Digital Approaches to Text Reuse in the Early Chinese Corpus

DONALD STURGEON

Abstract Observed textual similarities between different pieces of writing are frequently cited by textual scholars as grounds for interpretative stances about the meaning of a passage and its authorship, authenticity, and accuracy. Historically, identifying occurrences of such similarities has been a matter of extensive knowledge and recall of the content and locations of passages contained within certain texts, together with painstaking manual comparison by examining printed copies, use of concordances, or more recently, appropriate use of full-text searchable database systems. The development of increasingly comprehensive and accurate digital corpora of early Chinese transmitted writing raises many opportunities to study these phenomena using more systematic digital techniques. These offer the promise of not only vast savings in time and labor but also new insights made possible only through exhaustive comparisons of types that would be entirely impractical without the use of computational methods.

This article investigates and contrasts unsupervised techniques for the identification of textual similarities in premodern Chinese works in general, and the classical corpus in particular, taking the text of the *Mozi* 墨子 as a concrete example. While specific examples are presented in detail to concretely demonstrate the utility and potential of the techniques discussed, all of the methods described are generally applicable to a wide range of materials. With this in mind, this article also introduces an open-access platform designed to help researchers quickly and easily explore these phenomena within those materials most relevant to their own work.

Keywords text reuse, citation, quotation, similarity, classical Chinese

Many of the works comprising the transmitted corpus of classical Chinese—particularly those believed to in part predate the Han dynasty—are known to have endured long and complex histories of transmission. Some of these texts are clearly not the work of a single author but, rather, products of a protracted

The Journal of Chinese Literature and Culture • 5:2 • November 2018
DOI 10.1215/23290048-7256963 • © 2019 by Duke University Press

process of textual production and editing, in some cases likely spanning many centuries. This frequently results in complex relationships of textual production existing within and between early texts in their transmitted forms, many of which remain largely obscure. While in certain cases such relationships can be determined with high confidence—for example, if a work references a historical individual known to have been active in a particular time period, this excludes the possibility that this piece of writing was entirely composed prior to that date—in general unraveling different strands of authorship and textual history is an extremely complex task frequently hampered by a lack of unambiguous evidence. With the partial exception of excavated manuscripts for which precise dates can be determined, often little external evidence is available to support or refute hypotheses about textual histories. The majority of evidence ultimately derives from insights that can be gleaned from within the classical corpus itself, either from within the same text or from material recorded elsewhere in the corpus.

Scholars of these texts have long used formal similarities between passages in constructing specific arguments about textual histories. This includes both characteristically similar word usage across texts,[1] as well as more extended similarities in which two pieces of writing contain what are essentially the same sentence, passage, or significant fragment of text. Often, especially in early works, this latter type of similarity is not acknowledged or highlighted in any way in either location and so, without the aid of critical apparatus, comprehensive reference works, or digital tools, in many cases would be noticed only by a careful and attentive reader familiar with both of the connected passages.

This type of similarity is generally referred to as *text reuse*. This term, though standard terminology for describing the observed phenomenon of these types of similarity within such materials, is quite distinct from direct or indirect quotation; in this article, it refers to the observed phenomenon of extended textual similarity, regardless of the explanatory factors lying behind its occurrence. This type of similarity has long been referenced by critical commentators and annotators of historical Chinese works. In early Chinese works, such similarities are very often of such a strong degree that, having been identified, there remains little or no doubt that two pieces of writing exhibiting this property have some historical relationship to each other; the difficulties lie, first, in identifying the existence of such parallels and, second, in interpreting the causal factors that explain their existence.

Concrete examples of this application of text reuse can be found easily by examining critical editions and commentaries of early texts. The influential Qing dynasty commentary on the text of the *Mozi* 墨子 by Sun Yirang 孫詒讓, titled *Mozi jiangu* 墨子閒詁 (Exposing and Correcting the *Mozi*), along with work by

other Qing dynasty scholars, is credited with transforming the frequently corrupt received text of the *Mozi* and rendering it readable through textual emendations based largely on careful analysis of the text and extensive comparison of its contents both with other parts of the *Mozi* and with other pre-Han works. An example of the type of similarity frequently cited in this and similar commentaries is the following:

> 愛盜非愛人也，不愛盜非不愛人也，殺盜人非殺人也，
> （註：）「盜」下「人」字衍。《荀子·正名》篇云「『殺盜非殺人也』，此惑於用名以亂名者也」。
> Caring for thieves is not caring for people; not caring for thieves is not not-caring for people; killing thieving people is not killing people.
> [Commentary:] The word "people" below "thieves" is incorrectly repeated. The *Zhengming* chapter of the *Xunzi* states: "'Killing thieves is not killing people'—this is the confusion of using names to disorder names."

> 無難盜無難矣。
> （註：）據下文，疑衍「盜無難」三字。[2]
> There is no difficulty; thieves; no difficulty.
> [Commentary:] According with the text below, it is suspected that the three characters "thieves; no difficulty" are incorrectly repeated.

In this fragment of his commentary, Sun draws attention to two textual parallels, one between the *Mozi* and the *Xunzi*, and the other between the *Mozi* and another part of the same text. The first of these is easily seen from the passage cited above: the received text of the *Mozi* contains the line "sha daoren fei sha ren ye" 殺盜人非殺人也 (killing thieving people is not killing people), while the *Xunzi* contains the very similar line "sha dao fei sha ren ye" 殺盜非殺人也 (killing the thieving is not killing people), differing only in its omission of the first *ren* 人 (people). Given the unusualness of this line, a critical reader of the *Mozi* will benefit from being made aware of its existence—neither received text explicitly cites the other here, and in fact the *Xunzi* does not directly mention the Mohists at all in the chapter containing that line. It also provides some evidence for Sun's claim that the text of the *Mozi* should be emended to remove the first *ren*.

The second parallel alluded to in the passage above can be understood only by looking, as the comment suggests, at what occurs later in the same chapter of the *Mozi*. Later on in the text—though not directly below—another passage contains wording that, once located, is unmistakably parallel to the occurrence that Sun emends:

> 殺盜人，非殺人也，無難盜無難矣。此與彼同類，世有彼而不自非也，墨者有此而非之，無故也焉，所謂內膠外閉與心毋空乎？[3]
>
> Killing thieving people is not killing people—there is no (difficulty thieves no) difficulty in it! This and those are of the same kind; the world have those and don't oppose themselves, [yet] we Mohists have this and they oppose us. There can be no reason: they are stopped up within and closed without; their heart-minds are all filled in!
>
> . . .
>
> 非執有命，非命也，無難矣。此與彼同，世有彼而不自非也，墨者有此而罪非之，無也故焉，所謂內膠外閉與心毋空乎？[4]
>
> Opposing those who think there is fate is opposing fate—there is no difficulty in it! This and those are the same; the world have those and don't oppose themselves, [yet] we Mohists have this and they oppose us. There can be no other reason than that they are stopped up within and closed without; their heart-minds are all filled in!

The underlined portions of the Chinese text above correspond to the differences between the two passages; once placed side by side in this manner, it is clear even at a glance that the passages are substantially the same, though they do contain certain important differences. In fact, both of the other minor differences within the parallel—the omission of *lei* 類 (kind) in the second passage, and the swapped order of *ye* 也 (depending on the ordering, read either as a grammatical particle, or as equivalent to *ta* 他 meaning *other*) and *gu* 故 (reason)—are also cited by Sun's commentary as evidence for textual emendations of the second and first passages, respectively.

In both of these cases, the prerequisite to making the emendation in the first instance was being aware of the existence of such a textual parallel. This prompts the questions considered in the remainder of this article: how can we more efficiently identify such important similarities within and between texts without relying on either encyclopedic textual knowledge of the reader or comprehensive commentaries, and how can we start to make sense of this information on a scale greater than might be feasible to consider without the benefits of digital techniques?

Comparison of Text Reuse Metrics

A variety of digital approaches to identify similarities between textual items have been created, with several distinct but related objectives in mind. These include the development of document-focused systems and search engines—systems intended to identify, given any new document, which of a large number of preexisting documents the new document most closely resembles or relates to on the basis of its contents. In this approach, textual data are first segmented into individual units or documents (in this context, the term *document* can refer

to any predefined textual unit, such as a sentence, passage, chapter, or entire work), and the goal is to make comparisons at the level of entire documents, by answering the following question: given a new (unseen) document, which of the existing (seen) documents is most similar to it in some relevant sense?

