Preprint of paper accepted to the ACL Workshop LaTeCH 2008: Language Technology for Cultural Heritage Data.

The Logic and Discovery of Textual Allusion

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 describe here a method for discovering imitative textual allusions in a large collection of Classical Latin poetry. In translating the logic of literary allusion into computational terms, we include not only traditional IR variables such as token similarity and n-grams, but also incorporate a comparison of syntactic structure as well. This provides a more robust search method for Classical languages since it accomodates their relatively free word order and rich inflection, and has the potential to improve fuzzy string searching in other languages as well.

1 Introduction

> Five score years ago, a great American, in whose symbolic shadow we stand today, signed the Emancipation Proclamation ...

Thus begins Martin Luther King Jr.'s "I Have a Dream" speech of 1963. While the actual text of the Gettysburg Address is not directly quoted here, it is elicited by means of an allusion: King's audience would immediately have recognized the parallels between his first four words and the "Four score and seven years ago" that began Lincoln's own speech. By opening with this phrase, King is aligning Lincoln's invocation of human equality with "the greatest demonstration for freedom in the history of our nation" for which he was then speaking.

While the term "allusion" is commonly applied to any reference to a person, place, or thing already known to the reader, we are using it here in the specific context of an *imitative textual allusion* – a passage in one text that refers to a passage in another. When Willy Loman calls each of his sons an "Adonis" in *Death of a Salesman*, there is no doubt that this is an allusion to a Classical myth, but it does not point to a definable referent in the record of written humanity (as King's allusion refers specifically to the first six words of the Gettysburg Address).

The discovery of these allusions is a crucial process for the analysis of texts. As others have pointed out,[1] allusions have two main functions: 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.[2] Allusions of this specific variety are perhaps most widely known as a trope of modernist authors such as Eliot and Joyce, but they are common in the Classical world as well – most strongly in the Greek poetry of the Hellenistic era, in the Roman poetry of the republic and early empire and in New Testament texts (which allude to prophecies recorded in the Old Testament). Given the long history of Latin literature, we must also keep in the mind a text's *Nachleben* – how it has been received and appropriated by the generations that follow it.[3]

Uncovering allusions of this sort has long been the task of textual commentators, but we present

[1] For an overview of the function and interpretive significance of allusions, see Thomas (1986).

[2] Cf. Bloom (1973).

[3] Cicero, for example, was widely admired by Renaissance humanists after Petrarch and provided a model for textual imitation. Cf. Kristeller (1979).

a method here to automatically discover them in texts. Our approach has many similarities with research on text reuse (Clough et al., 2002), paraphrase and duplicate detection (Dolan et al., 2004), and locating textual reference (Takeda et al., 2003; Lee, 2007), but while these methods generally focus on string comparison and document structure, we include variables for considering the abstract structure of a sentence as well, as represented by its syntax. This enables a more robust search method since it is not restricted by word order or inflection. Our test corpus is a collection of Latin poetry, but the methods we describe are language independent.

2 Types of Textual Allusion

While others have categorized textual allusion into a number of types dependent on their function (e.g., Thomas (1986) distinguishes between "casual reference," "single reference," "self-reference," etc.), we are concerned only with a practical distinction in terms of the ease of locating them: an allusion is either direct (equivalent to a quotation) or indirect.

2.1 Direct reference

The most explicit and unambiguous type of allusion is direct reference in the form of a verbatim quotation. We see this form of allusion most often in the long afterlife of a text, as for instance in the reception of this line from Ovid's *Amores*.

(1) At si, quem mavis, Cephalum conplexa teneres / Clamares: **lente currite, noctis equi!** (Am. 1.13)[4]

While Ovid's line comes from the mouth of the mythic Aurora (dawn) pleading with her chariot to pull her more slowly across the sky to give her more time with her lover before returning to her husband, Christopher Marlowe sixteen centuries later appropriates it for Faust, who voices it in the final minutes before midnight in a plea to prolong his life.

(2) Stand still, you ever-moving spheres of heaven, That time may cease, and midnight never come: Fair Nature's eye, rise, rise again and make Perpetual day; or let this hour be but A year, a month, a week, a natural day, That Faustus may repent and save his soul! **O lente, lente, currite noctis equi!** (Act V, Scene 2)

And again, four centuries later, Vladimir Nabokov appropriates it for *Lolita* as his protagonist is chased along a highway.

