Discovering Multilingual Text Reuse in Literary Texts

David Bamman
The Perseus Project
Tufts University
Medford, MA
david.bamman@tufts.edu

Gregory Crane
The Perseus Project
Tufts University
Medford, MA
gregory.crane@tufts.edu

Abstract

We present here a method for automatically discovering several classes of text reuse across different languages, from the most similar (translations) to the most oblique (literary allusions). Allusions are an important subclass of reuse because they involve the appropriation of isolated words and phrases within otherwise unrelated sentences, so that traditional methods of identifying reuse including topical similarity and translation models do not apply. To evaluate this work we have created (and publicly released) a test set of literary allusions between John Milton's *Paradise Lost* and Vergil's *Aeneid*; we find that while the baseline discovery of translations (55.0% F-measure) far surpasses the discovery of allusions (4.8%), its ability to expedite the traditional work of humanities scholars makes it a line of research strongly worth pursuing.

1 Introduction

While recent work in discovering text reuse has focussed on tracking information flow in newswire, web pages and blogs (Seo and Croft, 2008; Bendersky and Croft, 2009; Bernstein and Zobel, 2006; Henzinger, 2006), we focus here on another important genre: literary texts. Authors refer to the texts of others (in the form of *imitative textual allusions*) largely for two main reasons: to express similarity between two passages, so that the latter can be interpreted in light of the former; and to simultaneously express their dissimilarity as well, in that the tradition they recall is revised. This reuse tends to be more oblique than resampling information from news stories and any individual instance is often the subject of vigorous debate. Discovering these allusions, however, is crucial for the act of criticism.

While an author's most immediate literary environment may be comprised of works written in the same language, this relationship of course extends across languages as well. The English poet T. S. Eliot refers often to the Italian works of Dante Alighieri; Vergil often refers in his Latin poems to the Greek epics of Homer; but one of the most prolific examples has been John Milton's use of the *Aeneid* in *Paradise Lost*.

Milton's use of Classical material extends far beyond simple reference to Greco-Roman subjects (such as figures from Classical mythology): his appropriations also include etymological wordplay (playing on the English sense of a word and the Latin sense from which it is derived)[1] and imitative textual allusion, where he samples and translates earlier material.

We can view this reuse on a continuum from most similar to least. At the far end are entire syntactic phrases that are translated between texts, as in examples 1 and 2.

(1) **the moon's resplendent globe** (Paradise Lost [PL] 4.723).

(2) **lucentemque globum Lunae** (Aeneid [Aen.] 6.725) ["and the shining globe of the moon"]

Since Latin is a highly inflected language, its word order is not bound directly to syntax and is hence much more free (especially in poetry) to serve other purposes such as discourse and meter. A syntactic representation of the sentences under (e.g.) a dependency grammar, however, would reveal them to be identical. In the middle of the continuum are sentences that share some structural similarity but are predominantly joined by topic, as in sentences 3 and 4.

[1] E.g., "the hastening Angel ... / Led them direct, and down the cliff as fast / To the subjected plain" (12.637-40) where *subjected* means both "under the authority of" and "lying below" (the original sense of the Latin *subiectus*).

(3) ... or faery elves, / Whose midnight revels, by a forest-side / Or fountain, some belated peasant **sees, / Or dreams he sees**, while overhead **the Moon** / Sits arbitress ... (PL 1.781-785).

(4) qualem primo qui surgere mense / **aut videt, aut vidisse putat** per nubila **lunam** ... (Aen. 6.453-454). ["Just as one who sees, or thinks he sees, the moon rising in the early month."]

In this example, while part of the reference is syntactically identical, the relationship to the other lexical similarity (*moon*) is different in each sentence (it is the subject of an adjunct clause in *Paradise Lost* and forms part of a complex object in the *Aeneid*). At the least similar end of the continuum are pairs that hold only a topical similarity to each other, as in sentences 5 and 6.

(5) Long is the way / And hard, that out of Hell leads up to light. (PL 2.432-433)

(6) Tros Anchisiade, **facilis descensus Averno** (Aen. 6.126) ["Trojan son of Anchises, easy is the descent to the underworld."]