A point of particular relevance when considering document similarity of classical Chinese materials is that mainstream techniques for document similarity comparison and text reuse identification have typically been developed to operate on documents that are (or can easily be) divided into sequences of distinct *tokens*—linguistic objects that will be modeled as indivisible semantic units, typically words. This allows equivalence between tokens to be determined by a straightforward identity comparison: *cat* and *catch* are two distinct and unrelated tokens, despite the latter containing the former, because they are not identical sequences of symbols—no further analysis of the structure of the tokens themselves is required to determine this. For alphabetic languages such as English, in which words are explicitly delimited by punctuation (spaces, periods, commas, etc.) and, additionally, individual characters in isolation typically have no semantic content, using words as these indivisible tokens is both a natural approach and one that presents few technical challenges—it is a relatively straightforward task to take a sequence of characters composing an English text and transform this into a sequence of words. For Chinese, this task in general is nontrivial due to the lack of any explicit delimiters between words. While there are easy cases, there are also many hard cases in which segmentation depends on the meaning of a sentence in ways that are extremely hard to model computationally. While much progress has been made toward solving this task computationally for modern Chinese writing, there are currently no adequate mechanisms for performing this task on classical Chinese materials. At the same time, while certainly not all words in classical Chinese are composed of a single character (proper names being an obvious example), characters in classical Chinese in general do have semantic content, unlike characters of an alphabetic language. The combination of these two factors frequently makes processing classical Chinese text by taking characters rather than words as the indivisible tokens an attractive option, and this approach is generally assumed for the rest of this article—that is, tokens will be assumed to be Chinese characters.[5]

The remainder of this article investigates three distinct digital approaches to measuring textual similarity and explores to what extent these can help identify meaningful trends and relationships within classical Chinese materials.

TF-IDF and Cosine Similarity

An established baseline method in comparing similarity of documents is cosine similarity.[6] In this approach, each document to be compared is represented by a

vector in n-dimensional space, where *n* is the total number of distinct token types occurring anywhere across all documents.[7] Each component of each vector is a value representing the frequency with which a particular term occurs in that document (referred to as term frequency, TF). Typically, these values are also weighted by multiplying them by the inverse document frequency (IDF; defined below) of the corresponding term, which provides a means for compensating for the intuition that globally infrequent terms shared by two documents are a stronger indicator of similarity than globally frequent terms. For example, in comparing two short documents containing only the sentence "zao fu zhe, tianxia zhi shan yu zhe yi" 造父者，天下之善御者矣 (Zaofu was, indeed, the best charioteer in the world) in the first and "zao fu shan yu" 造父善御 (Zaofu was the best charioteer) in the second purely by examining the terms they contain (while ignoring the order in which the terms occur), we might say that one formal property explaining their similarity is their shared use of the four terms *zao* 造, *fu* 父, *shan* 善 (best), and *yu* 御 (charioteer). Another document, such as "fan zhe yi you tianxia zhi ban yi" 反者已有天下之半矣 (indeed, the rebels already possess half the world), might easily contain an even higher number of shared terms with the first document—in this case, *zhe* 者 (a particle that appears twice in the first document), *tian* 天 (heaven), *xia* 下 (under), *zhi* 之 (a particle), and *yi* 矣 (a particle)—without being similar in any particularly significant way. IDF attempts to account for this by weighting the contribution of shared terms according to their global infrequency in the corpus being compared. When comparing a large number of documents, common terms such as *tian* 天, *xia* 下, *zhi* 之, and *yi* 矣 will normally be globally frequent within the corpus as a whole, so will be assigned a much lower IDF, and thus will contribute a lesser amount to the overall similarity score.[8]

IDF for a term *t* is computed by dividing the total number of documents to be compared, *N*, by the total number of documents in which that term occurs, and taking the logarithm of the result:

$$idf_t = \log \frac{N}{|\{d \in D \mid t \in \mathrm{d}\}|}$$

An important consequence of this is that any term that occurs in every document within a corpus will always have an IDF of o; thus, when using IDF weighting, any term that occurs in every document under consideration will always be assigned a weight of zero and so be excluded from comparison.

Cosine similarity produces a similarity score calculated between two vectors, each of which represents a single document. Element *i* in each vector—A_i and B_i, respectively—corresponds to the weight assigned to type *i* of all *n* token types considered in the comparison (i.e., of all tokens appearing anywhere in any of the set of all documents being compared). Cosine similarity is easily

computed from vectors A and B by summing the products of each pair of elements A_i and B_i and then dividing the total sum by the Euclidean norms of both vectors (||A|| and ||B||):

$$\text{similarity}_{\text{cosine}} = \cos\theta = \frac{\sum_{i=1}^{n} A_i B_i}{||A||\ ||B||}$$

Mathematically, the cosine similarity is the cosine of the angle in Euclidean space between the two document vectors; hence, it has a value ranging from 0 (corresponding to orthogonal or perpendicular vectors) to 1 (vectors pointing in an identical direction). A cosine similarity score of 0 implies that two vectors have no components in common (i.e., no shared vocabulary at all, excluding terms with an IDF of 0); a score of 1 implies that all components appear in both documents in identical ratios (again excluding any terms that have an IDF of 0).

Measuring cosine similarity between TF-IDF vectors is an example of a "bag of words" method: it completely ignores word order and considers only TFs within each document; thus, for example, "bu yan zhi jiao" 不言之教 (teaching without speaking) and "bu jiao zhi yan" 不教之言 (speaking without teaching) are treated as identical according to this model. Nevertheless, this metric of similarity has been practically successful in many domains and is sufficient to identify certain types of interpretatively meaningful similarity between textual objects in the classical Chinese case.

In what follows, I take the transmitted text of the *Mozi* 墨子 as a concrete example. While the received text has been transmitted in the guise of a single work and contains no explicit acknowledgment of multiple authorship or subdivisions into units of different origins (other than organization into chapters collected in fifteen *juan* 卷 [fascicles]), the text is known to have a complex history of transmission and believed to be composed of several contiguous sections with distinctive writing styles and concerns, likely of different authorship, and in some cases almost certainly created at different times.[9] Specifically, the text is thought to be divisible into main sections as shown in table 1.

The received text of the *Mozi* is notoriously corrupt, having suffered from many centuries of neglect by the tradition. In investigating textual similarities this article intentionally works directly with the uncorrected text of the *Mozi* as recorded in the *Zhengtong daozang* collection, which forms the basis for most modern editions. The reason for using this edition, rather than a modern corrected version of the text, is to avoid circularities that would otherwise occur when evaluating possible textual emendations motivated by closely related versions of the same piece of writing. Modern editions of the *Mozi* typically include large numbers of textual emendations, many of which have at some point been justified partly on the basis of precisely the same types of textual

Table 1. Sections of the transmitted text of the *Mozi*

Chapters	*Content*
1–7	Essays and dialogues.
8–37	Core chapters, composed of triads of texts with the same title followed by *shang* 上 (upper), *zhong* 中 (middle), *xia* 下 (lower). In some triads, only one or two chapters are extant.
38–39	"Fei Ru" 非儒 (Anti-Confucianism). Only chapter 39 is extant.
40–45	Dialectical chapters.
46–51	Dialogues between Mozi and others.
52–71	Military chapters.

parallels explored later in this article; thus, to properly evaluate the evidence for or against such emendations, it is necessary to begin first with the uncorrected text.

To investigate the utility of cosine similarity of TF-IDF vectors as a metric of text reuse, this metric was applied to every pair of chapters composing the received text of the *Mozi*. The results of this comparison are summarized in figure 1, in which each x,y-location represents the cosine similarity of the pair of chapters *x* and *y*, on a gradient from white (corresponding to 0 similarity) to red (1). The table is therefore diagonally symmetric; the solid red diagonal corresponds to the comparison of each chapter with itself, which by definition yields the value 1 on this metric. Thus, darker-shaded regions correspond to greater cosine similarity, suggesting greater commonality of distinctive vocabulary use between the pairs of chapters involved.