(3) We were many times weaker than his splendid, lacquered machine, so that I did not even attempt to outspeed him. **O lente currite noctis equi!** O softly run, nightmares! (Nabokov 219)

Following Irwin (2001), we can distinguish an allusion from a mere quotation in the level of context required to understand it. A quotation is self-contained; an allusion calls forth the original context in which it's found. Direct allusions like these are easier to find than their adapted counterparts (it is essentially a simple string search) but they reside on the same continuum as the others.

2.2 Indirect reference

Most of what we would consider allusions involve some transformation of the referent text. An example of this can be found in the first line of the first poem of Ovid's *Amores*, an imitation (and revision) of the first line of Vergil's *Aeneid*.

(4) Arma gravi numero violentaque bella parabam / Edere (Am. 1.1-2)[5]

(5) Arma virumque cano (Aen. 1.1)[6]

Vergil's *Aeneid* is an epic poem focussed on the figure of Aeneas (an ancestor of the Romans), written in dactylic hexameter, the same "heavy" meter as Homer's epics the *Iliad* and *Odyssey*. Ovid, in contrast, is a love poet, and elicits Vergil's famous opening to motivate his genre (the line continues with Cupid stealing one of the line's metrical feet, leaving it an elegaic couplet, a common meter of Roman love poetry).

This type of common allusion clearly presents much more difficulty in being found: any variety of simple string search (either exact or fuzzy) will not be successful, since only two word forms – *arma* ("arms") and the enclitic *-que* ("and") – are common to both strings.

[4] "But if you held Cephalus in your arms, whom you prefer, you would shout 'run slowly, horses of the night!'"

[5] "I was planning to write about arms and violent wars in a heavy meter."

[6] "I sing of arms and the man."

3 The Logic of Allusion

Clearly we need to add new methods for establishing similarity between two lines beyond simple string matches. This begs the question, however, of how it is we know (as humans) that one passage in a text is an allusion to another. The ultimate criterion of course involves higher-order reason (an allusion must make interpretive sense) but we can identify a number of explicit surface variables that give notice to the presence of an allusion in the first place.

Identical words. A quotation is an allusion where the edit distance between two strings is effectively 0: i.e., all word forms in one span of text are identical with those in another. In sentences 4 and 5, only *arma* and *que* are the same, but they nevertheless provide a necessary anchor for establishing a link between the two passages. While *arma* in both examples here in is the same grammatical case (accusative), many times an alternation occurs as well (e.g., transforming a word from the accusative to the nominative case). We can therefore define "identical" to mean both token identity (*arma* = *arma*) and root form (lemma) identity (*ego* = *me*).

Word order. Syntax in projective languages like English is strongly tied to word order (an adjective, for example, generally modifies the noun that immediately follows it), but for non-configurational languages like Latin and Greek, word order is much more free, especially in the genre of poetry in which allusion is so common. For this reason we treat syntax as a separate variable (see below) and isolate word order as its own phenomenon. For our example above, word order is another cue to the presence of an allusion since both lines begin with the same word, *arma*.

Syntactic similarity. When considering syntax we begin to see the strongest parallels between the two passages. In both sentences, *arma* is involved in coordination as a direct object of a verb. While the head verbs differ (*edere* vs. *cano*) as does the other object involved in coordination (*bella* vs. *virum*), the two structures are syntactically identical.

Figures 1 and 2 present a syntactic tree of each sentence under the formalism of dependency grammar.[7] In both of these trees, the two direct objects of the verbs are headed by the coordinator *que* via the syntactic relation OBJ_CO, while the coordinator is headed by the verb via the relation COORD. While the words themselves vary, the structure is the same.

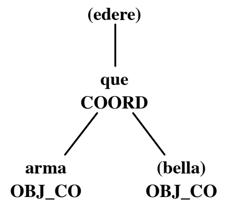

Figure 1: Dependency tree of *arma -que bella edere* ("to write about arms and wars").

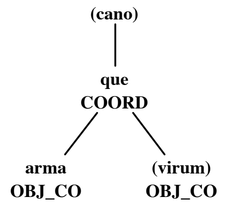

Figure 2: Dependency tree of *arma virumque cano* ("I sing of arms and the man").

Metrical and phonetic similarity The first lines of both of these poems are both written in dactylic hexameter, but the similarity between the two is much closer than that, since the first seven syllables of both lines are metrically identical – two dactyls followed by a stressed syllable and caesura. Additionally, the final long syllable before the caesura is the same in both sentences ("o"), eliciting a further phonetic similarity as well.

(6) Ārmă grăvī nŭmĕrō ‖ ...

(7) Ārmă vĭrūmquĕ cănō ‖ ...

