects the availability of all Tesseræ code on gitub (<https://github.com/tesseræ/tesseræ>) and as a virtual machine image (http://www.tesseræ-dev.org/tesseræ_v4_reference.vdi) will allow other users to experiment with and adapt its approaches. One major consequence of existing Tesseræ work and any future efforts should thus be to bring to students and a larger public a deeper and richer understanding of language, literature, and intellectual heritage.

CONTINUATION OF THE PROJECT

In addition to the work described above, the Tesserae team has taken several steps in the grant period to lay the foundation for future developments. These include creation of a prototype Tesserae Version 4 built on a Solr database that will allow for faster searching and easier addition of text and language search capabilities (Appendices 2, 3). Additional benchmark sets of scholarly parallels have been compiled, within ancient Greek and across Greek and Latin, for further performance testing. And a documentation project has been initiated to provide information on the workings of Tesserae to users and developers beyond comments in the code.

This foundation will support efforts toward the further objectives of the project. These include refinement of individual search features, leading to the combination of these features into one multi-dimensional search that more closely approximates a fully informed scholarly reading of textual relations. They also include creation of semi-assistive user-initiated text addition and language extension so that users can study the texts and literatures of their choosing without having to download and modify computer code. With these capacities in place, Tesserae can proceed toward the ultimate goal of enabling the tracing of linguistic, literary, and intellectual movements over time across multiple languages. The Tesserae team is pursuing funding to support these efforts, including further opportunities via the NEH Office of Digital Humanities and other sources.

APPENDICES

Appendix 1. Main Tesseræ Search Page (Latin)

TESSERÆ

SEARCH
HELP
BLOG

BASIC SEARCH ADVANCED FEATURES ENGLISH GREEK OTHER TOOLS SOURCES


Basic Search


The Tesseræ project aims to provide a flexible and robust web interface for exploring intertextual parallels. Select two poems below to see a list of lines sharing two or more words (regardless of inflectional changes).

SOURCE TEXT

TARGET TEXT



[Compare Texts](#)


University at Buffalo
The State University of New York


**NATIONAL ENDOWMENT FOR THE
Humanities**

Tesseræ is a collaborative project of the University at Buffalo's Department of Classics and Department of Linguistics, and the VAST Lab of the University of Colorado at Colorado Springs.

This project is funded by the Office of Digital Humanities of the National Endowment for the Humanities and by the Digital Humanities Initiative at Buffalo.

Inquiries or comments about this website should be directed to Neil Coffee | Department of Classics | 338 MFAC | Buffalo, NY 14281
 tel: (716) 645-2154 | fax: (716) 645-2225

Appendix 2. Prototype Tesseract Version 4 search page

The image shows a web browser window with the following elements:

- Browser title: Tesseract NG - Basic Latin
- Address bar: tesserae.vast.uccs.edu/search/latin/basic
- Navigation menu: Tesseract NG, Home, Search, About, Contact
- Authentication buttons: Login, Password, Sign in
- Main content area: A light gray box containing the heading "Basic Latin Search" and a search form.

The search form includes:

- Source: A dropdown menu with the text "Choose a source text".
- Target: A dropdown menu with the text "Choose a target text".
- Compare: A button labeled "Compare".

Appendix 3. Prototype Tesserae Version 4 results page

Basic Latin Search Results

Source: "Bellum Civile" (Part 1 / 10) by Lucan
Target: "Aeneid" (Part 4 / 12) by Vergil
Query time: 1634 ms
Stop list: ego, et, hic, in, neque, non, qui, quis, sum, tu
Found: 432 matches (displaying 1 - 50)

1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 »

Rank	Target	Source	Common Terms	Score
1	et Numidae infreni cingunt et inhospita Syrtis;	Duc age per Scythiae populos, per inhospita Syrtis	inhospita-inhospitus, syrtis	8.229
2	dum memor ipse mei, dum spiritus hos regit artus.	Pallida regna petunt: regit idem spiritus artus	artus, rego, spiritus	7.794
3	pallentis umbras Erebi noctemque profundam,	Non tacitas Erebi sedes Ditisque profundi	erebus, profundo-profundus	7.710
4	"Inveni, germana, viam---gratare sorori---	Invenere viam, magnoque aeterna parantur	inuenio, uia	7.651
5	retia rara, plagae, lato venabula ferro,	Haereat, aut latum subeant venabula pectus,	fero-latus, uenabulum	7.530
6	Stant arae circum, et crines effusa sacerdos	Crinibus effusis toti praelate Comatae:	crinis, effundo-effusa	7.495
7	Ascanione pater Romanas invidet arces?	Sedibus exsiluere Patres, invisaque belli	inuideo, pater	7.450
8	cur mea dicta neget duras demittere	Et coelum Mars solus habet. Cur signa	cur, meo	7.442

Appendix 4: Representative publication: Forstall, C., N. Coffee, T. Buck, K. Roache, and S. Jacobson. (2014). "Modeling the Scholars: Detecting Intertextuality through Enhanced Word-Level N-Gram Matching." *Literary and Linguistic Computing (LLC)*. doi: 10.1093/llc/fqu014. Additional publications listed in bibliography above.

Modeling the scholars: Detecting intertextuality through enhanced word-level n-gram matching

Christopher Forstall, Neil Coffee, Thomas Buck, Katherine Roache and Sarah Jacobson

Department of Classics, University at Buffalo, SUNY, Buffalo, NY, USA

Abstract

The study of intertextuality, or how authors make artistic use of other texts in their works, has a long tradition, and has in recent years benefited from a variety of applications of digital methods. This article describes an approach for detecting the sorts of intertexts that literary scholars have found most meaningful, as embodied in the free Tesseract website <http://tesseract.caset.buffalo.edu/>. Tests of Tesseract Versions 1 and 2 showed that word-level n-gram matching could recall a majority of parallels identified by scholarly commentators in a benchmark set. But these versions lacked precision, so that the meaningful parallels could be found only among long lists of those that were not meaningful. The Version 3 search described here adds a second stage scoring system that sorts the found parallels by a formula accounting for word frequency and phrase density. Testing against a benchmark set of intertexts in Latin epic poetry shows that the scoring system overall succeeds in ranking parallels of greater significance more highly, allowing site users to find meaningful parallels more quickly. Users can also choose to adjust both recall and precision by focusing only on results above given score levels. As a theoretical matter, these tests establish that lemma identity, word frequency, and phrase density are important constituents of what make a phrase parallel a meaningful intertext.

