Unsupervised identification of text reuse in early Chinese literature

Donald Sturgeon
Fairbank Center for Chinese Studies, Harvard University, USA

Abstract

Text reuse in early Chinese transmitted texts is extensive and widespread, often reflecting complex textual histories involving repeated transcription, compilation, and editing spanning many centuries and involving the work of multiple authors and editors. In this study, a fully automated method of identifying and representing complex text reuse patterns is presented, and the results evaluated by comparison to a manually compiled reference work. The resultant data are integrated into a widely used and publicly available online database system with browse, search, and visualization functionality. These same results are then aggregated to create a model of text reuse relationships at a corpus level, revealing patterns of systematic reuse among groups of texts. Lastly, the large number of reuse instances identified make possible the analysis of frequently observed string substitutions, which are observed to be strongly indicative of partial synonymy between strings.

Correspondence:
Donald Sturgeon Fairbank Center for Chinese Studies, Harvard University, Room S126, CGIS South Building, 1730 Cambridge St., Cambridge, MA 02138, USA.
E-mail:
djs@dsturgeon.net

1 Introduction

Text reuse occurs in many different forms, from the verbatim copying of extended passages through to indirect allusions that—though unmistakable to a reader with the appropriate background knowledge—may otherwise appear entirely opaque on a surface level. This study focuses primarily on the former type: text reuse of phrases, sentence fragments, or longer passages in which a side-by-side comparison of two or more sections of text, even in isolation of detailed background knowledge of the text, its authorship, and other factors, can lead one to the conclusion that the origins of the two passages likely share some relevant causal relation: either one was copied from the other, or both share a common origin in some earlier text, oral tradition, or well-known saying.

The first technical aim of this study was to implement an algorithm that can accurately detect occurrences of this type of text reuse as well as determine the meaningful extents of such matches with a minimal instance of false positives within the pre-Qin and Han corpus,[1] and use this algorithm to build a comprehensive model of text reuse within this corpus. A second important practical objective was to make the results of the study available online in an easily accessible form, such that specialists and non-specialists alike can immediately take advantage of the data produced by the study and incorporate its findings into their own research.

2 Defining Parallel Passages

Automated identification of parallel passages in early Chinese texts presents challenges that are interestingly different to those in many other languages. For example, where lemmatization or stemming typically plays an important role in the task for inflected languages, for Chinese it is completely unnecessary; by contrast, however, a somewhat similar problem arises with respect to variant characters. Due to many factors—including scribal error,

Digital Scholarship in the Humanities, Vol. 33, No. 3, 2018.
doi:10.1093/llc/fqx024 Advance Access published on 6 November 2017

imperial edict,[2] regional variation, and historical evolution of the language—words appearing in multiple versions of what is essentially the same text or passage may be written using differing characters. Sometimes these differences may be minor and have no influence upon the interpretation of the text, and sometimes they may completely alter the apparent meaning.[3] However the sheer range of characters available presents potential problems for any method of comparison that does not take similarity of characters into account[4]: it is quite possible, in principle and also in practice, to have two passages that have no identical characters in common at all, and yet are in fact paradigm instances of the type of text reuse considered here.[5] As a result, a satisfactory method of identifying parallel passages even of this relatively simple type for Chinese is likely to rely upon some notion of character similarity, and it seems clear that the identity relation—counting only identical characters or words as similar—will be less than ideal.

With this in mind, a more general characterization of parallel passages of the desired type seems useful in clarifying the aims of the study. For the purposes of this study, 'parallel passages' are defined as sections of text which contain significant common subsequences of characters having a high degree of similarity, and in addition appear likely to be causally related in terms of word choice by something more than the writers' shared linguistic competence of the language.[6] These may occur within the same text or in two different texts, and with or without any form of explicit citation.

3 Identifying Text Reuse

3.1 Background

A significant question in identifying text reuse in general and Chinese texts in particular is what unit of text should be taken as basic to the study. Many studies of text reuse focus on similarity between documents, paragraphs, or sentences—that is, between predefined textual units, each of which either does or does not share a reuse relation with each other unit of the same kind.[7] These characterizations significantly simplify the problem, but are problematic for pre-modern Chinese materials for a number of reasons. First, the grammar of classical Chinese is such that sentences can often be correctly delimited in different ways. When combined with the fact that pre-modern Chinese texts typically included no punctuation or sentence delimiting markers, this has the effect that even identical passages occurring in two different modern editions of the same early text will often disagree on exact sentence delimitation. The same is also true of paragraphs—though some transmitted texts do explicitly break content up into paragraphs, the majority do not, and even those that do often vary according to edition. If the units to be compared are sentences or paragraphs, these factors would have a negative impact on both accuracy and visualization of results, since even identical passages may be delimited differently in two distinct cases.

Secondly, a rather different issue arises as a result of factors explaining some instances of non-identical text reuse: the editing of fragmentary or damaged texts, and errors introduced accidentally during transcription and editing. Many transmitted texts as we have them today are in part the result of sometimes extensive reconstruction by Han dynasty editors, who—in addition to other challenges—were often faced with the task of reconstructing a text from characters written on bamboo strips, with a single column of characters on each strip, but which had over time lost some or all of the ties that once bound these strips into a particular order. Particularly given the flexible grammar of classical Chinese, in which the same word can typically function as a noun, verb, or adjective in different contexts without any explicit marker to indicate its grammatical category, and in which the majority of the words consist of a single character, the reordering of such strips would often present a challenging task, and different editors might easily decide upon different orderings and—particularly where strips were broken or partially illegible—different interpolations to make the same passage of text intelligible. As a result, fragmentary reuse that arbitrarily crosses sentence boundaries is likely to occur frequently in the corpus.

In contrast to concerns about delimiting sentence boundaries however, there is a much greater degree of agreement among interpreters about boundaries

between phrases in classical Chinese—indeed, until the relatively recent introduction of modern punctuation, the most common method of punctuating classical Chinese texts used a single type of punctuation to indicate both phrase and sentence boundaries without making any distinction between the two. In view of the definition of text reuse given in Section 2, this study therefore ignores sentence boundaries completely in identifying text reuse. Instead, reuse is identified between contiguous strings of phrases, where a phrase is any unit that would be delimited by any modern phrase-delimiting punctuation symbol (i.e. ◦ , ： ； etc.): that is, reuse is identified between n phrases of text A and m phrases of text B for some n, $m > 0$.