Many interesting observations can be made directly from this visual summary of the data. Perhaps the most striking feature is the clear distinction between the military chapters (labeled "8" on fig. 1) and the rest of the text, resulting in a shaded square region coinciding exactly with the military chapters, contrasting with much lighter bands to the left of and above this region. This result will not be surprising to those with prior knowledge of the text, given the radically different style and content of these chapters compared with the remainder of the work, but is a positive indication that even a simple metric of document similarity may be sufficient to provide useful insights about such texts. Looking more closely at the results, numerous other clusters of adjacent shaded regions are clearly visible along the diagonal (numbers below correspond to labels in fig. 1), indicating strong similarity within the following groups:

1. Within and between the first two triads: the three texts of the "Shang xian" 尚賢 (Exaltation of the Virtuous) triad, and the three texts of the "Shang tong" 尚同 (Identification with the Superior) triad. While each group is most closely related to the other members of its own group, all six have substantial similarities. Of

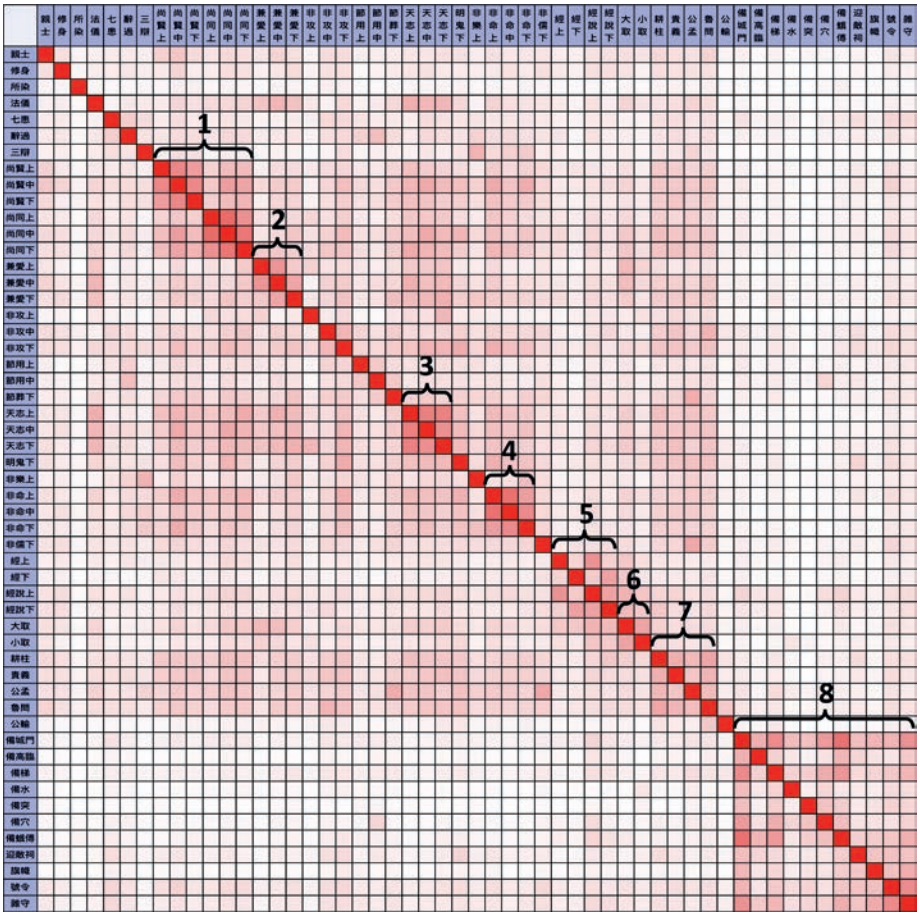

Figure 1. Cosine similarity between all pairs of chapters of the *Mozi*. For colors and notations, see text.

note, the same is *not* true of the "Fei gong" 非攻 (Condemnation of Offensive War) triad; while the second two parts of this group have significant mutual similarities, the first of the three texts has a low similarity score with both of the latter two.

2. The "Jian ai" 兼愛 (Universal Love) triad.
3. The "Tian zhi" 天志 (Will of Heaven) triad.
4. The "Fei ming" 非命 (Anti-Fatalism) triad.
5. Between the "Jing shang" 經上 (Canon I) and "Jing shuo shang" 經說上 (Exposition of the Canon I), between "Jing xia" 經下 (Canon II) and "Jing shuo xia" 經說下 (Exposition of the Canon II), and between "Jing shuo shang" and "Jing shua xia." (The "checkerboard" pattern here reflects the fact that each of the "Jing shuo" texts is a commentary on the corresponding "Jing" text.)

6. The Xiao qu 小取 (Minor Illustrations) and Da qu 大取 (Major Illustrations).
7. The first four of the five "dialogues."
8. The "military chapters."

Intriguingly, while the first seven chapters of the text do have substantial similarities with other chapters, they are unique among chapters of the *Mozi* in comprising a contiguous group within which there is very little similarity. While other individual chapters—notably "Fei gong shang" 非攻上 (Condemnation of Offensive War I) and "Fei yue shang" 非樂上 (Condemnation of Music I)—also exhibit this property, these are isolated cases (and, in the case of "Fei yue shang," this is partially explained by the absence of the two adjacent chapters whose existence is mentioned in the transmitted texts but have been lost to the tradition). Surprisingly, despite this, the first seven chapters do have some significant overlap with the core Mohist chapters—especially the "Fa yi" 法儀 (On the Necessity of Standards) chapter.

A second way to assess what is unusual within these data is to quantify to what extent individual chapters have word usage similar to any other part of the *Mozi* corpus, by computing the maximum cosine similarity score for a given chapter with any other chapter in the text (fig. 2). The most obvious outlier on this view is the third chapter, "Suo ran" 所染 (On Dyeing), which is exceptional in its striking lack of similarity with any other chapter of the text. Since this comparison has been made using TF-IDF, this implies that the term usage in this chapter is unusual in the context of the *Mozi*: its patterns of observed word frequency differ from those seen elsewhere in the corpus.

To investigate why this is the case, we can examine the contents of the vector calculated by TF-IDF on these data for the "Suo ran" chapter. Each vector contains one value for every term occurring anywhere in any part of the *Mozi*; its contents represent a weighted value of the relevance of this term to the particular chapter. If we sort the component values in the vector from highest to lowest, we can examine the terms that might be expected to contribute the most in similarity calculations between "Suo ran" and other chapters generally; the top ten terms are listed in table 2.

The highest value—unsurprisingly to those familiar with the text—is *ran* 染 (to dye), a term that is repeated rhetorically thirty-three times throughout this chapter. That this should be the most strongly weighted term is not in itself surprising or unusual, as it is both the theme and title of the chapter; the term *xian* 賢 (worthy, able), for instance, is similarly the most strongly weighted term in the vectors for both of the "Shang xian shang" 尚賢上 and "Shang xian zhong" 尚賢中 chapters. What is surprising, however, is the magnitude of this

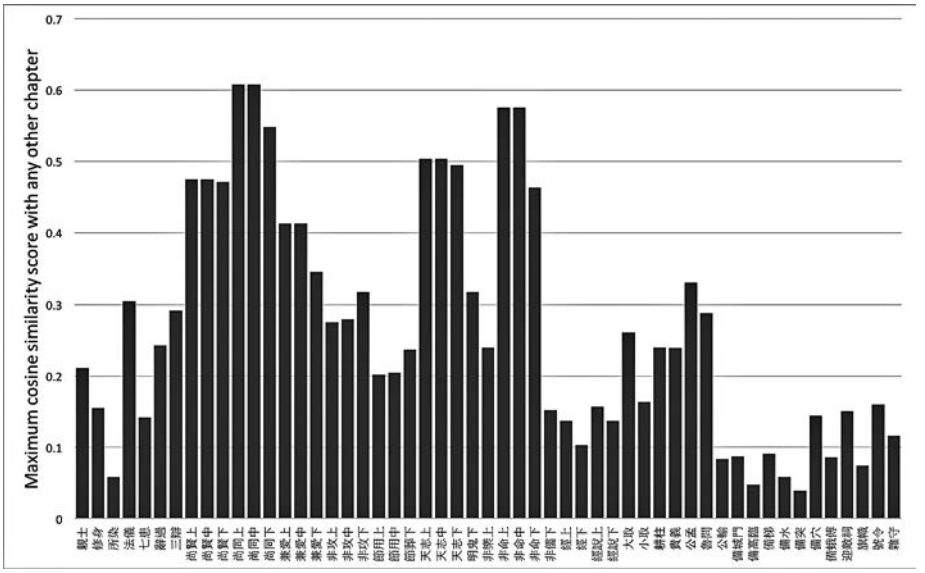

Figure 2. Maximum cosine similarity score between chapters of the *Mozi* and any other chapter of the text

component of the vector: as these vectors have been normalized by length, the highest possible value for any component weight is 1, and here *ran* 染 has a weight of over 0.9.[10] The explanation for this is in turn clear from examining the complete text of the *Mozi* for occurrences of the term: of thirty-six occurrences, thirty-four are in the "Suo ran" chapter. The remaining two both occur in the military chapters of the text, which are generally thought to be of much later authorship than the core chapters—hinting that the "Suo ran" itself may be of later construction. This hypothesis is further strengthened when looking at the

Table 2. Terms with highest TF-IDF scores in the "Suo ran" 所染 chapter of the *Mozi*

Term	*Term Frequency*	*Document Frequency*	*Inverse Document Frequency*	*TF-IDF*
染	33	3	2.87	0.924
辱	3	3	2.87	0.084
理	4	7	2.02	0.0789
蒼	2	2	3.28	0.0639
范	2	3	2.87	0.056
殘	2	3	2.87	0.056
堪	2	3	2.87	0.056
偃	2	3	2.87	0.056
稱	3	9	1.77	0.0518
逾	2	4	2.58	0.0504

Values are rounded to three significant figures. TF-IDF scores have been normalized to length 1; this has no effect on the cosine similarity scores.

pre-Qin corpus more broadly: *ran* 染 is an exceptionally uncommon term in texts believed to date to before the Han and does not, for instance, occur even once in the *Analects*, *Mengzi*, *Zhuangzi*, or *Xunzi*, texts thematically related to both the *Mozi* in general and this chapter in particular, and the latter three of which in places explicitly respond to Mohist ideas. Of the other terms, *cang* 蒼 (blue-green) occurs only in one other place in the *Mozi*, again in the military chapters; *fan* 范 (here and elsewhere in the *Mozi* appearing only in proper names), *can* 殘 (injure), *kan* 堪 (to bear), and *yan* 偃 (here used as a proper name) in only two other chapters.

