The Journal of Asian Studies Vol. 77, No. 1 (February) 2018: 19–57.
© The Association for Asian Studies, Inc., 2018 doi:10.1017/S0021911817000973

Modeling the Contested Relationship between *Analects*, *Mencius*, and *Xunzi*: Preliminary Evidence from a Machine-Learning Approach

RYAN NICHOLS, EDWARD SLINGERLAND, KRISTOFFER NIELBO, UFFE BERGETON, CARSON LOGAN, AND SCOTT KLEINMAN

This article presents preliminary findings from a multi-year, multi-disciplinary text analysis project using an ancient and medieval Chinese corpus of over five million characters in works that date from the earliest received texts to the Song dynasty. It describes "distant reading" methods in the humanities and the authors' corpus; introduces topic-modeling procedures; answers questions about the authors' data; discusses complementary relationships between machine learning and human expertise; explains topics represented in Analects, Mencius, *and* Xunzi *that set each of those texts apart from the other two; and explains topics that intersect all three texts. The authors' results confirm many scholarly opinions derived from close-reading methods, suggest a reappraisal of* Xunzi's *shared semantic content with* Analects, *and yield several actionable research questions for traditional scholarship. The aim of this article is to initiate a new conversation about implications of machine learning for the study of Asian texts.*

Mencius has been considered the philosophical heir to the moral philosophy and theory of human nature presented in *Analects*. *Analects* contains sayings and ideas attributed to Confucius (551–479 BCE) and his followers. Mencius (early fourth c. BCE – late fourth c. BCE) and Xunzi (c. 310 – c. 235 BCE / c. 314 – c. 217 BCE) both explicitly stated that they followed the teachings of Confucius. However, recent scholars argue that *Xunzi* is closer in content to *Analects* than *Mencius*. This article contributes to the debate by introducing a machine-learning approach to supplement traditional modes of inquiry. We make use of a technique known as *topic modeling* to provide a new perspective in ongoing conversations about Confucianism and the relationships between some of the most important source texts in early Chinese thought.

Ryan Nichols (rnichols@fullerton.edu) is Associate Professor in the Department of Philosophy at California State University, Fullerton, and Research Affiliate of the University of British Columbia's Centre for Human Evolution, Cognition, and Culture. Edward Slingerland (edward.slingerland@ubc.ca) is Professor in the Department of Asian Studies, University of British Columbia. Kristoffer Nielbo (kln@cas.au.dk) is History Researcher in the datakube, University of Southern Denmark, and Researcher at the Interacting Minds Centre, Aarhus University. Uffe Bergeton (bergeton@email.unc.edu) is Assistant Professor in the Department of Asian Studies, University of North Carolina. Carson Logan (carsonklogan@gmail.com) is Developer at Tuangru in Vancouver, Canada. Scott Kleinman (scott.kleinman@csun.edu) is Professor in the Department of English, California State University, Northridge.

Topic modeling has already become a complementary source of knowledge and information for scholars across the humanities who are accustomed to using close-reading methods for the extraction of meaning from texts. Topic models identify groups of words (called *topics*) that are statistically likely to co-occur in a text or corpus. Insofar as traditional studies prompt the scholar to bring ideas, themes, and assumptions *to* texts, topic modeling reverses this process. In this way, topic modeling supplements, confirms, or, in some cases, challenges conclusions from close-reading traditions. We understand our effort here as preliminary and one of the first of its kind. Nonetheless we aspire to combine knowledge of the contents of topics, contents of texts, and expertise in classical Chinese language, culture, and thought, and so bring a pioneering navigational tool to the exploration of historically important Chinese documents of deep and wide interest to a readership across Asian studies, philosophy, literature, religion, and more.

Below we explain what topic modeling is, introduce our corpus of ancient and medieval Chinese texts, and discuss the preliminary results of our topic-modeling research as applied to questions about the relationships between *Mencius* and *Xunzi* and *Analects*. As an authorship team composed of experts in pre-Qin Chinese religion and philosophy, Warring States Chinese language and linguistics, and humanities computing, we have used and will continue to use traditional close-reading techniques for understanding Chinese thought. Yet advocates of close-reading techniques are reluctant to question dubious hermeneutic assumptions and break out of tunneled interpretations (see Nichols 2015). So machine learning provides a valuable supplement to traditional methods. We treat the results that follow as the first machine-learning steps in a wider interdisciplinary effort to gain deep knowledge of the meaning of Chinese texts. Our primary goal is to present information capable of starting a new, exciting thread in a millennia-long conversation about the interpretation of a few of the world's most influential texts.

Mixed Methods: Machine Learning + Experimental Text Analysis + Close Reading

Understanding the literary, intellectual, and cultural history of ancient and medieval Chinese literature presents the traditional scholar with imposing challenges. The authenticity, authorship, and dates of composition of texts are often either unknown or widely contested (Loewe 1993). Furthermore, since many early Chinese texts are compilations of texts composed by different authors at different times put together by later editors, just what qualifies as a single text is debatable (Boltz 2007). Except for recently excavated manuscripts, most extant early Chinese texts are the products of scribal copying, censorship, redaction, loss of books, and other forms of textual corruption. These documents rarely received study independent of traditional commentary. On top of these concerns, the sheer size and complexity of the ancient and medieval Chinese corpus prevents any one individual from mastering all its texts.

