Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections

Russell Horton¹, Mark Olsen², and Glenn Roe²

¹University of Chicago, The University of California at San Diego, ²University of Chicago

russ@artfl.uchicago.edu

markymaypo57@gmail.com | http://artfl-project.uchicago.edu

glenn@artfl.uchicago.edu

Abstract

The following article describes a simple technique to identify lexically-similar passages in large collections of text using sequence alignment algorithms. Primarily used in the field of bioinformatics to identify similar segments of DNA in genome research, sequence alignment has also been employed in many other domains, from plagiarism detection to image processing. While we have applied this approach to a wide variety of diverse text collections, we will focus our discussion here on the identification of similar passages in the famous 18th-century *Encyclopédie* of Denis Diderot and Jean d'Alembert. Reference works, such as encyclopedias and dictionaries, are generally expected to "reuse" or "borrow" passages from many sources and Diderot and d'Alembert's *Encyclopédie* was no exception. Drawn from an immense variety of source material, both French and non-French, many, if not most, of the borrowings that occur in the *Encyclopédie* are not sufficiently identified (according to our standards of modern citation), or are only partially acknowledged in passing. The systematic identification of recycled passages can thus offer us a clear indication of the sources the *philosophes* were exploiting as well as the extent to which the intertextual relations that accompanied its composition and subsequent reception can be explored. In the end, we hope this approach to "Encyclopedic intertextuality" using sequence alignment can broaden the discussion concerning the relationship of Enlightenment thought to previous intellectual traditions as well as its reuse in the centuries that followed.

KEYWORDS / MOTS-CLÉS

sequence alignment, intertextuality, text recycling, plagiarism detection, document similarity, Encyclopédie, Diderot, French Enlightenment

Contents

- Introduction
- The Shared Passage Problem
Introduction

Relationships between texts are multifaceted and complex, ranging from the authority of directly attributed quotations to the subtlest hints of allusion and influence. Indeed, some theorists have gone so far as to suggest that these relationships may in fact lie outside of the conscious articulation of the writer, in the assumptions of shared language and culture. This is demonstrated most notably in Roland Barthes' seminal article "Texte (théorie du)" first published in the *Encyclopædia Universalis* (Barthes). Tracing the connections between texts at all levels is a core element of humanistic inquiry, illuminating the meanings and receptions of a text by placing it in a larger intellectual and cultural context. We might even assert, to borrow a phrase from Harold Bloom, that understanding in humanistic textual scholarship is the art of knowing the hidden roads that go from text to text (Bloom).

Finding these "hidden roads" of intertextuality, however, has always been subject to the limitations of human reading and recollection. One tantalizing promise of emerging digital libraries is that computer technology may augment the scholarly functions of reading and recollection by identifying related passages in very large collections, thus allowing for the discovery of relationships between textual passages on a far greater scale than any individual could reasonably manage. The ideal result of such processing might be envisaged as an "allusion" detector, wherein the sense, meaning, and content of related passages move beyond any singular dependence upon specific lexical constructions. Of course, machines do not read texts at all, if by reading we mean the evocation of mental images implied by allusions or other literary techniques to focus a reader's attention in some way. Nevertheless, there are more prosaic forms of intertextuality that can be identified through the use of machine-assisted reading. At its most rudimentary level, this sort of intertextuality would allow scholars to isolate networks of quotations, borrowings, plagiarism, or textual reuse across large collections of texts, providing them with useful ways to think about the intellectual origins of a particular document.

The following article describes a simple technique to identify lexically-similar passages in large collections of text using algorithms borrowed from the field of bioinformatics. This class of algorithms—generally referred to as "sequence alignment"—is used to identify similar segments of DNA in genome research, among other applications, most notably in the BLAST algorithm (Altschul) and its variants and successors. A subset of the more general "longest common string or substring" problem in computer science, these techniques are also used in many other domains, from plagiarism detection to image processing. While we have applied this technique to a variety of large humanistic text collections over various languages and time periods, we will focus our discussion here on the identification of borrowed passages in the famous eighteenth-century *Encyclopédie* of Denis Diderot and Jean d'Alembert from a wide range of outside sources, both French and non-French. Reference works, such as encyclopedias and dictionaries, are generally expected to "reuse" or "borrow" passages from many sources. In the case of Diderot and d'Alembert's *Encyclopédie*, many, if not most, of these borrowings are not sufficiently identified (according to our standards of modern citation), or are only partially acknowledged in passing.
Identification of recycled passages would thus offer us a clear indication of the sources the *philosophes* were drawing upon while also allowing for a systematic discussion of the relationship of Enlightenment thought to previous intellectual traditions.

### The Shared Passage Problem

Text recycling, or "plagiarism"—to use a more loaded term—has been in the news in recent years, owing in part to changing notions of intellectual property fostered by the Internet. Many well respected authors, including historians Stephen Ambrose and Doris Kearns Goodwin, have faced charges of plagiarism, the ramifications of which include serious damage to their reputations (see Bridgewater College Writing Center for a full discussion of recent plagiarism cases). Such concerns are not limited to academic writers. Chris Anderson, for example, was recently, and rather ironically, accused of pillaging Wikipedia (among other sources) in his latest book, *Free: The Future of a Radical Price*, which was itself an examination of the economics of internet publication (see Jaquith, Champion, and Anderson). Davis Schneiderman, in an op-ed article in the *Chicago Tribune*, gives a more nuanced reaction to this sort of critical backlash against "plagiarists," declaring that "[t]he concept of original content is pure fiction" (Schneiderman). He goes on to argue that "[w]e cannot start crediting every word we write or speak, because our entire culture is based upon recycling and plagiarism, and it's not just a function of the Internet."

Following several examples of situations in which artists reworked and recycled previous texts, Schneiderman concludes with the suggestion that modern notions of plagiarism derive from legal notions of copyright and protection of publishers' economic interests rather than authorial genius, creativity, or property.

Over the years, acknowledgments concerning the function and importance of "recycling" in literature have become more common, from Thomas Mann's notion of "higher cribbing" (*Adorno and Mann*) to Lethem's "ecstasy of influence." Indeed, some recyclers are known to be better than others. As T.S. Eliot famously reminds us:

> Immature poets imitate; mature poets steal; bad poets deface what they take, and good poets make it into something better, or at least something different. The good poet welds his theft into a whole of feeling which is unique, utterly different from that from which it was torn; the bad poet throws it into something which has no cohesion. A good poet will usually borrow from authors remote in time, or alien in language, or diverse in interest. (Eliot 125)

Micheal Maar makes a convincing case that Nabokov's *Lolita* related clearly to Heinz von Lichberg's 1916 short story with the same title and some similar plot elements. But, Maar suggests, this is not a case of "plagiarism" or even cryptomnesia. Rather, he says, "literature has always been a huge crucible in which familiar themes are continually recast. ... Nothing of what we admire in *Lolita* is already to be found in the tale; the former is in no way deducible from the latter" (*Maar 59*). Identification of "recycled" passages should not, we suggest, necessarily impugn the reputation of the borrower, but may well be the reflection of a vital literary tradition and culture.