Semantic similarity We can also note that on a semantic level, both of these passages are "about" similar things, at least in this first line (before the arrival of Cupid in Ovid) – in both lines, the author is communicating (via writing or singing) about war (*violenta bella*) and the instruments of war (*arma*).

[7] This is the structural representation of syntax as found in the Latin Dependency Treebank (Bamman and Crane, 2007) and the Prague Dependency Treebank of Czech (Hajič, 1998).

With semantic similarity we can also group another very important variable – cross-language semantic information in the form of translation equivalents. This is extremely important given the reception of these texts across cultures and distant eras. Classical Roman poets themselves are especially fond of borrowing from Homer and Hellenistic poets, but we see the same phenomenon in English as well – one only need to look at Milton's use of the *Aeneid* in *Paradise Lost* to see the level of appropriation, which in its simplest form approaches exact translations of fixed phrases, such as sentences 8 and 9 below, and in its more complex form also involves the host of other variables outlined above.

(8) The Moons resplendent Globe (PL 4.723)

(9) Lucentemque globum lunae (Aen. 6.725)

These five categories represent broad classes of similarity, but of course we must expect others on an ad hoc basis as well – in sentences 4 and 5 from above, we have the additional similarity that both passages come from the privileged first lines of both poems, suggesting a larger structural similarity. While these variables do not illuminate the interpretive significance of an allusion (we can leave that contentious task to critics), they do provide a means by which to discover them in the first place.

4 Discovering Allusions

Our task in automatically discovering allusions is to transform the variables listed above into ones that are computationally tractable. We need to be able to define the precise degree to which two passages are similar in order to quantitatively compare which pairs of passages are more similar to each other than others.

Information retrieval has produced a number of metrics for judging the similarity of two documents. The most widely used of these generally assign a relevance score based on some variation of tf/idf weighting: two documents are similar if they both contain words that occur less frequently in the collection of texts at large. The more uncommon words they share, the greater their similarity.

To establish the similarity between two sentences, we can use the cosine measure as a means of judging their vector similarity.

$$cos(\vec{s},\vec{t}) = \frac{\sum_{i=i}^{n} s_i t_i}{\sqrt{\sum_{i=i}^{n} {s_i}^2}\sqrt{\sum_{i=i}^{n} {t_i}^2}}$$

Here s_i is the tf/idf score for the term i in the source sentence s and t_i is the tf/idf score for that same term in the target comparsion sentence t. We measure each tf/idf by the following formula.

$$(1 + \log(tf_{i,j})) \log \frac{N}{df_i}$$

Here tf_i = the count of term i in sentence j, N = the total number of sentences in the collection, and df_i = the number of sentences in that collection that contain the term i.

The closer this cosine is to 1, the more similar two sentences are. We will use this general framework to inform all of the following variables: the difference between them will be in what exactly constitutes a "term."

4.1 Identical words

Given Latin's rich inflection, we will define two variables for establishing identity between words, token similarity and lemma similarity.

Token similarity. Here we define *term* to be the overt (i.e., inflected) form of the word. This measure reflects a typical search engine query in that it compares two documents (here, sentences) based on how closely their words match each other. More common words between the two documents leads to a greater level of similarity.

Lemma similarity. Here we define *term* to be the uninflected lemma from which the token is derived. In this variable, *omnia vincit amor* ("love conquers all") is identical to *omnia vincuntur amore* ("all things are conquered by love") since the lemmas underlying both are *omnis1 vinco1 amor1*. A measure for lemma similarity addresses the fact that many allusions are not simple quotations – the words that constitute the reference are not bound to their original case as they were used in the target text, but are often given a different grammatical role in the allusion.

4.2 Word order

We can measure the explicit order of words (as distinct from their abstract syntax) with the use of n-grams – specifically bigrams and trigrams, which measure how frequently two or three words appear in linear order. Using the beginning and end of sentences as distinct words of their own (in order to measure when a word begins or ends a line), the phrase *omnia vincit amor* has 4 bigrams (*[start] omnia, omnia vincit, vincit amor, and amor [end]*) and three trigrams: (*[start] omnia vincit, omnia vincit amor, and vincit amor [end]*).

This will let us capture, for instance, that *arma virumque cano* is similar to *arma gravi numero* in that both begin with the bigram *[start] arma*. We can again account for Latin's rich inflection with the use of lemma bigrams and trigrams in addition to tokens. This results in four total word order variables: token bigram, token trigram, lemma bigram and lemma trigram.

