The Sense of a Connection: Automatic Tracing of Intertextuality by Meaning

Walter J. Scheirer

Harvard University

Chris Forstall

University of Geneva

Neil Coffee

State University of New York at Buffalo

The Sense of a Connection: Automatic Tracing of Intertextuality by Meaning

1 Introduction

The recognition that poetic texts are often significantly linked to their predecessors through shared or similar language has been an important part of the reading and study of literature since antiquity. More recently, however, scholarly interest has broadened beyond the verbatim reuse of specific phrases[1] to take in the great scope and subtlety of intertextual connections[2]. The term *intertext*, coined by Julia Kristeva (1986, p. 37), has come to be used widely in literary studies to indicate linguistic similarity that, in presenting to the reader a marked connection between two texts, generates new meaning or novel stylistic effects. Emerging digital methods now make it possible to trace this sort of intertextuality with some success. Most typically, computational approaches search for the type of lexical correspondences that can be loosely described as paraphrase[3]. The process closely resembles the manual identification of intertextuality still commonly practiced by literary scholars, including the writers of commentaries (Coffee *et al*., 2012).

This same work has demonstrated, however, that meaningful connections between texts occur not only via lexical similarity but also through a broader similarity of meaning in the absence of words that have the same form or stem[4]. The classicist Stephen Hinds describes this phenomenon as a "poetic of corresponding inexactitude, which draws on but also distances itself from the rigidities of philological and intertextualist fundamentalisms alike" (Hinds, 1998, p. 50). One study indicated that, among a certain set of meaningful parallels between two texts, some 33% were made up by similarity of meaning in the absence of more than one shared word (Coffee *et al.*, 2012, p. 415).

Quite remarkably, human readers are rather adept at identifying text reuse when faced with such "inexactitude," where a predefined formula for lexical matching drawn from textual criticism would simply fail. For instance, consider the following lines from the Roman poet Lucan, which, in an epic simile, characterize the once-great general Pompey on the eve of the Roman Civil War as a tottering, but still venerated, oak[5]:

> qualis frugifero quercus sublimis in agro,
> exuvias veteres populi sacrataque gestans
> dona ducum, nec iam validis radicibus haerens,
> pondere fixa suo est; nudosque per aera ramos
> effundens, trunco, non frondibus, efficit umbram;
> et quamvis primo nutet casura sub Euro,
> tot circum silvae firmo se robore tollant,
> sola tamen colitur.
> (Lucan, *Civil War* 1.136–143)

> Just as a lofty oak in a productive field, bearing the ancient spoils and consecrated gifts of leaders, but no longer clinging with healthy roots, is fixed in place by its own weight; and spreading out bare branches through the air, it casts a shadow from its trunk rather than its leaves; and, although it sways, ready to fall at the first easterly wind, while so many of the surrounding trees bear themselves up on sturdy hardwood, it alone is honored.

Commentators[6] on this poem have noted intertextual connections to several passages in Vergil's *Aeneid*. Among them is another simile, this one comparing the doomed city of

Troy, finally penetrated by the besieging Greeks, to a moribund ash-tree toppled by industrious peasants:

> ac veluti summis antiquam in montibus ornum
> cum ferro accisam crebrisque bipennibus instant
> eruere agricolae certatim,—illa usque minatur
> et tremefacta comam concusso vertice nutat,
> volneribus donec paulatim evicta, supremum
> congemuit, traxitque iugis avolsa ruinam.
> (Vergil, *Aeneid* 2.626–531)
>
> Just as when farmers vie to uproot an ancient ash-tree high in the mountains, hacked at with a rain of blows from their iron axes—it keeps threatening to fall, and, with its foliage trembling, its crown shaken, it sways, until, overcome little by little with its wounds, uprooted from the ridge, it at last gives a groan and heaves forward its own collapse.

As readers, how do we recognize that these two texts are related, when they share just one distinctive word, "sway" (*nuto*)[7]? We see a resemblance of theme: both texts describe a tottering old tree. Both passages also share a narrative function. In each case the tottering tree foreshadows the downfall of a hitherto stalwart bastion: the Trojan citadel in the *Aeneid*, the republican general Pompey in the *Civil War*. Indeed, the two events appear intertextually connected: the capture and destruction of mythological Troy anticipates the historical defeat and death of the Roman leader[8].

Theme, narrative structure, historical and mythical events: the ability of poetic language to forge connections simultaneously among such different sign-systems is precisely what Kristeva's original broad notion of intertextuality as "an intersection of textual surfaces" (Kristeva, 1986, p. 37) was meant to encompass. This view of intertextuality leads us to think of words, even different but related ones, as part of a continuum of reuse and repurposing, and so to see in our examples of epic collapse, and countless others, the potential for thematic material from one context to be redeployed in another to new effect.

Given the complexity of literary meaning that arises when readers encounter such instances of intertextuality, how can we capture it adequately with a computer model? What we need, in the words of Hinds, is a "fuzzy logic" that is flexible enough to identify highly inexact matches often based in thematic similarity (Hinds, 1998, p. 50). The technique we employ for identifying such semantic intertextuality is the popular natural language processing strategy of *semantic analysis*. Algorithms for semantic analysis are typically designed around the notion of word co-occurrence. That is, they start from the assumption, possibly counterintuitive but well-demonstrated, that words that occur in the same contexts have related meanings. This assumption, coupled with the cognitive matching process described above, motivated the design of Latent Semantic Indexing (LSI), an early and still powerful approach (Deerwester *et al.*, 1990). The use of algorithms for semantic analysis, including topic modeling (Blei, 2011), has spread from the practical applications of natural language processing to become a popular tool for literary studies among digital humanists. Recent work has used semantic analysis to

distinguish between genres, produce an algorithmic historiography of classical scholarship, and characterize sentiment in political writing[9].