Consider the following two passages which were identified using PhiloLine, our sequence alignment extension to PhiloLogic, in the main ARTFL database. On the left is an extract from Louis Le Comte's *Nouveaux mémoires sur l'état présent de la Chine* and on the right is an aligned passage from *La crise de la conscience européenne*, by Paul Hazard, who was an influential French intellectual historian of the first half of the 20th century.
que se trouve le véritable saint. Et cette sentence estoit tellement gravée dans l'esprit des sçavans, que soixante-cinq ans après la naissance de nostre seigneur, l'empereur Mimti touché de ces paroles, et déterminé par l'image d'un homme qui se presenta à luy durant le sommeil venant de l'occident, envoya de ce costé-là des ambassadeurs, avec ordre de continuer leur voyage jusqu'à ce qu'ils eussent rencontré le saint que le ciel luy avoit fait connoistre. C'estoit à peu près le temps auquel Saint Thomas preschoit dans les Indes la loy chrétienne; et si ces mandarins eussent suivi leurs ordres, peut-estre que la Chine auroit profité de la prédication de cet apostre. Mais les dangers de la mer, qu'ils craignirent, les obligea de s'arrester à la premièr isle, où ils trouveront l'idole Fo ou Foe, qui avoit déja corrompu les Indes plusieurs siecles auparavant, de son execrable doctrine. (LeComte [1696])

soyvent, tel un prophète : Dans l'Occident se trouve le véritable saint. Soixante-cinq ans après la naissance du Christ, l'Empereur Mimti, interprétant cette parole du Maître, et sollicité par un songe, envoya vers l'Occident des ambassadeurs, avec ordre de continuer leur voyage jusqu'à ce qu'ils eussent rencontré le saint. En ce temps-là, saint Thomas prêchait dans les Indes la foi chrétienne; et si ces mandarins s'étaient acquittés de leur mission, au lieu de s'arrêter dans la première île, à cause du danger de la mer, peut-être la Chine aurait-elle fait partie de l'Église romaine... (Hazard [1935])

Hazard's reuse of this passage is unattributed, forming part of a "panoramic survey of ideas" of late 17th-century France, in which the mysteries of China are purported to play an important role. Hazard does mention elsewhere another work by Le Comte, Des cérémonies de la Chine, as well as other Jesuit reports from Asia during the same period. The recycled passages tell the rather implausible story of how, but for the failings of will or nerve of two Chinese mandarins, China could well have been converted to Christianity in the time of the apostles. A perfectly reasonable sentiment for the 17th-century Jesuit missionary, but less comprehensible for a 20th-century intellectual historian, unless by this recycling Hazard wished to provide the reader with a taste of the sensibilities of a previous age. While one might be tempted to think this is conscious "plagiarism" or simply an artifact of sloppy note-taking, a more charitable take on this may be that of inadvertent cryptomnesia, an echo of a hidden textual reminiscence, or even a quotation or paraphrase from memory. Such echoes of a possible common place, theme, or borrowed passage, across more than 250 years, had one more hop to reach us. We selected this example because one of us (Olsen) read Hazard's work in English translation as an undergraduate, and was immediately able to identify the passage in his long-forgotten copy of the work.

Of course, text recycling may well be more direct in nature and used for either commercial or entertainment purposes. It is well known, for example, that Shakespeare's rather racy 1593 poem Venus and Adonis, itself a borrowing from Ovid's Metamorphoses, was endlessly alluded to in the decades following its publication. In the two following passages, again detected using sequence alignment, we see a passage from the poem being interspersed in a dialogue of a much less well known play written several years later.

She locks her lily fingers one in one. "Fondling," she saith, "since I have hemmed thee here Within the circuit of this ivory pale, I'll be a park, and thou shalt be my deer; Feed where thou wilt, on mountain or in dale: Graze on my lips; and if those hills be dry, Stray lower, where the pleasant fountains lie." Within this limit is relief enough....

Pre. Fondling, said he, since I haue hem'd thee heere, VVithin the circuit of this Iuory pale, Dra. I pray you sir help vs to the speech of your master.
Pre. Ile be a parke, and thou shalt be my Deere: He is very busie in his study. Feed where thou
wilt, in mountaine or on dale. Stay a while he will come out anon. Graze on my lips, and when those mounts are drie, Stray lower where the pleasant fountains lie. Go thy way thou best booke in the world.

Ve. I pray you sir, what booke doe you read?

In this instance, the borrowing is acknowledged in the answer to the final question in Markham: "A booke that neuer an Orators clarke in this kingdome but is beholden vnto: it is called maides philosophie or Venus and Adonis: Looke you gentlemen, I haue diuers other pretty booke" (Shakespeare; Markham). Nothing like a juicy tidbit to enliven a comedy, or better yet, an academic paper.

Both passages give some indication of the technical requirements needed for similar passage identification in literary and historical texts, which add significant complications that may not be encountered in similar plagiarism detection schemes or other models of common substrings. Clearly, this is a long way from block copying passages from a website. From a cursory examination of the passages one can see that there are significant differences in orthography. In the Shakespeare example, we find a fairly modernized version of the text, while in Markham the original orthography has been maintained. In both passages, there are a significant number of insertions and deletions, even of entire lines and sentences. And one should further note that there are some slight differences in word order. As we shall see, many borrowings are often quite inexact, taking the form of reworked passages, extracts from various parts of a document compiled in a different order, or simply differences in wording. These variations result from the multifaceted nature of text recycling, which is often comprised of passages quoted from memory, paraphrases to situate a passage into another work, or various other forms of modification. Of course, even more precise borrowings may come from long lost or rare editions, intermediate works, or as we shall see, translations by an author or a contemporary translator. We might add one more layer of complication: many of the works found in current digital libraries are based on the output of uncorrected optical character recognition (OCR). Thus, faced with the intricacies of text recycling in historical and literary works, along with the frequently degraded status in which these texts are currently made available, we are required to formulate computational approaches that are flexible and tolerant of variation.