Correspondence:

Neil Coffee, Department of Classics, 338 MFAC, University at Buffalo, SUNY, Buffalo, NY 14261, USA.

Email:

ncoffee@buffalo.edu

Intertextuality is an important part of linguistic and literary expression, and has consequently been the object of sustained scholarly attention from antiquity onward. The definition of intertextuality has been much debated, but it is commonly understood as the reuse of text where the reuse itself creates new meaning or has expressive effects, distinct from the unmarked reuse of language.¹ In recent years, digital humanists have taken various approaches to detecting forms of intertextuality.² This article reports on an advance in automatic detection of a subset of intertextuality, namely,

instances of text reuse determined by scholars of classical Latin to bear literary significance. This work was carried out by the Tesseract Project research group, whose approach is distinctive for combining (1) efforts to use digital methods to emulate scholarly intertextual reading, (2) corresponding procedures for testing results against scholarship, and (3) an evolving free website for intertextual detection and analysis (<http://tesseract.caset.buffalo.edu/>).³

Tesseract Version 1 matched exact word strings within moveable word windows. Version 2 added

the capacity for lemma matching by line or sentence. Deployment of these versions on the Tesseract website provided scholars with a means of automatically finding phrase parallels that were candidates for instances of intertextuality. A previous test comparison of two Latin epic poems demonstrated that the word-level n-gram matching used by both versions could detect the majority of intertexts identified by scholars.⁴ Word-level n-gram matching means matching by a certain number of words, in our case a minimum of two in each text that may or may not be adjacent, as opposed to matching by longer strings of characters including spaces, or other possible units. The search lacked precision, however, so intertexts lay undifferentiated in long lists of candidate parallels, the vast majority of which were not meaningful. Version 3 now provides a filtering function that ranks parallels by significance, making it substantially easier to find those of greater potential interest. The Version 3 search algorithm is now the default method for searching the newly expanded corpus of Latin, ancient Greek, and English available on the Tesseract site. This article describes the performance of a Version 3 search.

1 Methodology

A Tesseract search proceeds in two stages. In the first stage, the search identifies all instances where a given unit in one selected text shares at least two words with a unit in another selected text. The units can be either lines of poetry or ‘phrases’, where a phrase is equivalent to a sentence or text demarcated by a semicolon or colon. Words can be matched by the exact word form (for Latin, ‘canit’, ‘she sings’ = ‘canit’) or dictionary headword (‘canit’, ‘she sings’ = ‘cecini’ ‘I sang’, as forms of the headword ‘cano’). Users can choose to exclude common words using a stop list, the size and source of which (one text, both texts, or the corpus) can be adjusted. This first stage of the Version 3 search is conceptually identical to that of previous versions, but incorporates some modifications to the code that produce a greatly increased number of phrase matches.

$$\text{score} = \ln \left(\frac{\sum \frac{1}{f(t)} + \sum \frac{1}{f(s)}}{d_t + d_s} \right)$$

Fig. 1. Equation for Tesseract Version 3 scoring system

To achieve better precision than that provided by the stop list alone, Version 3 introduces a second stage scoring system that ranks results by two additional criteria: the relative rarity of the words in the phrases shared by the two texts (‘word frequency’), and the proximity of the shared words in each text (‘phrase density’). We privileged word frequency because we observed that, with notable exceptions, phrases identified by scholars as intertexts consist of words that are relatively rare in their contexts. We privileged phrase density because we observed that scholars generally found intertexts to consist of compact rather than diffuse collocations. The equation given in Fig. 1 represents our attempt to express the relationship of these criteria as a measure of intertextual significance. The inputs to this equation are the frequency of each matching word in its respective text and the distance between the two most infrequent words in each of the two phrases. The output is a prediction of interpretive significance generally falling between 2 and 10. The effect of the equation is that, for a given parallel, the rarer the shared words are, and the closer together in their respective texts, the higher its score will be.

2 Testing

2.1 Search stage 1: Phrase matching

To assess the Version 3 search, we conducted a test that compared our results with a benchmark set of scholarly parallels between two Latin epic poems considered to have a high level of intertextual relation, Vergil’s *Aeneid* (9,896 lines of hexameter verse) and book 1 of Lucan’s *Civil War* (695 lines of hexameter verse). We performed the search using the Tesseract Corpus-wide search interface (<http://tesseract.caset.buffalo.edu/multi-text.php>, Fig. 2). The interface allowed us to generate a list of parallel passages with common phrases, and also to see

SEARCH
HELP
BLOG

BASIC SEARCH ADVANCED FEATURES ENGLISH GREEK OTHER TOOLS SOURCES

Corpus-wide Search

This experimental search cross-checks your results against all other texts in the corpus. This will allow you to see whether a particular parallel is unique to your two selected works, or whether there is a broader precedent for the repeated expression.

SOURCE:

TARGET:

UNIT:

FEATURE:

NUMBER OF STOP WORDS:

STOPLIST BASIS:

MAXIMUM DISTANCE:

DISTANCE METRIC:

DROP SCORES BELOW:

FILTER MATCHES WITH OTHER TEXTS:

TEXTS TO SEARCH: All Prose Verse

Caesar - De Bello Gallico
Caesar - De Bellum Civile
Caesar Augustus - Res Gestae Divi Augusti
Catullus - Carmina

Tesseract is a collaborative project of the University at Buffalo's Department of Classics and Department of Linguistics, and the VAST Lab of the University of Colorado at Colorado Springs.

This project is funded by the Office of Digital Humanities of the National Endowment for the Humanities and by the Digital Humanities Initiative at Buffalo.