These types of tasks fall into what Jockers terms *macroanalysis* (Jockers, 2013), which applies the tools of machine learning to collect quantifiable evidence of literary phenomena over large corpora, which might consist of the collected works of an author, whole genres, and entire literatures. When instead used for close reading, semantic analysis has the potential to reveal the characteristics and behavior of the language elements that participate in intertextual connections. In this work, we are concerned with texts from antiquity where intertextuality takes the form of similar small phrases or passages, as opposed to corpora of large documents where semantic analysis is more commonly applied.

Let us begin with an example of how semantic analysis can be applied to this sort of small collection. Consider the following lines of Latin from Lucan's *Civil War* as a simple corpus:

1. bella per Emathios plus quam civilia campos iusque datum sceleri canimus
2. post Cilicasne vagos, et lassi Pontica regis proelia, barbarico vix consummata veneno, ultima Pompeio dabitur provincia Caesar
3. sed non in Caesare tantum nomen erat, nec fama ducis: sed nescia virtus stare loco: solusque pudor, non vincere bello
4. turba minor ritu sequitur succincta Gabino, Vestalemque chorum ducit vittata sacerdos, Troianam soli cui fas vidisse Minervam

5. certe populi, quos despicit Arctos, felices errore suo, quos ille timorum maximus haud urget, leti metus
6. quodque (nefas) nullis inpune apparuit extis, ecce, videt capiti fibrarum increscere molem alterius capitis
7. iam gelidas Caesar cursu superaverat Alpes, ingentisque animo motus, bellumque futurum ceperat. ut ventum est parvi Rubiconis ad undas
8. rupta quies populi, stratisque excita iuventus deripuit sacris adfixa penatibus arma, quae pax longa dabat
9. non, si tumido me gurgite Ganges summoveat, stabit iam flumine Caesar in ullo, post Rubiconis aquas

Now suppose that we would like to find which of the above lines, if any, have some thematic similarity to the phrase *Rubiconis aquas* ("the waters of the Rubicon"). Given that Caesar is famously associated with crossing the Rubicon, if a semantic analysis approach were effective, we would expect a search for thematic material related to the Rubicon to turn up phrases in which Caesar appears. To test this hypothesis, we applied LSI to search for content similar to *Rubiconis aquas*. An in-depth look at the LSI algorithm, including a description of the relevant parameters, follows in the next section. For now, let us simply consider the top three results returned when we perform this test with an LSI approach, using two topics and cosine distance:

1. post Cilicasne vagos, et lassi Pontica regis proelia, barbarico vix consummata veneno, ultima	After defeating roving Cilician pirates and after battles on the Black sea with the fading king,

Pompeio dabitur provincia Caesar . . . ? (*Civil War* 1.336-8)	scarcely ended by barbaric poison, will Caesar now be handed over to Pompey as his last charge?
2. iam gelidas Caesar cursu superaverat Alpes, ingentisque animo motus bellumque futurum ceperat. ut ventum est parvi Rubiconis ad undas. (*Civil War* 1.183-5)	Already Caesar had overcome the frozen Alps with speed, and in his heart he had anticipated the great upheavals and war to come, when he arrived at the waters of the slender Rubicon.
3. sed non in Caesare tantum nomen erat, nec fama ducis: sed nescia virtus stare loco: solusque pudor, non vincere bello. (*Civil War* 1.143-5)	But Caesar had not only a name and renown as a general, but also a courage incapable of standing still, and shame only at conquering without war.

As the results indicate, the test was successful: the search for thematic content similar to "waters of the Rubicon" turned up passages referring to Caesar as the top three results. In one of these phrases, the search for meanings similar to those of *Rubiconis aquas* detected mention of the Rubicon itself, along with Caesar, but the two others did not. The results also show substantial precision. The algorithm did not recall everything related to Caesar, but only hits rich in the martial language that also co-occurs with the word Rubicon, an emblem of the civil war.

This simple test suggests that material likely to be thematically associated in the mind of the reader (Caesar and Rubicon) can also be identified through semantic analysis. The remainder of this article will address in greater detail a more complicated task. Whereas we have just demonstrated a search that finds passages matching a phrase, we turn now to

detecting semantic similarity between two whole passages. The goal of this sort of search is that the reader interested in finding instances of textual similarity absent verbal repetition will ultimately not need to input a search term, as we have just done, but will be able to simply search all passages of one given work against all of those in another.

With this basic understanding of our goals and approach in place, we can summarize the contributions we describe in the remainder of this article:

1. A methodology for applying semantic analysis to the problem of detecting instances of intertextuality without strict lexical correspondence (Sec. 2).
2. An extensive experimental analysis that compares the results of semantic analysis to human analysis, *i.e.* scholarly commentaries that compare two texts (Sec. 3).
3. A publically accessible web tool that allows non-experts to apply our semantic analysis methodology to a large corpus of Latin writers (Sec. 4).
4. The discovery of thematic matches between Lucan's *Civil War* and Vergil's *Aeneid* not previously recorded by commentators that were detected by our tool (Sec. 4).

2 Methods

2.1 LSI Approach

To find the passages that best match a particular query phrase by context, we need to not only generate a semantic model, but also assess similarity within that model space. For this purpose, we chose to use the LSI module of the Gensim[10] framework (Rehurek and Sojka, 2010) in a custom Python program. The underlying algorithm performs a transformation on a set of document vectors to draw out latent structure in the corpus, and to reduce dimensionality for computational efficiency. This is accomplished via Singular

Value Decomposition (SVD), a matrix factorization technique in linear algebra. A similarity search is then performed in the resulting low rank transformation space.