Sequence Alignment

Identifying text reuse is a specific case of the more general problem of sequence alignment; that is, the task of identifying regions of similarity shared by two strings or sequences, often thought of as the longest common substring problem. This technique is widely applied in the field of bioinformatics, where it is used to identify repeated genetic sequences (Gusfield). Sequence alignment is also the basis for many plagiarism detection applications that attempt to identify borrowings in running text or even computer programs and code (Lyon; Bourdilliet). Moreover, sequence alignment can also generate similarity scores used as metrics in machine learning models of text reuse, as demonstrated by the METER project on the reuse of wirefeed copy in newspapers (Clough).
Our adaptation of these techniques for detecting literary text reuse is available as open source software releases of two related systems: PAIR (Pairwise Alignment for Intertextual Relations), a one-to-many comparison system that allows the user to submit a document or part of a document to be aligned against a pre-indexed corpus; and PhiloLine (PhiloLogic Alignment), a batch mode many-to-many aligner that compares all documents or document parts to all other documents or parts in the same database or between two collections. PhiloLine generates output either as static alignment reports or as structured data for subsequent search and analysis and is lightly dependent on PhiloLogic, our primary text analysis system. While functionally similar, the two systems are quite distinct in implementation, given the significant differences in processing requirements between the interactive and batch mode methods. When discussing the sequence alignment approach in general, we shall refer to PAIR.

PAIR works by treating documents as ordered sets of \( n \)-grams or "shingles" formed by each overlapping sequence of \( n \) words in the document. Preprocessing, such as the removal of function words and short words and the reduction of orthographic variants (accents, spelling changes, case folding, etc.), is performed during shingle generation. This has the effect of folding numerous shingles into one underlying form for matching purposes, thus eliminating minor textual variations, which makes matching more flexible or "fuzzy." It also somewhat reduces the overall number of unique shingles, which aids speed of search. Rousseau's famous declaration in the *Contrat social* provides a good example:

L'homme est né libre, et partout il est dans les fers. Tel se croit le maître des autres, qui ne laisse pas d'être plus esclave qu'eux.

rendered as trigrams (\( n \)-grams with \( n = 3 \)), with short and function words removed and accents and case flattened, would look like:

<table>
<thead>
<tr>
<th>trigram</th>
<th>doc</th>
<th>sequence</th>
<th>bytes</th>
</tr>
</thead>
<tbody>
<tr>
<td>homme_libre_partout</td>
<td>755</td>
<td>208-213</td>
<td>5084-31</td>
</tr>
<tr>
<td>libre_partout_fers</td>
<td>755</td>
<td>211-218</td>
<td>5098-38</td>
</tr>
<tr>
<td>partout_fers_croit</td>
<td>755</td>
<td>213-221</td>
<td>5108-46</td>
</tr>
<tr>
<td>fers_croit_maitre</td>
<td>755</td>
<td>218-223</td>
<td>5132-33</td>
</tr>
<tr>
<td>croit_maitre_laisse</td>
<td>755</td>
<td>221-228</td>
<td>5149-42</td>
</tr>
<tr>
<td>maitre_laisse_esclave</td>
<td>755</td>
<td>223-233</td>
<td>5158-58</td>
</tr>
</tbody>
</table>

The shingles are indexed with a document identifier, word sequence range, and the source file byte position and size of the corresponding section of the text. An ordered list of shingles is generated for each document. Each shingle also becomes a key in an index of all shingle occurrences for the entire collection of documents, so that all occurrences of a given shingle can be readily retrieved. The first step in identifying shared text sequences between documents is finding a single shared shingle. All shared shingles within the corpus are identified by finding all shingles with more than one occurrence. A new document can be shingled and then compared with the entire corpus by finding all occurrences of the shingles contained in that document in the master shingle index for the corpus (Kolak; Schilit).
Once a shingle match is identified, the shingles within a defined window surrounding the shared shingle in each document are retrieved and examined for other matches. PAIR allows the user to set criteria for the acceptability of matches, such as the minimum overlap in shingles between the two sets, the minimum length of a shared shingle sequence, or the maximum number of consecutive gaps allowed between matching sequences in either set. If the criteria are met, the match is expanded, examining wider contexts in each document; once the criteria are violated, the match is terminated and recorded. The user-configurable parameters for match retention and expansion allow for fine tuning of results, balancing the scholar's desire for comprehensiveness with his or her tolerance for sifting through short, trivial, or tenuous associations. The maximum gap parameter, in particular, allows for flexibility in matches, reflecting possible differences in spelling, errors in data capture, insertions or deletions, and other variations. Other measures could also be included, such as assigning high scores to relatively rare n-grams in a particular data set or assigning a similarity threshold (based on a measure such as string edit distance) to the matching of n-grams rather than requiring an exact match.

As we have seen, the most obvious use for this technique is the identification of quotations, whether they are clearly cited by the borrowing author or simply used without attribution. The French Revolutionary theorist Jean-Paul Marat, for example, appears to have borrowed the following passage without attribution in his *Les chaînes de l'esclavage* (1792):

... prévaut d'un silence qu'il empêche de rompre, ou des irrégularités qu'il fait commettre, pour supposer en sa faveur l'aveu de ceux que la crainte fait taire, et pour punir ceux qui osent parler ...

(93)

from Rousseau's *Du contrat social* (1762):

... prévaut du silence qu'il les empêche de rompre, ou des irrégularités qu'il leur à fait commettre, pour supposer en sa faveur le vœu de ceux que la crainte à fait taire, ou punir ceux qui osent parler ...

(316)

That Marat borrows from Rousseau—although only once in this text—is hardly surprising, given the centrality of Jean-Jacques' thought in revolutionary ideology. Of potentially greater interest are the numerous citations and borrowings from *Du contrat social* by counter-revolutionary thinkers, such as Louis de Bonald, Joseph de Maistre, and l'Abbé Barruel. As might be expected, there can be significant variations between passages that may require human inspection to confirm as a valid match, or at least one of interest. The alignment found in the following pair, for example, is probably quite reasonable:

Rousseau, Jean-Jacques, 1712-1778, [1762], *Emile, ou, De l'éducation*:

... prive de leurs facultés spirituelles, mais non de leur qualité d'homme ni par consequent du droit aux bienfaits de leur créateur. Pourquoi donc n'en pas convenir aussi pour ceux qui sequestrés de toute société dès leur enfance auroient mené une vie absolument sauvage, privés des lumières qu'on n'acquiert que dans le commerce des hommes? Car il est d'une impossibilité démontrée qu'un pareil sauvage pût jamais elever ses reflexions jusqu'à la connoissance du vrai dieu. La raison nous dit qu'un homme n'est punissable que par les fautes de sa volonté, et qu'une ignorance invincible ne lui sauroit être imputée à crime. D'où il suit que devant la justice éternelle tout homme qui croiroit, s'il avoit les lumières nécessaires, est rejeté croire ... (556)
... vous-même sur Jean-Jacques, habitant d'une île déserte, n'ayant jamais vu d'autre homme que lui, par la raison seule découvrant l'être-suprême, remplissant tous ses devoirs envers Dieu, et sur l'impossibilité démontrée qu'un être privé des lumières qu'on n'acquiert que dans le commerce des hommes, pût jamais s'élever à la connaissance du vrai dieu. J'espère aussi que, dans le second texte, comme dans le premier, vous verrez très-bien qu'il ne s'agit pas seulement des attributs de Dieu et de sa nature, mais de son existence; qu'ainsi l'affirmation et la négation tombent précisément sur le même objet. (64)