4.3 Syntax

The two variables outlined so far form the backbone of information retrieval applications. By considering syntax, we can get beyond simple string resemblance metrics and begin to consider similarities in abstract structure as well.

With syntactic relations, we can specify the true syntactic distance between two phrases (as distinct from simple word order). Several measures of syntactic distance have recently been proposed: Spruit (2006) presents a method for classifying dialects based on previously human-curated variables (e.g., the presence of personal vs. reflexive pronouns etc.); Nerbonne and Wiersma (2006) approximate syntactic distance using part of speech trigrams, which works well for classifying different language groups (adults vs. child) in English (a language with strict word order); and Sanders (2007) measures distance using Sampson's (2000) leaf-ancestor paths, in which each word in a sentence is identified as its path from itself to the top of the syntactic tree (e.g., in a phrase structure grammar: "The"-Det-NP-S/"dog"-N-NP-S/"barks"-V-VP-S). Given Latin's non-projectivity, we have adopted this third measure and augmented it along three dimensions to make it suitable for a dependency grammar.

Figure 3 presents a syntactic tree annotated under a dependency-based grammar. Since dependency grammars do not have intermediate phrase structures such as NP or VP, we take our basic syntactic structure to be a child-parent relationship between words themselves. How we represent those words constitutes the first dimension:

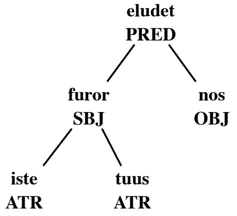

Figure 3: Dependency tree of *furor iste tuus nos eludet* ("that madness of yours will play with us"), Cicero, *In Catilinam 1.1*.

- part of speech: adj:noun:verb
- token: iste:furor:eludet
- lemma: iste1:furor1:eludo1

The second dimension is the length of the path. While Sanders' metric identifies each word by its full path from itself to the top of the tree, we can use a number of intermediary paths to assert similarity as well. Since a full path from the word itself to the top of the tree is very unlikely to be repeated across sentences, we approximate it by considering only paths of lengths 2 and 3 (bigrams and trigrams): a path of length 2 would for instance be comprised of "adj:noun"/"iste:furor"/etc. while a path of length 3 would be comprised of "adj:noun:verb" (as above).

The third dimension is the presence or absence of the syntactic label. Dependency grammars differ from phrase structure grammars by providing an explicit relation between words (where phrase structure grammars often imply it by position – a subject, for example, is an NP that c-commands a VP). Using the syntactic labels specified in the Latin Dependency Treebank (Bamman and Crane, 2007), a labeled path would be comprised of "iste:ATR:furor:SBJ:eludet" for token trigrams, while an unlabeled path would leave this information out (as above).

These three dimensions provide 12 distinct syntactic variables for each word in a sentence, ranging from least explicit (unlabeled part of speech bigrams ["adj:noun"]) to most (labeled token trigrams ["iste:ATR:furor:SBJ:eludet"]). The most explicit variables will have the lowest inverse document frequencies and will therefore be the most informative for judging similarity if present, while the least explicit variables will still provide a back-off means to provide some similarity in the event of a more explicit non-match.

4.4 Metrical/phonetic similarity and semantic similarity

While we do not implement metrical/phonetic or semantic similarity measures in what follows, we can address the means by which we could do so in the future.

We can measure metrical and phonetic similarity in a manner similar to the term frequencies used in the variables above, by comparing the meter of two passages (this of course requires metrically annotated texts). Meter in this case can be seen as a language with two letters, *long*(_) and *short*(◡), and we can judge the similarity between two meters as a simple string comparison of that representation.

We can judge the semantic similarity between two words using either monolingual clustering techniques such as latent semantic analysis (which notes, for example, that an *apple* is semantically close to an *orange* since both appear often with words such as *eat* and *tree*) (Deerwester et al., 1990), or by cross-language translation equivalents (such as those induced in the course of parallel text alignment (Och and Ney, 2003)), which notes the frequency with which a word in one language (such as *oratio* in Latin) is translated by different terms (e.g., *speech* vs. *prayer*).

5 Evaluation

We evaluated the first three variable classes above (word identity, word order and syntax) on a collection of 14 texts from 5 Latin poets – Catullus (*Carmina*), Ovid (*Metamorphoses*, *Amores*, *Epistulae*, *Medicamina Faciei Femineae*, *Ars Amatoria*, *Remedia Amoris*), Vergil (*Aeneid*, *Eclogues*, *Georgics*), Propertius (*Elegies* I), and Horace (*Carmina*, *Satyrarum libri*, *De Arte Poetica liber*).