In order to provide enough contextual information for the models and still keep the input highly localized to specific phrases, a window of approximately 500 characters around and including a target line of text was always selected to form a passage considered a "query." Similarly, a window of approximately 1,000 characters around and including a line from the text we wanted to match against formed a passage considered a "document." Note that each line from the text was used as a basis to create a document, resulting in a large measure of overlap between documents as the window moved across the text. A collection of all such documents from a text represents a training corpus. During pre-processing, the most common 250 words from the Tesserae corpus[11] were removed from consideration. This list contains function words, as well as the most common nouns, verbs, adjectives, and adverbs. Each passage was then processed into a bag-of-words representation, with the inflected form of each word replaced with the set of all possible stems. This was done in lieu of typical lemmatization to increase the amount of text available for training (see the discussion of small sample sizes below).

Each LSI model for a corpus was trained using a user-specified number of topics (*i.e.* the dimensions retained after SVD is applied by the algorithm). Similarity queries proceeded by projecting a query passage and a corpus into the transformed model space, and assessing cosine similarity between the query passage and each document in the corpus to produce a set of match scores (in the range -1 to 1, where a higher score indicates a better match). These scores were then sorted to provide a ranked list of

potential matches. The source code for this algorithm is available publicly on Github as part of the Tesserae web tool[12].

A mathematically inclined reader might ask why we opted for LSI instead of a more flexible topic modeling approach such as Latent Dirichlet Allocation (LDA) (Blei, 2003). During the course of this work, we evaluated several LDA implementations including the online learning technique provided by Gensim, and the efficient sampling-based implementation provided by MALLET (McCallum, 2002). For text samples as small as our passages, these algorithms were not numerically stable, *i.e.* they produced radically different match scores for the exact same input across multiple trials. This is a significant problem for the scholar attempting to search for instances of textual reuse with some degree of confidence. The cause is an artifact of random bootstrapping (*i.e.* initializing the algorithm with different random data each time it is run) with limited sampling. The minimum sampling of text required for the statistical estimators to converge is something greater than what we are providing – LDA is most typically applied to long-form documents and any implementation must make certain assumptions on its input. This is a key open issue in machine learning for the digital humanities: textual analysis for forms such as poetry, song, or epigraphy will nearly always involve small samples[13].

Our testing revealed drift in only the least significant digit of the scores produced by Gensim's LSI implementation[14], giving us enough stability to reliably replicate our results over any number of trials. The sizes of the query and document passages described above were determined experimentally with numerical stability in mind. 500 characters for the query and 1000 characters for the document represent the smallest passage sizes that form a highly localized window around their respective target lines

(ensuring that matches are not too broad), while providing enough numerical stability for the LSI algorithm.

For comparison, we also considered a simpler semantic analysis approach without the rank lowering of the LSI algorithm on the same texts. Again using the Gensim module, we computed the cosine distance between just the bag-of-words representations for a query passage and each passage in the corpus to produce a second set of match scores. The goal of this comparison was to see what LSI adds beyond the basic language model. According to Deerwester *et al.* (1990), rank lowering helps us find all words that are related to each document. This is typically a much larger set than the plain bag-of-words representation because it accounts for synonymy across the corpus. If LSI is indeed exploiting the "semantic structure" of our corpus via low-rank approximation, we should observe better match scores for relevant parallels compared to this simple approach.

2.2 Experiment Design

Our baseline for experimentation is the n-gram matching capability that forms the core of the Tesserae search engine, which is freely available on the web[15]. Briefly summarized, the matching algorithm operates in two distinct stages[16]. In the first stage, all instances where a given unit (*e.g.* verse line or phrase) in one text shares at least two words with another unit in a different text are identified. The words may be exact forms or lemmata. In the second stage, the candidate matches are ranked by the relative rarity and proximity of their shared words. The final result is a score that reflects the overall strength of the match, if some word reuse is present.

We validated our approach with reference to two Latin epic poems, Lucan's *Civil War*, book 1, and Vergil's *Aeneid*. *Civil War* 1 consists of 695 hexameter lines, while all of the *Aeneid* consists of 9,896 hexameter lines. These epics are generally considered to have a deep and remarkable intertextual relationship[17]. This relationship is attested in the work of scholarly commentators, who, as expert readers, document a variety of forms of intertextual relationship, among them instances of shared meaning. We therefore tested our results against a benchmark data set assembled by the Tesserae Project comprising all intertexts between the two texts recorded by four major commentaries. The ability of the algorithm to replicate commentator decisions is used as the measure of performance.

From previous experiments with the Tesserae search engine, we know that it is possible to identify the majority of known intertexts by searching for sentences that share two or more lemmata. In a test on a set of given samples, the word-based algorithm missed 35 of the commentator parallels, however, which accounted for 1/3 of the benchmark set. Analysis of such missed samples suggests that they consist wholly or partially of instances of similar meaning, without shared words (Coffee *et al.*, 2012b, 2014). This subset of the overall benchmark represents a union of parallels described in the four commentaries. Of these, individual commentators identified 30 distinct parallels, while two commentators independently identified each parallel in the remaining five. With due allowance for the subjectivity of the commentators, the objective of this work was to see how many of the 35 missed intertexts could be recovered by automatic matching by semantic context rather than words.

3 Lucan-Vergil Benchmark Results

Our first test case was the following excerpt from book 1 of the *Civil War*, where Lucan uses metaphorical language to describe the abandonment of Rome by its military age men (left panel below).

qualis, cum turbidus Auster
reppulit a Libycis inmensum Syrtibus aequor
fractaque veliferi sonuerunt pondera mali,
desilit in fluctus deserta puppe magister
nauitaque et nondum sparsa conpage carinae
naufragium sibi quisque facit, sic urbe relicta
in bellum fugitur. *nullum iam languidus aevo*
evaluit revocare parens coniunxve maritum
fletibus, aut patrii, dubiae dum vota salutis
conciperent, tenuere lares; nec limine quisquam
haesit et extremo tunc forsitan urbis amatae
plenus abit *visu*: ruit inrevocabile volgus.
o faciles dare summa deos eademque tueri
difficiles!
(*Civil War* 1.498 – 511)