The resulting text reuse detection problem differs significantly from those considered in many previous studies of text reuse. One key difference between this study and prior work on text reuse is that the primary goal of this study is to identify individual reuse instances which will be of practical utility to scholars engaged in close reading of historical texts. This contrasts first with a large body of research on document-level text reuse, which has many contemporary applications in addition to those in the study of literary text reuse, such as in search engine duplicate removal and plagiarism detection. In document-level text reuse identification, the primary goal is to identify pairs of documents between which significant levels of reuse occur, without any requirement to establish precise bounds or more specific locations of this reuse within these documents. This type of approach is directly applicable to some types of study of historical text reuse, particularly to those in which reuse is to be identified among a set of predefined units. For example, in the METER project studying text reuse of newspaper articles (Clough *et al.*, 2002), the desired goal was to determine for each individual article whether or not that article was similar to each of the other articles in the study, rather than identify which specific parts of each article were similar to which parts of which others, and thus standard document similarity techniques could be directly applied.

Secondly, the problem considered in this article also contrasts with many previous computational studies of historical text reuse, including some in which the type of reuse being sought is local reuse—that is, reuse between parts of documents rather than between entire documents. In many such studies, the ultimate objective of identifying such local text reuse is to establish that the documents within which it occurs are themselves related. This is a sub-problem in which the use of many simplifying approximations may be acceptable, and even desirable as they can make feasible studies on larger data sets. For example, Smith *et al.* (2013) only attempted to identify cases of reuse of 100 words or more between documents, as shorter reuse instances might not be sufficient to conclude that the pair of documents in which they appeared were related; in the present study, any meaningful reuse of four or more characters is sought, as in close reading short reuse instances of the desired type are very often of significant interest. For the same reason, many studies accept approximate solutions to the reuse identification problem, because while these approximations will cause some instances of local reuse to fail to be identified, this will not significantly affect the corpus-level results—particularly if the missed instances are shorter cases, which are both weaker signals of document-level similarity, and harder to reliably identify without an increased risk of incurring false positives. By contrast in this study, the objective is to accurately identify *all* reuse instances, because the individual instances do not merely have instrumental value as a means by which to identify relationships between larger textual units, but also intrinsic value as parallels with utility in close reading of the passages they appear in. For this same reason, in this study it is also necessary to identify the precise boundaries of reuse instances, because these will be presented directly to the researcher; in an aggregate study ultimately focusing only on the relationships between larger units as implied by local text reuse instances, this step is not required.

3.2 Algorithm

Due to the relatively high degree of similarity and length of many parallel passages in early Chinese texts, even a naïve string-matching algorithm (or equivalently, a naïve character n-gram matching or shingling algorithm) is sufficient to detect the *existence* of many similar passages, at the expense of

including many false positives in cases of short strings and not making a useful determination as to the full extents of reuse.[8] In this study, likely instances of text reuse were initially identified using a modified version of naïve string matching which treats closely related characters as identical. The aim of this initial stage was to identify pairs of strings having a relatively high probability of belonging to a genuine parallel pair as compared with arbitrarily chosen pairs; this would allow a subsequent phase of analysis to use more computationally expensive comparison methods without making the entire process prohibitively expensive in terms of required computation time. To perform the initial matching efficiently, the entire data set was first normalized to remove the radical from compound characters and take account of a number of common variants.[9] A naïve string-matching procedure was run on this modified data set, producing a list of around 3.5 million pairs of strings of four characters or longer that were identical under the chosen normalization.[10] As was expected, the pairs generated by this procedure included many matches that did not meet the relevant definition of text reuse, and the majority of matches did not have the desired extents, since any single variation not explicitly taken account of by the normalization process would limit the extent of the match.

The second phase of the comparison considered only those pairs of strings that had been identified in the first phase and the text surrounding them. This phase had the dual aims of determining the optimal extents of each match in both texts, and deciding whether or not each match so determined met the relevant definition of text reuse. To accomplish this, a metric was defined to score the significance of a match based on several factors. These included the number of matching characters occurring in the two (original, unnormalized) strings in the same order versus the number differing, where 'matching characters' were those that were either identical or shared a common structural component as well as one identical reading.[11] Non-matching grammatical particles (虛辭) were weighted with lower significance than non-matching content words. An additional factor that was included in the calculation was the relative frequency of matched characters in the corpus as a whole; the metric was increased when generally infrequent characters occurred in both strings. This allowed the algorithm to distinguish between uninteresting identical matches such as '此之謂也' and short but statistically unusual matches such as four-character quotations from the *Book of Poetry*. Matches were scored on the number of words they contained—though most words in classical Chinese consist of a single character, proper names often do not; without taking this into account, two occurrences of a name-title combination such as '越王勾踐' (King Goujian of Yue) would be indistinguishable from a genuinely significant four-character phrase.

With this metric defined, the second phase of the algorithm consisted of examining in turn each candidate pair of strings A and B identified in the first phase, and maximizing the value of the metric over contiguous sequences of phrases a_n, a_{n+1}, … a_{n+x} and b_m, b_{m+1}, … b_{m+y}, where a_n, a_{n+1}, … a_{n+x} overlaps some part of string A, and b_m, b_{m+1}, … b_{m+y} overlaps some part of string B. The extents of the match which maximized the value of the metric were taken as the optimal extents, and matches which did not meet a minimum score using the metric were discarded. The weightings of the various parameters in the metric as well as the cut-off value for discarding low-quality matches were manually tuned to give an acceptable trade-off between precision and recall (see Section 6).

Finally, the accepted matches were integrated into the Chinese Text Project (CTP) parallel passage database, making it possible to both browse and search the data in a meaningful way.[12] This process included grouping sets of similar parallel pairs together for more convenient visualization.