Since one can identify similar passages that both predate and postdate a particular work, one can at a glance examine borrowings in that text and subsequent citations. Even in the case of attributed citations, linkage via sequence alignment provides an invaluable technique for navigating between documents in a far more reliable fashion than attempting to identify cited passages from citation information alone, which can be incomplete for earlier works or tied to older and/or unavailable editions. Matching sequences can have a significant degree of variation, depending on the match parameters. This variability is required to identify possible borrowings which have significant errors, insertions, deletions, or other modifications. A trigram sequence aligner using a 5 n-gram maximum gap, for example, detected the following match:

Source: ... traitent avec le dernier mépris: les uns les chargent d'injures, et les autres de coups.
Comment, chien d'esprit, luy disent-ils quelquefois, nous te logeons ...

Target: ... traitent avec le dernier mépris: comment, chien d'esprit, lui disent-ils quelquefois, nous te logeons ...

The source is Louis LeComte's *Nouveaux mémoires sur l'état présent de la Chine* (Paris, J. Anisson, 1696), a description of China by a Jesuit missionary. The target passage, again an unattributed borrowing, is found in a work with little relationship to China, the Marquis d'Argens' *Lettres juives* (La Haye, P. Paupie, 1738). Comparing the two passages in greater context shows that the relationship between the two passages extends beyond what the aligner detected, because of variations in language and word choice:

| Car il arrive assez souvent qu'après avoir esté bien honorez, si le peuple n'obtient pas d'eux ce qu'il demande, il se lasse enfin et les abandonne comme des dieux impuissans; d'autres les traitent avec le dernier mépris: les uns les chargent d'injures, et les autres de coups. Comment, chien d'esprit, luy disent-ils quelquefois, nous te logeons dans un temple magnifique, tu es bien doré, bien nourri, bien encensé, et après tous ces soins que nous prenons de toy, tu es assez ingrat pour nous refuser ce qui nous est necessaire? (Le Comte) | ... vouloir excuser la ridicule conduite des chinois envers leurs dieux. Ils les honorent et les respectent autant qu'ils croient en recevoir du bien: mais, dès qu'ils n'en obtiennent pas ce qu'ils leur demandent, ils les traitent avec le dernier mépris: comment, chien d'esprit, lui disent-ils quelquefois, nous te logeons dans un fort beau temple, nous te nourrissons à gogo, tu es bien doré, bien encensé: et tu ne nous accordes pas les graces, que nous te demandons! (Marquis d'Argens) |

Relaxing the maximum shingle gap to 8 or more allows for the identification of the full borrowing.

The available parameter adjustments for PhiloLine, and to a lesser degree for PAIR, are designed to balance the
overall number of matches detected against the number of matches that a user would consider similar enough and salient enough to be of interest. If the matching is set too loosely, the user will have to wade through a large collection of short matches or common, stock phrases. If the settings are too strict, however, interesting matches that are short or heavily reworked may be missed. Speed and memory-use constraints also play a role in tuning the parameters for an optimal run. For PhiloLine, there are two sets of parameters to be adjusted. The first is for \(n\)-gram or shingle generation and the second for the matching process itself. Shingle generation parameters include the shingle size, stop words and/or short word deletion, omission of numerals, accented character flattening, virtual normalization—which can be look-up tables or rules—definitions of what are considered to be words, apostrophe handling and the like. The most important parameters for shingle generation are the sizes of the shingles; a larger shingle size will result in fewer repeated occurrences in a given collection, but the resolution of the alignment will be lower, as will the word filtering settings.

The parameters for the PAIR alignment process are even more numerous than those for PhiloLine. Some of these are a matter of personal preference, such as whether to generate flat (HTML) or delimited (for loading in MySQL) output, where to find shingle databases, and other important, but uninteresting, details. The most important parameters include the minimum number of shingles to be considered a match (span) and the maximum number of un-matched shingles allowed within a match (gap). Generating larger shingles would suggest aligning with a small span and longer gap, since longer \(n\)-grams (say 8 words) would be far rarer than short \(n\)-grams. One may also set a near duplicate document identification scheme, to eliminate the (time-consuming and output-flooding) alignment of documents that are duplicates or contain very significant duplications, by indicating the overall maximum matching shingle count. We have also found that a simple approach to eliminate "banal matches" is effective. By setting the appropriate parameters, the system checks to see if short matches contain a set number of highly frequent \(n\)-grams in the entire collection. These may be formulaic constructions, such as nomine Patris et Filii et Spiritus Sancti, amen or vostre tres-humble et tres-obeïssant et tres-affectionné. Of course, not all short matches are uninteresting. One interesting example is erreurs en desirs, Les mortels insensés promènent leur folie, a common citation of a short passage of Dryden in translation found in texts by Voltaire and Madame de Genlis. It should be clear that parameter setting is, to a certain degree, a matter of trial and error, and depends on the task at hand, the kinds of documents in question, the patience of the investigator, what might considered to be banal, and so on (for a full description of parameters, please consult the PAIR/PhiloMine site: 
http://code.google.com/p/text-pair/).

We believe that PAIR is largely effective in identifying similar passages with extensive variations. As an experiment, we used PhiloLine to align citations of Montesquieu's De l'esprit des lois found in Emile Littré's Dictionnaire de la langue française (1872-1877) with the text of Montesquieu's works found in the ARTFL database. The Littré is well suited to systematic assessments, since it contains over 290,000 citations from some 3,900 authors in a well structured data set. This example is a Littré citation from the thirteenth book of Montesquieu's Esprit des lois:

En Angleterre, l'administration de l'accise a été empruntée des fermiers

In our first experiment, we attempted to align the 2,344 Montesquieu citations found in the Littré with the all works by Montesquieu contained in the ARTFL-Frantext database. The parameters for this comparison are using trigrams, span=2, gap=10, banal=top 20, stopwords= top 30. Of the 2,344 citations, we were able to identify
1,560 aligned passages between the Littré and our subset of Montesquieu texts. Using the same parameters, we found 894 aligned passages of the 1,211 Littré references to Montesquieu's *De l'esprit des lois* alone. The passages that were not aligned were determined to be very short; 266 out of over 300 were fewer than 12 words long, including stop words. A full citation of this sort, "Le supplice de la honte," for example, would not generate a single trigram. Small passage size significantly amplifies the variants that would be spanned in larger matches. Accordingly, the following pair may or may not be aligned depending on the variant parameters one sets before performing the alignment task:

Il mit peu de chose au hasard (Littré). Il mit peu de chose au hazard (Montesquieu).