Inquiries or comments about this website should be directed to Neil Coffee | Department of Classics | 338 MFAC | Buffalo, NY 14261 tel: (716) 645-2154 | fax: (716) 645-2225

Fig. 2. Screenshot of Tesseract Corpus-wide Search interface used in testing

where else in the corpus those phrases appeared, as an aid to the hand-ranking process described below. We selected relatively unrestricted settings for our search to capture the greatest number of meaningful results. We compared texts by phrases rather than

lines because phrases were generally longer and so could find a broader range of intertexts. We searched by lemma rather than exact word, at the cost of some false matches,⁵ to allow for the detection of intertexts with identical roots but different

Table 1 Tesserae scale for ranking significance of intertextual parallels, from *Coffee et al., 2012*, pp. 392–8

Type	Characteristics	Significance categories	
5	High formal similarity in analogous context.	Meaningful	Interpretable
4	Moderate formal similarity in analogous context; or High formal similarity in moderately analogous context.	Meaningful	Interpretable
3	High/moderate formal similarity with very common phrase or words; High/moderate formal similarity with no analogous context; or Moderate formal similarity with moderate/highly analogous context.	Meaningful	Not interpretable
2	Very common words in very common phrase or Words too distant to form a phrase.	Not-meaningful	Not interpretable
1	Error in discovery algorithm, words should not have matched.	Not-meaningful	Not interpretable

forms, a measure necessary for a highly inflected language like Latin. We chose a stop list that excluded only the ten most common lemmata in *Civil War 1* and the *Aeneid* taken together. The stop list words were ‘et’, ‘qui’, ‘quis’, ‘in’, ‘hic’, ‘sum’, ‘tu’, ‘per’, ‘neque’, and ‘fero’.⁶ The resulting search generated a list of 23,617 phrase parallels between the *Aeneid* and *Civil War 1*, each with an automatically assigned score. Comparison of these parallels with the benchmark set showed that the search captured 62% of the intertexts recorded by scholars.⁷

We further attempted to determine if the search had revealed new meaningful intertexts. This required assessing the quality of the parallels returned in the search that had not been noted by scholars. For the assessment, we used a hand-ranking scale we had previously developed for this purpose (given in *Table 1*).⁸ The scale has five ranks, from least to greatest significance for the literary interpreter. For testing purposes, we concentrated principally on whether parallels passed one of the two thresholds. To clear the first threshold, a phrase parallel needed to have marked language, and therefore be of potential interest for its artistry. This standard excluded both erroneous matches (type 1) and instances of unmarked, ordinary language (type 2). The determination as to whether a given phrase parallel had marked language was made in part through consideration of how often it appeared elsewhere in the corpus, as indicated by results from the Corpus-wide Search function. All other things being equal, a phrase parallel between the two texts that was rare in the corpus was considered of greater

Table 2 Total number of Version 3 results and number hand-ranked

Automatic Tesserae score	Total in test set	Number sampled (≈5%)
10	1	1
9	32	3
8	342	19
7	1,721	86
6	6,314	316
5	10,004	507
4	4,942	243
3	259	17
2	2	2

interest than a parallel common in the corpus.⁹ Parallels passing this threshold were awarded a minimum score of 3 and deemed, in our terms, ‘meaningful’. To clear the second threshold, a phrase parallel needed, in addition to marked language, sufficient contextual analogy between its two passages that a reader could interpret significance in their interaction.¹⁰ Parallels passing this threshold were awarded a minimum score of 4 and deemed, in our terms, ‘interpretable’.

Evaluating all the parallels in the test set was prohibitive, so we chose instead to rank a random sample consisting of 5% of the results at each automatic score level, amounting to 1,194 parallels, distributed as shown in *Table 2*.¹¹ The resulting quality distribution of the sample set was as follows, from most to least meaningful: type 5: 7 (1% of results sampled), type 4: 39 (3%), type 3: 145 (12%), type 2: 879 (74%), and type 1: 124 (10%). *Fig. 3* shows

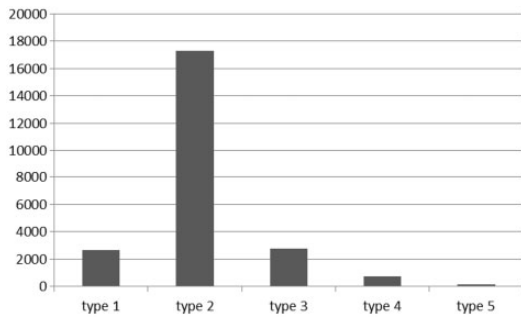


Fig. 3. Projected distribution by type of all 23,617 *Aeneid*–*Civil War* candidate parallels, prior to application of scoring algorithm

these proportions projected onto the full set of 23,617 results returned. Based on this projection, between Lucan’s first book and the *Aeneid* we should expect to find 2,770 instances of phrase parallels that constitute more or less distinctive generic language (type 3) and 899 interpretable intertexts (739 type 4 and 160 type 5). Although this may appear to be an unduly large number of intertexts to be found in 695 hexameter lines, two considerations make it seem less so. First, we counted every set of parallel loci between the two texts separately. So when a given locus in the *Civil War* had parallels with multiple passages in the *Aeneid*, each of these was counted as separate parallels. The 899 interpretable intertexts are thus constituted by fewer than 899 separate loci in the *Civil War*. Second, a high level of interaction is not surprising for verse (hexameter) and genre (epic) traditions generally regarded as densely intertextual.