Effectively, what this metric of textual similarity has highlighted in this case is that the language of the "Suo ran" chapter is highly unusual in terms of its word choice compared with the rest of the *Mozi*. On inspection of the data, we can confirm that this is indeed the case, and the algorithm has also provided us with a list of terms that explain its result: the *Suo ran* chapter is unusual in its extensive use of terms like *ran* 染, *ru* 辱 (dishonor), and *li* 理 (inherent pattern).

Similar methods of detailed comparison are useful in understanding what similarities have been identified in cases where cosine similarity indicates that two documents are related. For example, one extant group of three chapters from the core section of the text, titled "Fei gong" 非攻 (Condemnation of Offensive War), is noticeably different from all other extant sets of three texts, in that the three sections do not have strong mutual similarity: "Fei gong shang," which unlike "Suo ran" remains consistent with word usage in the *Mozi* generally, lacks significant similarities with "Fei gong zhong" and "Fei gong xia," as might be expected. In this case, it is useful to start by exploring what is similar between the latter pair of texts. One way to do this is to look at what key factors explain the high cosine similarity score between these documents.

The formula by which cosine similarity is calculated from document vectors is in fact very simple: the corresponding weighted components (i.e., TF-IDF weights for each term in the corpus) in a pair of vectors are multiplied, then all of these products added together, and lastly, this sum is divided by the magnitudes of the two vectors. It is therefore straightforward to calculate the relative contribution that each individual term makes towards the similarity score for any pair of documents (the product A_iB_i in the cosine similarity calculation, i.e., the product of the TF-IDF weights corresponding to that term in the two vectors)—this can be used to provide an ordering of terms explaining specifically which terms, through their combination of TF and IDF weights, contributed the most to any given cosine similarity score.

Table 3 lists the ten terms contributing the largest amount to the cosine similarity score between "Fei gong zhong" and "Fei gong xia." Of these terms, only *gong* 攻 (to attack), *guo* 國 (state), and *wang* 往 (to go) occur in "Fei gong

Table 3. Terms contributing the most to cosine similarity between "Fei gong zhong" 非攻中 and "Fei gong xia" 非攻下

Term	*Fei gong zhong*	*Fei gong xia*	Product	Document Frequency
攻	0.209	0.188	0.0393	25
數	0.197	0.0563	0.0111	25
師	0.0617	0.1330	0.00821	14
萬	0.0816	0.0878	0.00716	22
地	0.0716	0.0925	0.00662	21
國	0.0616	0.103	0.00634	39
賓	0.0674	0.0816	0.00550	6
往	0.0823	0.0665	0.00547	14
伐	0.0301	0.1700	0.00512	20
勝	0.1710	0.0297	0.00508	24

shang" at all; hence, the others contribute nothing to the similarity between "Fei gong shang" and either of the other two texts. Most of these terms are not especially uncommon in the *Mozi* as a whole, as evidenced by their relatively high document frequency; rather, they contribute significantly to the cosine similarity score by virtue of being repeated several times in each chapter. On examining the appearance of these terms in the text, it becomes clear why this should be the case: as is well known, the chapters of each triad frequently repeat one another, not always identically, but in ways that are clearly related. For example, part of the explanation for some of these terms appearing in this list is the repeated appearance of identical or closely related phrases in both chapters (table 4). This type of repetition is modeled only implicitly by the cosine similarity TF-IDF metric (or any other bag-of-words model): it is treated no differently than would be the occurrence of these same terms elsewhere in the document; it merely happens that, in this case, examining the cause of these similarities leads us to identify longer fragments of locally similar text. This points toward the potential utility of alternative metrics of text reuse that do not use a bag-of-words model and that take account of word order as well as similarities in word usage.

Table 4. Examples of similar phrasing in the "Fei gong zhong" and "Fei gong xia" chapters

Fei gong zhong	Count	*Fei gong xia*	Count
不可勝數(也)	9	不可勝數	2
誰敢不賓服哉	1	而天下諸侯莫敢不賓服	1
九夷之國莫不賓服	1	而天下莫不賓	1
則是棄所不足，而重所有餘也	1	然則是虧不足，而重有餘也	1

N-gram Overlap Models of Text Reuse

An alternative and in some respects even simpler method capable of identifying meaningful similarities between pieces of writing is the n-gram overlap method.[11] An n-gram is simply a sequence of *n* tokens (in this article, a sequence of *n* characters), where *n* is some integer 1, 2, 3, . . . ; *n-gram overlap* simply refers to the shared occurrence of sequences of *n* tokens in different locations within a body of writing. The intuition behind this approach is that it can capture similarities of the kind appearing in table 4, in which word sequence is important: similar, though not necessarily identical, phraseology. By modifying the value of *n*, it becomes possible to allow greater or lesser flexibility in terms of what types of similarities are identified, in other words, what formal threshold is necessary for two expressions to be considered "similar." In its simplest form, a comparison of n-gram overlap simply involves exhaustively identifying the sequences of tokens of length *n* that are shared between (and possibly also within) the documents to be compared. For example, comparing the two short pieces of text in the last row of table 4, "ze shi qi suo bu zu, er zhong suo you yu ye" 則是棄所不足，而重所有餘也 (this is to give up what is needed and to value that which is already in abundance) and "ran ze shi kui bu zu, er zhong you yu ye" 然則是虧不足，而重有餘也 (this is to neglect what is needed and to value what is already in abundance), we find that these two expressions share the following n-grams for various possible values of *n*:[12]

$n = 1$	則 是 不 足 而 重 有 餘 也
$n = 2$	則是 不足 足而 而重 有餘 餘也
$n = 3$	不足而 足而重 有餘也
$n = 4$	不足而重
$n > 4$	[No overlapping n-grams exist for any value of $n > 4$.]

Aside from being able to capture the sequential similarity ignored in bag-of-words models, this approach has the advantage of potentially directing our attention to much more specific similarities between texts: not just reuse of the same terms within large units of text, but reuse of similar expressions and phrases, which are likely to be much more localized and occur in many fewer locations overall. How specific will depend on the text itself and, of course, the value of *n* chosen: a small value will match many things, the repeated occurrence of which may not be interpretatively interesting; for example, when using characters as tokens, "Zi Mozi yue" 子墨子曰 (Mozi said) is a 4-gram that will be present in many places in the text. A large value of *n*, such as $n = 8$, will result in a high confidence that any match identified will have some significance, purely by virtue of its length. However, choosing a high value of *n* will also

eliminate the possibility of locating more subtle matches; in the specific example given above, any value of *n* greater than 4 will prevent the parallel listed in the last row of table 4 from being identified.

In addition to identifying individual instances of overlap, for any given value of *n* we can also calculate a similarity score between any chosen pair of documents somewhat analogous to the cosine similarity score. Since we would expect that for a given rate of text reuse the number of shared n-grams will increase as the lengths of each text increases, we can divide the total number of overlapping n-grams by the combined lengths of the two textual items being compared to derive a similarity score:[13]

$$\text{similarity}_{\text{ngram}} = \frac{\text{ngrams}_{\text{AB}}}{\text{length}_{\text{A}} + \text{length}_{\text{B}}}$$

Having done this, we can again visualize reuse between pairs of texts in the form of a symmetric grid for any given value of *n*, such as $n = 7$ (fig. 3). On this metric of similarity, fewer clear patterns are visible than in the corresponding matrix using cosine similarity. Two patterns however are visible: first, the core chapters, with the exception of the three chapters "Jian ai shang," "Fei gong shang," and "Jie yong shang," all have similarities with the remainder of the core chapters, and with little outside of them; second, the same four of the group of five texts identified as the Mohist dialogues are again similar to one another. Unlike cosine similarity, on this metric, documents need not necessarily be similar to themselves; rather, similarity within a document is a function of the number of n-grams that are repeated within that same document. Hence, the depth of color in the diagonal of figure 3 indicates the strength of reuse within a chapter of the text—something that bag-of-words methods cannot identify due to their lack of consideration of word order. Looking at specific results is also more intuitive than with cosine similarity; for example, the square in the matrix corresponding to the comparison between "Ci guo" 辭過 (Indulgence in Excess) and "Jian ai zhong" 兼愛中 (Universal Love II) corresponds to matching 7-grams in this pair of texts, several of which overlap, and belong to the following similarities (identical portions shared between these two texts underlined; the first three are from "Ci guo," the fourth from "Jian ai zhong"):

> 君實欲天下之治，而惡其亂也，當為宮室，不可不節。[14]
> If the rulers sincerely desire to have the empire orderly, and hate to see it in disorder, they must not indulge in building houses and palaces.