These varieties of reuse are related in principle to the monolingual word and fact reuse found in other studies, and we may be tempted to categorize them accordingly (using, for example, the sentence-level judgment classes found in Metzler et al. (2005)). Literary allusions, however, present one strong difference, which influences how we can find them: in all of the examples above, it is only discrete fragments of a sentence that are similar between the source text (the *Aeneid*) and the derived text (*Paradise Lost*) – at the level of the entire sentence, both texts are generally on very different topics. The first three examples listed above have been excerpted from their full sentences for the sake of brevity; consider, however, an allusion within the context of its entire sentence, as in 7 and 8 below.

(7) Of Man's first disobedience, and the fruit / Of that forbidden tree whose mortal taste / Brought death into the World, and all our woe, / With loss of Eden, till one greater Man / Restore us, and regain the **blissful seat**, / Sing, Heavenly Muse, that, on the secret top / Of Oreb, or of Sinai, didst inspire / That shepherd who first taught the chosen seed / In the beginning how the heavens and earth / Rose out of Chaos; (PL 1.1-10)

(8) His demum exactis, perfecto munere divae, / devenere locos laetos et amoena virecta / fortunatorum nemorum **sedesque beatas**. (Aen. 6.637-639) ["Now, these things done and tribute paid to the goddess, they came to the cheerful spots, the pleasant grasslands of fortunate bowers and **blessed seats**."]

Here indeed *sedesque beatas* and *blissful seat* are the only elements of similarity within the much wider context of the two sentences – the source text of the *Aeneid* makes no mention, however oblique, to disobedience, a forbidden tree, or any of the other topics laid out in that opening invocation of *Paradise Lost*. This broader topical dissimilarity means we must reduce the window of reuse to a much finer level, to a granularity of individual words and phrases.

2 Background

The window of reuse for most other studies has been either the document level (Brin et al., 1995; Hoad and Zobel, 2003; Shivakumar and Garcia-Molina, 1995) or the sentence level (Metzler et al., 2005; Bendersky and Croft, 2009; Hatzivassiloglou et al., 1999), though any technique that uses fingerprinting or n-grams essentially establishes similarity at the subsentence level as well, even if in the service of comparing larger documents.

Successful methods for the efficient detection of duplicate or near-duplicate documents – such as relative frequency measures, fingerprinting, or even simple word overlap – work best of course when the documents to be compared are sizable enough to generate effective word distributions or fingerprints of high enough resolution.

For smaller granularities, such as detecting reuse on the sentence level, successful methods have used TF-IDF measures to highlight the overlap of less-frequent words and a statistical translation model based on IBM Model 1 (Brown et al., 1993) to locate the probability of a source sentence being "translated" as a target one (even if both are in the same language). Metzler et al. (2005) test a number of these methods for detecting sentence-level similarity, and find that simple word overlap, IDF-weighted overlap, query likelihood, and a statistical translation model all perform best (though with query likelihood achieving the highest precision for topically related sentences, and the translation model performing best

with exact duplicates).

Other sentence-level methods deal with the problem of data sparseness by supplementing standard TF-IDF term weights with additional information, including WordNet synonyms and semantic classes (Hatzivassiloglou et al., 1999), positional information of the source and derived sentences with respect to each other and within the larger document (Lee, 2007) and syntactic information (Uzuner et al., 2005; Bamman and Crane, 2008).

Detecting multilingual allusions tests the limits of existing methods in three ways.

Data sparseness. Since we are attempting to uncover similarities on a phrasal and even individual word level, the problem of data sparseness is exacerbated.

No fixed boundaries. Document-level detection and sentence-level detection both compare sections of text with previously fixed boundaries (e.g., delimited by periods or question marks) but phrasal similarity has a fluid window. While examples 1 and 2 could in principle be found with a sliding window four words long, examples 5 and 6 contain phrases of much more disparate lengths (14 and 3 words long, respectively). If we increase a fixed window size to 14, we essentially have the same problem we would have if we consider entire sentences: in literary allusions, only a relatively small fragment of the sentence is allusive.

Multiple languages. With the notable exception of the statistical translation model used in Metzler et al. (2005), almost all of the current methods for detecting text reuse apply only to documents of the same language. This means that some of the most effective measures for determining similarity between two documents (such as overlapping n-grams) cannot apply given the variability of word order between different languages.

3 Methodology

Our approach to discovering text reuse at a phrasal level (without prior sentence delimiting and across languages), first involves quickly identifying possible translations between a source text and derived text, clustering those instances together to determine phrasal boundaries, and then applying more elaborate comparisons between the source and derived phrasal pairs that result.