And finally, not all of the passages "quoted" in Littré are in fact passages from Montesquieu's *De l'esprit des lois*. The passage "Les agnats et les cognats" (cited in Littré) is not a sequence found in Montesquieu directly, but is more a reference to a longer passage in which these words occur:

La loi de la division des terres demanda que les biens d'une famille ne passassent pas dans une autre: de là il suivit qu'il n'y eut que deux ordres d'héritiers établis par la loi; les enfans et tous les descendans qui vivaient sous la puissance du père, qu'on appella héritiers-siens; et, à leur défaut, les plus proches parens par mâles, qu'on appella *agnats*. Il suivit encore que les parens par femmes, qu'on appella *cognats*, ne devoient point succéder ... (9)

While, as this experiment suggests, our approach is generally quite productive, success in sequence alignment is dependent upon using appropriate settings for the size of the passages to be aligned. One may set parameters to permit matches on small passages, but this increases the likelihood of aligning unrelated passages or generating too many trivial alignments. Should one wish to build such alignments, it would be better to set very loose matching parameters and apply these to small subsets of larger collections.

The generation of too many alignments is, somewhat counter-intuitively, a significant problem in some domains and collections where one encounters a heavily referenced "common text." In our examination of a 1,900-work collection of Reformation and Counter-Reformation documents,[2] we have found extensive citation of the Latin Vulgate in theological texts and works of Biblical exegesis. PhiloLine detected some 54,607 common passages alone in a text by Petrus Canisius, *Authoritatum sacrae scripturae, et sanctorum patrum...* (1569), which were all mainly the same Vulgate citations present in other texts. This appears to be related to both genre and period, since we have not encountered similar problems in our more general collection of French works in the ARTFL database or other collections with which we have worked, such as EEBO-TCP. Our proposed solution, but one that we have not implemented, is simply to perform the alignments in two phases. The first pass would align all of the documents in a collection to the "common texts" and eliminate the matching passages; the second pass would align all of the remaining texts to each other.

As noted, PAIR and PhiloLine function in rather different ways and have significantly different design and implementation implications. PAIR is a system which builds a database of shingles to be aligned against an unknown document, submitted by users; this is a one-to-many alignment. PhiloLine is a many-to-many alignment system, which performs cross-alignments of entire collections or across collections. The key engineering distinction between the two implementations is the problem of "hapax shingles." A hapax shingle is one that occurs exactly once in a dataset or collection. Of course, hapax shingles are by far the most numerous type of
shingle, regardless of corpus size. For a many-to-many alignment, hapax shingles can simply be discarded, as they clearly cannot be shared between any two documents. This trims the shingle database substantially, enough so that even for our larger collections, it can be loaded into memory (as a hash) and processing may be done largely in memory.

This happy heuristic is not available for PAIR’s one-to-many comparison; here, every shingle must be retained because it could occur in an unseen future query document. This unpruned index is large, too large to fit into memory in most cases, and as this is a real-time application, we need a fast way of determining whether a given shingle occurs in the index without searching on disk. For this purpose, we use a Bloom Filter (Ceglowski), which is a highly space- and time-efficient method of probabilistically checking a value to see if it exists in a lookup table. The Bloom filter will occasionally return a false positive (saying that a value exists in the index when it does not), but will never return a false negative. The shingles generated from the query document are compared to the data stored in the Bloom filter, and only those that match are searched for on disk.

Both the one-to-many and many-to-many approaches have been tested on 10,000 or more documents and should be scalable by at least an order of magnitude. It is also clear that the many-to-many alignments can be parallelized across many processors since they are, in the final analysis, the accumulation of many document-to-document alignments.

Sourcing the Encyclopédie

We hope it is now evident that there are many use cases for sequence alignment in humanities applications, and we now turn our attention to a more extensive discussion of one particular use case: our ongoing work to identify the sources of Diderot and d’Alembert’s Encyclopédie. The Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers, edited by Denis Diderot and Jean le Rond d’Alembert, was one of the crowning achievements of the French Enlightenment. Published in Paris between 1751 and 1772 in 17 volumes of text and 11 volumes of plates, this monumental work contains some 77,000 articles written by more than 130 contributors. As with all reference works, the authors and editors of the Encyclopédie made extensive use of a vast array of contemporary reference works and scholarship to complete their massive compendium of enlightened knowledge. Identification of the sources used by the philosophes is a massive undertaking in itself, as the authors rarely acknowledged the works upon which they relied in writing their contributions. It is our expectation that systematic identification of the sources of the Encyclopédie will shed considerable light on the relationship of Enlightenment thought to French intellectual traditions, and also to currents of thought from classical antiquity to contemporary Western thinking.

In a previous work (Allen, et al.), we examined the relationship of the Encyclopédie to a single contemporary reference work, the Jesuit Dictionnaire universel français et latin (colloquially known as the Dictionnaire de Trévoux). It was widely assumed during the 18th century that the philosophes made extensive use of the Dictionnaire de Trévoux in the compilation of their work. Indeed, Jesuit critics of the Encyclopédie complained loudly of the extent to which entries were copied from earlier works, although among the possible sources of plagiarism the Trévoux dictionary was never explicitly mentioned. In order to attempt to detect possible borrowings, we used a general document similarity measure—the Vector Space Model (VSM)—to identify articles in the Encyclopédie which may have been borrowed, in whole or in part, from the Trévoux. The work used the text mining and machine learning extensions to PhiloLogic called PhiloMine (For documentation, examples,
The VSM uses a "bag-of-words" approach to calculate the similarity between two documents—documents being in this case articles in both sources. This simple technique is surprisingly effective. Our procedure involved enlisting several researchers familiar with the period and materials to evaluate the possible borrowed passages and to select those objects that were, in their judgment, most probably borrowed. At the end of the process, we had determined that 5.32 percent of the examined articles in the Encyclopédie were either borrowings or shared passages from the Jesuit Dictionnaire de Trévoux. Additionally, we found that VSM works best when comparing already identified blocks of text (in this case articles), but was less effective when borrowings occurred in smaller, unidentified passages, such as sentence groups or selected paragraphs. This conclusion led us to experiment further with an early implementation of sequence alignment as an alternative technique to identify not only similar "documents," but also borrowings and shared passages at finer levels of granularity.