> 君實欲天下之治而惡其亂，當為衣服，不可不節。[15]
> If the rulers sincerely desire the empire to have order and hate to see it in disorder, they must not indulge in making clothing excessively.

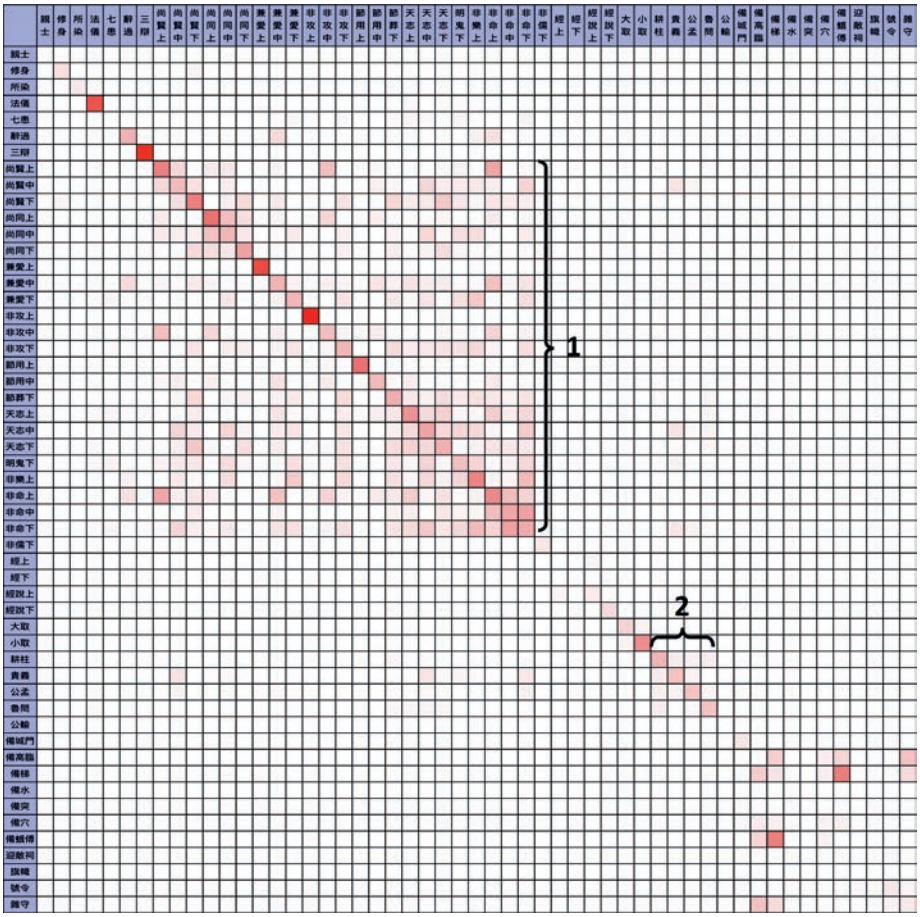

Figure 3. N-gram similarity, $n = 7$. Highlighted regions: (1) the core chapters of the *Mozi*, with the notable exceptions of "Jian ai shang" 兼愛上, "Fei gong shang" 非攻上, and "Jie yong shang" 節用上 (Economy of Expenditures I), all have similarities with other core chapters on this metric, as well as few similarities outside of this group; (2) the "Geng zhu" 耕柱, "Gui yi" 貴義 (Esteem for Righteousness), "Gong meng" 公孟, and "Lu wen" 魯問 (Lu's Question) chapters also have mutual similarities on this metric.

君實欲天下之治而惡其亂，當為舟車，不可不節。[16]
If the rulers sincerely desire the empire to have order and hate to see it in disorder, they must not indulge in constructing boats and carts excessively.

欲天下之治，而惡其亂，當兼相愛，交相利 。[17]
If [the rulers] desire to have the empire to have order and hate to see it in disorder, they should bring about universal love and mutual aid.

Another useful way of visualizing these data to get an overview of the relationships represented is using a network graph. In this type of visualization,

entities (in this example, chapters of the *Mozi*) are represented by "nodes" (shown here as filled ellipses); relationships between these entities are represented by edges, each of which joins exactly two nodes (here these lines represent similarity scores calculated between two chapters of the text, with thicker lines corresponding to higher values). In figure 4, a force-directed layout algorithm has been applied to determine the locations of each node on the two-dimensional page, distributing the nodes so as to reduce overlapping nodes and edges and highlight patterns of connections between nodes. To draw attention to the clear connections between these relationships and the hypothesized divisions within chapters of the text described in table 1, the nodes in figure 4 have been colored according to these divisions; these divisions, however, do not form part of the analysis itself or otherwise affect any aspects of the visualization.

While based on exactly the same data as the heat map of figure 3, figure 4 highlights different features of the patterns within the results. In particular, the network highlights clusters of reuse relationships: groups of chapters that have strong mutual text reuse relationships within their group but little or no reuse relationships with other parts of the text. The clearest example of this is the orange disjoint cluster of chapters at the top of figure 4. This contains all eight of the eleven military chapters that appear in the graph (the remaining three military chapters lack any shared 7-grams with any other part of the *Mozi* and therefore do not appear). While none of these chapters has direct relationships with all others in the group, all of them have *some* such relationship, and none of them has any relationship with any chapters outside the group: all reuse relationships occurring between any of these chapters and any other part of the *Mozi* fall within this same group of texts. This is reflected in the graph by the precise coincidence between the orange coloring (representing the hypothesized grouping of these texts, not included in the analysis) and the clustering (based solely on text reuse relationships measured through n-gram overlap). The same is true on a smaller scale for two pairs of chapters belonging to the set of six dialectical chapters: the "Xiao qu" and "Da qu" form a disjoint cluster, related to each other but to nothing else, as do the "Jing shang" and "Jing shuo shang" chapters.

The other main feature that can be identified from figure 4 is the strong interconnectedness of the core chapters (dark green nodes) versus all other parts of the text. While there are relationships between these chapters and several chapters from the first and second groups of dialogues (light blue and light green, respectively), these latter chapters all appear on the periphery of the strongly interconnected segment, the remainder of which is composed entirely of the core chapters only. This reflects the fact that these chapters, unlike many

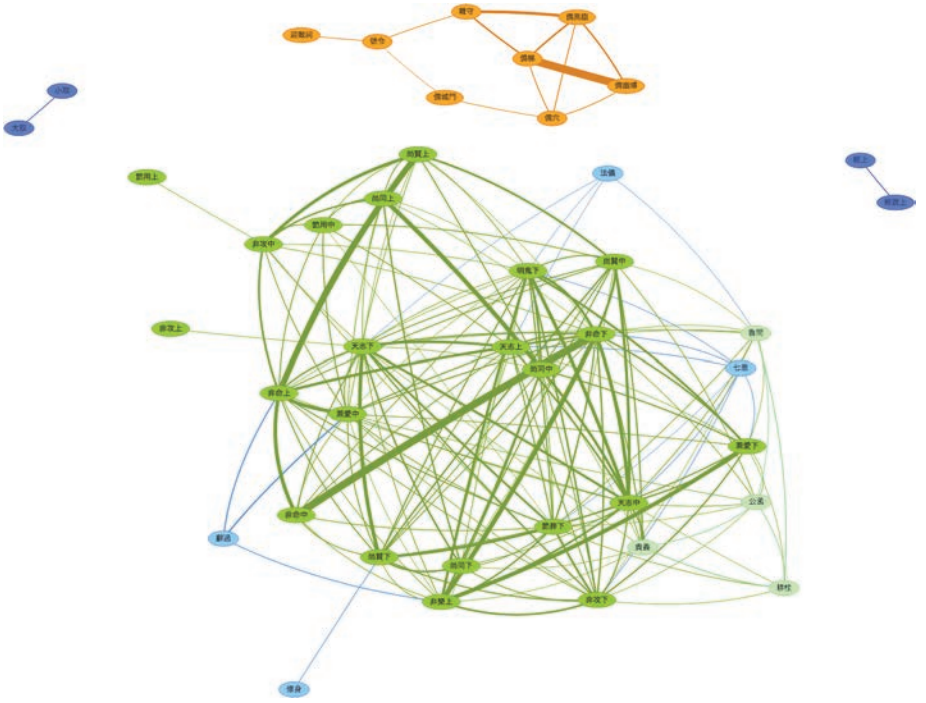

Figure 4. N-gram similarity, $n = 7$, visualized as a network graph. Colorization follows the grouping of table 1: light blue, chapters 1–7; dark green, 8–37; dark blue, 40–45; light green, 46–51; orange, 52–71. Only nodes with nonzero similarity relationships on this metric are shown.

of the core chapters, have text reuse relationships with only a limited number of core chapters.