To situate our method, we will distinguish between three approaches to texts. The first is *distant reading*, increasingly popular across the humanities due to contexts in which the size and complexity of a corpus precludes its mastery. Coined by Franco Moretti (2000), the portmanteau "distant reading" refers to a method using computational tools to analyze texts and overcome these challenges. This makes distant reading a form of machine learning that leverages the power of programming to address canonical research questions in the humanities. Distant-reading and machine-learning methods compute relationships between texts, terms, and topics via mathematical algorithms rather than expert judgments. Distant-reading methods differ from *experimental text analysis*. In experimental text analysis, scholars code terms, classify synonyms, track associations between texts, or examine the contexts of keywords. These procedures occur in the context of the scientific method, but without the help of machine-learning algorithms that find patterns in texts. Formalizing interpretive procedures and testing specific hypotheses means that experimental text analysis takes a huge step toward the scientific study of literature. For example, Slingerland and Chudek (2011) used expert coders to track changes to the meaning of *xin* 心 or "heart-mind" in pre-Qin texts; Clark and Winslett (2011) used one expert coder to determine whether terms for high gods and deities co-occurred with terms for morality. Experimental text analysis formalizes scholars' interpretations of parts of texts, which separates this method from a third, traditional *close reading*. Close readings of texts by experts produce unparalleled insights into the meaning, subtlety, beauty, and power of historical texts in a way that neither of the other methods can hope to replicate.

Each method has its challenges. But suppose that our goal is to infer the meaning of a text from what it says? Here a *mixed method* combining elements of all three approaches stands head and shoulders above the individual methods as the most promising way forward. Experimental text analytics exclusively uses human coders to determine meanings from words and sentences. This may allow flexibility, but studies that rely on coders are subject to human error and bias. Traditional close reading faces a number of challenges in determining the meanings of texts from their sentences. These include in-group biases, fallacies, self-deception, cognitive limits (when corpora are large), and social pressures.[1] Researchers have argued that close-reading methods also make very little cumulative progress in the understanding of a text, given that continual interpretive disagreement is a feature of the humanities (Dietrich 2011). Yet machine learning and distant reading may provide a means of side-stepping some forms of human error and bias. They are not free of bias (Goldstone and Underwood 2014, 364); they cannot infer meaning from words without the help of area expertise earned through years of close reading, and the form of results in a topic model often means the data are difficult to interpret. Yet no method is better able to identify patterns that are often hidden from the view of scholars not because of scholarly bias but because these patterns only appear at scale, or involve word usage that does not typically catch the human eye.

In the present case, we have designed our study as taking the best from all three approaches. We start with a robust *distant-reading and machine-learning* method. This provides us data to work with. How do we interpret the data? Three of us are experts in ancient Chinese thought, so we interpret the topic models in light of many *close readings* of relevant texts. How then do we control for our own biases and foibles? Since we did not trust ourselves to deliver error-free interpretations, we enlisted about sixty other experts in ancient Chinese thought to independently interpret our topic models. This

[1] For evidence of such problems as they arise in philosophy, see Draper and Nichols (2013).

experimental text analysis work provides a validation check on our interpretations of the data.

We feature the machine-learning component of our method because no one has used it on a corpus as we have. This method converts words into data and uses algorithms to find patterns among those words and their relationships. At the root of this process is computation. To get a sense for what computation involves, consider the following example. If we are given an unordered, random list of whole numbers, we might map the variable *larger than* onto *integers* in order to compute the largest number in the list. This mapping is algorithmic. An algorithm is a bit-by-bit recipe for implementing a computation. Familiar processes like tying one's shoes and baking a cake are processes that can be described algorithmically. In these cases, the user of the algorithm identifies the data to which the algorithm will apply (the ingredients), writes a set of instructions that structure the iteration of a step-by-step process (the recipe), and has a specific outcome in mind (the cake). In other cases, algorithms are exploratory and used without this sort of supervision. For example, we might have no prior idea about how many prime numbers there are between 2,576 and 6,509,322. Despite not knowing the outcome in advance, we can still write an algorithm to give us this information.

This leads to three takeaway points for what follows. First, at the most basic level, our modeling activity represents a simple algorithmic mapping of the distance between character frequencies across sentences, chapters, and texts within our corpus. Second, just like the prime number example, we undertake this modeling activity without knowing what relationships between characters we will uncover. This is often described as an "unsupervised" analysis. Third, algorithms operate on diverse types of data, and the outcomes of computations are purely mathematical constructions. In other words, the *meanings* of the units of data—physical movements of an assembly robot, changes in velocity in an orbital reentry, Chinese characters—are irrelevant to the *computation*. Understanding the meanings of our data is left for experts in the area of inquiry.

Topic modeling has supplemented and invigorated a number of other humanities research areas, including history, philosophy, journalism, and literary studies. Literary and historical studies have benefited the most from topic modeling, as is apparent in the work and influence of Matthew Jockers and his remarkable study of nineteenth-century novels in the United Kingdom, Ireland, and America (Jockers 2013). He explores major themes and, having created sub-corpora at the level of national literature, often contrasts emergent themes in national corpora. For example, landlord-tenant relations become a significant topic in Irish novels while race becomes a significant topic in American novels. In history, Robert Nelson topic modeled the archives of the *Richmond Daily Dispatch* newspaper from November 1860 to December 1865 during the American Civil War. Nelson tracked changes in relationships between words about the Confederate military draft, fatalities, and patriotism by using the algorithm to compute a mapping of words to words and words to dates. Combining knowledge of dates of movements of the Union army, Nelson found that ads for fugitive slaves spiked on the two occasions when the Union army came closest to Richmond. These results work in harmony with research by historians by providing correlational evidence for a theory: a minority of civil war historians have argued for greater appreciation of the role of the Union army in the destabilization of slavery in the Confederate south, independent of the Emancipation Proclamation (Nelson 2015).