3.3 Example

To illustrate the procedure, consider the following two fragments of text from the *Huainanzi* (A) and *Lüshi Chunqiu* (B):

A …則庸人能以制勝。今使烏獲、藉蕃從後牽牛尾，尾絕而不從者，逆也；若指之桑條以貫其鼻，…
B …此論不可不熟。使烏獲疾引牛尾，尾絕力勷，而牛不可行，逆也。使五尺豎子引其棬，…

A naïve string matching algorithm with a four-character threshold and ignoring punctuation will identify within this pair a single string of identical characters: '牛尾尾絕', this being the longest identical substring of the pair. The second phase of the algorithm attempts to determine what the maximally similar pair of phrase-aligned substrings of A and B overlapping the string '牛尾尾絕' are, and whether or not this pair constitutes a parallel passage. Maximizing the similarity metric, the following pair is identified:

A′ 今使烏獲、藉蕃從後牽牛尾, 尾絕而不從者, 逆也;
B′ 使烏獲疾引牛尾, 尾絕力勩, 而牛不可行, 逆也。

While clearly far from identical, the common subsequences of these strings indicate a strong similarity: both strings contain the identical subsequences '使烏獲', '牛尾尾絕', '而', '不', and '逆也', as can be seen by omitting punctuation and aligning the strings:

A′ 今 使烏獲 藉蕃從後牽 牛尾尾絕 ...
B′ 使烏獲 疾引 牛尾尾絕 ...
A′ ... 而 不 從者 逆也
B′ ... 而 牛 不 可行 逆也

By contrast, there are no non-trivial similarities between the phrases immediately before ('則庸人能以制勝' versus '此論不可不熟') and after ('若指之桑條以貫其鼻' versus '使五尺豎子引其棬') this match, demonstrating why increasing the match extents within A and B to be larger than those of A′ and B′ would in all cases decrease rather than increase the similarity metric, and therefore why this should be the pair of spans which maximizes the metric, as well as the formal extents of the parallel.[13]

4 Results

Meaningful quantification of the volume of text reuse found is complicated by the fact that text reuse as defined in this article includes reuse through explicit and implicit citation, and thus, for certain phrases and quotations that came to be well-known and widely cited, counting individual *pairs* of parallels leads to excessively high numbers as each citation is parallel to both the source (if known) as well as all of the other citations of it. As a result, it is more meaningful to consider the number of *groups* of parallel passages and their members rather than numbers of parallels *per se*—for instance, if ten versions of a line of text exist that are all similar to one another, this can be counted as a single group with ten members (one of which may be a text explicitly cited by some or all of the others). Enumerating the data in this way, the automated procedure identified 202,000 memberships in 66,000 parallel groups.[14]

As these numbers are affected by the precise choice of extents of parallel passages and their groupings, a more objective measure of text reuse which can be considered is the total number of characters belonging to one or more parallel shared with some other text, or with the corpus as a whole. This gives a concrete and intuitive measure of the amount of text reuse within individual texts or with the corpus generally. This number can be normalized by the length of text (or more generally, fragment of the corpus) being compared to give the fraction of one text belonging to at least one parallel group shared with another text or corpus fragment[15]:

$$par(A, B) = \frac{characters\ of\ A\ in\ A : B\ parallels}{characters\ in\ A}$$

For the pre-Qin and Han corpus as a whole, 40% of all writing as measured by character had at least one parallel as defined above with some other segment of text in the corpus. This implies that within the corpus, for any given text, line, or passage there is a substantial probability that some of it can be found repeated in some form elsewhere within the corpus—once we know where to look. High levels of reuse were as expected found in a number of texts throughout the corpus, including both canonical texts that subsequently became widely quoted as well as texts which share large amounts of unattributed material with passages appearing elsewhere. There was considerable variation in reuse among texts; those with the greatest proportion of reuse

included those widely recognized as borrowing from or otherwise duplicating content found elsewhere such as the *Kongzi Jiayu* and *Qian Han Ji*, both of which shared the highest observed proportion (79%) of their content with other parts of the corpus. Widely cited canonical texts such as the *Book of Poetry* (62%),[16] *Analects* (61%), and *Laozi* (51%) also scored highly. Texts with very low scores included early dictionaries and dictionary-like writings, such as the *Jijiupian* (1.1%) and *Shi Ming* (3.5%), as well as other compositions like the *Lie Xian Zhuan* (5.3%). The *Shuowen Jiezi* dictionary (13%) contained significantly more reuse than other dictionary-like writings due to its relatively frequent inclusion of citations from other texts.

5 Representing the Data

In addition to placing concrete values on the amount of text reuse in early Chinese texts, this study also aimed to make the data accessible to scholars working with the individual texts which comprise the corpus itself. To achieve this, the data have been integrated into an existing online database of early Chinese texts, the Chinese Text Project (http://ctext.org/), which provides access to the data primarily in two ways. First, for any paragraph of any text in the database which contains any parallel to any other passage anywhere in the database, the system displays an icon indicating this to the left of the passage, which links to a page containing the corresponding results. The results page (Fig. 1) begins by displaying the selected paragraph itself; this is highlighted to indicate the locations of all parallels identified within the passage. Where multiple parallels overlap at the same location, this is indicated using successively deeper shades of red to give a heat-map effect highlighting precisely where text reuse occurs within the passage. When the mouse cursor is placed over a shaded region, the system lists the titles of the texts and chapters in which the matches were detected. Clicking on a highlighted segment moves down the page to the corresponding results, which are listed in turn below the original paragraph.

As the present study only considers reuse involving a high degree of formal similarity and not cases of subtler allusion, it has also been possible to highlight the specific differences between alternative versions of a text on the results page. Key applications of text reuse in textual criticism and interpretation rely upon the identification of local variations within generally agreeing sections of text, which can be automatically highlighted in a digital resource once the extents of the similar segments have been determined precisely. In the current implementation, the user nominates any one version of the passage (the default being the originally selected passage) by moving the mouse cursor over it; the system highlights in red within the selected version those characters that are absent or different in any other of the parallels, and in green within each other version those characters that disagree with the selected parallel version.

A second means of interacting with the data is provided through the search function, which makes it possible to search for parallels between arbitrary chapters, texts, or categories of text. The search results list each matched paragraph in the source domain together with an excerpt of the corresponding parallel in the target domain, highlighting in each case the parallel segment or segments (Fig. 2). The full results pages are also linked to from each paragraph, as is also the case during all other supported types of search (such as full-text search).

By integrating the parallel passage function into this existing database, it becomes much easier for researchers to ask 'what if' questions involving potential parallels in cases where consulting a standalone database of parallel passages might otherwise not have been considered worthwhile. Having already searched for a particular usage of a word or phrase in the database, additionally finding out whether or not there are interesting parallels to it in other texts now requires merely clicking a button and examining the visual summary.