While the word identity and word order variables can be calculated on unstructured text, we need syntactically parsed data in order to measure syntactic similarity. To create this, we trained McDonald et al.'s dependency parser (McDonald et al., 2005) on the manually curated data in the Latin Dependency Treebank and used it to parse all of the texts in our collection.[8]

Author	Words	Sentences
Ovid	141,091	10,459
Vergil	97,495	6,553
Horace	35,136	2,345
Catullus	14,793	903
Propertius	4,867	366
	293,382	20,626

Table 1: Composition of the test corpus by author.

After finding the most similar sentences for each of the 20,626 sentences in our collection, we filtered the results to require a lower limit for sentence length in order to find meaningful pairs (short sentences such as *Quid est?* can be found across many authors and are not allusions even though they match exactly) and to avoid sentence pairs that are both found in the same immediate context (e.g., Catullus' poem 61, where a chorus of the same 7 words is exactly repeated 11 times throughout the poem).[9]

The results are encouraging: while a detailed quantitative evaluation must await the creation of a test corpus of canonical allusions, we can at least now provide a list of the closest matches for all sentences in our collection. For any given sentence, further research will of course be necessary to discern whether it represents a real allusion, but the highest scoring pairs in our experiment tend to be strong examples. Sentences 10 and 11, for instance, present one such pair from Ovid and Vergil with a similarity score of .173.

(10) **Innumeras urbes atque aurea tecta** videbis, / Quaeque suos dicas templa decere deos (Ov. Ep. 16)[10]

(11) Iam subeunt **Triviae lucos atque aurea tecta** (Verg., Aen. 6.13)[11]

[8] In a tenfold test on the treebank data itself, we measured the parser's unlabeled accuracy to be 64.99% and its labeled accuracy to be 54.34% (Bamman and Crane, 2008).

[9] *o Hymen Hymenaee io, o Hymen Hymenaee.*

[10] "You will see innumerable cities and golden roofs, and temples that you would say are fitting to their gods."

[11] "Already they enter Trivia's groves and golden roofs."

Sentences 12 and 13 likewise present a pair from Ovid and Catullus with a score of .141.

(12) **nulli illum iuvenes, nullae tetigere puellae** (Ov., Met. 3.353)[12]

(13) idem cum tenui carptus defloruit ungui / **nulli illum pueri, nullae optavere puellae** (Cat., Carm. 62)[13]

The strongest matches, however, came within authors, who often sample their own work in other contexts. This occurs most often by far in Vergil, where the re-appropriation involves exactly repeating complete sentences (9 instances), exactly repeating substantial sentence fragments (23 instances),[14] and more significant modifications.

Additionally, since our weights are based on preset variables, the process by which we come to the most similar match is transparent. Table 2 presents the term weights for several of the highest and lowest variables at play in establishing the similarity between sentences 12 and 13 above.

This table presents the clear importance of using syntax as a method for establishing the similarity between sentences – three of top four variables that have linked these two sentences to each other involve syntax (e.g., *nullae* depends on *puellae* in both sentences as an attribute).[15]

Our search for *loci similes* to our original allusion from above – Ovid's *Arma gravi numero violentaque bella parabam* – illustrates well the importance of bringing a variety of information to the search. The closest sentences to Ovid's original line all bear some similarity to it on both a lexical and syntactic level (as sentences 1 and 2 demonstrate below). Our target sentence of Vergil (*Arma virumque cano* ...), however, only shows up in 11th place on the list.

Variable	tf/idf
nullae:puellae:ATR	9.24
nullae:puellae	9.24
nulli/illum	9.24
p:SBJ_EXD_OBJ_CO:u:COORD:v	9.24
,/nullae	8.84
nullus1:puella1	8.55
nullus1:puella1:ATR	8.32
nullae	8.55
...	...
nulli	6.30
puellae	5.55
illum	5.34
a:n:ATR:v:SBJ	1.67

Table 2: Sample of variable contribution. Components separated by a colon represent syntactic relations; those with slashes are n-grams.

1. Arma procul currusque virum miratur inanes (.059) (Aen. 6.651)[16]

2. Quid tibi de turba narrem numeroque virorum (.042) (Ov., Ep. 16.183)[17]

11. Arma virumque cano, Troiae qui primus ab oris Italiam, fato profugus, Laviniaque venit litora, multum ille et terris iactatus et alto vi superum saevae memorem Iunonis ob iram (.025) (Aen. 1.1)[18]

This is understandable given the variables we have implemented – the first three sentences do indeed bear a closer similarity to the original without being diluted by extra words (since our cosine value normalizes for sentence length). We hope in the future to be able to include other important variables (such as metrical similarity) as well.