Asian studies has not yet caught up with other humanities areas in the use of topic modeling, though this may be changing (Chen et al. 2014; Hou and Frank 2015). We hope that these results—and others' to follow—can inspire Asian studies researchers with concrete questions, or even testable hypotheses motivated by secondary literature. For example, using metadata about dates of texts, one might test a hypothesis that in the later Han dynasty topics associated with trade and commerce peak; or one might predict that the *Yìjīng* 易經 has had much more influence on Daoism than on Confucianism, and test that hypothesis by examining the relative weights of topics loading heaviest in *Yìjīng* with those loading heaviest in the set of Daoist or Confucian texts; or using metadata about dates, one might explore (rather than test) whether opinions in secondary literature about the relative dates of chapters of *Shangshu* can be confirmed on the basis of their linguistic similarity.[2]

Here we apply a topic-modeling algorithm to a corpus of 5.74 million characters across ninety-six ancient and medieval Chinese works, including many of the most important texts in the tradition.[3] We selected this corpus because of the scope of the texts it includes, its accessibility, its familiarity, and its temporal breadth. The corpus spans several eras of historical Chinese literature. It includes the pre–Warring States *Book of Poetry* (*Shījīng* 詩經), the Warring States *Dào Dé Jīng* 道德經, the short treatise on philosophy of language *Gōngsūnlóngzi* 公孫龍子, the lengthy history text *Hàn Shū* 漢書, Han medical texts like *Huángdì Nèijīng* 黃帝内經, and pre-Qin encyclopedic texts like *Lǚ Shì Chūnqiū* 呂氏春秋. (See appendix 1, "Texts, Genres, and Dates," for the complete list of texts and table 1 for era classifications of the corpus.)

Topic Modeling

We do not duplicate the comprehensive and friendly introductions to topic modeling for humanists already written (see Blei 2012b; Mohr and Bogdanov 2013; Underwood 2012a; Weingart 2012). Yet we see broad benefits in directly providing researchers across subfields of Asian studies with hands-on knowledge of the topic-modeling process, since many scholars of texts of any kind will soon benefit from—or need to acquire—the ability to interpret topic models in their research area.

Topic modeling was developed for search and retrieval in large collections of text-heavy data, but topic models efficiently sum, visualize, and explore the semantics of any kind of text corpus. Words are assigned to topics based on their tendency to co-occur in texts with other terms found in the topic.[4] For topic modeling we use an

[2] We are working on this last one.

[3] The texts in this corpus were processed with generous permission of Dr. Donald Sturgeon from the Chinese Text Project (http://ctext.org/).

[4] In order not to interrupt the article's narrative with technical detail of little interest to the majority of readers, we use footnotes to present more formal or technical features behind our study. A topic is a mapping or a probability distribution over terms. "Terms" refers to countable linguistic forms such as words or Chinese characters. A number of algorithms and tools can be used to calculate such distributions. We use a sampling-based algorithm for latent Dirichlet allocation known as "LDA." LDA is a generative probabilistic model that extracts a set of latent variables (i.e., topics) in large collections of documents. This is implemented in a software environment called

Table 1. Corpus composition by era.

Era	*Dates*	*Character Count*	*Percent of Corpus*
Pre–Warring States	Before 480 BCE	30,447	0.50
Warring States	479–222 BCE	1,424,080	24.80
Han	221 BCE–220 CE	3,501,256	61.00
Post-Han to Song	221 CE–1044 CE	786,546	13.70
Totals		5,742,329	1.00

algorithm that maps, that is, computes probabilistic values for, the relations amongst all the terms in the corpus to all the other terms in the corpus. Through this process, the model extracts topics in large collections of documents. The model is *probabilistic* because the topics consist of words that have a high probability of occurring together in documents (Blei, Ng, and Jordan 2003). The model is *generative* because topics are formulated from latent relationships amongst terms in documents. Unlike an algorithm for tying one's shoes, but like an algorithm for discovering a set of unknown prime numbers, our model works without supervision. This means that the algorithm discovers the topics without its being fed prior knowledge about genre or date or any other information about the texts. The upshot is that before seeing the results, we do not know what the topics will be or which topic will have the biggest representation in the corpus.

Topic models produce several different types of data, including word weights, corpus weights, and text weights. In practice, many digital humanities papers using topic modeling neglect much of these data in preference for focusing on the resulting topics. Since we attempt to exploit the full range of these data to address our research question about the relationships between *Analects*, *Mencius*, and *Xunzi*, we now introduce these types of data. The number of times the topic-modeling algorithm has assigned a given term to the topic determines its *word weight*. "Weight" in this context refers to the relative size of the contribution that a word makes in a topic. The centrality of the word (or character or term or glyph) to the topic can be determined by rank-ordering the word weights. Topics are customarily split into short lists of the top-ranked words, which we refer to as the topic's *keywords*. Keywords serve as a metonym for the long list of words in the entire topic. The plot for Topic 29, included in figure 1, shows the character *tiān* 天 (heaven or God) as having the largest weight of any keyword in that topic. In this context, the large word weight of *tiān* results from its nearly 12,000 occurrences in Topic 29.

Next is *corpus weight*. Table 2 and figure 2 illustrate a topic's weight in the corpus, or *corpus weight*. A topic's corpus weight is the ratio of the sum of words in a topic over the total number of words in the corpus. Corpus weights are not standardized (ours sum to 14.9) because they are based on the total occurrence of words within each topic. Instead of representing the weights of individual characters, table 2 depicts keywords (in the

MALLET, an acronym for "MAchine Learning for LanguagE Toolkit." MALLET has proved effective in modeling humanities data (McCallum 2002). The LDA topic model employed in this study uses a Gibbs sampler, which is a Markov chain Monte Carlo algorithm for obtaining observations from the multivariate Dirichlet distribution. For the underlying mathematics, see Blei (2012a).