In addition to these instance-level visualizations of text reuse, it is also possible to use the same results to visualize text reuse on a much larger scale. The method of quantifying reuse between arbitrary pairs of texts in the corpus described in the previous section makes possible visualization of text reuse on

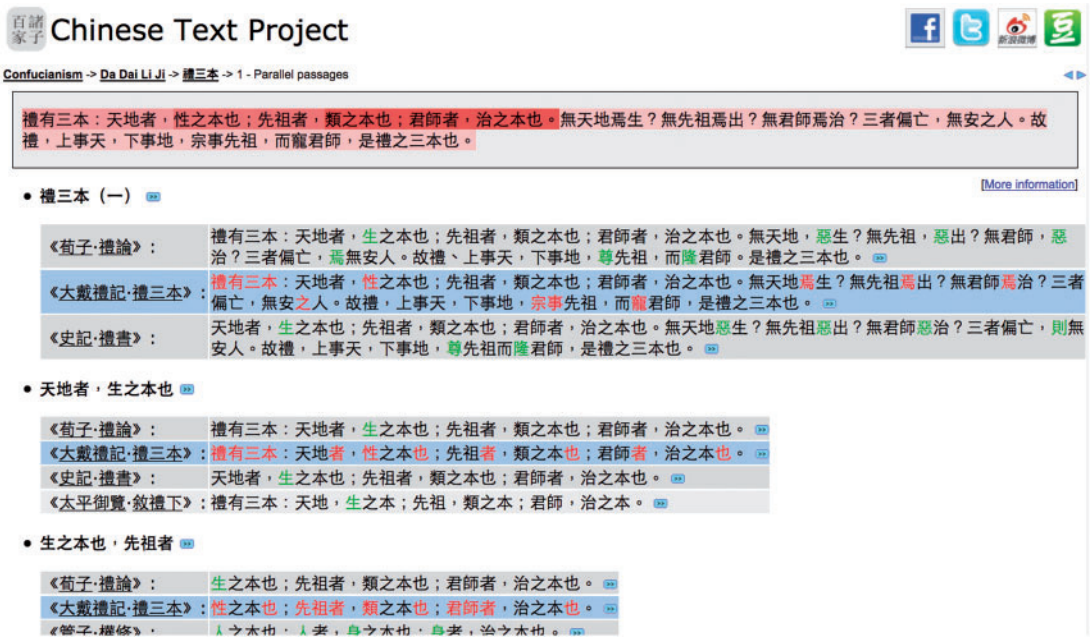

Fig. 1 Automatic highlighting of differences between parallels

a corpus scale as a network graph, in which each node represents a work from the corpus, and each edge a reuse relationship between two such works. Taking the edge weights as the fraction of two texts contained in parallels between them and applying a force-based network layout algorithm (Jacomy *et al.*, 2014) results in the network graph shown in Fig. 3, created using Gephi (Bastian *et al.*, 2009). As highlighted in the figure, many regions of this graph correspond closely to standard categorizations of these texts: for example, Daoist texts are localized in the bottom-right quadrant of the graph, while texts cataloguing historical events form a cluster at the top; several isolated communities correspond exactly to distinct categories of texts. As a result, community analysis of the same graph (using Gephi's implementation of an algorithm described in Blondel *et al.*, 2008) identifies communities which in several cases exactly match standard classifications (Fig. 4).

Several aspects of this agreement with traditional classifications of these texts are particularly surprising given that many of these—including the class of Daoist texts—are widely acknowledged to be classifications imposed upon these works by later historians, rather than affiliations recognized by the original authors of such texts (Smith, 2003). Community analysis of text reuse relationships provides useful objective data regarding one respect in which certain groups of texts can be said to be mutually related: on the basis of text reuse relationships alone, Daoist texts arguably form a natural class of texts.

6 Evaluation and Comparison with Previous Work

Previous work on text reuse in the pre-Qin and Han corpus has been most prominently conducted by the Institute of Chinese Studies (ICS) at the Chinese University of Hong Kong. The ICS has developed an annotated corpus of early Chinese texts that is made available by subscription to individuals and institutions under the title 'CHANT (CHinese Ancient Texts)'.[17] The same data were used in the development of a series of printed concordances to the texts, and also as the basis for a number of

Fig. 2 Search results—parallels to the first chapter of the Analects found anywhere within the Analects

printed publications focusing on parallel passages to selected texts in the early corpus as a whole. Citations and parallel passages were identified by a combination of automated methods and manual work; in the case of the *Parallel Passages Found in Pre-Qin and Han Texts* series most directly comparable to this study, this involved similarities being identified by automated comparison followed by manual editorial review.[18]

In comparison with the ICS publications, the results of the present study differ due to the diverging objectives and methodologies of the two projects. First, the ICS series occasionally include cases of more subtle allusion that are of interest to scholars but are not considered in this study due to the difficulty of reliably detecting them using automated software without producing significant numbers of false positives. Additionally, in many cases they also identify longer segments of text than are strictly parallel under the definition given here where these may be of interest to the reader in a printed reference work—for instance, they may include a complete anecdote even though only part of it contains a formal structural parallel, since the complete passage will often be of interest to the reader. In a born-digital resource this contextual information is less important, since such a resource can link directly to the location in the full text within which the similarity occurs. Secondly, the printed nature of the ICS series places practical limits on the number of parallels that can realistically be included. In a born-digital resource, there is little reason to select some similar citations of a passage as canonical to the exclusion of others[19]; in a printed resource, space considerations are an important factor. Even in cases of large numbers of parallels to a given segment of text, in a born-digital resource the more natural impulse is to store the full data and provide an accurate summary of it where appropriate rather than to omit portions of it unnecessarily. To give a common type of example, the ICS series typically cite the *Book of Poetry* when it is quoted (with or without citation) in a passage; the data presented in this study and integrated into the CTP instead cites *all* known parallel versions of the passage, including the *Book of Poetry* itself as well as all other texts which contain substantially the same quotation. This can provide useful information regarding whether the same passage is cited differently elsewhere, but would likely be cumbersome in a printed work because of the sheer number of citations available in many such cases.