3.1 Inducing translation equivalents

Since we're trying to detect similarities between the English text of *Paradise Lost* and the Latin text of the *Aeneid*, we need a translation lexicon between the two languages. While several machine readable bilingual dictionaries exist for Latin (including the Lewis and Short Latin Dictionary (Lewis and Short, 1879) and the Elementary Lewis (Lewis, 1891)), we also need to have a translation likelihood estimate that a given term X in Latin is translated as term Y in English, and vice versa. To create this probabilistic lexicon, we induce the English senses for all Latin words using a corpus of parallel texts. The texts released under a Creative Commons license from the Perseus Digital Library include 3.5 million words of Latin source texts and 2.4 million words of corresponding English translations. Aligning these texts proceeds in three phases, from a coarse chunk level to a final granularity of individual words. Since the printed editions of all Latin texts are organized according to their logical citation scheme, aligning a source text with its translation at this chunk level is trivial: Book 1, chapter 1 of Tacitus' *Annales* corresponds to Book 1, chapter 1 of Church and Brodribb's translation of it.

At this finer resolution, we align sentences using Moore's Bilingual Sentence Aligner (Moore, 2002), which aligns sentences that are 1-1 translations of each other with a high precision (98.5% for a corpus of 10,000 English-Hindi sentence pairs (Singh and Husain, 2005)). This process aligns approximately 30% of the sentences (since most are not 1-1 translations), but we use those high-precision alignments as anchors for the 1-2, 2-1 and many-to-many alignments that fall between.

All of these sentences – both the 1-1 aligned ones and the sentences in between – are then aligned at the level of individual words using MGIZA++ (Gao and Vogel, 2008), a multi-threaded version of GIZA++ (Och and Ney, 2003), which significantly reduced the alignment time on an 8-core CPU. Prior to alignment, all of the tokens in the source text and translation are stemmed (to account especially for Latin's rich inflection). After alignment, the original Latin word forms were restored and then lemmatized, using the English sense as a feature for lemma disambiguation.[2] After aggregating the alignment results, we

[2]E.g., if a word such as *est* is aligned to the English word

have a translational distribution for each word in each language such as that shown in tables 1 and 2.

English	Probability
speech	61.0%
prayer	16.7%
language	8.8%
oration	7.9%
word	5.5%
talk	5.0%
argument	4.1%

Table 1: Translational distribution for the Latin word *oratio*.

Latin	Probability
oratio	63.2%
dico	11.9%
verbum	10.0%
sermo	9.1%
contio	5.8%

Table 2: Translational distribution for the English word *speech*.

3.2 Clustering similar phrases

Discovering literary allusions requires a different approach than uncovering topical text reuse, since isolated islands of allusiveness can be embedded within sentences that bear no other similarity to each other. An allusion of three words is equally allusive within a sentence of 10 words as it in within a sentence of 100. While some information retrieval methods (such as cosine similarity) normalize for sentence length, they tend to favor short sentences as a result; what we need is a method that discounts it altogether.

What sentence-level comparison does offer, however, is relative computational efficiency; in the worst-case scenario, n^2 sentences must be compared. When comparing a fixed window of 5 words between two documents, however, n must be calculated in words, not sentences, and if we compare a range of variable windows (e.g., a 3-word window in document A with a 14-word window in document B and so on), then the complexity becomes exponential and intractable.

"eat," it is more probably derived from the lemma *edo* ("to eat") than *sum* ("to be") since many unambiguous inflections of *edo* (such as *edisti*) also align to "eat."

To avoid this explosion, we have taken a two-step approach to identifying reuse on a phrasal level. The first involves quickly establishing segments of both the source and derived texts that are likely to be related to each other. In the second step we calculate a more thorough similarity score.