Fig. 4 illustrates the projected recall of meaningful parallels (types 3–5) from our test in relation to those recorded by commentators, showing that Version 3 is projected to substantially increase the number of recognized meaningful intertexts. Figs 5 and 6 illustrate the recall of interpretable parallels (types 4–5) produced by the Versions 1 and 2 combined (Fig. 5) and the projected recall produced by Version 3 (Fig. 6), both again in relation to those recorded by commentators. Comparison of Figs 5 and 6 illustrates the significant improvement in recall of Version 3 over even the combination of the two previous Tesseract versions. Overall, the projections from our sample suggest that Version 3

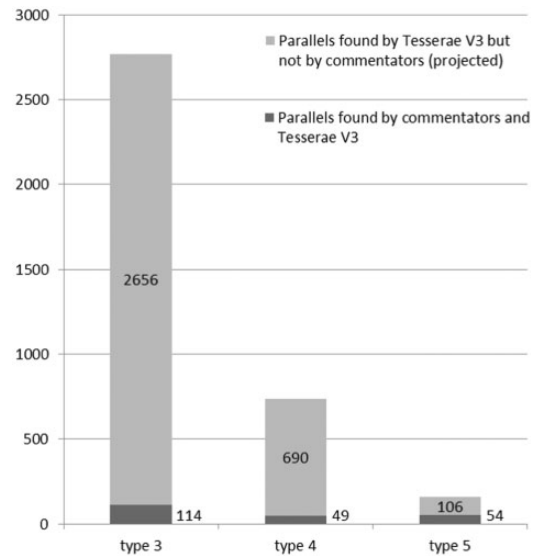


Fig. 4. Numbers of meaningful (types 3–5) parallels between Lucan’s *Civil War* 1 and Vergil’s *Aeneid* found by Tesseract Version 3 (projected) and by commentators. Projected figures are produced by projecting the quality scores for a test sample over the entire test set

improves considerably upon previous versions in discovering meaningful and interpretable intertexts, including many that have not previously been recorded.¹²

An example of these results is a parallel found in our Tesseract Version 3 test sample, but neither noted by commentators nor discovered with previous Tesseract versions, which was assigned an automatic score of 7 and a hand-rank of 5. In *Civil War* 1, Lucan narrates the abandonment of Rome at the advent of Caesar, comparing the panicked reaction of Romans with the fear of Hannibal generations earlier:

non secus ingenti bellorum Roma tumultu
concutitur, quam si Poenus transcenderit
Alpes
Hannibal.

(*Civil War* 1.303–5)

Rome was rocked by the massive upheaval of war,
no less than if the Carthaginian should cross the Alps.

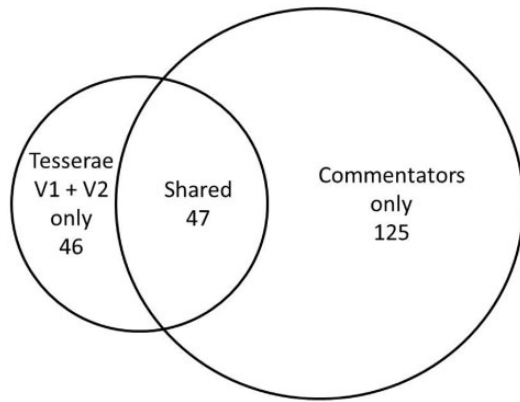


Fig. 5. Unique interpretable (types 4 and 5) parallels between Lucan's *Civil War* 1 and Vergil's *Aeneid* found by Tesseractae Versions 1 and 2, commentators, and both, as reported in *Coffee et al.*, 2012, p. 398

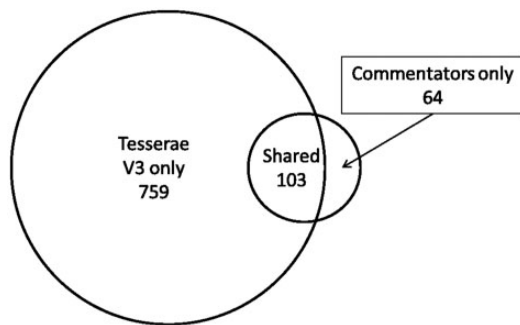


Fig. 6. Unique interpretable (types 4 and 5) parallels between Lucan's *Civil War* 1 and Vergil's *Aeneid* found by Tesseractae Version 3 (projected), commentators, and by both combined

This passage bears some similarity to an episode in the underworld narrative of *Aeneid* book 6. In the *Aeneid* episode, set in Rome's mythical prehistory, Aeneas's father Anchises looks forward over the centuries to the birth of the great general Marcellus, who saved Rome from the Carthaginians in the First Punic War and fended off Gallic incursions:

hic rem Romanam, magno turbante tumultu,
sistet, eques sternet Poenos Gallumque
rebellem,
tertiaque arma patri suspendet capta Quirino.
(*Aeneid* 6.857-9)

This [Marcellus] will keep Roman affairs
standing
When it is threatened by great upheaval,
He will lay low the Carthaginian horsemen,
the rebellious Gaul,
He will offer a captured general's arms to
Father Quirinus,
For only the third time ever.

There are other sources, beyond this Vergilian passage, that Lucan may be drawing upon and alluding to, including some with lines that also end with the word 'tumultu'.¹³ But several features make for a distinctive recollection of the description of Marcellus by Anchises: the pairing of Rome and upheaval (tumultu) in the same line, the enjambement of the verb for the first line at the beginning of the second, and the placement of a form of the word "Carthaginian" ('Poenus/-os') in the same metrical position before a caesura, in a line with identical metrical rhythm.¹⁴

The similarity of language features in the two passages meets our requirements for a meaningful intertext. There is also sufficient analogy in context to make the parallel interpretable. Both passages deal overall with the possibility of the destruction of Rome through foreign invasion and the corresponding Roman response (or lack thereof). The analogy invites the reader's interpretation. We can thus observe that the echoing of *Aeneid* 6 in this *Civil War* passage figures Romans as not only fleeing from Caesar as they might have done from Hannibal, but also fleeing as Marcellus did not when faced with an earlier Carthaginian threat in the First Punic War. The resonance compounds Lucan's criticism of Romans for deserting their city.¹⁵

2.1.1 Digital discovery and intertextual theory

Having offered a reading of an intertext discovered by a digital method, we have entered into a contested area of literary theory. We will, therefore, pause for a moment the course of explanation of our testing to provide a theoretical justification of our approach. Although the formal and thematic similarities between the passages above are clear, we may still ask whether they are sufficient to prompt the reader of Lucan to recall the Vergilian

lines. This example was chosen from among others, some with closer parallelism, precisely to raise this issue. A full treatment would require addressing long-standing questions about the nature of intertextuality: how reader recollection works; what formal features, at what thresholds, prompt recollection; and how these might have varied between ancient and modern reading cultures. In the longer view, we believe these questions can best be answered through continuing research involving formal modeling with digital methods, the exercise of scholarly judgment, and consideration of cognitive processes. For the moment, however, we attempt to provide a brief explanation of how we understand the problem of reader recognition of intertexts, and how that guides the work explained here.