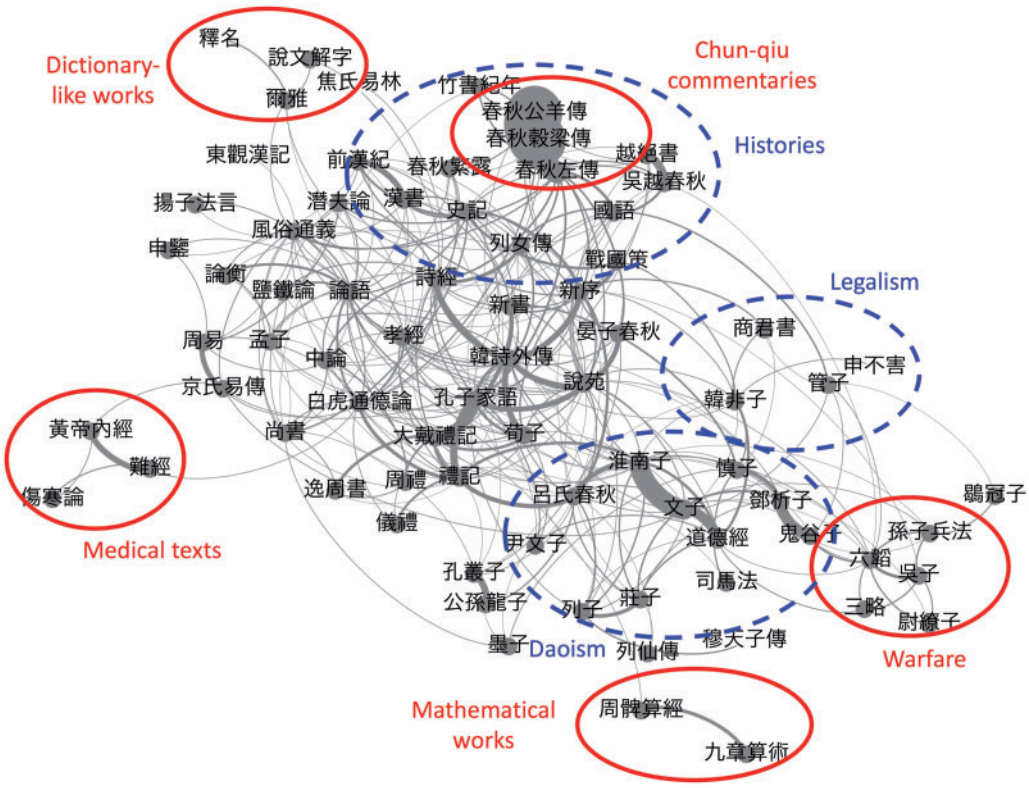

Fig. 3 Textual relationships in pre-Qin and Han texts visualized as a network graph. Communities corresponding to traditional classifications of texts by content are circled in solid lines (outlying groups) and dashed lines (less clearly defined areas of the graph)

To give a more concrete indication of the differences between the two sets of results, a manual comparison was conducted between a chapter of one ICS text (the first chapter of the *Xunzi*) and the equivalent data in the CTP.[20] To offer a meaningful comparison, the following procedure was used. First, in the CTP data, many additional citations were given where quotations from well-known or often-cited texts such as the *Book of Poetry* are made—not only is the original passage cited, but so are all other references to it. Since there are large numbers of these in the CTP data, and the ICS concordances do not include matches of this kind, they are excluded from the figures given here. Secondly, ICS matches may correspond to multiple CTP matches, for example, when a match consists of three sections in which the first and last are clearly parallel, but the middle sections otherwise differ significantly—in ICS this is often given as a single contiguous match, whereas in the CTP it may be given as two disjoint matches; here any number of CTP matches that correspond to a single ICS match are counted as a single match.

With these caveats, in total, ICS listed fifty-one matches for this text, whereas the CTP listed eighty-seven. Of these, seven ICS matches did not appear in the CTP, and forty CTP matches did not appear in ICS. Of the seven ICS matches not found in the CTP, two were due to the text in question being a reconstructed text not appearing in the CTP; four were due to low formal similarity as defined above, and one was a case of higher formal similarity that was not identified by the algorithm. Of the forty CTP matches not appearing in ICS (all of

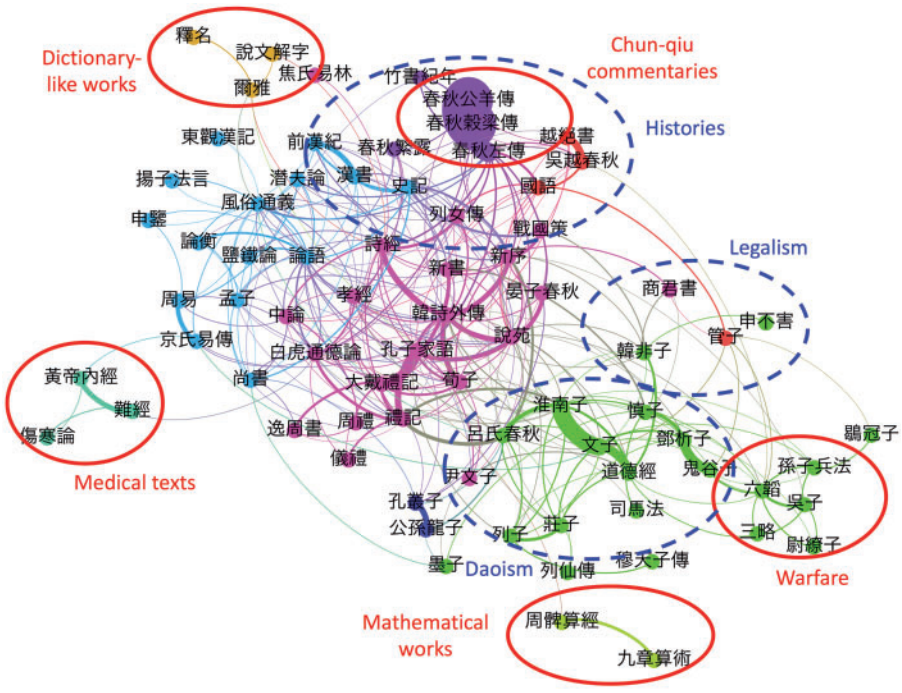

Fig. 4 Textual relationships in pre-Qin and Han texts visualized as a network graph, colorized using the results of community analysis

which met the formal criteria), twenty-three were clear cases of significant and meaningful parallels, and seventeen were less clear cases, the significance of which might be contested. The former included a number of curious omissions from the ICS data, for instance, the lack of inclusion of an unattributed reference to the *Analects* consisting of twelve identical characters in sequence[21]; perhaps the editors of the ICS series felt that this would be obvious to a reader familiar with such a well-known text—yet it seems a strange omission from what is clearly intended to be a comprehensive reference work.