More Sophisticated Methods of Identifying Text Reuse

With appropriately chosen values of n, the n-gram overlap method described above has the ability to draw attention to many complex reuse relationships at the expense of simultaneously identifying commonly occurring sequences the presence of which is of little interpretative significance or, alternatively, to identify only clear cases of text reuse at the expense of overlooking many cases that would be considered clear parallels if presented side by side to a reader of the text but that do not contain identical n-grams of sufficient length. It cannot, however, do both at the same time—in many cases it would be much more desirable to be able to identify exactly those instances of reuse that human readers would consider meaningful, even where these may be short and non-identical, without simultaneously returning false-positive matches where certain formal criteria are met, yet a reader would not consider this to represent a meaningful parallel.

To give a simple example, the following pair of lines appear in two separate chapters of the *Mozi*:

> 其直如矢，其平如砥，[18]
> His straightness is like an arrow and his smoothness like a whetstone, . . .
> 其直若矢，其易若底，[19]
> His straightness resembles an arrow and his ease resembles a whetstone, . . . [20]

This pair of lines cannot be identified as having any similarity at all using a simple n-gram overlap algorithm, unless n is at most 2; the only identical sequences of more than one character shared between these two lines are the initial *qi zhi* 其直 (his straightness) and the *yi, qi* 矢，其 (arrow, his). Nevertheless, there are formal similarities between the two lines, and a reader having been shown them side by side would likely agree they are parallel to each other. Performing an n-gram comparison with $n=2$ on a typical text will, however, identify very large numbers of similarities that are likely of much less interest—any two lines sharing a common construction such as *er yi* 而已 (and that's all) or *yu ci* 於此 (from this) would also be identified as matches.

To distinguish between these two types of cases, a more sophisticated metric of text reuse is needed. A metric designed to reliably identify these types of parallels within the pre-Qin and Han corpus is described in Sturgeon, "Unsupervised Identification of Text Reuse," which identifies parallel passages of this type using an algorithm based on maximizing the value of a similarity metric over pairs of text strings in the corpus. An evaluation of the data compared the automated results with those of a manually reviewed reference work using the metrics of precision, recall, and F-score usually used for this type of information retrieval task and demonstrated that the results obtained match or exceed the accuracy of those of human editors for this task; in other words, the results identified correspond closely to human intuition for what constitutes a meaningful parallel.[21]

This metric is quite different in kind from the two other reuse metrics considered so far. Cosine similarity and n-gram overlap are both strictly formal methods, which perform relatively straightforward calculations according to particular formulas and produce results that require careful examination of identified instances to establish what, if any, relevant text reuse features have been identified in each case. Their strengths are their simple, mechanical nature, which makes them simple to apply and also makes it easy to interpret the literal meaning of an identified instance, namely, that the instance meets the formal requirements of the algorithm, such as possessing an identical n-gram. The data produced, however, are noisy not just because of the nature of the textual

materials but also because these algorithms often capture a mixture of different syntactic and semantic elements of the texts they are applied to, as interpretatively different cases frequently lie behind the same formal features.[22] By contrast, the algorithm for identifying parallel passages aims very specifically to correctly identify one particular type of text reuse relationship and to do so just as reliably, if not more so, than human experts attempting to perform the same task. The ability to do this comes at a price of increased complexity: in accurately locating parallel passages in a large body of text, many different factors have to be weighed to capture the complexity of language use. This explains why, for example, repeated identical sentence fragments consisting only of strings like "Yue wang Gou Jian" 越王勾踐 (King Gou Jian of Yue, r. 496–94 BCE) or "ci zhi wei ye" 此之謂也 (is what this refers to) are not parallel passages in the relevant sense, while nonidentical pairs such as "chuo yu tu xing" 啜於土鉶 (drank out of an earthen *xing*) and "chuo hu tu xing" 啜乎土型 (drank from an earthen mold) do in fact meet the definition despite their superficial differences.[23]

Conceptually, the approach is straightforward: define a mathematical function taking two arbitrary strings of text as inputs, and compute a similarity score on the basis of both their contents as well as facts about language use in the relevant time period, for example, the interchangeability in many contexts of the terms *ru* 如 and *ruo* 若 (which both mean "such as" or introduce conditional phrases), *yu* 於 and *hu* 乎 (which are interchangeable locative particles), and so on. Maximize the value of this function over all possible pairs of fragments of the corpus of all possible lengths; then, for each pair identified, use the similarity score as a cutoff to decide whether or not the pair constitutes a genuine parallel; and finally, evaluate the performance of the procedure and cutoff against human-edited data to confirm the accuracy of the results. Practically this task is difficult because of both the complexity involved in defining an adequate function and the extremely large number of comparisons that would need to be made in principle to maximize its values over a corpus.

The results produced by this procedure are somewhat similar to those of an n-gram overlap comparison, in that they relate specific regions of two texts together. The key differences are, first, that the regions need not be identical and, second, that specific "meaningful extents" of reuse are identified as a result. The significance of these two aspects lies primarily in the specificity of the relationships identified; this is most easily illustrated by looking at a concrete example (here modern additions to the text shown within the marks 【 . . . 】 have been added for clarity and are not present in the *Daozang* text nor used in the actual comparison):

A. 若有美善則歸之上，是以美善在上，而所怨謗在下，寧樂在君，憂慼在臣，故古者聖王之為政若此。[24]
When there were any excellences and virtues they were attributed to the emperor. Thus excellences and virtues belonged to the emperor while complaints and slanders were directed against the subordinates. Peace and joy abode with the king while worries and sorrows were lodged with the officials. This was how the ancient sage-kings administered the government.

B. 外匡其邪，而入其善，尚【同】而無下比，【是】以美善在上，而怨讐在下，安樂在上，而憂慼在臣。此翟之【所】謂忠臣者也。[25]
He should correct irregularities and lead in goodness; he should identify himself with the superior and not ally himself with subordinates, so that goodness and excellences will be attributed to the superior and complaints and grudges lodged against the subordinates; so that ease and happiness be with the superior and trouble and worry with the ministers. This is what I call a loyal minister.

Comparing the uncorrected text of these two short passages using n-gram overlap, assuming a sufficiently small value of *n* is chosen to identify nonidentical parallels, will identify multiple short similarities. For example, choosing $n = 3$ and highlighting each matched n-gram as a shade of red (with text corresponding to multiple overlapping 3-grams shaded in progressively darker shades) gives the result shown in figure 5.

This expresses the knowledge that in this pair of sentences two parts have clear similarity—in fact, both contain the identical strings "yi meishan zai shang er" 以美善在上而 and "you qi zai chen" 憂慼在臣; additionally, the fragment "zai shang er" 在上而 occurs in both (in A, toward the start of the sentence, and in B, toward the end). However, no particular information has been produced about possible relationships between any other parts of these sentences. In fact, comparing them side by side, we see that, in addition to the parts identified by a 3-gram overlap analysis, the sentence fragments occurring between these two near-identical fragments also have a high degree of formal similarity. This is easily seen by aligning the identical parts within the pair:

A′. 以美善在上，而所怨謗在下，寧樂在君， 憂慼在臣，
B′. 是以美善在上，而 怨讐在下，安樂在上，而憂慼在臣。

Thus, a more natural thing for us to say about sentences A and B is not that they share six identical 3-grams but that they contain one extended nonidentical parallel, specifically, the one parallel indicated above between fragments A′

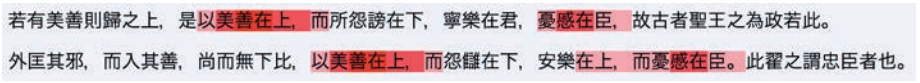

Figure 5. Shared 3-grams between two lines of the *Mozi*

and B′. This more abstract form of information offers a number of advantages. First, it becomes possible to employ visualizations based on the entirety of what we, as interpreters, would be likely to consider as related fragments, rather than what is easily identifiable using a simple algorithm. For example, we might want to automatically align the parallel fragments as shown above or, alternatively, highlight within each parallel segment which parts are *not* identical in these two versions of the text (or, more generally, between however many parallels there are to this piece of text within a given corpus), as shown in figure 6.