Just as when the swirling south wind drives the vast sea back from the Libyan Syrtes, and the shattered mass of the mast, with its sail, groans, the helmsman abandons the stern and leaps into the waves; and though the fittings of the hull are not yet strewn

postquam res Asiae Priamique evertere gentem
immeritam visum superis, ceciditque superbum
Ilium et omnis humo fumat Neptunia Troia,
diversa exsilia et desertas quaerere terras
auguriis agimur divum, classemque sub ipsa
Antandro et Phrygiae molimur montibus Idae,
incerti quo fata ferant, ubi sistere detur,
contrahimusque viros. vix prima inceperat aestas
et pater Anchises dare fatis vela iubebat,
litora cum patriae lacrimans portusque relinquo
et campos ubi Troia fuit. *feror exsul in altum*
cum sociis natoque penatibus et magnis dis.
(*Aeneid* 3.1-12)

After the gods saw fit to overturn the affairs of Asia and visit undeserved punishment on the race of Priam, after proud Ilium had fallen and all of Troy, built by Neptune, was a smoking ruin, we were driven by signs from the gods to seek exile far away

apart, each sailor fashions his own personal shipwreck; so too they desert the city and flee into war. Parents, frail with age, cannot call back their sons, nor wives, by their tears, their husbands; nor the ancestral homes, so long as they place their hopes on an unlikely salvation. No one hesitated on his threshold, to depart, perhaps, with a final look, filled with the love of his city. The crowd rushed on, heedless. How easily the gods give everything, how little they care to preserve it.

(*Civil War* 1.498 – 511)

and find vacant lands. Near Antander and the mountains of Phrygian Ida we constructed a fleet, though we were unsure where the fates were taking us, where we were to settle, and we gathered our men together. Summer had only just begun and my father Anchises ordered us to give sail for our destiny. I wept as I left the shores and harbors of my fatherland, and the plains where once was Troy. I was cast, an exile, onto the high seas, together with my companions, my son, the spirits of my household and the great gods above.

(*Aeneid* 3.1-12)

The major theme of these lines is abandonment, in this case of the city of Rome, (*sic urbe relicta in bellum fugitur*), articulated in part through a simile of shipwreck (*desilit in fluctus, puppe magister, naufragium*).

These lines are thought to be richly intertextual with the *Aeneid*. The commentator Paul Roche, author of the most recent and extensive commentary on this part of Lucan's epic, notes numerous parallels[18] in lines 504-7 alone (italicized in the left panel above), particularly with book 2 of the *Aeneid* 2. But Roche also remarks on a similarity with part of *Aeneid* 3 that has no shared words, making it a good test for detecting resemblance of meaning alone. The relevant passage from *Aeneid* 3 comes at the opening of the book, where Aeneas begins the story of his wanderings. It is marked in italics in the right panel above.

This entire passage was in fact included in a top match returned by our algorithm for the comparison above between *Aeneid* 3 and the Lucan passage. Roche observes the contrast between Aeneas's concern for his family in flight and the disregard for their families shown by Romans fleeing their city in Lucan's epic (Roche 2009, pp. 504-7). Our LSI method responds to related themes over a longer stretch of text. As in Lucan's description of citizens' flight from Rome, in the opening of *Aeneid* 3 we find pronounced themes of abandonment (*diversa exsilia et desertas quaerere terras, litora cum patriae lacrimans portusque relinquo*) intermingled with naval imagery (*classem, vela, portus*). This thematic similarity creates a connection between the texts despite the absence of any significant lexical overlap of the kind targeted by Tesserae lexical search and other text reuse search engines. The infrequent words common to both texts are underlined above, illustrating that the passages share none of the compact, word-level n-grams typically picked out in scholarly commentaries[19]. The passages could, in theory, be identified as similar based upon this sparse collection of shared words, but only by a search so minimally restrictive as to produce a flood of results. Matching via semantic analysis thus brings us much more directly to the thematic resemblance identified by Roche.

Taking this approach further, we experimented with the LSI modeling to see how many of the 35 missing commentator parallels between *Civil War* book 1 and the *Aeneid* we could return in the top 50 results, on the assumption that this was a highly manageable number for scholars to check. Passages (queries) from book 1 of *Civil War* were matched against all passages (documents) found in individual books of the *Aeneid*, and the results were ranked in descending order by LSI score. This search involved setting one arbitrary parameter, the number of topics (or dimensions) into which the passages would be

categorized by content. For our experiment, we evaluated each query at 10, 15 and 20 topics, and reported the parameter at which a valid parallel was found.

To provide the reader with a more thorough analysis of the proposed approach, we also computed precision, the fraction of retrieved instances relevant to a valid parallel, for each result. This was done by counting the number of matches that contained text from a valid parallel and dividing by the 50 total matches we always considered to be candidates. Recall from Sec. 2.1 that our approach generates a large sampling of overlapping windows, meaning that it is possible to have multiple valid matches per search instance. This is a useful feature for a scholar, in that we have good coverage of the context surrounding a target line of interest from a set of windows that overlap, but not completely. We exploit this behavior in our user interface (described below in Sec. 4) to highlight relevant passages of text.

Of the 35 missing parallels, the LSI approach returned 12, listed in Table 1. Several of these results were ranked in the top five returned by the algorithm for a given number of topics, indicating very strong thematic links. One additional parallel also found by the n-gram matching algorithm of the Tesserae search engine was returned as a rank-3 result. Comparing the methods of analysis, we found that lower ranks tend to be correlated with higher precision.