[12] "No youths, no girls touched him."

[13] "This same one withered when plucked by a slender nail; no boys, no girls hope for it."

[14] Here "substantial" means at least seven consecutive identical words.

[15] Note that the labeled syntactic bigram nullae:puella:ATR has the same tf/idf score as the unlabeled nullae:puellae since all instances of *nullae* depending on *puella* in our automatically parsed corpus do so via the relation ATR.

[16] "At a distance he marvels at the arms and the shadowy chariots of men."

[17] "What could I tell you of the crowd and the number of men?"

[18] "I sing of arms and the man, who first from the borders of Troy, exiled by fate, came to the Lavinian shores – much was he thrown about on land and sea by force of the gods on account of the mindful anger of cruel Juno."

6 Conclusion

Allusion is by nature an oblique art; its very essence – referring to something that the audience already knows – gives it the opportunity to be highly economical in its expression. Since even a single word or structure can refer to another text, we must leverage as many different varieties of information as we can in order to discover them, from lexical information to syntax and beyond. We have defined five different variable classes that contribute to the surface realization of allusion, and have implemented a system that includes three of those five. By considering the abstract structure of sentences, we are able to effectively search Latin without being encumbered by its flexible word order and rich inflectional morphology, which allows similar sentences to be expressed in a variety of ways. While we have designed this method for a collection of Classical texts, we expect that it can also be used to improve the robustness of searches in any language.

7 Acknowledgments

Grants from the Digital Library Initiative Phrase 2 (IIS-9817484), the National Science Foundation (BCS-0616521) and the Andrew W. Mellon Foundation (#40700635) provided support for this work. Thanks are due also to Meg Luthin and Molly Miller for their invaluable research assistance.

References

David Bamman and Gregory Crane. 2007. The Latin Dependency Treebank in a cultural heritage digital library. In *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, pages 33–40, Prague.

David Bamman and Gregory Crane. 2008. Building a dynamic lexicon from a digital library. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2008)*.

Harold Bloom. 1973. *The Anxiety of Influence; A Theory of Poetry*. Oxford University Press, New York.

Paul Clough, Robert J. Gaizauskas, Scott S. L. Piao, and Yorick Wilks. 2002. METER: Measuring text reuse. In *Proceedings of the ACL*, pages 152–159.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In *COLING '04*. Association for Computational Linguistics.

Jan Hajič. 1998. Building a syntactically annotated corpus: The Prague Dependency Treebank. In Eva Hajičová, editor, *Issues of Valency and Meaning. Studies in Honor of Jarmila Panevová*, pages 12–19. Prague Karolinum, Charles University Press.

William Irwin. 2001. What is an allusion? *The Journal of Aesthetics and Art Criticism*, 59:287–297.

Paul Oskar Kristeller. 1979. *Renaissance Thought and Its Sources*. Columbia University Press, New York.

John Lee. 2007. A computational model of text reuse in ancient literary texts. In *Proceedings of the ACL*, pages 472–479, Prague, Czech Republic.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the HLT-EMNLP*, pages 523–530.

Vladimir Nabokov. 1991. *The Annotated Lolita. Edited, with preface, introduction, and notes by Alfred Appel, Jr.* Vintage Books, New York.

John Nerbonne and Wybo Wiersma. 2006. A measure of aggregate syntactic distance. In *Proceedings of the Workshop on Linguistic Distances*, pages 82–90.

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

Geoffrey Sampson. 2000. A proposal for improving the measurement of parse accuracy. *International Journal of Corpus Linguistics*, 5(1):53–68.

Nathan C. Sanders. 2007. Measuring syntactic difference in British English. In *Proceedings of the ACL2007 Student Research Workshop*, pages 1–6.

Marco Rene Spruit. 2006. Measuring syntactic variation in Dutch dialects. *Literary and Linguistic Computing*, 21:493–506.

Masayuki Takeda, Tomoko Fukuda, Ichiro Nanri, Mayumi Yamasaki, and Koichi Tamari. 2003. Discovering instances of poetic allusion from anthologies of classical Japanese poems. *Theoretical Computer Science*, 292(2):497–524.

R. F. Thomas. 1986. Vergil's *Georgics* and the art of reference. *Harvard Studies in Classical Philology*, 90:171–98.