To cluster similar phrases, we first build an inverted index for the words in the derived document (e.g., *Paradise Lost*), grouping together all word positions by their word form (e.g., "speech" is the token at position 1001, 2803, 3335 etc.). Iterating through each possible lemma for every word in the source document (e.g., the *Aeneid*), we recall its possible translations from the induced translation inventory and find all positions from the inverted index where each translation appears. At this point we calculate the average IDF of the Latin term (s) and the English translation (t) and normalize it by the cross-probability that the Latin word is a translation of the English and that the English is a translation of the Latin, as in the following:

$$\frac{1}{2}(\frac{|T|}{df_t} + \frac{|S|}{df_s}) \cdot P(s|t) \cdot P(t|s)$$

This mutual probability allows us to filter out high-frequency words for which we would otherwise need stoplists. For ambiguous words possibly derived from several lemmas, the word score is the individual lemma score with the highest value. Relying on an inverted index for this allows us to create a word-by-word matrix as in table 3 without having to make $|T| \cdot |S|$ comparisons. With this two-dimensional matrix, we can now apply standard clustering techniques to partition the space into coherent units. Rather than specify a fixed number of clusters to be found, our stop criterion is a fixed distance between elements to be clustered: if no further elements can be added to any cluster beyond some fixed radius r, then the clustering is complete.

We also adopt a canopy-based approach (McCallum et al., 2000) to avoid having to compute the distance between all matrix elements: for any element in position $[i, j]$, we consider only those within a fixed window of $[i - d, j - d]$ to $[i + d, j + d]$. If $d = 1$, for instance, the element $[mutatus, changed]$ in table 3 would successfully link to the elements at $[quantum, how]$ and $[ab, from]$ (i.e., one position away in either the source text or the derivation).[3] The cluster as a whole would then be comprised of all of those elements linked to each other, each of them at most

[3] In the experiments which follow, $d = 4$.

	ei [ei1, Eos1, is1]	mihi [ego1]	[,]	qualis [qualum1, qualus1, qualis1]	erat [sum1]	[,]	quantum [quantum1, quantus1]	mutatus [muto1, mutatus1]	ab [ab1]	illo [ille1]	Hectore [Hector1]	[,]	qui [qui2, quis1, quis2, qui1]
if													
thou													
beest													
he										0.92			
-													
but													
o													
how							0.30						
fallen													
!													
how							0.30						
changed								3.74					
from									2.15				
him	2.47									0.40			
who													1.41

Table 3: Translation matrix for *Paradise Lost* 1.84-85 ("If thou beest he - but O how fallen! how changed / From him who") and *Aeneid* 2.274-275 (*Ei mihi, qualis erat, quantum mutatus ab illo Hectore, qui* – "Ah me, what sort he was, how changed from that Hector, who ...").

1 word position away from another. With $d = 1$, this would identify *quantum mutatus ab illo* and *how changed from him* as a potential allusion; with $d = 3$, this would identify *quantum mutatus ab illo Hectore, qui* and *how changed from him who*.

3.3 Identifying reuse

Step one provides us with a set of pairs, each containing one substring from the source text and one substring from the derived text. In step two we calculate the degree of overlap between the two substrings and use this as a score for determining reuse. In principle we can use features here that are more intensive to compute, such as syntactic and semantic similarity or the higher IBM translation models, but for the sake of this experiment we simply use the sum of all aligned words in the cluster. In the example above, for instance, the alignment score for *quantum mutatus ab illo* and *how changed from him* would be 6.59 $(= 0.30 + 3.74 + 2.15 + 0.40)$.

There are two reasons why we want to use an aggregate score here rather than, for example, a more optimal solution such as the emission probability of a derived sentence given a source sentence and alignment. First, a straight statistical translation probability is only interpretable when compared against other potential translations of the same sentence. We can, for instance, state with confidence that *how changed from him* is a better translation of *quantum mutatus ab illo* than *Hello world* is, but not whether that first pair is an optimally "better" translation than another unrelated pair (such as *magna cum laude* and *with great praise*). Secondly, all other things being equal, shorter translation pairs have higher translation likelihoods than longer pairs. What we are interested in is specifically the discovery of those longer pairs.

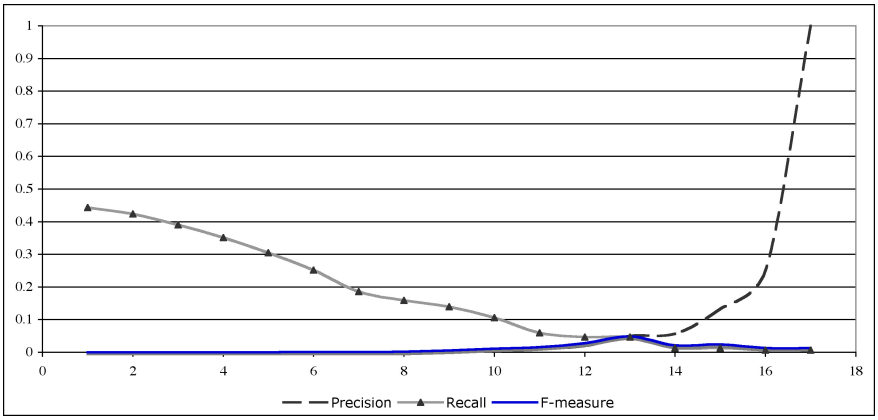

Figure 1: Precision, recall and F-measure for allusion discovery.