Modern intertextual theory provides essentially two options for how we view our interpretive efforts.¹⁶ Either we are trying to understand how some original reader would have read a text (the traditional philological project),¹⁷ or if we believe the original perspective is fundamentally inaccessible, we are instead constructing a reading that is avowedly from our current perspective, but addresses and incorporates the history of other readings.¹⁸ In either case, for an intertext to exist, there must be a moment of recognition, actual or supposed, on the part of a reader. If our exegesis of the Lucan and Vergil passages sufficed to induce that moment of recognition in the modern reader, or suggest that it would have occurred for an ancient reader, the parallel can be said to constitute an intertext.

Let us assume for the moment, however, that no such recognition took place. The passages indeed share a distinctive set of features: Rome, Carthage, and the ending tag *tumultu recur* as a constellation in the Roman tradition, and the metrical expression stamped on this idea by Vergil brings an additional specificity that apparently set a template for Lucan. But suppose, nonetheless, that neither the ancient nor the modern reader would have seen a parallel here unprompted, and our explanation has failed to animate a moment of felt recognition. What do we do with a textual similarity that is real but unrecognized?

The traditional answer has been that the burden still lies with the interpreter. Un(re)marked

repetition is regular language, and marked repetition has meaning and literary significance. For an instance of text reuse to be meaningful, the critic must show that it is marked and ‘make it mean’ to the reader.¹⁹ So much is certainly true. But it is also true that what counts as marked has changed over time, at least partly in response to technology. So, for example, it is commonly observed that ancient literary criticism lagged ancient literary practice. As Paul Oscar Kristeller writes,

We have to admit the conclusion . . . that ancient writers and thinkers, though confronted with excellent works of art and quite susceptible to their charm, were neither able nor eager to detach the aesthetic quality of these works of art from their intellectual, moral, religious or practical function or content.²⁰

The rich thematic interplay between Horace’s varied Odes seems of great significance to modern classical scholars, but, as far as we can tell, was not the sort of thing that drew critical attention in antiquity, however advanced critical practice was in other ways. The change in the modern era would seem to be at least partly attributable to the advent of print, which allowed for many more readers to read texts and study them closely. Subtleties in the interplay of Horace’s Odes were now discerned, with the understanding that they had always had effects on Horace’s readers, even if they previously lay unexplained.

If perceptions of what is of interpretive interest have changed over the millennia, perceptions of what constitutes an intertext have changed in just the last few decades. A previous study has indicated that, in response to the availability of simple computer text searches, classical commentary writers have been expanding their definition of intertextuality to include less salient correspondences.²¹ We would suggest that the more developed automated methods described here represent a further step in this direction. They have the potential to bring into the realm of interpretation, and so register as ‘marked’, intertexts that were always meaningful within both the text and the larger literary tradition, but did not rise to the level of conscious acknowledgement on the part of readers and critics.

Table 3 Rates of precision for various sources in *Civil War 1–Aeneid* test search

Quality (Rank)	Commentators (%)	Version 1 (exact form match) (%)	Version 2 (lemma match) (%)	Version 3 (lemma match) (%)
Meaningful (3–5)	86	53	11	17
Interpretable (4–5)	41	27	2	5

Version 3 precision rates are prior to application of the secondary scoring system.

The sample parallel we have given above is in fact one we have considered an intertext, though it lies toward the subtle end of the section of the language spectrum we have considered marked. At the other end are the sorts of parallels that are more universally acknowledged to be intertexts or allusions. The relatively broad conception of intertextuality represented by this band of the language spectrum underlies the work presented here.

2.2 Search stage 2: Scoring

Having demonstrated that Tesseract Version 3 can capture intertexts with some success, let us return to how recognized intertexts were identified among all the phrase parallels returned, the majority of which were not meaningful. This part of the testing involved evaluating how the scoring system developed for Version 3 could improve precision.

Our procedure for calculating precision was to divide the number of meaningful (types 3–5) or interpretable (types 4 and 5) results in our test set by the total number of results of all types (1–5). To provide a baseline, we began by calculating precision for our sample set before engaging the automatic scoring system, with results illustrated in Table 3. The published commentaries that were our model naturally had a very high rate of precision: 86% of the parallels they record are meaningful, and the remaining 14% are instances of ordinary (metrically compatible) language (type 2). For interpretable parallels (types 4–5), Version 1 gave the highest precision among Tesseract versions, as it matched by exact words, whereas the lemma matching of Version 2 and Version 3 without the scoring system, though capturing a broader range of parallels, had lower precision.

We then tested how effective the automatic scoring system was at identifying the most meaningful

Table 4 Comparison of automatic scores and hand-ranks for Tesseract Version 3 sample set of parallels between *Civil War 1* and *Aeneid*

	Automatic score		Hand rank type				
	Total		5	4	3	2	1
10 (highest)	1			1			
9	3			1	2		
8	19	2	3	6	8		
7	86	5	10	20	44	7	
6	316			20	79	184	33
5	507			4	31	412	60
4	243				7	214	22
3	17					15	2
2 (lowest)	2						2
Total	1,194	7	39	145	879	124	

parallels. Table 4 shows how automatic scores in our sample set correspond to hand-rankings. If we average the automatic scores at each hand-rank level, we find the correlation illustrated in Fig. 7. As this figure shows, the scoring system succeeds in distinguishing the more meaningful intertexts given higher hand ranks by assigning them higher scores. In other words, the automatic scoring system replicated the trends in assessment of intertexts performed by human readers.