Our initial success with this preliminary experiment resulted in the current implementations of both PhiloLine and PAIR. One of the main advantages to this sort of approach is that, rather than being limited to comparing pre-identified blocks of text such as articles or documents, sequence alignment can identify arbitrary sequences, which may or may not be recognizable blocks of text. This is particularly relevant for a work of such textual complexity as the Encyclopédie, which sought to compile and then rationally organize the inherited summa of human knowledge. Contributors such as the Chevalier de Jaucourt—author of some 17,000 Encyclopédie articles and a known text recycler—were particularly adept in knitting together passages from various sources (such as Montesquieu), often changing the order of borrowed passages even when taken from the same work. Jaucourt's article "République," for example, is composed of no less than 58 separate passages from Montesquieu's Considérations sur les causes de la grandeur des Romains et de leur décadence and De l'esprit des lois, many of which are significant extracts of the works. Thirty-nine of the aligned passages are drawn from the Considérations, and found largely in the subarticle "République romaine," which begins by acknowledging its debt to Montesquieu: "portons nos regards avec M. de Montesquieu sur les causes de sa grandeur & de sa décadence, & traçons ici le précis de ses admirables réflexions sur un si beau sujet" (V14, p. 154).

While Montesquieu is indeed mentioned as a primary source for the article, the extent to which Jaucourt is borrowing and reworking the author's work to fit his text, and the exact location of each borrowing, would be impossible to surmise without PhiloLine's sequence alignment.

As might be expected, the results of our various alignment experiments suggest that the philosophes made extensive use of previous reference works. Indeed, as Diderot himself admitted in the foreword of the third volume of the Encyclopédie (1753), textual borrowings often were, among other things, not actual "borrowings" in the modern sense of the word, but simply represented the fact that dictionaries cannot write very differently about certain topics: "[I]ls ne sauroient faire autrement" 'Indeed they cannot do otherwise,' thus guaranteeing a certain level of similarity between any two reference works that treat the same topic. At the top of the list of possible encyclopedic sources is the previously mentioned Dictionnaire universel de Trévoux, with some 11,430 total aligned passages, 6,621 of which have 20 or more words. The authors who borrowed the most from the Trévoux were anonymous authors (2,395 passages), Jaucourt (1,909), and the Abbé Mallet (544). Other reference works frequently used in the Encyclopédie include Le Grand dictionnaire historique de Moréri (1759) with 2,606 total aligned passages, 1,407 with 20 or more words, and Savary des Brûlons's Dictionnaire universel de commerce.
We have similarly identified aligned passages in the over 900 texts that predate the publication of the Encyclopédie in the ARTFL-Frantext collection. As shown in the list below, which displays the most frequently borrowed-from authors and works, there are some expected sources (Montesquieu and Voltaire) and some that are more surprising, such as Arnauld, Rollin, and perhaps even Condillac:

- 3,610 Shared Passages:
  - Montesquieu - 681 passages
    - De l'esprit des lois (1746) - 477 passages
    - Considérations sur les Romains... (1734) - 173 passages
  - Voltaire - 528 passages
    - Essai sur l'histoire générale... (1756) - 415 passages
  - Jean-Baptiste Dubos - 229 passages
    - Réflexions critiques sur la poésie et sur la peinture (1719) - 227 passages
  - René Aubert de Vertot - 122 passages
    - Histoire des révolutions arrivées dans le gouvernement romain (1727) - 122 passages
  - Antoine Arnauld & Pierre Nicole - 107 passages
    - La logique, or l'art de penser (1662) - 107 passages
  - Charles Rollin - 100 passages
    - Histoire ancienne des Égyptiens (1738) - 94 passages
  - Montaigne - 91 passages
    - Les Essais (1595) - 91 passages
  - Condillac - 91 passages
    - Essai sur l'origine des connaissances humaines (1746) - 91 passages

Systematic identification of the most frequently borrowed sources in the French collection at ARTFL is of great value. It is, however, limited by the relatively restricted number of available texts and the fact that the collection does not reflect the well-known influences of other intellectual traditions, most notably that of England, in French Enlightenment thought.

Before addressing the issue of works in translation, however, we wanted to verify whether PhiloLine could also function adequately with uncorrected OCR output. The following example shows the alignment between Jaucourt's article "Monnoie" and the entry for "Denier" in Savary des Brûlons's Dictionnaire de commerce:


les frais de la fabrication, qu'on nomme brassage, y doivent être ajoutés. A l'égard des qualités moins essentielles, le volume de la monnoie n'est autre chose que la grandeur & l'épaisseur de chaque piece. La figure, c'est cette forme extérieure qu'elle a à la vue; ronde en France; irrégulière
& à plusieurs angles en Espagne; quarrée en quelques lieux des Indes; presque sphérique dans d'autres, ou de la forme d'une petite navette en plusieurs. Le nom lui vient, tantôt de ce que représente l'empreinte, comme les moutons & les angelots; tantôt du nom du prince, comme les Louis, les Philippes, les Henris; quelquefois de leur valeur, comme les quarts d'écus & les pieces de douze sous; & d'autres fois du lieu où les especes sont frappées, comme autrefois les parisis & les tournois. Legrenetis est un petit cordon fait en forme de grain, qui regne tout - au - tour de la piece, & qui enferme les légendes des deux

Target: Dictionnaire de commerce, "DENIER:"

les frais de la fa- brication qu'on nomme Braffage, y doivent atre a- jOtitéS. A l'égard des qualités moins effentielles, le vo- lume de la ilonniiioie n'el autre chofe que la gran- deur & l'épailleur de chaque piece. La fignrp cer cette forme extérieure qu'elle a à la vue, ronde en France, irrégulière & à plusieurs angles en Efpa- gnie , quarrée en quelques lieux des Indes , préfque sphérique dans d'autres, ou de la forme d'une petit- te navette en plufieurs. Le nom lui vient, tantôt de ce que représente l'eminreinte , comme les Moutons & les Angelots; tantôt du nom du Prince, comme les Louis, les Philippes, les Henris ; quelquefois de leur valeur, comme les quarts d'écus & les pièces de quatre fous; & d'autres fois du lieu où les efpéces font frappées, comme anciennement les Parifis & les Tournois. Le grenetis efl un petit cordon fait en forme de grain , qui régne tout autour de la pièce, & qui en- ferme les llUgendes des deux côtes.

Examination of the differences in these two passages suggests that, given the parameters used in the alignment, PhiloLine can handle so-called "dirty OCR" relatively well. It is important to note, however, that such alignments fail when either the errors are too dense or there are other changes in word order and word segmentation. Again, adjusting parameters can increase the detection, but may also increase the number of unrelated alignments. It is difficult to design an experiment that would determine the overall effectiveness of our alignment approach, as source data of uncorrected OCR can vary significantly from page to page and even across a single page. These reservations are further compounded given the well-known limitations of current OCR systems and the variations in print, paper, and reproduction (usually images from microfilm) quality.