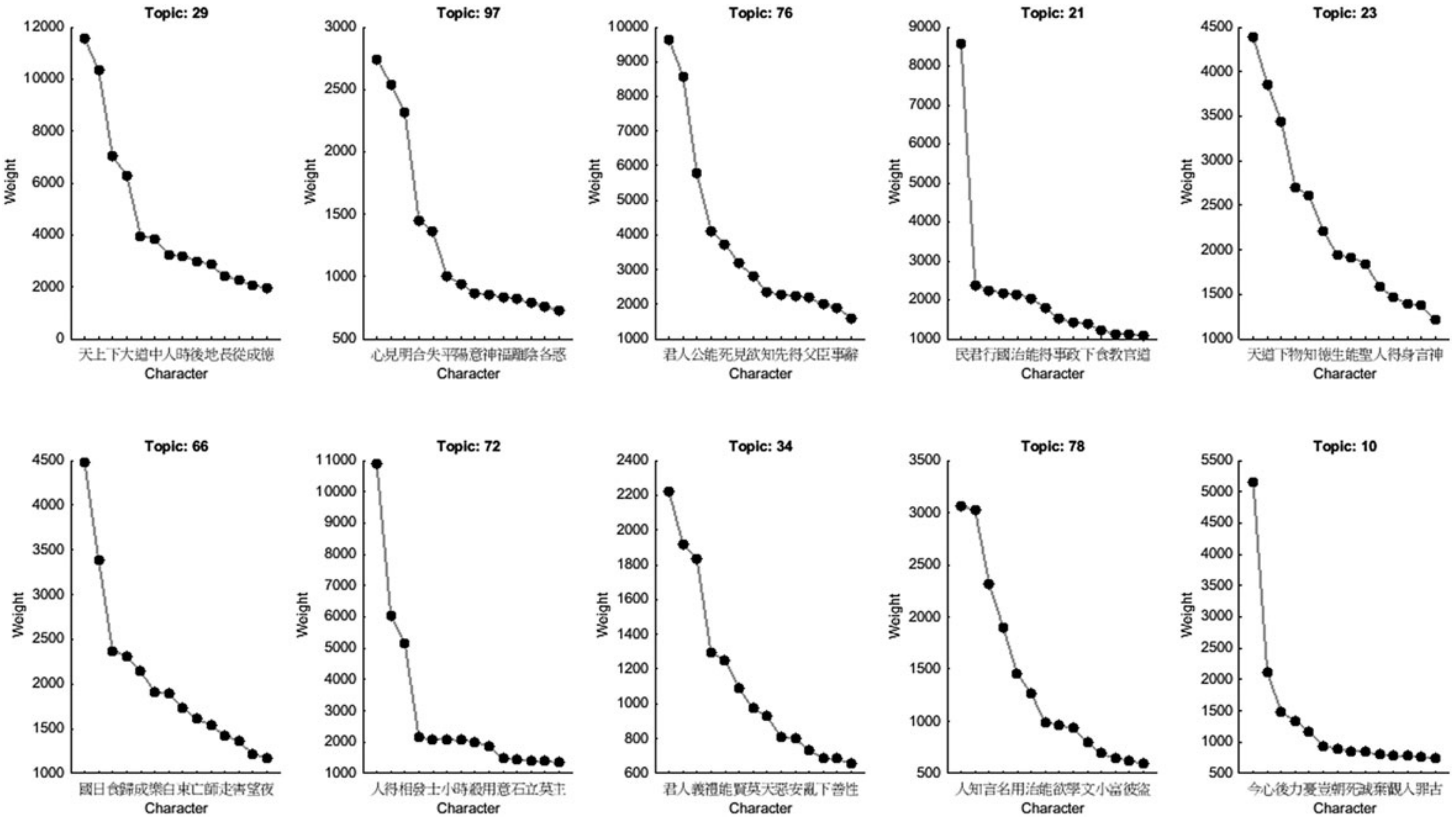

Figure 1. Keyword loading in highest weighted ten topics in our corpus. Individual plots in this figure represent the distribution of heaviest keywords within the target topic. Characters along the horizontal axis represent central characters in the target topic, with the most central character nearest the vertical axis. The numbers along the vertical axis represent the number of occurrences of each character. Typically keyword weights approximate a discrete power law distribution, with the weights being inversely proportionate to the keyword's rank for any given topic. This describes Topic 21 because *mín* 民 (people) has nearly four times the word weight as 21's second-ranked character *jūn* 君 (prince). Contrast Topic 23, which is almost linearly distributed.

Table 2. Highest weighted ten topics in the corpus.

Topic #	*Corpus Weight*	*Label*	*Topic Keywords in Descending Order of Weight*
29	0.600	Heaven, Earth, Man, & The Way	天 上 下 大 道 中 人 時 後 地 長 從 成 德
97	0.475	Cognition, Perception, & Fortune	心 見 明 合 失 平 陽 意 神 福 離 陰 各 惑
76	0.471	Rulers, Ability, Knowledge,	君 人 公 能 死 見 欲 知 先 得 父 臣 事 辭
21	0.459	Political & Social Order	民 君 行 國 治 能 得 事 政 下 食 教 官 道
23	0.452	Moral-Cosmic Attunement	天 道 下 物 知 德 生 能 聖 人 得 身 言 神
66	0.446	Ritual Sacrifice	國 日 食 歸 成 樂 白 東 亡 師 走 害 望 夜
72	0.428	Political Roles, Political Affairs	人 得 相 發 士 小 時 殺 用 意 石 立 莫 主
34	0.373	Ethical Rulership	君 人 義 禮 能 賢 莫 天 惡 安 亂 下 善 性
78	0.364	Learning & Governance	人 知 言 名 用 治 能 欲 學 文 小 富 彼 盜
10	0.348	Cognition & Planning	今 心 後 力 憂 豈 朝 死 誠 棄 觀 入 罪 古

order of their word weight within the topic) along with the corpus weight of the topic. Table 2 represents our topic model's findings as to the ten most weighty themes in ancient and medieval Chinese writing. Figure 2 visualizes the corpus weights of all 100 topics in our model.