Other examples of clear matches found by this study, but not present in ICS include:

Xunzi: 順風而呼，聲非加疾也，
Lüshi Chunqiu: 順風而呼，聲不加疾也;
Shiji: 比如順風而呼，聲非加疾，

Xunzi: 青取之於藍，而青於藍;
Shiji: 青采出於藍，而質青於藍

Several instances were notable in that the ICS listed one or more other parallels to the same part of a passage, yet failed to list certain instances in other texts that are clearly comparable. This points to their lack of inclusion being an accidental omission rather than a conscious choice made by the ICS editors deeming these passages not sufficiently similar.

Examples of borderline matches found by this study, but not in ICS include the following:

Xunzi: 南方有鳥焉，名曰蒙鳩，
Han Shi Wai Zhuan: 南方有鳥，名曰鶼，

Fragments such as these—each of the form 'in the south there is a bird called the …' might or might not share the required type of causal connection. It might be argued that this phraseology is causally related, in that (following some influential prior use) it came to be repeated and reused by other writers; but equally one could argue that tales of there being birds in the south having particular names are most naturally expressed in

classical Chinese with this construction, and thus it is merely a coincidental similarity and not a genuine parallel.

One case of significant formal similarity present in the ICS text was not identified:

> *Xunzi*: 君子之學也，入乎耳，著乎心，
> 布乎四體，形乎動靜。端而言，
> 蝡而動，一可以爲法則。小人之
> 學也，入乎耳，出乎口；口耳之間，
> 則四寸耳，曷足以美七尺之軀哉！
> *Han Shi Wai Zhuan:* 君子之聞道，
> 入之於耳，藏之於心，察之以仁，
> 守之以信，行之以義，出之以遜，
> 故人無不虛心而聽也。小人之聞道，
> 入之於耳， 出之於口，苟言而已，

In this case, there is a relatively clear parallel between the '君子之學也，入乎耳，著乎心，' and '君子之聞道，入之於耳，藏之於心，', and also and perhaps more strongly between '小人之學也，入乎耳，出乎口；'and '小人之聞道，入之於耳，出之於口，'. This hints at a type of similarity not directly accounted for in this study: similarity of grammatical structure. Nevertheless, the method described above would be expected to identify at least the latter pair if '乎' and '之於' were ascribed sufficiently strong similarity when defining the metric.[22]

Evaluating performance of parallel passage identification on a corpus of this type in terms of the standard metrics of precision, recall, and F-score is made difficult by the fact that recall can only be calculated when the total number of parallels in existence within the corpus is known. Since it is possible that there exist further parallels within the corpus of which we are currently unaware, we cannot be certain of this figure. To facilitate a comparison of the results of this study with the ICS results, here the set of all valid parallels identified by either the ICS or CTP method is taken as an approximation to the set of all parallels in existence. Furthermore, as there may be disagreement as to whether 'borderline matches' as described above ought to be considered legitimate matches, the results for each method are evaluated against two separate standards: first, a 'permissive' standard, upon which borderline parallels do count as legitimate matches; secondly, a 'strict' standard, on which borderline parallels do not count as legitimate matches. Evaluating both methods against this standard gives the results shown in Table 1. These results indicate that regardless of whether a permissive or strict standard regarding parallels is adopted, the automated CTP method results in a higher F-score than the manual ICS method, demonstrating that its results are competitive with—and perhaps superior to—the manually produced reference data.

Finally, while the present study has focused on the pre-Qin and Han corpus, the ICS has also published a series of books titled *Citations of Ancient Texts Found in the Leishu Compiled in the Tang and Song Dynasties*.[23] Each of these again focuses on a particular early text and systematically identifies passages parallel to it within various other texts. In the case of this series, a complete systematic comparison with the data presented here was not possible because this ICS series includes a greater range of *Leishu* than were available for use in this study. However, the methods used on the pre-Qin and Han corpus have also been applied successfully to several *Leishu* including the *Taiping Yulan* and *Yiwen Leiju*; informal comparison suggests the results largely agree with those of the ICS series in a similar way to the pre-Qin and Han corpus. An interesting anomaly observed was that the ICS *Leishu* series appear to identify *only* correctly cited text reuse, which means that parallel passages between a text and a particular *Leishu* may be identified only where the *Leishu* correctly cites by name that particular text. The results of the comparison presented here show that this is not always the case: in some cases a passage now belonging to an extant text A is cited as coming instead from text B, where B may or may not be an extant transmitted text, and even in some cases where B is an extent text, the extant version of B does not contain the text cited. Given the multiplicity of editions of early texts and their variation over time, this demonstrates that methods based solely upon similarity can produce interesting results that may be overlooked if correct *citation* is relied upon in addition to similarity.

Table 1 Precision and recall scores for parallel passage identification

Method (evaluation)	Precision	Recall	F score
ICS (permissive)	1.00	0.54	0.71
CTP (permissive)	1.00	0.94	0.97
ICS (strict)	0.92	0.65	0.76
CTP (strict)	0.80	0.99	0.88

7 Observed Substitutions Within Parallels

The large number of parallel passages present in the pre-Qin and Han corpus (and likely also present in other corpora of historical Chinese texts) together with the high level of accuracy of automated identification of such parallels means that a significant volume of machine-readable data describing parallel passages can be obtained programmatically. As shown in Section 5, these data can be used directly to aid in the close reading of texts, as well as be usefully aggregated to give a picture of text reuse at a corpus level—one form of distant reading. However, this data set also presents opportunities for other types of corpus-level study. One such example concerns the patterns of substitution observed within individual parallels. While parallels are identified on the basis of overall similarity, within this broader similarity there are very frequently local differences. These can be extracted automatically by aligning each pair of parallels and counting the observed substitutions between similar but non-identical versions of the same passage. For example, consider the following pair of parallels (differences highlighted using [. . .]):

天地者，[生]之本也；先祖者，類之本也；君師者，治之本也。無天地[惡]生？

天地者，[性]之本也；先祖者，類之本也；君師者，治之本也。無天地[焉]生？

Within this pair of parallels, two substitutions are observed: '生' for '性', and '惡' for '焉'.[24] Though little can be concluded from a single instance such as this, aggregating this type of data over hundreds of thousands of instances reveals trends that are clearly meaningful (Table 2).