Building upon our experiences with uncorrected OCR, we next sought to examine the possibilities of identifying passages from English works in translation as potential Encyclopédie sources. To this end, we assembled a small database of roughly contemporary French translations drawn from the Gale Eighteenth Century Collections Online (Gale-ECCO) database. French titles, including English works in translation, are taken from the uncorrected OCR from Gale-ECCO: http://www.gale.cengage.com/DigitalCollections/products/ecco/. These 90 texts are again examples of uncorrected OCR and have the added complication of bearing, as many eighteenth-century translations did, a less than faithful relationship to the original texts; a situation that we hope is not the case with modern translations. It is also at least possible, and in some instances quite likely, that the Encyclopédie contributors translated passages themselves where necessary, recalling that the original conception of Diderot's work was as a translation of Ephraim Chambers's Cyclopedia. As with the uncorrected OCR of the French dictionaries mentioned above, the system here performs alignments where the source data is not too unrecognizable and the wording is at least somewhat similar; for example, Jaucourt appears to have borrowed the following passage from a translation of David Hume's Political Discourses:

Quoique le Christianisme dans sa pureté primitive ne soit pas défavorable à la société, on abuse des meilleures institutions; & il ne seroit peut-être pas aisé de justifier tous les édits des empeures chrétiens à ce sujet. Ce qu'il y a de sûr, c'est qu'on regarde la quantité de moines, & celle des personnes du sexe qui dans les couvens font voeu de virginité, comme une des principales causes de la disette de peuple dans tous les lieux soumis à la domination du souverain pontife. On ne doit pas être surpris que des auteurs protestans tiennent ce langage, lorsque les écrivains catholiques les plus judicieux & les plus attachés à la religion, ne peuvent s'empêcher de former les mêmes plaintes.

Si l'Espagne, autrefois si peuplée, est aujourd'hui deserte, c'est sur-tout à la quantité de monasteres qu'il faut s'en prendre, selon les auteurs espagnols: « Je laisse, dit le célèbre dom Diego de Saavedra dans un de ses emblèmes, à ceux dont le devoir est d'examiner si le nombre excessif des ecclésiastiques & des monasteres est proportionné aux facultés de la société des laïques qui doit les entretenir, & s'il n'est pas contraire aux vûes mêmes de l'Eglise. Le conseil de Castille, dans le projet de réforme qui fut présenté à Philippe III

Bolingbroke, *Discours politiques, par M. Hume, traduit de l'anglois* (p. 404)

vrai-femblable qu'anciennement chaque grai ,,digieux de Prêtres non mariés dans tous les Pa) 
,,Catholiques, qui font une fi grande parriec ,,l'Europe, & celui dcs personnes du Sexe, qi ,, dans les Couvens font Vu de virginité, comm ,,une des principales caufes de la dilette dePeuP ,,dans les Pays qui font fous la domination d ,, Souverain Pontife." On ne doit pas être surpri que des Auteurs Protestans tiennent ce langag, lorsque lesEcrivains Catholiques les plus judicieu & les plus attachés à la Religion ne peuvent s'en pecher de former les mêmes plaintes.

Si l'Espagne, autrefois fi peuplée, est aujourd'hui deferte, c'et fur-tout au trop grand nombre d Couvens, qu'il faut s'en prendre felon les Auteur Espagnols. ,,Jelaisse, dit le celebre DON DIE ,,GO DE SAAVEDRA, dans fou EmbléimfeLXII.I , ceux dont c'efl le devoir a examiner fi le nombre ,,excelF;f des Eccléfiastiques & des Couvens, l ,, proportionné aux Facultés de la Société des Lai ,,ques qui doit les entretenir, & s'il n'est pas con ,,traire aux vues mêmes de l'Eglife. Le Confil ,, de Castille dans le projet deRéforme qui fut pr fenté à Philippe III.

In all, we identified approximately 165 aligned passages in the *Encyclopédie* that appear in the subset of English works in translation. As a caveat, however, we note that multiple passages may only represent a single borrowing due to errors in the OCR, and that, in general, quantitative and qualitative results from uncorrected OCR output should be further verified by hand. With this in mind, the most frequently sourced English works from our limited collection include:

- Locke, *Du gouvernement civil* - 33 passages
- Hume, *Discours politiques* - 30 passages
- Bolingbroke, *Lettres sur l'esprit du patrioctisme* - 16 passages
- Locke, *Abregé de l'essay sur l'entendement humain* - 10 passages
- Pope, *Essai sur l'homme* - 6 passages
While the *philosophes* would periodically mention the original authors from whom they were borrowing, often to call on the authority of a Montesquieu or Voltaire, they appear to have systematically avoided mentioning any English author as a source. The following table shows each *Encyclopédie* article, its author, the number of borrowed passages that appear in the article, and whether any of these have acknowledged their English forbears:

<table>
<thead>
<tr>
<th>Article</th>
<th>Author</th>
<th>Borrowed Passages</th>
<th>Source(s)</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gouvernement</td>
<td>Jaucourt</td>
<td>10</td>
<td>6 from Locke's <em>Du gouvernement civil</em>; 4 from Hume's <em>Discours</em></td>
<td>no citation</td>
</tr>
<tr>
<td>État de nature</td>
<td>Jaucourt</td>
<td>10</td>
<td>from Locke's <em>Du gouvernement civil</em></td>
<td>no citation</td>
</tr>
<tr>
<td>Patriote</td>
<td>Jaucourt</td>
<td>7</td>
<td>from Bolingbroke's <em>L'esprit du patriotisme</em></td>
<td>no citation</td>
</tr>
<tr>
<td>Monastere</td>
<td>Jaucourt</td>
<td>6</td>
<td>from Hume's <em>Discours politiques</em></td>
<td>no citation</td>
</tr>
<tr>
<td>Prérérogative royale</td>
<td>Jaucourt</td>
<td>5</td>
<td>from Locke's <em>Du gouvernement civil</em></td>
<td>no citation</td>
</tr>
</tbody>
</table>

These preliminary—and admittedly rudimentary—results nevertheless suggest the potential utility of including contemporary translations of important non-French works in our alignment collection. We are thus currently creating a larger collection for alignment purposes, including German and Italian works in translation as well as Latin and Greek classics.