These results also provide a basis for comparing our LSI method with the alternative approach of cosine distance between bag-of-words representations. When testing the latter, we observed much higher ranks (indicating worse performance) and lower precision values for most of the parallels in Table 1. In many cases, the ranks fall outside of the top 50 results, and are not considered valid matches by the matching criteria of this

paper. Scores produced by this simple model were also significantly lower than those generated by the LSI approach. In every instance LSI outperformed the simpler bag-of-words approach. Thus, for this corpus, we can conclude that by making use of low-rank approximation to capture the broader synonymy of the corpus, the LSI approach yields stronger matches that appear higher up in the rank order.

This is not to say, however, that the simpler model has been rendered useless. Table 2 lists an additional set of missing *Civil War* 1 – *Aeneid* commentator parallels found in the top-50 results returned by the cosine distance between bag-of-words representations. These parallels are not found by the LSI approach. Similar to the results in Table 1, we again observe higher ranks and lower precision values for each parallel – not a single one of these matches falls within the top-10 results. This indicates that even as a weak approach, the simple bag-of-words model could be useful in combination with other, more powerful approaches via fusion (using a reasonably intelligent score analysis algorithm) to improve the rank position of a match. We are investigating this possibility in our ongoing work.

4 A New Tool for the Study of Intertextuality

Based on the satisfactory benchmark results, we designed an accessible front-end to the proposed algorithm for more traditional scholars of the classics. Those interested in trying out the algorithm have free access to an easy-to-use web-based tool via the Tesserae Project website[20]. Figs. 1 and 2 show the interface, which provides simple drop-down menus for all parameters (author, work, book and number of topics), and a point-and-click mechanism to allow the user to explore the texts while reading. Scholars

without significant training in machine learning will find this tool to be a convenient starting point for conducting studies related to intertextuality and semantics at a large scale. At the time of this writing, 61 different Latin poets and prose writers are available for comparison.

An important question is whether this tool (and the underlying LSI algorithm) can be useful in revealing new instances of text reuse. Ideally, the approach should produce results beyond those in our benchmark set that were noted by commentators but missed by lemma matching. To this end, we used our web interface to visualize other strongly matching passages between *Civil War* 1 and the *Aeneid*, using the lines from *Civil War* 1 in Tables 1 & 2 as "targets" (*i.e.* queries) against the passages from all of the "source" books of the *Aeneid*. This search turned up significant thematic correspondences not recorded by commentators, listed in Table 3. These included another passage in the *Aeneid* that shares the themes of abandonment and the sea with the lines around *Civil War* 1.504 quoted above. Here sailors flee from the shores of Polyphemus, and the related words are concentrated in the first two lines:

sed fugite, o miseri, fugite atque ab litore funem rumpite.	But flee, you wretches, flee and slash the cables from the shore. For as great and tall as Polyphemus
nam qualis quantusque cavo Polyphemus in antro lanigeras claudit pecudes atque ubera pressat,	is who lives in his hollow cave, keeps wooly flocks, and milks their udders, a hundred such other
centum alii curva haec habitant ad litora vulgo infandi Cyclopes et altis montibus errant.	monstrous Cyclopes live together along the curved shore, and wander the steep mountains.
(*Aeneid* 3.639-44)	

Other passages were related by different common themes. The LSI algorithm identified the following two passages as highly related, and both in fact describe the god Bacchus, though in almost entirely different terms (they share just one word, *vertice*):

nam, qualis vertice Pindi	nec qui pampineis victor iuga flectit habenis
Edonis Ogygio decurrit plena Lyaeo . . .	Liber, agens celso Nysae de vertice tigris
(*Civil War* 1.674-5)	(*Aeneid* 6.804-5)
For just as a Thracian bacchant, filled	Nor did Bacchus, who in victory guides his chariot
with Theban Bacchus, rushes down from	with reins of vine, leading his tigers from the
the summit of Mount Pindus . . .	summit of lofty Nysa, [traverse as much land as
	Augustus will rule].

We also found a similar correspondence between text surrounding *Civil War* 1.676 and *Aeneid* 4.300-3. This instance contained both identical words (*qualis*, *per urbem*) and (near) synonyms (*attonitam* ~ *excita*, *urguentem* ~ *stimulant*).

In sum, then, our employment of LSI proved successful for the needs of users, in that it can bring them swiftly to significant instances of semantic similarity not previously recorded. And as a computational method, in every case the LSI algorithm again outperformed the simpler bag-of-words approach.

5 Discussion

Our experiment demonstrates that LSI can be used to detect intertextual relationships of meaning where few or no words are shared by the two texts. The same approach can in principle be extended to discover common themes and generic material, though

computational constraints currently make it impossible to conduct a rapid search for such material over very large-scale corpora. The distinction between intertext and non-intertext has always been fundamentally a heuristic one[21] that can shift and change. If this sort of searching can be brought to larger scales, it will likely begin to dissolve the border between the instances of intertextuality most frequently noted by scholars – tight verbal correspondences – and the traditional understanding of similarities of mood and theme.

6 Acknowledgements

This work was supported by the National Endowment for the Humanities [Start-Up Grant Award #HD-51570-12]. We thank Prof. Neil Bernstein of Ohio University, who provided valuable feedback on an early draft of this work.

Table 1. List of missing *Civil War* 1 (BC) – *Aeneid* (AEN) commentator parallels found by the LSI approach in the top 50 results returned by the algorithm for each query. Both rank and precision are reported. An asterisk denotes a parallel outside the missing parallel set also found by the lexical matching algorithm of the Tesserae search engine. For comparison, the corresponding ranks and precision values are also provided for a cosine distance between bag-of-words representations for each parallel. In nearly every instance, LSI outperformed the simpler bag-of-words approach. Comparing rank to precision, it can be seen that lower ranks tend to be correlated with higher precision.