The third and final type of data produced by a topic model is the *text weight*. This term refers to the proportion of a text's vocabulary that is assigned to a given topic, which represents how saturated a text is by a topic. Text weights are normalized and sum to 1. In each text in the corpus, some of the 100 total topics will have greater representation than others. For example, in *Xunzi* experts would expect that topics having to

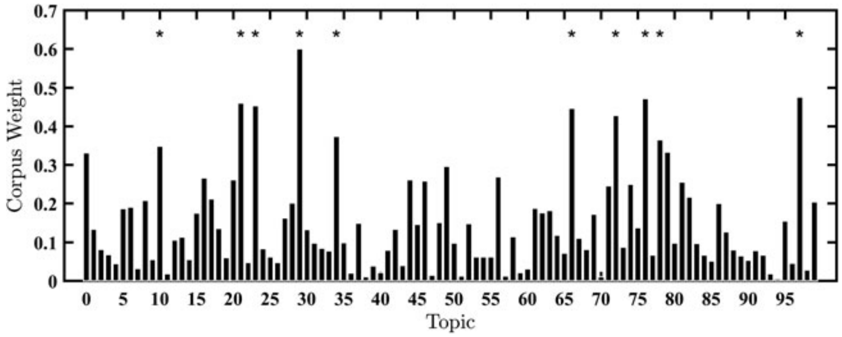

Figure 2. Corpus weights for Topics 0–99.[5]

[5] Corpus weights are calculated from Dirichlet distributions that serve as hidden or latent variables responsible for the allocation of words to topics (see Blei 2012a, n4; 2012b, 79–81). Since MALLET outputs the Dirichlet parameter, which is "roughly proportional to the overall portion of the collection assigned to a given topic" (McCallum 2002), we use this number as a measure of corpus weight.

do with ritual matters will have bigger representation, and so larger text weights, as compared to *Mencius*.

Let us illustrate text weight, word weight, and corpus weight. Consider the topic that has the heaviest corpus weight in *Xunzi*, Topic 34, which we call "Ethical Rulership." First, Topic 34 has a text weight of 0.256 in *Xunzi*. In contrast, its text weight in *Analects* is only 0.043 and half of that in *Mencius* at 0.023. This alone represents a discovery in terms of our research question, since the distribution of Topic 34 into *Xunzi* is six times greater than its distribution in *Analects* and eleven times greater than in *Mencius*. This warrants a practical inference for scholars of ancient Chinese documents, namely, Topic 34 sets *Xunzi* apart from *Mencius*.

To understand the significance of this discovery, we turn to look at the characters in Topic 34 and information about them. (See table 3 for the keywords and word weights of Topic 34.) To avoid misapprehending topic model results, it is important to understand information about characters that make up topics. Person (*rén* 人) has 219 occurrences in *Analects*, 611 in *Mencius*, and 1,241 in *Xunzi*. Frequencies of terms are often relevant to answer research questions, but for purposes of comparison the use of frequencies neglects a couple of issues. *Mencius* is 2.3 times the size of *Analects*, and *Xunzi* is 5.3 times the size of *Analects*, facts that hamper one's ability to interpret semantic importance from character frequencies alone. Zipf's law has the same effect (Zipf 1949). Zipf's law states that in any given text in a natural language, a word's frequency is inversely proportional to its rank in the corpus. This means that, in a given text, the most frequent word is typically twice as frequent as the second most frequent word, three times as frequent as the third most frequent word, and so forth. A better way of understanding the importance of a character *in a set of texts* is to examine its rank within and across the texts, and to look at its rate of occurrence per 1,000 characters. Raw frequencies do not disclose that, once common stopwords are removed (see below), *rén* is the most frequent character in each of *Analects*, *Mencius*, and *Xunzi*, and has a rate of occurrence per 1,000 characters of 28.7, 34.6, and 30.6, respectively. To understand the importance of a character *in a topic* rather than in a text, however, we must consult its word weight (see table 3, column 3). By doing so, for example, we see that with a word weight of 0.037, nobleman (*jūn* 君, occurring for example in *jūnzǐ* 君子) is three times as important to Topic 34 as is peace (*ān* 安). The algorithmic mapping at the heart of topic modeling allows us to go beyond information about simple frequencies to discover much more robust and reliable relationships between terms and texts.

Corpus weight is not a helpful statistic unless a topic's corpus weight is put in comparison with others. The corpus weight of Topic 34 is 0.375. Of 100 topics in our model, this is a very large corpus weight, ranking Topic 34's corpus weight eighth of 100 topics (see table 3). This fact justifies the inference that, despite the fact that it was not nearly as representative in *Mencius*, *Xunzi*'s particular focus on ethical rulership is well-distributed in the corpus.[6]

While topic models are a form of unsupervised machine learning, human decisions play some role in what topics are generated. At many junctures, we pooled our expertise

[6] We thank an anonymous reviewer for several comments that led to substantial improvements in our presentation of these types of data throughout the article.

Table 3. Word weight & character-level data for Topic 34.

			Analects			*Mengzi*			*Xunzi*		
Term	*English*	*Word Weight*	*Occurences*	*Per 1,000 Characters*	*Term Rank*	*Occurences*	*Per 1,000 Characters*	*Term Rank*	*Occurences*	*Per 1,000 Characters*	*Term Rank*
君	nobleman	0.037	160	21.0	2	253	14.3	5	547	13.5	4
人	person	0.032	219	28.7	1	611	34.6	1	1241	30.6	1
義	righteousness	0.031	24	3.1	63	107	6.1	25	315	7.8	13
禮	ritual	0.022	75	9.8	9	68	3.8	40	343	8.5	10
能	able	0.021	69	9.0	12	135	7.6	12	519	12.8	5
賢	virtuous	0.018	25	3.3	60	74	4.2	37	152	3.7	44
莫	none, do not	0.016	18	2.4	89	58	3.3	53	257	6.3	18
天	day, heaven	0.016	49	6.4	23	293	16.6	4	598	14.7	3
惡	evil	0.014	39	5.1	38	80	4.5	36	190	4.7	30
安	peace	0.013	17	2.2	94	23	1.3	167	190	4.7	29

in programming, in preprocessing, and in classical Chinese language and thought to make decisions that influence the quality of the topics generated by the algorithm. The effects of the subset of decisions that are made prior to the application of a topic-modeling algorithm to a corpus is referred to as *preprocessing*. Due to the nature of our texts in Chinese, we removed punctuation, tokenized, and applied a stopword list. Classical Chinese manuscripts do not include much punctuation at all, but the Chinese Text Project (CTP) texts include punctuation. Therefore we removed all but sentence-ending punctuation from the corpus. Tokenization refers to the management of word boundaries. We used a procedure that rendered each character separated by spaces before and after from every other character. This allowed us to treat each character as a unit of semantic meaning.