4 Evaluation

Our goal with this experiment is the discovery of multilingual text reuse, that is, the production of an ordered list of substrings from a source text S and a derived text D in which higher ranked pairs present "stronger" instances of reuse. We test this on two classes of reuse: the more difficult task of uncovering literary allusions, and a baseline task of identifying passages in one text that are translations of some passage in another.

4.1 Discovering allusions

To evaluate the performance for uncovering allusions, we created a test corpus of 151 known allusions between *Paradise Lost* and the *Aeneid*. The scholarly source for this work is Verbart (1995), which methodically lists the line numbers of "verbal parallels" between the two texts. Using these line numbers as an index to the Creative Commons licensed *Aeneid* (Greenough, 1882) released by the Perseus Project and the open source *Paradise Lost* released by Project Gutenberg (Gutenberg, 1992), we created a publicly available test corpus that references these texts.

The algorithm is embarassingly parallelized by segmenting either the source text or the derived text along fixed breakpoints across which we are certain no referential substring can pass. For this experiment we divided *Paradise Lost* into each of its twelve books and processed them in parallel.

Figure 1 presents the precision, recall and F-measure by score threshold for discovering these literary allusions. It is immediately clear that this method is only able to find a very small subset of the allusions documented in the test set at a reasonable score threshold – only 3, in fact, within the highest scoring 150 (the highest F-measure is a mere 4.8%).

An analysis of the highest scoring pairs, however, suggests that this outcome may be more due to the oblique nature of literary allusions themselves, both in the human judgment of what actually constitutes one, and in our computational modeling of what quantitative features best describe it. Table 4 presents a ranked list of the sentence pairs with the highest reuse scores. While the first and fifth pairs were successfully identified as allusions found in the test set, none of the others are. All, however, retain the same characteristics of known allusions – specifically, the reuse of multiple less common words. In light of this similarity we suspect that the more important measure of this work may be the recall – presenting a list of these discovered *possible* allusions to critics to decide if a potential allusion is in fact an actual one.

4.2 Discovering translations

While interpreting the discovery of allusions may be as oblique as allusions themselves, we can still provide a meaningful quantitative evaluation by measuring our ability to find an unambiguous class of reuse: translations.

While an optimal approach to finding translations between the best sentences in a source document S and a derived document D may be a brute-force method of comparing each sentence in S to each sentence in D, we are operating here under the same assumption as above, that the two documents may be almost entirely unrelated to each other, and our job is to find those fragments of D

Score	*Paradise Lost*	*Aeneid*
*17.43	so he departing gave command (10.429-30)	ita discedens praeceperat (9.40) [“so departing he commanded”]
16.19	see the hubbub strange , and hear [the din](12.60)	videt volitantia miris et varias audit (7.89) [“he sees flying things of wondrous (shapes) and hears various (sounds)”]
16.00	double Janus (11.129)	duplex ianumque (12.198) [“the double (offspring of Latona) and janus”]
15.97	second, or oppose (2.419)	secunda aut adversa (9.282-283) [“favorable or adverse”]
*15.71	who first, who last (1.376)	quem telo primum, quem postremum (11.664) [“who first, who last with your spear”]
15.55	tawny lion (7.464)	fulva leonis (8.552) [“the tawny [skin] of a lion”]
15.39	Heaven to us perhaps he brings, and (5.311-312)	caeloque animum fortasse ferebat canitiemque sibi et (10.548-549) [“he brought his heart to heaven and perhaps (foretold) his old age”]
15.16	when night darkens the [streets] (1.500-501)	nox cum terras obscura (4.461) [“when dark night (held) the lands”]

Table 4: Strongest similarities between *Paradise Lost* and the *Aeneid*. An asterisk denotes a found known allusion.

that are translations of fragments of S.