While our primary objective in this use case was the identification of sources used to construct the *Encyclopédie*, it became immediately evident that sequence alignment could also be used to assess, in a systematic manner, the reception of the French Enlightenment's philosophical *machine de guerre*. Aligning the *Encyclopédie* against the almost 1,800 documents in the ARTFL-Frantext database that post-date the appearance of the final volumes in 1772, we found some 430 shared passages. It is perhaps not surprising, in a collection of canonical literary works such as Frantext, that there would be relatively few borrowings from a general reference work that was rapidly outdated. In the following table, quantifying borrowed passages broken down by authors and titles, we see an interesting split between borrowings by conservative critics of the *philosophes* (de Maistre and Barruel) and later intellectual historians of eighteenth-century France (Taine, Hazard, and Mornet):

<table>
<thead>
<tr>
<th>430 Shared Passages:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paul Hazard</td>
</tr>
<tr>
<td>La pensée européenne au XVIIIe siècle (1946)</td>
</tr>
<tr>
<td>La crise de la conscience européenne (1935)</td>
</tr>
<tr>
<td>Chateaubriand</td>
</tr>
<tr>
<td>Génie du christianisme (1802)</td>
</tr>
<tr>
<td>Jean-Paul Marat</td>
</tr>
<tr>
<td>Éloge de Montesquieu (1785)</td>
</tr>
<tr>
<td>Joseph de Maistre</td>
</tr>
<tr>
<td>Des constitutions politiques (1810)</td>
</tr>
<tr>
<td>L'abbé Barruel</td>
</tr>
<tr>
<td>Les helviennes (1781)</td>
</tr>
</tbody>
</table>
o Sainte-Beuve - 17 passages
  - *Port-Royal* (1859) - 15 passages

o Daniel Mornet - 14 passages
  - *Les origines intellectuelles de la révolution française* (1933) - 11 passages

o Hippolyte Taine - 12 passages
  - *Les origines de la France contemporaine. L'ancien régime* (1875) - 10 passages

This point is reflected nicely in the Figure 1 below, which suggests that borrowings from the *Encyclopédie* were relatively infrequent during the nineteenth century, when it was no longer at the center of intellectual debate or a reflection of contemporary knowledge, rising again in importance as it became an object of study itself.

*Figure 1: Borrowings from the Encyclopédie by year*

![Graph showing borrowings from the Encyclopédie by year]

Thus, by way of the identification of similar passages—whether attributed citations, borrowings, or shared passages—one can begin to conceive of a method that would systematically trace the "influence" or "reception" of a particular work or author across thousands of documents and hundreds of years.

The almost intractable problems of identifying the origins of the *Encyclopédie*, and of tracing its later use and influence, is therefore greatly aided by our simple sequence alignment scheme. Of course, more detailed examination of the passages cited or borrowed is required to establish stronger conclusions and to give a more nuanced sense of the importance of certain borrowings compared to others. Reading is therefore not replaced by the machine, but algorithms like sequence alignment may help to direct our reading, situating it systematically within larger contexts, and thus leading to more rigorous study than could be achieved through reliance on the limited capacity of the human mind to read and recollect.

**Future Work**

We believe that there are a number of extensions to our simple alignment approach that may enhance the system: a standard distance measure for similar n-gram matching that would significantly alter shingle generation and matching algorithms; sliding gap measures that would improve the identification of longer passages that are often broken by insertions or OCR errors; match scoring by inverse shingle frequency in which matching regions
would be identified by a score rather than a binary yes/no; secondary string matching that would include regions of text around the generated n-gram matches; collaborative tagging of named entities for exclusion/inclusion; and the improved identification of "banal" matches for the system to ignore.

We will also work to improve the handling of results and reporting of the system. First, we will build an encoding mechanism that will flag passages related to other texts, possibly through the use of embedded XML tagging, which could then be used to generate links between documents. Secondly, we will offer a global perspective of the alignments through the further enhancement of quantitative reporting and the inclusion of clustering and/or graphing approaches.

We have adopted a very simple approach to sequence alignment and have made several modifications that correspond to the typical use cases found in the digital humanities. We are aware, however, that there are many more complex algorithms in use—in bioinformatics, for example—that we intend to explore further with the promise of significant improvement in performance. In his thoughtful essay on text ownership, Yorick Wilks concludes, "[i]t may be that ... ownership of text is only a transitory thing at best, more like the lifetime ownership of a piece of genetic code than we might want to think" (127). Perhaps it is unsurprising, then, that the same tools that are used to identify shared sequences of DNA can be pressed into service to track shared text.

Conclusion

Similar passage identification is a general approach that can be tailored to enhance a broad variety of digital humanities applications. Above, we have shown that it is an effective means in identifying borrowings across large collections of text data, independent of language and encoding scheme. Using this approach, we believe that scholars will be able to discover shared passages, borrowings, plagiarisms, and other forms of text recycling, whether alluded to in the source data or not. Indeed, even in cases in which source documents have clear indicators of a borrowing or a direct citation, this approach can significantly improve the manner in which these relationships are linked from text to text. Rather than parsing a reference and link using citation data or outside referencing schemes—which can be highly variable, inconsistent, and typically keyed on page numbers or other rather arbitrary attributes—links could be identified and contextualized using the alignment techniques outlined above.

PhiloLine and PAIR, and other sequence alignment approaches, may also introduce a higher level of systematicity to the analysis of the complex and multifaceted problem of "intertextuality." By this we mean that scholars can assume a level of certainty about connections drawn between documents and develop arguments based on quantitative measures, rather than more general impressions about the importance of a particular work over time or where a particular author drew inspiration. Indeed, these relationships can be shown quite reasonably as a directed graph and may be useful in visualizing dependencies that refer to particular intellectual schools or traditions. It is equally clear, however, that our implementations form only a rudimentary first step in the exploration of the nebulous connections of intertextual relations inherent to any exercise of textual analysis. This approach is still based on simple and evident forms of lexical matching, even if we have succeeded in introducing a certain degree of flexibility in various areas. One might imagine extensions of this simple model that would allow for the identification of other types of similarities based on more abstract representations. Rather than perform alignments on n-grams of specific words, for example, it may be possible to map specific lexical constructions to
semantic fields using the facilities provided by WordNet. In this case, all references to felines—such as cat, lion, tiger, or puma—would be mapped accordingly, making n-grams far more powerful indicators of semantic diversity. We believe that experimental work on an "allusion" detector could be based not only on the specific algorithmic improvements suggested above, but also on an effort to generalize the problem to a greater level of semantic awareness. Indeed, while the identification of the "hidden roads" running between texts will remain an important challenge for the digital humanities, constrained perhaps only by algorithmic limitations, there are also promising approaches to be found in other disciplines, providing further avenues for significant progress in the future.

Notes

[1] PAIR/PhiloLine is our implementation of sequence alignment under discussion in this article. Please consult http://code.google.com/p/philomine/ for documentation, demonstrations and source code. PhiloLogic is the main text analysis package developed by the ARTFL Project and is available at http://philologic.uchicago.edu/. The ARTFL-FRANTEXT database (access restricted to ARTFL subscribers) is available at: http://artfl-project.uchicago.edu/node/23. Unless otherwise stated, all primary sources referred to in the text of this article are from the ARTFL collections.


[3] Only selected individual works are listed for each author. For this reason there can be a discrepancy between the total number of passages listed beside the author name and the sum of the number of passages from the author's listed works.

Works Cited