In a second preprocessing step, we used experts' knowledge to generate a stopword list. A stopword is a high-frequency word that tends to be highly ranked in topics but that also tends to make the topics less valuable for interpretation. Stopwords typically consist of common function words. Applying a stopword list means removing those common characters from the corpus prior to analysis. Examples of terms on our stopword list are *zhī* 之, a grammatical term used as a pronoun and subordination particle, and *yě* 也, a grammatical particle used to indicate noun predication (among other things). These and other stopwords were removed because during a series of pilot studies those words tended to blur the semantic coherence of topics. Applying a stopword list is standard procedure in topic modeling. We provide a full list of stopwords used in this study in appendix 2, "Stopwords."

In a third preprocessing step, we encountered problems with the software to implement our topic-modeling algorithm because that software was not designed to handle all the Chinese characters in our corpus. We scripted a method of encoding our input and decoding our output that allowed us to work around that problem.[7] Following common practice using LDA on texts, we did not chunk or split the texts in our corpus for analysis.

Moving from preprocessing to processing, the most important decision is the number of topics chosen to model. Too few topics may combine semantically unrelated material into so-called *chimera* topics; too many may cause related material to split into separate topics, redundancy between topics, or accumulation of irrelevant "junk" topics (Schmidt 2012). Topic quality is typically determined by semantic coherence of the keywords in the topic. Although significant strides have been taken in algorithmic determination of the ideal number of topics (Marshall 2013), the assessment of topic coherence is typically a product of the scholar's interpretation. There is ongoing discussion in digital humanities scholarship over the interpretive significance of topics—whether they constitute subjects, themes, or discourses—and topic models do not always produce topics that appear semantically coherent to the scholar (Underwood 2012b). To the extent that text corpora are composed of figurative language, such as that found in poetry (our corpus contains poetry), topic models produce higher rates of apparently incoherent topics

[7] MALLET's default tokenizing rules failed to process some characters in our corpus. To ensure all characters were counted correctly, we converted them to Unicode escape sequences, then to purely alphabetic equivalents, before importing the texts into MALLET. We then converted the MALLET output back to Chinese characters for analysis. A Python-based version of our conversion algorithm is available at https://github.com/scottkleinman/zcoder.

(Rhody 2012). After experimenting with a number of models in several pilot studies using different numbers of topics, we settled on 100 topics, which, after the removal of stop-words, seemed to offer a good balance of scope and granularity while yielding few junk or chimera topics.

After the topics are generated, researchers are faced with interpreting them and their relations to texts in the corpus. Some scholars in the field of ancient Chinese thought have argued that contemporary interpretations of ancient Chinese documents, especially philosophical, political, and religious documents, fall victim to debilitating biases and errors, for example, either Orientalizing or Westernizing the texts (Ames 2001). Since the texts were canonized long ago, a commentarial tradition two millennia long continues to structure the (presumed) central themes of the early Chinese source texts. But this tradition makes assumptions that are open to reexamination. Topic modeling has the potential to reveal the unexpected and even challenge canonical claims about themes and contents of these texts, opening up new avenues for our understanding of ancient and medieval Chinese thought.

At the same time that our results may challenge leading interpretations of certain texts, we are well aware that our interpretations of the topics may be subject to biases of which we ourselves are unaware. Since three of the six of us publish actively in early Chinese thought, we aimed to minimize scholarly biases of our own that, unbeknownst to us, might influence our interpretations of our topics. For this reason, we decided that interpretation of our topics should be informed by independent expert knowledge in historical Chinese thought and language. So we enlisted the help of over sixty experts in the field recruited from the *Warp, Weft, and Way* blog (http://warpweftandway.com) to independently code topics. We refer to these results frequently in what follows to demonstrate a partial validation of our interpretations. This process worked as follows.

Expert coders were presented with word clouds showing a target topic's keywords. First they were given an open-ended question reading, "Suppose you had to guess what is the theme of this word cloud. What are one to three English words you would use to describe this theme?" Second, experts were asked how confident they were about their judgment in the open-ended question. Third, experts received a forced-choice question

Table 4. Topic 27 keywords and weights.

Chinese	*Pinyin*	*English*	*Word Weight*
馬	mǎ	horse	0.049
白	bái	white	0.04
物	wù	thing	0.035
生	shēng	birth, life	0.033
汝	rǔ	you	0.031
無	wú	without, nothingness	0.028
見	jiàn	see	0.022
指	zhǐ	finger, point	0.022
色	sè	color	0.019
列	liè	column	0.019

with answers enabling us to probe their opinions about the contents of these topics. In response, they could inform us that the topic was about the military, politics, philosophy, the mind, etc. Due to the likelihood of chimera or junk topics, and limitations among our experts, we included an option of "uncategorizable" as well. Fourth and finally, if an expert coder responded to a top-level multiple-choice question by saying that, in his or her opinion, the topic was about military affairs, he or she would receive a supplemental forced-choice question inquiring whether the topic represented issues including weaponry, peace, the state, war, violence, order, and/or government. Experts were always able to select multiple answers. These three levels of answers allowed us to use the expertise of generous volunteers knowledgeable about ancient and medieval Chinese thought to partially confirm or contest our interpretations of specific topics. (See appendix 3, "Survey Given Independent Coders," for the survey text and an example word cloud.)

To take an example from our own corpus, consider Topic 27 in table 4. Traditional scholars skeptical of our methods may think that a topic as incoherent as 27 is evidence that our method is of little assistance in answering research questions about early Chinese thought. However, to experts of Warring States philosophical discourse on logic and language associated with Later Mohists and the School of Names, this topic makes perfect sense. These logicians focused on problems of reference (*zhǐ* 指) and how words are related to "things" (*wù* 物). They wanted to know whether a "white horse" (*báimǎ* 白馬) is a "horse," a famous example, and how attributes such as "hard and white" (*jiān bái* 堅白) relate to substances. Such was our initial interpretation, but to minimize our own bias and error, we took additional steps. We partially confirmed this interpretation of Topic 27 by reviewing its text weights in specific texts to determine in which documents the weight of Topic 27 is heaviest. The fact that its heaviest topic weight is in the School of Names text *Gōngsūnlóngzi* provides further justification of our interpretation. We then examined responses from our independent expert coders to determine whether their interpretations were supportive of our "Logic and Language" interpretation. One of three experts assigned Topic 27, presumably not as knowledgeable about philosophical materials in our corpus as about other materials, did not understand this topic. This was revealed in his or her answer to the top-level forced-choice question, which was "uncategorizable." The other two coders agreed that it was a coherent topic. Furthermore, these two knew precisely what this topic was about. In open-ended questions, they reported that Topic 27 concerned "logicians, philosophy," "disputation," and "appearance, language." This too provides further justification of our interpretation.