To test this category of reuse, we compiled all of the 1-1 sentence alignments between the *Aeneid* and its English translation generated by the sentence alignment procedure described above, a total of 2048 sentences in each language. In order to remove any potential bias in the translation inventory, we removed any translation equivalences induced from the *Aeneid* and calculated the Latin-English and English-Latin translation probabilities with only the remaining texts.

As with the allusion test, we calculate the precision, recall and F-measure at each score threshold for all pairs with scores above that threshold. The results are presented in figure 2. Here the highest F-measure (55.0%) comes with sentence pairs holding a score over 14 (with a corresponding precision of 83.2% and recall of 41.1%).

Here again we must note exactly what we are measuring. If our job were to consider each (previously delimited) sentence in a derived text and find the most likely sentence in a source text of which it is a translation, we would expect an ideal F-measure to be 100% and could use a statistical translation model to find an optimal solution. Here, however, we are generating a list of substring pairs and ranking them by how strong we consider the translation to be. At a score threshold of 14, this means that 83.2% of the sentences we attempt to pair are indeed translations of each other – the remaining 16.8% are not translations but have a distinctive enough verbal similarity and high IDF scores for the individual words they contain to rank them higher than legitimate sentence translations with, for instance, higher frequency words. For example, in trying to locate the source text for the translation *Here were her arms, her chariot* (Aen. 1.17), we do in fact find the correct source fragment as the strongest sentence pair associated with it (example 9 below).

(9) hic illius arma, hic currus fuit [“Here her arms, her chariot were”] (Aen. 1.17), score: 12.2

The translation, however, is also similar to another source sentence, that shown in example 10.

(10) **illis** omnibus **arma**, nec clipei **currusve** sonant: **hi** (Aen. 7.685-6), score: 10.6

By virtue of the high mutual IDF scores for the relatively infrequent *arms/arma* and *chariot/currus*, this incorrect pairing has a higher overall score (10.6) than a legitimate translation for a

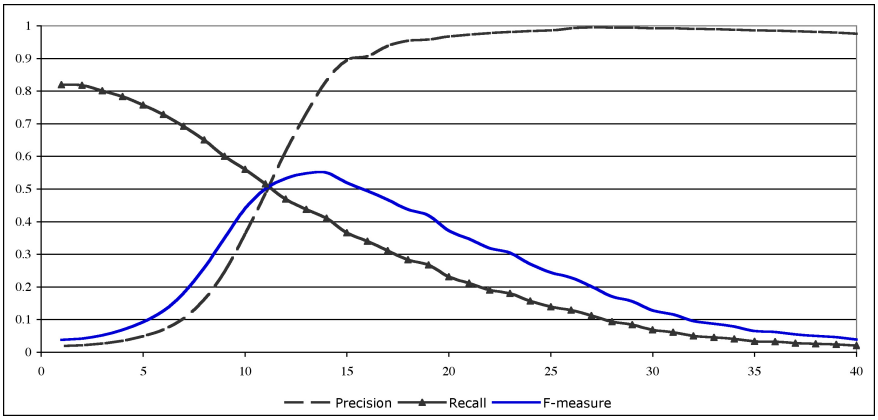

Figure 2: Precision, recall and F-measure for translation discovery.

Score	*Paradise Lost*	*Aeneid*
*18.01	divine interpreter! by favour sent (7.72)	**interpres divom**, Iove **missus ab** (4.356)
*18.01	divine interpreter! by favour sent (7.72)	**missus ab** ipso **interpres divom** (4.377)
*11.21	divine interpreter (7.72)	**interpres divom** (3.359)
*11.21	divine interpreter (7.72)	**divomque interpres** (10.175)
7.80	by favour sent (7.72)	**ab** alto aethere se **misit** (9.644-645)

Table 5: Strongest similarities to "Divine interpreter! By favour sent ..." (PL 7.72)

different sentence (e.g., *such anger* and *tantaene animis caelestibus irae* [Aen. 1.11], with a score of 8.6). Here, as with allusions, our attempt is to find those segments of the source and derived texts that are characteristic signals of reuse.