BC Line	AEN Line	Shared Context	Topics	LSI Rank	LSI Prec.	BoW Rank	BoW Prec.
1.60	1.291	Destiny of Caesar; peace	10	3	0.18	86	0.00
1.139	4.441	The blowing wind; tree	20	1	0.22	1	0.28
1.141	2.626	The blowing wind; tree	15	1	0.32	45	0.00
1.193	2.774	An apparition	20	33	0.08	47	0.00
1.193	3.47	An apparition	15	42	0.06	145	0.00
1.291	11.492	Horses	20	26	0.02	212	0.00
1.490	11.142	Flight	15	17	0.22	23	0.08
1.504	2.634	Abandonment	15	3*	0.52	70	0.00
1.504	3.11	Abandonment; Navy	15	4	0.16	215	0.00
1.673	2.199	Omens; terror	15	31	0.02	162	0.00
1.676	4.68	Dido as Bacchant	15	1	0.40	2	0.08
1.676	6.48	Prophecy	15	39	0.44	148	0.00
1.695	6.102	Frenzied discussion	20	22	0.20	31	0.20

Table 2. List of missing *Civil War* 1 (BC) – *Aeneid* (AEN) commentator parallels found in the top 50 results returned by the cosine distance between bag-of-words representations. These results are not found by the LSI approach. Compared to the LSI approach in a general sense, we find that the ranks tend to be much higher and precision much lower for this baseline, with no result in this table placing in the top 10 of those returned. Low precision scores are also observed for this experiment.

BC Line	AEN Line	Shared Context	BoW Rank	BoW Prec.
1.1	4.628	War	48	0.02
1.8	12.313	Hostility	23	0.06
1.226	4.624	Broken treaty	13	0.10
1.226	12.435	Fortune	38	0.04
1.674	4.300	Dido as Bacchant	28	0.02
1.678	10.670	Questioning destination	17	0.16
1.685	2.554	Decapitation; shore	45	0.08

Table 3. Additional thematic matches found between *Civil War* 1 (BC) – *Aeneid* (AEN) by the LSI approach. These include highly specific parallels (a Bacchant in the passage around *Civil War* 1.676 and Bacchus around *Aeneid* 6.809 and *Aeneid* 4.304), weaker parallels with some lexical correspondence (*Civil War* 1.1 and *Aeneid* 4.98, *Civil War* 1.212 and *Aeneid* 1.647), and interesting contextual parallels (a metaphorical description of nautical abandonment around *Civil War* 1.291 and sailors fleeing the shores of Polyphemus around *Aeneid* 3.639). An asterisk denotes a parallel also found by the lexical matching algorithm of the Tesserae search engine. Ranks are also provided for a cosine distance between bag-of-words representations. As with the experiment shown in Table 1, our LSI method consistently outperformed the simpler bag-of-words approach.

BC Line	**AEN Line**	**Shared Context**	**Topics**	**LSI Rank**	**BoW Rank**
1.1	4.98	War	15	1*	1
1.141	2.252	The blowing wind	15	6	128
1.291	11.291	Conquest	20	1	4
1.353	1.647	City; nation	15	1*	26
1.504	3.639	Abandonment; nautical imagery	20	2	129
1.676	6.809	Bacchus	15	1	81
1.676	4.304	Bacchus	10	36	295

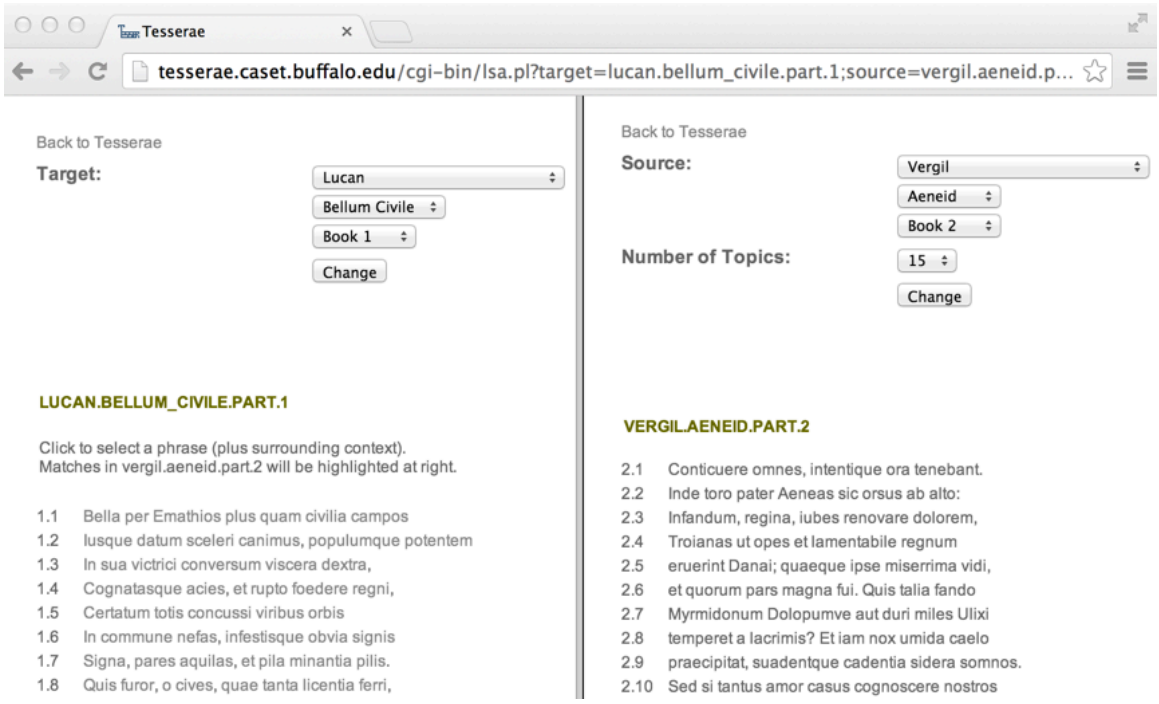

Fig. 1. The public web interface to the algorithm described in this article. Parameters are presented to the user as a series of drop-down menus. The user can click on any line in the "Target" frame, which will initiate the LSI matching process between the passage centered on the target line and all passages in the "Source" frame. The Tesserae Project's entire Latin corpus is available for search. The simple interface allows scholars with minimal training in machine learning to conduct sophisticated studies of semantic intertextuality at a large scale.