To get a more concrete sense of the performance of Version 3 search, we further assessed our results in terms of recall and precision. Figs 8 and 9 illustrate how recall and precision of meaningful (types 3–5, Fig. 8) and interpretable (types 4–5, Fig. 9) parallels vary when we discard results below certain score levels. In both cases, discarding results with increasingly higher score levels steadily increases the proportion of interpretable or meaningful intertexts in the remaining set, leading toward

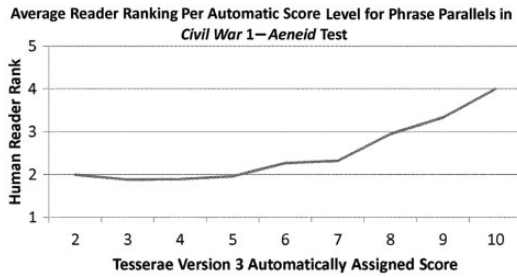


Fig. 7. Correlation of Tesseract automatic scoring system with hand ranking of intertextual significance

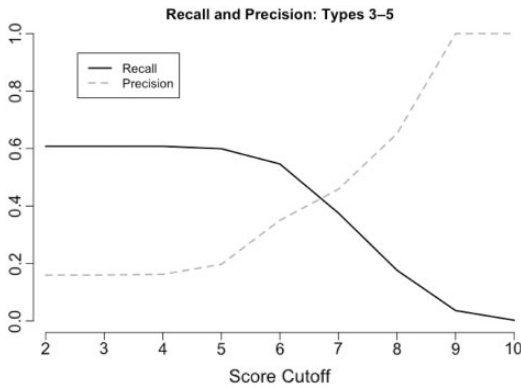


Fig. 8. The effects of score cutoff on recall and precision rates for meaningful (types 3–5) parallels

consistently higher precision. Raising the score threshold also reduces recall, however, by progressively eliminating meaningful and interpretable intertexts. At this stage of development, then, the scoring system may best be used to allow the user to filter results according to his or her needs. For example, by discarding all parallels below an automatic score level of 6 in our test set, the user can eliminate nearly three-quarters (727/1,003) of the non-meaningful types 1 and 2, yet retain some three-quarters of type 3 parallels (107/145), 90% (35/39) of type 4 parallels, and all type 5 parallels. On the other hand, those who wished to get only a high quality sample could choose to consider results only at a higher score level.

Another way to choose a score cutoff level would be to consider the combined measure of recall and precision known as an F-measure. F-measure is a

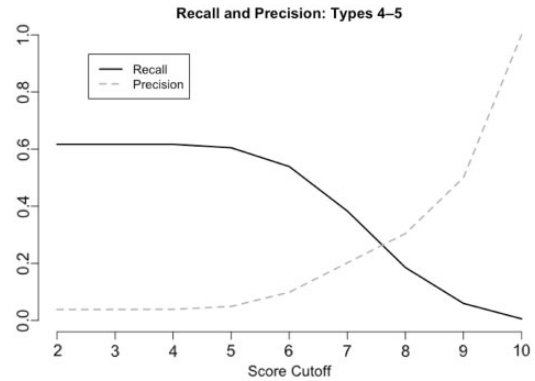


Fig. 9. The effect of score cutoff on recall and precision rates for interpretable (types 4 and 5) parallels²³

term for a combined measurement of recall and precision. For our F-measure assessment, we used the following equation:²²

$$F_1 = \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \times 2$$

Fig. 10 illustrates the F-measure scores produced when we progressively discard results below increasingly higher automatic score levels. Although the results fall considerably below the perfect F-measure of 1 at any score cutoff level, this measurement does suggest that those interested in a relatively economical investigation into meaningful parallels would be best served by investigating those at a score level of 6 or above, while those interested in a range more likely to be interpretable could investigate those at a score level of 7 or above.

3 Conclusions

The Version 3 algorithm behind the current default Tesseract search is designed to identify meaningful intertexts through word-level n-gram lemma matching, word frequency, and phrase density. Our tests demonstrate that Version 3 search has considerable success in identifying intertexts in a sample comparison from two Latin epic poems. It gives higher scores to phrase parallels of greater interest, pointing users to those more likely to constitute an intertext. With relatively unrestricted

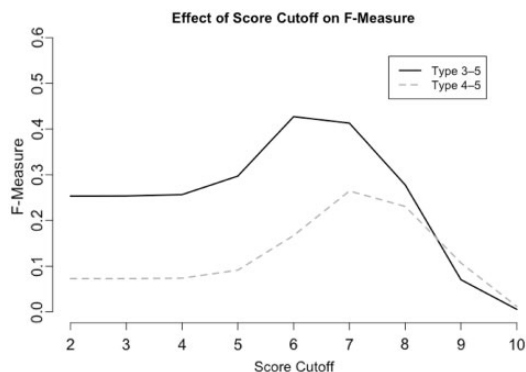


Fig. 10. Effects of score cutoff on F-measure for type 4–5 parallels and for type 3–5 parallels

settings, it can identify a majority of the intertexts recorded by scholars. These results, along with our further informal experimentation, suggest Version 3 can be similarly used for other comparisons of Latin texts in our corpus, as well as for comparisons of ancient Greek and English texts, making Tesseract search a substantial aid to intertextual study. Our results also suggest that the three criteria of lemma identity, word frequency, and phrase density are important formal components of what constitutes an intertext. When scholars identify two or more passages as intertextual, they may be using the presence or absence of these three features as implicit, if not explicit, criteria.