In this table, the right-hand column lists the most frequent substitutions for the string in the left-hand column. In each of these cases, all of the most frequent terms are either synonymous or partially synonymous with the substituted term: '與', '余', '我', '賜', '分', and '吾' are all ways of saying the same thing as '予' (in particular contexts). The significance of this result is not that these relationships are surprising or unknown, but rather that a quite straightforward statistical procedure can produce such rich semantic data, and that these data can therefore be produced automatically for large numbers of strings. Most obviously, these relationships can be used in future studies of parallel passages, allusion, and other types of text reuse, by providing a much more accurate model of synonymy than was previously available. Many other possible avenues for future research also exist, such as investigating different patterns of substitution within subsets of the corpus, for example, the directionality of substitutions. Further study of these patterns is needed, and future studies may benefit from extracting this type of data from even larger corpora.

To evaluate to what extent the most frequently occurring substitutions actually correspond to substitution of synonymous terms, the 100 most frequently occurring pairs were manually evaluated. Of the 100 most frequently observed pairs, ninety-three were cases of clear complete or partial synonymy (examples include: 弗 and 不; 人 and 民; 曰 and 云).[25] The clearest non-synonymous pairs observed were three pairs of numerals: 二 and 三; 一 and 二; 三 and 四. Given the grammar of classical Chinese (e.g. lack of inflection due to plurality), such substitutions would be unlikely to render a sentence ungrammatical, and would be easily introduced accidentally due to the graphical similarity of the characters. The remaining four pairs were all pairs of commonly occurring grammatical particles, for which a strict determination of partial

Table 2 Frequent substitutions within parallels

String	Most frequent substitutions in decreasing order of frequency
予	與、余、我、賜、分、吾
殺	煞、弒、伐、誅、焚
無	毋、无、亡、不
曰	爲、云、稱、言
天下	四海、四海之內、萬民、國、民、海內
大王	陛下、君、公
寡人	吾、孤、我、不穀

synonymy is less easy to make; while not usually considered synonymous, in some contexts these particles can be substituted for one another without significantly altering sentence meaning. These pairs were: 者 and 也; 而 and 之; 也and 之; 者 and 則.

8 Conclusions and Future Work

This study has focused on identifying clear instances of text reuse in early Chinese texts and making the results of an automated digital analysis of text reuse freely available to the scholarly community. As such, to the extent that this study has been successful, there exists much potential for others to take the results of this enquiry further forward in interesting and valuable directions. For example, no attempt has been made in this study to explain the causal factors behind the observed text reuse; it is hoped that the availability and accessibility of the data will encourage others to further investigate these questions.[26] Is such text reuse primarily due to the copying of sources widely distributed in early China? How much of the apparent reuse is actually due to the attempts of Han dynasty editors to reconstruct earlier texts from the fragmentary evidence available to them? Did such editors introduce errors in texts due to differing interpretations of archaic characters? Can specific instances of text reuse help in the challenging task of dating early Chinese texts? Many interesting and extremely important research questions regarding text reuse in the early Chinese corpus remain unanswered, and there is consequently enormous potential for further research in this regard.

A further observation emerging from this study is that a key advantage of digital systems over traditional printed forms of research lies in the possibility of allowing scholars to work with *all* data of a particular kind, rather than a useful and important subset selected by experts. With the sheer volume of parallels spread throughout the early Chinese corpus, even the most diligent of conventional studies risks omitting information that may prove relevant to the particular research questions that *someone* else may be interested in. The ICS parallel passage and citation series of reference books provide excellent data on important early texts such as the *Xunzi*, assembling all parallels from throughout the pre-Qin and Han, or the post-Han *Leishu* corpus. Yet some data are difficult to represent in this type of printed format—for instance, parallels within and between various cited texts that now exist only within the *Leishu* corpus itself are likely to be interesting to some scholars even when they do not correspond to any extant part of the pre-Qin corpus—indeed, one important use of this corpus is in the quotations it provides from texts which have since been entirely lost to the transmitted tradition. Given that the ICS printed work describing post-Han *Leishu* parallels to the *Zhuangzi*—a text of around 65,000 characters—consumes over 200 pages, it seems likely that a similarly conceived printed work for the much larger *Taiping Yulan* encyclopaedia (over 3.7 million characters including commentary) would have to be of a formidable length. In a digital resource, it is possible to tailor the representation of data more closely to a specific research question under consideration, excluding data not on the basis that it is uninteresting *per se*, but on the basis that it is not relevant to the specific research question being asked. For instance, a scholar interested in a particular text no longer extant in the transmitted corpus might wish to examine all passages parallel to all passages which mention that text by name—in the printed realm, this might be achievable using a lengthy set of volumes documenting every parallel in every *Leishu* together with an extensive set of indexes, and some time spent navigating through the printed work. In the digital realm, however, answers to this specific question can be generated on demand from the full data

set in response to a user request. Future research may benefit from leveraging this aspect of digital resources to provide new ways of exploring text reuse in large corpora that further transcend limitations inherent in the printed medium.

References

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media*. Menlo Park, California: AAAI Press.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, **2008**: P10008.

Büchler, M., Geßner, A., Eckart, T., and Heyer, G. (2010). Unsupervised detection and visualisation of textual reuse on Ancient Greek texts. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*, **1**(2). https://letterpress.uchicago.edu/index.php/jdhcs/article/view/60/71/ (accessed 01 June 2017).

Clough, P., Gaizauskas, R., Piao, S. S. L. and Wilks, Y. (2002). METER: measuring text reuse. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Philadelphia, USA: Association for Computational Linguistics, pp. 152–9.

Coffee, N., Koenig, J. P., Poornima, S., Forstall, C. W., Ossewaarde, R., and Jacobson, S. L.(2013). The Tesserae project: intertextual analysis of Latin poetry. *Literary and Linguistic Computing*, **28**(2): 221–8.

Ho, C. W. (2002). CHANT (CHinese ANcient Texts): a comprehensive database of all ancient Chinese texts up to 600 AD. *Journal of Digital Information*, **3**(2). https://journals.tdl.org/jodi/index.php/jodi/article/-view/81/80/ (accessed 01 June 2017).

Ho, C. W, Chu, K. F., and Fan, S. P. (eds). (2005). *The Xunzi with Parallel Passages from Other Pre-Han and Han Texts*. Hong Kong: Chinese University Press.

Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One*, **9**(6): e98679.

Seo, J. and Croft, W. B. (2008). Local text reuse detection. *Proceedings of SIGIR '08*. New York, USA: Association for Computing Machinery, pp. 571–8.

Smith, D. A., Cordell, R., and Maddock Dillon, E. (2013). Infectious texts: modeling text reuse in nineteenth-century newspapers. In *Proceedings of the Workshop on Big Humanities*, IEEE Computer Society Press.

Smith, D. A., Cordell, R., and Stramp, N. (2014). Detecting and modeling local text reuse. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 183–92.

Smith, K. (2003). Sima Tan and the invention of daoism, "Legalism", et cetera. *Journal of Asian Studies*, **62**(1): 129–56.

Notes

1 The primary scope of the initial study was the corpus of pre-Qin and Han (to 220 A.D.) transmitted texts. This was chosen for a number of reasons: firstly, this corpus is most representative of the classical form of writing that later writers consciously attempted to imitate; secondly, the corpus composed of all such transmitted texts is of a relatively manageable size (less than 6 million words) for computational study of the entire corpus. While the statistics mentioned in this article refer to this corpus alone except where noted, the methodology has also been successfully extended to many later texts, and data for these has also been incorporated into the online database described below.

2 It became common practice in later dynasties for rulers to forbid the writing of certain characters due to their appearance in various names of the ruler or his relations. This often led to character substitutions in transcriptions of earlier texts.

3 This is particularly the case with classical Chinese given its grammatical flexibility.

4 The chosen corpus contains around 15,000 distinct characters.

5 To give one example that was identified by the method described in this study, the first four characters of a quotation from the *Book of Poetry* as given by the *Han Shu* and the *Qian Han Ji* share no identical characters, but have clearly related character forms: '歙歙訛訛' in the former versus '翕翕訾訾' in the latter.

6 Though this characterization is far from an adequate formal definition, it highlights two key features in the parallels concerned here: they must be formally similar in some way, and they must also be non-trivial. Thus '歙歙訛訛' and '翕翕訾訾' are parallels, while '此之謂也' and '此之謂也' are not, despite the former pair consisting entirely of different characters and the latter pair being entirely identical, since the

occurrence of the former independently in two texts given only our knowledge of the language itself would be highly improbable, whereas the latter would not.

7 Examples include Büchler *et al.*, 2010, Clough *et al.*, 2002, Coffee *et al.*, 2013, Seo and Croft, 2008, and Smith *et al.*, 2013, 2014.

8 'Naïve string matching' here refers to any algorithm that identifies pairs of strings containing identical sequences of characters (ignoring any differences in punctuation) of a certain minimum length without regard to word segmentation, character similarity, or any other factors. A key advantage of such matching is that as it is an established and widely researched problem in computer science, efficient and scalable implementations of it are available. Characters rather than words are a natural choice for processing classical Chinese due to the nature of the language (high per-character semantic content—most words consist of a single character) as well as the technical difficulty of tokenization (classical Chinese is written with no explicit delimitation between words and is difficult to reliably tokenize).

9 Removing the radical component was chosen as an initial normalization because it captures many similarities between characters that were originally written without any radical, and subsequently had a radical added to distinguish between different usages. In many cases radicals were added by editors as an unavoidable part of interpreting a text written in obsolete characters into the characters of the day, but different editors faced with the same manuscript might choose to add different radicals to such a character depending upon their interpretation of the text.

10 By way of example, removing the radicals from the pair of quotations from the *Book of Poetry* cited in footnote 5 gives the normalized string '翕翕此此' in both cases, which was thus detected as a match by naïve string matching since the normalized strings are identical.

11 An important feature of character similarity so defined is that it is not transitive: if character A is similar to B, and B to C, it is not necessarily the case that A is similar to C. For example, of the three characters 註, 貯, and 貯, all of which share a common Mandarin reading 'zhu4', 註 is similar in this sense to 貯, and 貯 to 貯, but 註 is not similar to 貯. Were this relation transitive, it could be incorporated into the initial normalization step, but forcing it to be transitive would greatly reduce its specificity and greatly decrease the ratio of useful matches identified by the naïve string matching process.

12 http://ctext.org/tools/parallel-passages

13 The algorithm considers multiple adjacent phrases around the located match, thus it is possible for an entirely dissimilar pair of phrases to be incorporated within a pair of parallels, provided that this dissimilar pair of phrases is itself surrounded on both sides by text that is similar.

14 These groups can overlap, for instance, if texts A and B share a long parallel 'DEFGHIJK', and text C contains only 'HIJK', this will result in two parallel groups: a shorter one for the 'HIJK' with all three texts as members, and a longer one corresponding to the full string whose members are only A and B.

15 This equation can be read as 'the fraction of A which belongs to some parallel with B'; thus the parameters in the function are not interchangeable, and in general $par(A, B) \neq par(B, A)$.

16 In addition to being widely cited, the *Book of Poetry* also incorporates extensive internal reuse, for instance where fragments are intentionally repeated verbatim or with minor modifications in several stanzas of the same poem.

17 http://www.chant.org/, described in Ho, 2002.

18 http://www.cuhk.edu.hk/ics/rccat/en/series2.html

19 This is particularly so in the case of early Chinese texts, as there is often little scholarly consensus on the dating of texts and even their chronological ordering; it is thus a non-trivial task to determine which if any of a group of parallel passages ought to be considered earliest or most canonical.

20 Ho *et al.* (2005), and http://ctext.org/xunzi

21 古之學者爲己, 今之學者爲人。

22 A method by which a list of such equivalences can be obtained automatically is described in Section 7.

23 '*Leishu*' refers to a class of Chinese texts typically consisting of quotations cited from earlier canonical texts arranged by topic.

24 The examples given here ignore directionality, which can also be considered as a factor: if we know that one text is earlier than another, we may want to consider such substitutions as being directional in nature, for example, because we may want to model how later authors modified earlier material.

25 Partial synonymy was determined for non-trivial cases by using definitions given in the *Hanyu Da Zidian* dictionary. Characters sharing at least one synonymous usage were considered partially synonymous.

26 In addition to the publicly searchable user interface for navigating the results, raw statistical data created as part of this study will be made available here: http://ctext.org/tools/parallel-passages#analysis