The foregoing discussion about how we arrived at our interpretation of Topic 27 provides a self-contained illustration of the mixed methods we champion in this article: our *close-reading* knowledge of the Later Mohists and School of Names prompted our initial understanding of the topic; our *machine-learning* outputs revealed that Topic 27 was heavily represented in just the texts that we would hypothesize it to be; and the *experimental text analysis* that enlisted our experts' opinions in the process further confirmed the interpretation.

What Topics Make *Analects*, *Mencius*, and *Xunzi* Each Unique?

Longstanding debate surrounds the relationship between Confucius of *Analects* and his two declared successors, Mencius and Xunzi (Lau [1970] 2005; Van Norden 1992).

The notion that it is *Mencius*, rather than *Xunzi*, which is the true inheritor of the teachings of Confucius contained in *Analects* has deep roots in the late imperial Chinese commentarial tradition. (For the distribution of all topics in our model across these three texts, see figure 3.)

Tang dynasty scholar Han Yu (768–824) first asserted that authentic transmission of the teachings of Confucius ended with Mencius. The point was reiterated by Song dynasty (960–1279) neo-Confucians and canonized by Zhu Xi's (1130–1200) inclusion of *Mencius* in the collection of Confucian texts referred to as *Four Books*—along with *Analects*, *Great Learning* (Dà xué 大學), and *Doctrine of the Mean* (Zhōng yōng 中庸). The Four Books were a central part of the core curriculum memorized by students and examination candidates from the early 1300s to the abolition of the examination system in 1905. Xú Fùguān 徐復觀 (1904–82) reinforced the traditional idea that when Confucius speaks of "human nature" he is expressing the same idea that Mencius later formulated, namely, that human nature is good (Xú 1969, 89; see also Zhang 2012, 197). Following in Xú's footsteps, influential contemporary scholar Fù Pèiróng 傅佩榮 (1950–) argues that Mencius's theory of the potential for goodness latent in human nature "is an excellent expression of Confucius's thought" (Fù 2011, 872).

Others argue that this traditional interpretation is problematic and that analysis of the language of self-cultivation, including craft metaphors, indicates closer affinities between *Analects* and *Xunzi*. If human nature is like a raw piece of jade, values have to be carved into it by an outside force (*Analects* 1.15, tr. adapted from Slingerland 2003, 6–7; *Xunzi* 27.514–523, tr. Hutton 2014, 309). Knowledge of normative values is not innately present in human nature; it has to come from an external source. In contrast, *Mencius* contains an internalist theory that assumes normative values to be innate. Human nature is a seed with the potential to grow into a fully developed plant (Ivanhoe [1993] 2000, 2008; Slingerland 2003).

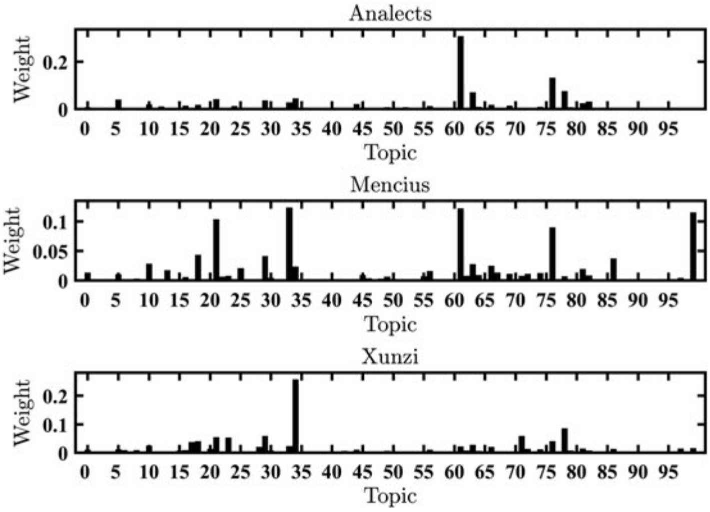

Figure 3. Text weights in *Analects*, *Mencius*, and *Xunzi* across the corpus.

Table 5. Highest weighted ten topics in each of *Analects*, *Mencius*, and *Xunzi*.

Topic	*Label*	*Keywords*	*Text Weight in* Analects
61	*Analects* Stylistics	孔 問 仁 言 人 禮 行 聞 道 貢	0.307
76	Rulers, Ability, Knowledge	君 人 公 能 死 見 欲 知 先 得	0.130
63	Ritual, Family & Governance	禮 君 人 喪 士 父 樂 母 侯 廟	0.069
78	Learning & Governance	人 知 言 名 用 治 能 欲 學 文	0.074
21	Political & Social Order	民 君 行 國 治 能 得 事 政 下	0.040
34	Ethical Rulership	君 人 義 禮 能 賢 莫 天 惡 安	0.043
5	Sacrifice, Ritual, Etiquette	大 祭 食 門 婦 先 入 既 服 出	0.038
33	Knowledge, Rulership, & Heaven	人 大 天 知 王 得 世 一 心 已	0.026
29	Heaven, Earth, Man, & the Way	天 上 下 大 道 中 人 時 後 地	0.034
82	Rulers, Virtue & Governing the People	公 王 德 成 事 民 告 用 聞 既	0.029