Fig. 2. An example of a match between *Civil War* 1.141 and *Aeneid* 2.262. The entire passage highlighted on the left represents the query centered on *Civil War* 1.141. To reduce visual clutter, we only highlight the lines matching passages are centered upon on the right. The matches provide the scholar with an indication of the general neighborhood where semantically similar text can be found. Color intensity in the right-hand frame indicates the strength of the match (brighter colors mean a stronger match).

Notes

1. In classical scholarship sometimes called *loci similes*, or "similar passages," and typically consisting of (near) verbatim reuse of a two-word phrase.

2. Two useful surveys of practices within classical philology can be found in Pucci (1998, ch. 1) and Schmitz (2002, ch. 5); see also Coffee (2012).

3. Examples include: global linear models for assessing verse similarity in the New Testament Gospels (Lee, 2007); unsupervised detection of Greek quotation (Büchler *et al*., 2010), and hash coding to detect reuse and citations in Lautréamont and Balzac (Ganascia *et al.*, 2013). More flexible sequence alignment approaches (Horton *et al*., 2010; Roe, 2012; Wolff, 2012; Smith *et al.*, 2013), inspired by related analysis techniques in genetics, are prevalent as well. Most closely aligned with the goals of this present work are the eTRACES (Bamman and Crane, 2008; Büchler *et al.*, 2011; Büchler *et al.*, 2013; Geßner *et al*., 2013) and Tesserae (Forstall *et al*., 2011; Forstall and Scheirer, 2012; Coffee *et al*., 2012a; Coffee *et al*., 2012b; Coffee *et al*., 2014) projects.

4. Wills (1996, ch. 1) gives an extensive set of textual features that can serve as the basis for intertextual connections, with examples of each.

5. Latin texts cited here are from the Perseus Digital Library (see also Note 11 below); translations are our own.

6. For example, Paul Roche (2009, *ad loc.*).

7. Excluding extremely common function words such as *et* (“and”) and *in* (“in/on”).

8. The process of recognition does not necessarily proceed in such an orderly fashion, however, from the concrete to the abstract; rather, the alert reader is often sensitized to the possibility of intertext in advance. This potential for an intertextual relationship to prime the reader to see further connections is described in detail by Wills (1996, pp. 26–27). In general terms, “a poetic sign signals first to the other signs within the poetic system . . . before signaling its specific sense in a precise context” (Conte 1986, p. 44). For example, Vergil himself seems already to have foreshadowed Pompey’s death in his description of the death of Priam, patriarch of the Trojans (Hinds 1998, p. 8). Lucan’s readers might well have recognized the intertext first on this basis and only subsequently (or never) noticed the reuse of *nuto*.

9. Allison *et al.* (2012), Mimno (2012), and Nelson (2013), respectively.

10. http://radimrehurek.com/gensim/intro.html

11. The Tesserae Latin corpus currently comprises just under 250 texts, evenly divided between prose and verse, principally from the first century BCE to the third century CE. Most of these texts are sourced from the Perseus Digital Library

(http://www.perseus.tufts.edu; G. Crane, Editor). For further information, see http://tesserae.caset.buffalo.edu/sources.php.

12. https://github.com/tesserae

13. On difficulties associated with small samples in literary applications of text analysis, see, *e.g.*, Eder (2014).

14. The Gensim module implements scalable truncated Singular Value Decomposition in Python to calculate the low-rank approximation of a matrix. While there is no specific property of LSI that makes it more suited to small corpora, this particular SVD solver is stable for small sample sizes, making it useful for the kinds of searches demonstrated here. With a lack of good alternatives, we recommend that other researchers consider this implementation for semantic analysis problems that are constrained to small sample sizes.

15. http://tesserae.caset.buffalo.edu

16. Additional detail can be found in the "Methodology" section of Coffee *et al*. (2014).

17. The Lucan commentaries we consulted for this information were Heitland and Haskins (1887), Thompson and Bruére (1968), Viansino (1995), and Roche (2009).

18. (Roche, 2009), 312-313 records parallels between *Civil War* 1.504-7 and *Aeneid* 2.635f., 651-3, 657-70, 673-8, 707-25, 747-51; 3.11f.; 7.757f; 11.160f. Most are contrastive, evoking the difference between Aeneas's concern for keeping his loved ones together while fleeing Troy, and the disregard for family ties shown by those fleeing Rome in *Civil War*.

19. Not underlined are the function words *cum*, *et*, and *in*, which are extremely common and typically excluded by even the shortest stop lists.

20. http://tesserae.caset.buffalo.edu/cgi-bin/lsa.pl

21. So, for example, novelist and semiotician Umberto Eco, in reflecting on the various intertextual relationships between his own fiction and that of Jorge Luis Borges, notes links of several distinct types, including "cases where I was not aware of it, but subsequently readers . . . forced me to recognize that Borges had influenced me unconsciously," as well as others in which a reminiscence of Borges in Eco's writing is due rather to a mutual debt to "preceding sources and the universe of intertextuality." (Eco 2002, p. 121).

References

Allison, S., Heuser, R., Jockers, M. L., Moretti, F., and Witmore, M. (2012). Quantitative Formalism: An Experiment. *n* + 1, 13: 81-108.

Bamman, D. and Crane, G. (2008). The Logic and Discovery of Textual Allusion. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data* (LaTeCH 2008), Marrakesh, Morrocco.

Blei, D. (2011). Probabilistic Topic Models, Comminications of the ACM, 55(4): 77-84.

Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3: 993-1022.