Second, the results obtained using this approach can be used to evaluate the validity of other techniques such as n-gram overlap as an approximation for identifying parallels of this type. Using data obtained through this more sophisticated type of comparison, we can recreate the grid and graph summaries of figures 3 and 4 to see to what extent the results of the more accurate approach agree qualitatively with the simpler n-gram analysis (figs. 7 and 8).

While the details of the results are not identical, they broadly confirm the trends identified using n-gram overlap. The main conclusions drawn from figure 4 also apply to figure 8, with several minor caveats; in particular, several chapters that were not previously connected are now connected, such as "Jie yong shang" and "Hao ling" (for which the parallel use of the term "yin shi bu shi" 飲食不時 [eating and drinking irregularly) falls beneath the threshold of $n = 7$ for identification by n-gram overlap); "Da qu" is now connected to "Shang tong zhong" (by means of a similarly short parallel, "xing li chu hai" 興利除害 [create benefits and remove harms]). The broader conclusions, however, remain unchanged, suggesting that n-gram overlap can provide a useful approximation to more rigorous approaches to identifying textual parallels.

Identifying Text Reuse in Practice

Document similarity calculations using cosine similarity and TF-IDF weighting are widely used, mainstream techniques in the context of document search systems, including open-source platforms such as Apache Solr and

Figure 6. Automatic highlighting of differences between parallels (visualization from Chinese Text Project, ctext.org/text.pl?node=3898&if=en&show=parallel)

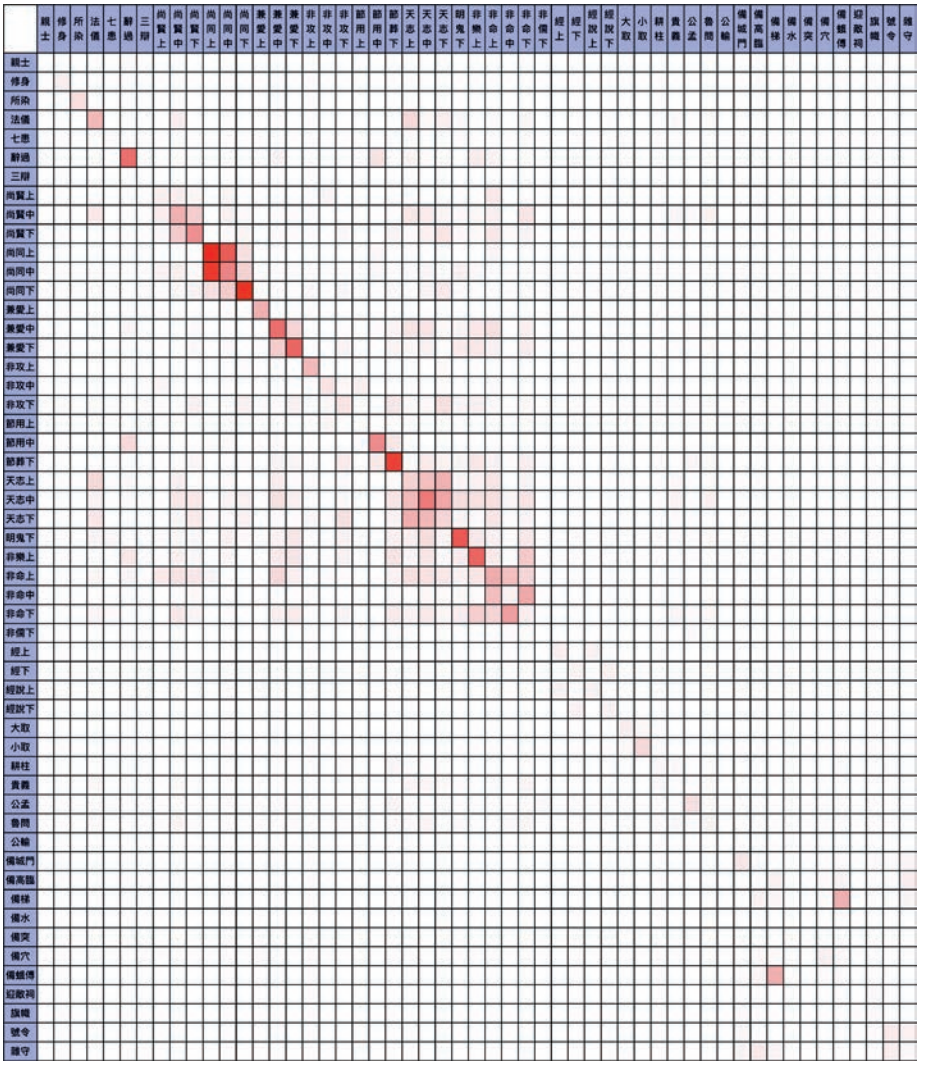

Figure 7. Parallel passage similarity (data from Sturgeon, "Unsupervised Identification of Text Reuse")

Elasticsearch. Nevertheless, such systems do not generally fit well with the workflow of a researcher of historical literature, being designed primarily for the goal of implementing effective and scalable document search rather than introspection into the details of the comparison process itself, and requiring some effort to configure and adapt to this task. Plagiarism detection systems have more in common with the task of identifying and highlighting specific instances of text reuse but by their very nature must operate at a different scale from that of the historical text reuse case, constantly incorporating new data as new works are published, and with the aim of identifying given a particular

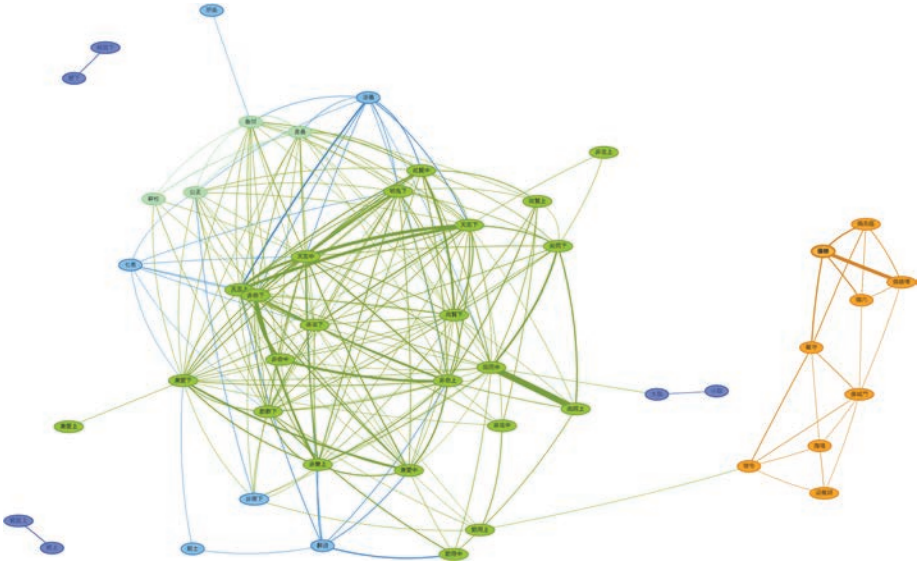

Figure 8. Parallel passage similarity visualized as a network graph. Colorization follows figure 4.

passage that *some* significant portion has been illegitimately reused, as opposed to locating all relevant reuse instances within a closed corpus—thus requiring extremely large and centralized systems to process and maintain the data.[26] In the humanities context, by contrast, corpus size is likely to be considerably smaller, less time may be available to commit to technical implementation and maintenance of complex systems, and last but most important, the methods used must be critically evaluated and facilitate direct examination of the evidence: the textual details always matter.

As part of the work described in this article, an online platform called Text Tools (ctext.org/plugins/texttools/#help) has been developed to better meet the needs of researchers dealing with historical textual materials in general, and premodern Chinese works in particular. This platform provides the ability to perform a number of basic analyses on arbitrarily chosen textual materials; most relevant to this article are its ability to perform n-gram overlap analysis and cosine similarity calculations, as well as a variety of appropriate visualizations using charts, heat maps, text highlighting, and network graphs. While capable of working with user-provided textual materials, the platform can also import texts directly from the Chinese Text Project digital library (ctext.org), which is also the source of the digital editions used in this article, and so can be used to easily reproduce most of the visualizations included in this article in interactive form.