5 Impact

The goal of this research has been the automatic discovery of several varieties of text reuse across languages, from the most explicitly similar (translations) to those that present only shadows of resemblance (allusions). The automatic discovery of translations on a sentence and phrasal level allows us to extend the traditional applications of general text reuse (tracking information flow, plagiarism detection) to much broader collections, letting us compare a possibly derived target text with sources in many languages. In the more nebulous domain of tracking allusions, human intervention is more likely necessary, but here we can still envision a strong impact upon humanities scholarship.

There are two ways to visualize the data created here: either as a ranked list of the strongest allusions over the entire derived document or, alternatively, as a ranked list presenting similarities for any individual phrase in either the source or derived text. Table 5 presents such a ranking for one line from *Paradise Lost*, which has (according to our test data) no fewer than four references to passages in the *Aeneid*. All four of these allusions were indeed found by our method (and hold a strong separation between the closest non-match). We believe that it is these targeted searches that will be of most use to traditional textual critics, by proposing a set of *possible* allusions for any passage under consideration that can afterwards be refined by an expert. In the end, our ultimate goal is the occasioning the discovery of new knowledge about the texts in our cultural heritage, and the most tangible benefit of this research may be in giving traditional scholars the tools they need to investigate this complex phenomenon.

References

David Bamman and Gregory Crane. 2008. The logic and discovery of textual allusion. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, Marrakesh.

Michael Bendersky and W. Bruce Croft. 2009. Finding text reuse on the web. In *WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 262–271, New York, NY, USA. ACM.

Yaniv Bernstein and Justin Zobel. 2006. Accurate discovery of co-derivative documents via duplicate text detection. *Inf. Syst.*, 31(7):595–609.

Sergey Brin, James Davis, and Héctor García-Molina. 1995. Copy detection mechanisms for digital documents. *SIGMOD Rec.*, 24(2):398–409.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19(2):263–311.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio, June. Association for Computational Linguistics.

J. B. Greenough, editor. 1882. *Bucolica, Aeneis, Georgica: The Greater Poems of Virgil*. Ginn and Co., Boston.

Project Gutenberg. 1992. John Milton's Paradise Lost. `http://www.gutenberg.org/etext/26`.

Vasileios Hatzivassiloglou, Judith L. Klavans, and Eleazar Eskin. 1999. Detecting text similarity over short passages: exploring linguistic feature combinations via machine learning. In *In Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 203–212.

Monika Henzinger. 2006. Finding near-duplicate web pages: a large-scale evaluation of algorithms. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 284–291, New York, NY, USA. ACM.

Timothy C. Hoad and Justin Zobel. 2003. Methods for identifying versioned and plagiarized documents. *J. Am. Soc. Inf. Sci. Technol.*, 54(3):203–215.

John Lee. 2007. A computational model of text reuse in ancient literary texts. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 472–479, Prague, Czech Republic, June. Association for Computational Linguistics.

Charles T. Lewis and Charles Short, editors. 1879. *A Latin Dictionary*. Clarendon Press, Oxford.

Charles T. Lewis, editor. 1891. *An Elementary Latin Dictionary*. Clarendon Press, Oxford.

Andrew McCallum, Kamal Nigam, and Lyle H. Ungar. 2000. Efficient clustering of high-dimensional data sets with application to reference matching. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 169–178, New York, NY, USA. ACM.

Donald Metzler, Yaniv Bernstein, W. Bruce Croft, Alistair Moffat, and Justin Zobel. 2005. Similarity measures for tracking information flow. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 517–524, New York, NY, USA. ACM.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *AMTA '02: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, pages 135–144, London, UK. Springer-Verlag.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Jangwon Seo and W. Bruce Croft. 2008. Local text reuse detection. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 571–578, New York, NY, USA. ACM.

Narayanan Shivakumar and Hector Garcia-Molina. 1995. SCAM: A copy detection mechanism for digital documents. In *In Proceedings of the Second Annual Conference on the Theory and Practice of Digital Libraries*.

Anil Kumar Singh and Samar Husain. 2005. Comparison, selection and use of sentence alignment algorithms for new language pairs. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 99–106, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Ozlem Uzuner, Boris Katz, and Thade Nahnsen. 2005. Using syntactic information to identify plagiarism. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, pages 37–44, Ann Arbor, Michigan, June. Association for Computational Linguistics.

André Verbart. 1995. *Fellowship in Paradise Lost: Vergil, Milton, Wordsworth*. Rodopi, Amsterdam and Atlanta.