Büchler, M., Geßner, A., Berti, M., and Eckart, T. (2013). Measuring the Influence of a Work by Text Re-use. *Bulletin of the Institute of Classical Studies Supplement*, 122: 63-79.

Büchler, M., Crane, G., Mueller, M., Burns, P., and Heyer, G. (2011). One Step Closer To Paraphrase Detection On Historical Texts. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*, 1(3).

Büchler, M., Geßner, A., Eckart, T., and Heyer, G. (2010). Unsupervised Detection and Visualization of Textual Reuse on Ancient Greek Texts. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*, 1(2).

Coffee, N., Koenig, J.-P., Poornima, S., Forstall, C. W., Ossewaarde, R., and Jacobson, S. (2012a). The Tesserae Project: Intertextual Analysis of Latin Poetry. *Literary and Linguistic Computing*, 28(2): 221:228.

Coffee, N., Koenig, J.-P., Poornima, S., Ossewarde, R., Forstall, C., and Jacobson, S. (2012b). Intertextuality in the Digital Age. *Transactions of the American Philological Association*, 142(2): 381-419.

Coffee, N. (2012). "Intertextuality in Latin Poetry." In *Oxford Bibliographies in Classics*. Ed. D. Clayman. New York, Oxford University Press.

Coffee, N., Forstall, C., Buck, T., Roache, K., and Jacobson, S. (2014). Modeling the Scholars: Detecting Intertextuality through Enhanced Word-Level N-Gram Matching. To appear in *Literary and Linguistic Computing*, Pre-print available at: http://tesserae.caset.buffalo.edu/blog/wp-content/uploads/2012/10/Modeling-the-Scholars-2013-11-6LLC-preprint1.pdf.

Conte, G. B. (1986). *The Rhetoric of Imitation: Genre and Poetic Memory in Virgil and Other Latin Poets*. Translated by Charles Segal. Cornell University Press, Ithaca, New York.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science* 41(6): 391-407.

Eco, U. (2002). Borges and My Anxiety of Influence. In, *On Literature*, pp. 118-135. Translated by Martin McLaughlin. Harcourt, Inc., Orlando FL.

Eder, M. (2014). Does size matter? Authorship attribution, small samples, big problem. *Literary and Linguistic Computing*, forthcoming. Published online November 2013, at http://llc.oxfordjournals.org/content/early/2013/11/14/llc.fqt066.full.

Forstall, C. W. and Scheirer, W. J. (2012). Revealing hidden patterns in the meter of Homer's *Iliad*. In *Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science*, Chicago, Illinois.

Forstall, C. W., Jacobson, S., and Scheirer, W. J. (2011). Evidence of Intertextuality: Investigating Paul the Deacon's *Angustae Vitae*. *Literary and Linguistic Computing* 26(3): 285-296.

Ganascia, J.-G., Glaudes, P., and DeLungo, A. (2013). Automatic Detection of Reuses and Citations in Literary Texts, In *Proceedings of Digital Humanities*, Lincoln, Nebraska.

Geßner, A., Kötteritzsch, C., and Lauer, G. (2013). Biblical Intertextuality in the Digital World: The Tool GERTRUDE. In *Proceedings of the 1st International Workshop on Collaborative Annotations in Shared Environment: Metadata, Vocabularies and Techniques in the Digital Humanities*, Bologna, Italy.

Heitland, W. E. and Haskins, C. E. (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.

Horton, R., Olsen, M., and Roe, G. (2010). Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections. *Digital Studies / Le champ numérique*, 2(1).

Jockers, M. (2013). *Macroanalysis: Digital Methods and Literary History*. Champaign: University of Illinois Press.

Kristeva, J. (1986). Word, Dialogue and Novel. In Moi, T. (ed.), The Kristeva Reader, New York: Columbia University Press, pp. 34-61.

Lee, J. (2007). A Computational Model of Text Reuse in Ancient Literary Texts. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, pp. 472-479.

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu (accessed 25 January 2014).

Mimno, D. (2012). Computational Historiography: Data Mining in a Century of Classics Journals. *Journal on Computing and Cultural Heritage* 5(1).

Nelson, R. K. (2011). Of Monsters, Men – And Topic Modeling. The New York Times. http://opinionator.blogs.nytimes.com/2011/05/29/of-monsters-men-and-topic-modeling (accessed 25 January 2014).

Rehurek, R. and Sojka, P. (2010). Software Framework for Topic Modeling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta, pp. 46-50.

Pucci, Joseph (1998). *The Full-Knowing Reader: Allusion and the Power of the Reader in the Western Literary Tradition*. Yale University Press, New Haven, CT.

Roe, G. H. (2012). Intertextuality and Influence in the Age of Enlightenment: Sequence Alignment Applications for Humanities Research. In *Proceedings of Digital Humanities*, Hamburg, Germany.

Roche, P., Ed. (2009). *Lucan: De bello civili. Book 1*. Oxford: Oxford University Press.

Schmitz, Thomas A. (2002). *Modern Literary Theory and Ancient Texts: An Introduction*. Blackwell Publishing, Malden MA.

Smith, D. A., Cordelly, R., and Dillony, E. M. (2013). Infectious Texts: Modeling Text Reuse in Nineteenth-Century Newspapers. In *Proceedings of the IEEE Workshop on Big Data and the Humanities*, Santa Clara, California.

Thompson, L. and Bruére, R. T. (1968). Lucan's Use of Vergilian Reminiscence. *Classical Philology*, 63: 1-21.

Viansino, G., Ed. (1995). *Marco Annaeo Lucano: La Guerra Civile (Farsaglia) libri I-V*. Milan: Arnoldo Mondadori.

Wills, Jeffrey (1996). *Repetition in Latin Poetry: Figures of Allusion*. Clarendon Press, Oxford.

Wolff, M. (2012). Surveying a Corpus with Alignment Visualization and Topic Modeling. In *Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science*, Chicago, Illinois.