REFERENCES

- Allen, G. (2011). *Intertextuality*. London: Routledge.
- Bamman, D. and Crane, G. (2008). The Logic and Discovery of Textual Allusion. *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*. Marrakesh.
- Barchiesi, A. (ed.) (2001). *Speaking Volumes: Narrative and Intertext in Ovid and Other Latin Poets*. London: Duckworth.
- Ben-Porat, Z. (1976). The poetics of literary allusion. *A Journal for Descriptive Poetics and Theory of Literature*, 1: 105–126.
- Berti, M. (2013). Collecting quotations by topic: Degrees of preservation and transtextual relations among genres. *Ancient Society*, 43: 269–88.
- Büchler, M., Geßner, A., Eckart, T., and Heyer, G. (2010). Unsupervised detection and visualization of textual reuse on ancient greek texts. *Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science*, 1: 2.
- Coffee, N. (2012). Intertextuality in Latin Poetry. In Clayman, D. (ed.), *Oxford Bibliographies in Classics*. New York: Oxford University Press.
- Coffee, N., Koenig, J. P., Poornima, S., Forstall, C. W., Ossewaarde, R., and Jacobson, S. L. (2013). The Tesseract Project: Intertextual analysis of latin poetry. *Literary and Linguistic Computing*, 28: 221–8.
- Coffee, N., Koenig, J. P., Poornima, S., Ossewaarde, R., Forstall, C., and Jacobson, S. (2012). Intertextuality in the digital age. *Transactions of the American Philological Association*, 142: 381–419.
- Conte, G. B. (1986). *The Rhetoric of Imitation: Genre and Poetic Memory in Virgil and Other Latin Poets*. Ithaca: Cornell University Press.
- Edmunds, L. (2001). *Intertextuality and the Reading of Roman Poetry*. Baltimore: Johns Hopkins University Press.
- Farrell, J. (2005). Intention and intertext. *Phoenix*, 59: 98–111.
- Ford, A. (2013). Review of Ineke. In Sluiter and Ralph M. Rosen (eds), *Aesthetic Value in Classical Antiquity*. *Bryn Mawr Classical Review* 2013.09.26.
- Fowler, D. (2000). *Roman Constructions: Readings in Postmodern Latin*. New York: Oxford University Press.
- Genette, G. (1997). *Palimpsests: Literature in the Second Degree*. Lincoln: University of Nebraska Press.
- Heitland, W. and Haskins, C. (1887). *M. Annaei Lucani Pharsalia*. London: G. Bell.
- Hinds, S. (1998). *Allusion and Intertext: The Dynamics of Appropriation in Roman Poetry*. New York: Cambridge University Press.
- Hutchinson, G. O. (2013). *Greek to Latin: Frameworks and Contexts for Intertextuality*. Oxford: Oxford University Press.
- Irwin, W. (2001). What Is an Allusion?. *The Journal of Aesthetics and Art Criticism*, 59: 287–97.
- Kristeller, P. O. (1951). The modern system of the arts: A study in the history of aesthetics. *Journal of the History of Ideas*, 12: 496–527. Reprinted in idem. 1980. *Renaissance Thought and the Arts: Collected Essays*: 174.

- Kristeva, J.** (1986). Word, dialogue and novel. In Moi, K. (ed.), *The Kristeva Reader*. New York: Columbia University Press, pp. 34–61.
- Martindale, C.** (1993). *Redeeming the Text: Latin Poetry and the Hermeneutics of Reception*. Cambridge: Cambridge University Press.
- Martindale, C.** (2013). Response to forum debate. *Classical Receptions Journal*, 5: 246–51.
- Martindale, C. and Thomas, R. F.** (eds), (2006). *Classics and the Uses of Reception*. Oxford: Blackwell.
- Pucci, J.** (1998). *The Full-Knowing Reader: Allusion and the Power of the Reader in the Western Literary Tradition*. New Haven: Yale University Press.
- Ricks, C. B.** (2002). *Allusion to the Poets*. Oxford: Oxford University Press.
- Roche, P.** (2009). *Lucan: De bello civili. Book 1*. Oxford: Oxford University Press.
- Russell, D. A.** (1981). *Criticism in Antiquity*. London: Duckworth.
- Thomas, R.** (1986). Virgil's Georgics and the art of reference. *Harvard Studies in Classical Philology*, 90: 171–98.
- Thomas, R.** (1999). *Reading Virgil and His Texts: Studies in Intertextuality*. Ann Arbor: University of Michigan Press.
- Thompson, L. and Bruère, R.** (1968). Lucan's use of Vergilian reminiscence. *Classical Philology*, 63: 1–21.
- Trillini, R. H. and Quassdorf, S.** (2010). A 'Key to All Quotations?': A Corpus-based Parameter Model of Intertextuality. *Literary and Linguistic Computing*, 25: 269–86.
- Viansino, G.** (1995). *Marco Annaeo Lucano: La Guerra Civile (Farsaglia) libri I-V*. Milan: Mondadori.
- Wills, J.** (1996). *Repetition in Latin Poetry: Figures of Allusion*. Oxford: Oxford University Press.
- surveying these and other works is provided by [Coffee 2012](#).
- 2 [Bamman and Crane, 2008](#); [Büchler et al., 2010](#); [Trillini and Quassdorf, 2010](#); [Berti 2013](#).
 - 3 The complete code is available at <https://github.com/tesseract/tesseract>.
 - 4 [Coffee et al., 2012, 2013](#).
 - 5 Lemmatization is at present unsupervised. In cases where an inflected form is ambiguous (e.g. Latin 'bello' could mean 'war' or 'handsome'), it is allowed to match on any of the possible lemmata.
 - 6 Users can replicate the search discussed here by using the following parameters on the Corpus-wide Search page. Source: Vergil's *Aeneid*; target: Lucan's *Bellum Civile* book 1; unit: phrase; feature: lemma; number of stop words: 10; stop list basis: target + source; maximum distance: 50 words; distance metric: frequency; drop scores below: 0; filter matches with other texts: no filter; texts to search: all. The original distance metric counted both words and non-word tokens such as spaces and punctuation marks. Because word and non-word tokens generally alternate, one should reduce this number by half to estimate the number of intervening words in the 'sparsest' parallels. The current, revised metric counts only words, and produces comparable results when set to a maximum of 23.
 - 7 Our list of scholarly parallels was compiled from the Lucan commentaries of [Heitland and Haskins, 1887](#); [Thompson and Bruère, 1968](#); [Viansino 1995](#); and [Roche 2009](#). These were supplemented by a list of parallels not recorded by scholars that had been generated in previous testing and graded according to the scoring system described below. Note that the 62% recall reported here excluded matches on the list of stop words, as well as phrases in which matching words were very far apart (see below). Without these restrictions, recall would be higher, around 72%, though at the expense of substantially decreased precision.
 - 8 For a full explanation of the scale, see [Coffee et al., 2012](#), pp. 392–8.
 - 9 This criterion is meant to exclude very common collocations. For example, forms of the expression 'lift oneself up' ('se tollere') occur at *Civil War* 1.142 and *Aeneid* 2.699, but also in 82 other texts in our corpus, confirming that it is a common expression and uninteresting in and of itself. At the same time, classicists have recognized instances where an intertext in fact becomes more meaningful by having been repeated, generally with variation, in multiple locations. A distinction is commonly made between a parallel