This system is also an example of the use of application programming interfaces in the humanities;[27] while designed to work with the Chinese Text

Project digital library, the tool is not itself a part of that library and is afforded no special access to the database upon which it is built, instead requesting textual data from the library through publicly available interfaces only. Taken together, these functions allow the comprehensive identification of text reuse in arbitrarily selected materials. A significant benefit of this type of digital platform is the interactivity possible in the digital medium; in the case of n-gram overlap, for example, in addition to producing summary statistics as well as comprehensive text highlighting of all specific overlaps identified, network visualizations of text reuse patterns can be produced by the platform that have their edges linked to the specific highlighted textual display. This means that visualizations such as those shown in figures 4 and 8, which were produced using this online platform, can also serve as direct navigational aids with which to locate textual parallels within a large corpus of material. Summary statistics can also connect directly to the underlying data, exposing, for example, the TF-IDF values and related vectors and enabling their immediate comparison, highlighting the types of features described in the examples discussed above. This has the effect of making efficient access to both techniques and data available to researchers who may have neither the time nor technical expertise to implement these methods themselves using a programming language. With these goals in mind, detailed overviews of the functionality as well as step-by-step tutorials have also been published online.[28] When used together with the Chinese Text Project application programming interface, most of the figures and tables included in this article can be reproduced directly using the online interface; it follows as a corollary that analogous results can also be generated and explored for arbitrarily selected texts.[29] In the case of parallel passages, a separate purpose-designed system has been created and published online to present the results and enable efficient access to the underlying data.[30]

Conclusions and Future Work

Text reuse provides a clear example of the huge potential benefits of computationally aided research in the humanities. Computational techniques are capable of greatly assisting human interpreters in locating similarities within a corpus with degrees of comprehensiveness and accuracy that would otherwise be prohibitively time-consuming and, for some types of task, can in practice even produce results superior to those of human editors. The speed at which such comparisons can be made makes possible new types of data-driven study, in which statistical evidence can be provided to back up claims about complex textual relationships and textual production histories.

As more sophisticated techniques continue to be developed for identifying specific types of text reuse of particular interest to researchers, entirely new

types of study become possible as a by-product of the large and comprehensive data sets these techniques are able to generate. For example, having identified all parallels within a particular corpus of texts, it becomes possible to ask many questions about the properties of these textual parallels in aggregate, such as, what is "typical" reuse within a given corpus, and what is atypical; where are there absences of any text reuse relationships that might be expected to occur; what substitutions are made between generally parallel passages, and do these occur in patterns suggesting further insights into processes of textual production? Some of these questions may require the development of further analytic techniques; others simply require that digital platforms be made accessible to researchers working with particular materials. While identifying text reuse is a task well-suited to computer software, interpreting its significance and the causal factors behind its occurrence is by contrast a task for expert readers familiar with the particular materials and their contents and context. A combination of automated methods, with their ability to scale to large bodies of writing, and detailed scholarly analysis of both individual instances and broader trends, which can be identified using these techniques, has the potential to significantly further our understanding of early transmitted texts and their origins.

DONALD STURGEON 德龍
Harvard University
djs@dsturgeon.net

Notes

1. Graham, *Composition of the Gongsuen Long Tzyy*, offers a systematic example of this approach.
2. *Mozi jiangu*, p.418.
3. *Mozi* 78/45/17–18. In this article, the text of the *Mozi* used generally follows the uncorrected *Zhengtong daozang* 正統道藏 edition of the text, except where indicated. English translations are based on those of Mei, *Ethical and Political Works of Motse*, with some modifications. References to the *Mozi* cite page, section, and line numbers in *A Concordance to Mo Tzu* 墨子引得, as quoted in the Chinese Text Project website (ctext.org). Locations of textual references given in this article can also be determined using the Chinese Text Project concordance tool (ctext.org/tools/concordance).
4. *Mozi* 78/45/21–22.
5. However, the methods described in this article all operate on tokens irrespective of whether these tokens are characters or words, so they can equally be applied to textual materials that have been segmented into words where this is feasible to do.
6. Examples of applications of cosine similarity to historical text reuse identification include Lee, "Computational Model."

7. In our case, this means that *n* is the number of distinct Chinese characters appearing anywhere within any of the textual objects being compared. A vector, in this context, can be understood as an ordered list containing *n* components, each of which is a number representing some information about one term from the corpus and its relationship with the particular document concerned.
8. Note that all of the comments on these examples would apply equally if we used words rather than characters as tokens, i.e. *tianxia* instead of *tian* and *xia*, and *Zaofu* rather than *zao* and *fu*, provided that this tokenization was applied consistently to all of the documents.
9. Graham, *Mo Tzu*, 336–38.
10. Normalization in this context means that all components of each vector have been divided by the magnitude (Euclidean length) of the vector. This has the effect of making all vectors have a length 1; because all components within each vector are divided by the same value, the ratios of components do not change, and this operation has no effect at all on cosine similarity between vectors.
11. See Clough et al., "METER"; Sturgeon, "Unsupervised Identification of Text Reuse"; Forstall et al., "Tesserae Project"; Smith, Cordell, and Stramp, "Detecting and Modeling Local Text Reuse"; and Seo and Croft, "Local Text Reuse Detection."
12. In what follows, it is assumed that punctuation is always ignored in making comparisons; however, punctuation is included in the quoted examples for the convenience of the reader.
13. Other possibilities include using an asymmetric metric of containment, as suggested in Clough et al., "METER."
14. *Mozi* 6/6/9.
15. *Mozi* 6/6/21.
16. *Mozi* 7/6/32–33.
17. *Mozi* 24/15/41–42.
18. *Mozi* 1/1/19.
19. *Mozi* 27/16/61.
20. The translation here follows Sun Yirang's argument that *di* 底 (bottom) is in fact a mistake for *di* 砥 (whetstone). Sun, *Mozi jiangu*, 124.
21. Analysis in Sturgeon, "Unsupervised Identification of Text Reuse," based on comparison with manually edited data based on Ho, Chu, and Fan, *Xunzi*.
22. A simple example of this is n-gram overlap with a small value of *n*; this may capture a mixture of genuinely reused or borrowed sentences, common phrases, proper names, and name-title combinations, because these can all result in sequences of writing that fit the formal criteria for n-gram overlap.
23. For details of how this can be addressed, see Sturgeon, "Unsupervised Identification of Text Reuse."
24. *Mozi* 10/9/21–22.
25. *Mozi* 90/49/1.
26. Examples include commercial plagiarism detection services such as Turnitin.
27. An application programming interface, or API, provides machine-readable access to data stored in a system—such as a database—primarily intended for human use. This helps facilitate the future use of such data by third parties.
28. See digitalsinology.org/text-tools/ and dsturgeon.net/texttools/.

29. In particular, all figures, graphs, diagrams, and tables relating to TF-IDF and n-gram shingling can be reproduced by a researcher using the tool either directly or in conjunction with basic spreadsheet skills.
30. This system is introduced and described in Sturgeon, "Unsupervised Identification of Text Reuse."

References

Clough, Paul, Robert Gaizauskas, Scott S. L. Piao, and Yorick Wilks. "METER: MEasuring TExt Reuse." In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 152–59. Stroudsburg, PA: Association for Computational Linguistics, 2002.

Graham, A. C. "Mo Tzu 墨子." In *Early Chinese Texts: A Bibliographical Guide*, edited by Michael Loewe, 336–41. Berkeley: Society for the Study of Early China; Institute of East Asian Studies, University of California Berkeley, 1993.

———. "The Composition of the Gongsuen Long Tzyy." *Asia Major* 11 (1956): 147–83.

Ho, Che Wah, Kwok Fan Chu, and Sin Piu Fan, eds. *The Xunzi with Parallel Passages from Other Pre-Han and Han Texts*. Hong Kong: Chinese University Press, 2005.

Lee, John. "A Computational Model of Text Reuse in Ancient Literary Texts." In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 472–79. Stroudsburg, PA: Association for Computational Linguistics, 2007.

Mei, Yi-Pao. *The Ethical and Political Works of Motse*. London: Probsthain, 1929.

Mozi 墨子. *Zhengtong daozang* 正統道藏 edition.

Seo, Jangwon, and W. Bruce Croft. "Local Text Reuse Detection." In *Proceedings of SIGIR '08*, 571–78. New York: Association for Computing Machinery, 2008.

Smith, David A., Ryan Cordell, and Elizabeth Maddock Dillon. "Infectious Texts: Modeling Text Reuse in Nineteenth-Century Newspapers." In *Proceedings of the Workshop on Big Humanities*. Silicon Valley, CA: IEEE, 2013.

Smith, David A., Ryan Cordell, and Nick Stramp. "Detecting and Modeling Local Text Reuse." In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, 183–92. New York: Association for Computing Machinery, 2014.

———. "Unsupervised Identification of Text Reuse in Early Chinese Literature." *Digital Scholarship in the Humanities* 33, no. 3 (2017): 670–84.

Sun Yirang 孫詒讓. *Mozi jiangu* 墨子閒詁 (Exposing and Correcting the *Mozi*). Beijing: Zhonghua shuju, 2001.