Notes

- 1 In the area of Latin literature, which we focus on here, key works on intertextuality include [Conte 1986](#), [Martindale 1993](#), [Wills 1996](#), [Hinds 1998](#), [Pucci 1998](#), [Barchiesi 2001](#), [Edmunds 2001](#), [Farrell 2005](#), and [Hutchinson 2013](#). More general studies include [Ben-Porat 1976](#), [Genette 1997](#), [Irwin 2001](#), [Ricks 2002](#), and [Allen 2011](#). The term 'intertextuality' was coined by [Kristeva 1986](#), originally to describe the full range of interactions between the sign systems of a culture. An annotated bibliography on intertextuality

- consisting of two (or few) textual loci, called an ‘allusion’ or ‘intertext’, and a set of multiple occurrences with close similarities, called a ‘topos’. Homer initiates the ‘many mouths’ topos by declaring that he could not name all the Greek forces at Troy even if he had ten tongues, ten mouths, an unstoppable voice, and a heart of bronze (*Il.* 2.488–90). The Roman poets Lucretius, Vergil, Ovid, Persius, Silius Italicus, Statius, and Valerius Flaccus later pick up and rework the conceit into a commonplace (Hinds 1998, pp. 34–47). Overall, it would seem that the sense of a continuum from fewest to greatest number of phrase repetitions underlies the qualitative labels allusion/intertext, topos, generic language, and ordinary language, even if there is more to these categories than phrase repetition. It may be possible to incorporate phrase frequency into a future scoring system, in which case, this issue would need closer examination. For this test, phrase frequency was considered by human evaluators, which allowed for the possibility of discrimination between these types.
- 10 Our criteria for meaningful and interpretable parallels draw upon existing theoretical distinctions. Fowler 2000, p. 122 has written that the two fundamental criteria for an intertext are ‘markedness and sense’. Markedness is the quality that makes a parallel ‘stand out’ and makes it ‘special’. We take Fowler’s criterion of markedness to refer principally, if not exclusively, to the sort of distinctive shared language features required to make a parallel ‘meaningful’ in our terms. Fowler further explains that for a parallel to have ‘sense’, the interpreter must ‘make it mean’. Fowler’s criterion of ‘sense’ corresponds to our requirement that an ‘interpretable’ parallel have a contextual similarity in the parallel passages that generates significance.
 - 11 Of the parallels thus selected, 1,078 had already been hand-ranked in previous testing. The remaining 116 were ranked for the first time in this study. The previously ranked and newly ranked results were then combined to make a sample set where each parallel had both an automatic score and a hand rank. All results were collated into a spreadsheet that is posted on the Tesseræ blog (<http://tesseræ.caset.buffalo.edu/blog/benchmark-data/> under ‘Tesseræ 2012 Benchmark’).
 - 12 The total number of commentator parallels is lower in the Version 3 test because review of the earlier commentator parallels for the current test found some that were judged duplicates.
 - 13 In his comment on the Lucan passage, Roche 2009, p. 248 ad 1.303–4 does not mention this possible Vergilian parallel, but observes that ‘the allusion to Hannibal is compounded by the intertextual allusion to Lucretius’ description of the effects of the Punic war at 3.834f. ‘omnia cum belli trepido concussa tumultu/horrida contremuere sub altis aetheris altis’. Horace’s *Carmina* 4.4.45–52 has a similar combination of thought and language: ‘Romana pubes crevit et impio/vastata Poenorum tumultu/fana deos habuere rectos,/dixitque tandem perfidus Hannibal...’ The ancestor of all expressions of upheaval in Africa with tumultu at line-end would seem to be Ennius’s ‘Africa terribili tremit horrida terra tumultu’ (*Annales* 309 Skutsch), a line that stuck in Cicero’s memory (*De oratore* 3.42).
 - 14 Among the variable first four feet, both lines have an initial dactyl and then spondees. ‘Poenus/-os’ takes up the end of the third foot and beginning of the fourth foot.
 - 15 We have chosen to focus on the *Civil War 1–Aeneid* comparison precisely because it is well-studied, and so allows comparison of automatic methods with existing scholarship. As is true in this case, therefore, any new parallels between the two poems revealed by Tesseræ contribute to, and must be interpreted within, a larger set of recognized connections.
 - 16 Hinds 1998 remains the indispensable guide to these positions.
 - 17 Thomas 1986 presents this point of view, which is reframed but not retracted in Thomas 1999. Despite the subsequent dominance in theoretical discussions of those advocating reading from a contemporary perspective, in practice, most interpreters tacitly assume the goal of reconstructing an original perspective. So Hutchinson 2013, the first major study of Latin intertextuality with Greek authors, avoids discussion of intertextuality in modern theoretical terms, and instead surveys related ancient critical discourse and offers a wealth of readings.
 - 18 Stronger and milder forms of this view are advanced by, respectively, Martindale 1993 and Edmunds 2001. Our formulation attempts to paraphrase the position of Martindale, who advocates acknowledging that the ‘reception’ of a work, or the legacy of its interpretation, is inescapably integral to how we read it (further discussion in Martindale and Thomas, 2006). Martindale 2013 acknowledges with frustration that his calls for greater theoretical development and practical application of reception approaches have gone unheeded.
 - 19 Fowler 2000, 122, as in n. 10 above.
 - 20 Kristeller 1951, p. 106, cited by Ford 2013. So, similarly, Russell 1981, p. 1: ‘The recorded critical

judgments . . . are puzzling. We find them often inadequate and unsatisfactory, if we compare them with our own responses to the same texts’.

- 21 Coffee and Koenig 2012, p. 402.
- 22 Rijsbergen, C. J. V. (1974). Foundation of Evaluation. *Journal of Documentation* 30: 365–73.
- 23 Note that the stop list and distance restrictions apply to all points on this and the following two graphs. If these constraints were removed, recall would be slightly higher and precision slightly lower, with little or no change to F-measure.