Topic	*Label*	*Keywords*	*Text Weight in* Mencius
21	Political & Social Order	民 君 行 國 治 能 得 事 政 下	0.102
61	*Analects* Stylistics	孔 問 仁 言 人 禮 行 聞 道 貢	0.121
33	Knowledge, Rulership & Heaven	人 大 天 知 王 得 世 一 心 已	0.122
99	*Mencius* stylistics	王 人 下 孟 取 相 或 士 他 好	0.114
76	Rulers, Ability, Knowledge	君 人 公 能 死 見 欲 知 先 得	0.089
29	Heaven, Earth, Man, & the Way	天 上 下 大 道 中 人 時 後 地	0.041
18	Kings, Heaven & Officials	下 王 詩 天 亡 士 得 侯 善 臣	0.043
86	Benefit & Moral Excellence	文 利 學 用 大 古 賢 義 能 今	0.036
63	Ritual, Family & Governance	禮 君 人 喪 士 父 樂 母 侯 廟	0.027
10	Cognition & Planning	今 心 後 力 憂 豈 朝 死 誠 棄	0.027

Topic	*Label*	*Keywords*	*Text Weight in* Xunzi
34	Ethical Rulership	君 人 義 禮 能 賢 莫 天 惡 安	0.256
78	Learning & Governance	人 知 言 名 用 治 能 欲 學 文	0.084
29	Heaven, Earth, Man, & the Way	天 上 下 大 道 中 人 時 後 地	0.057
71	Political Order vs. Disorder	人 治 事 法 世 行 功 明 主 亂	0.058
21	Political & Social Order	民 君 行 國 治 能 得 事 政 下	0.053
23	Moral-Cosmic Attunement	天 道 下 物 知 德 生 能 聖 人	0.052
76	Rulers, Ability, Knowledge	君 人 公 能 死 見 欲 知 先 得	0.038
18	Kings, Heaven & Officials	下 王 詩 天 亡 士 得 侯 善 臣	0.039
17	Statecraft, Laws, Punishments & Rewards	國 法 民 兵 賞 力 利 刑 重 上	0.035
63	Ritual, Family & Governance	禮 君 人 喪 士 父 樂 母 侯 廟	0.025

Our guiding research question in this section is "How do our topic models describe conceptual and linguistic differences that set each of these texts apart from the other two?"[8] So, which topics do we select for analysis in order to initiate a new conversation about this canonical issue? We look at the top ten topics in each document. We begin by looking at unique topics, those that show up in one text's top ten topics, but not in the other texts' sets of top ten topics (see table 5). In other words, in this section we discuss only topics that render each of these texts *unique and different from* one another. Given our results, this means we discuss Topics 5 and 82 in *Analects*; 10, 99, and 86 in *Mencius*; and 23, 71, and 17 in *Xunzi*. In the following section, we discuss those topics that our texts share *in common with* one another.

Analects

We focus first on *Analects*, which is a collection of sayings attributed to Confucius (551–479 BCE) and his followers and contains material likely dating predominantly to the early Warring States.[9] Two topics in the top ten differentiate *Analects* from other texts, including other texts within Confucianism. These are Topic 5, with a text weight in *Analects* of 0.038, which we label "Sacrifice, Ritual, Etiquette," and 82, with a text weight of 0.029, which we call "Rulers, Virtue and Governing the People." Since the text weight of Topic 5 in *Analects* is 0.038, this means that 3.8 percent of *Analects* is composed of the clustered terms representing Topic 5 (see table 6).

Keyword characters in Topic 5 include great (*dà* 大), sacrifice (*jì* 祭), feed or eat (*shí* 食), gate or school (*mén* 門), wife (*fù* 婦), first or before (*xiān* 先), enter (*rù* 入), submit or ritual garb (*fú* 服), exit or go out (*chū* 出), drink (*yǐn* 飲), assist or assist someone (*xiàng/xiāng* 相), and weep or cry (*kū* 哭). These terms describe a semantic space revolving around important rituals and sacrifices, particularly those involved in ancestor worship and mourning. Independent coders reported that this topic concerned ritual and religion. Topic 5 has heavy text weight in only a handful of the texts in the corpus, including in *The Classic of Rites* (Lǐjì 禮記, 0.175, and Yílǐ 儀禮, 0.170) and *The Rites of Zhou* (Zhōulǐ 周禮, 0.052), which contains the core of the Book of Changes or *Yìjīng* (see table 7). Together these three texts form a unit known in the Chinese commentarial tradition as the *Three Ritual Texts* (Sānlǐ 三禮). The bulk of each of these works consists of long lists of ritual prescriptions specifying the correct way of executing various rites and sacrifices, for example, specifying which clothes to wear and which color of accouterments to use. In form and content, parts of the *Analects*, especially Chapter 10, are strikingly similar to the *Three Ritual Texts*. This is a distinctive feature of *Analects* in comparison to

[8] Notice this concerns the *semantic contents* of the works rather than the *phylogenies* of the works. Phylogenetic analyses familiar from biology and genetics are increasingly used in text analytics to great success to determine a text's origins by tracking small variations in word use over increments of time. See, e.g., the remarkable phylogenetic study of the *Canterbury Tales* by Barbrook et al. (1998). We are strictly interested in the texts' conceptual, and sometimes linguistic, similarity, not in phylogeny, so we make no claims about the origins of these texts.

[9] The bulk of the textual material in *Analects* was composed over the span of at least several centuries from the early Warring States period to the third century BCE. See Brooks and Brooks (1998); Cheng (1993, 313–23); Makeham (1996); Qu (1983, 382–89); Slingerland (2000). As indicated by Hunter (2014), its compilation likely occurred in the Han dynasty.

Table 6. Topics differentiating *Analects*, *Mencius*, and *Xunzi* from one another.

Document	*Text Weight*	*Topic*	*Corpus Weight*	*Label*	*Topic Keywords in Descending Order of Weight*
Analects	0.029	82	0.22	Rulers, Virtue & Governing the People	公 王 德 成 事 民 告 用 聞 既 實 能 先 政
Analects	0.038	5	0.19	Sacrifice, Ritual, Etiquette	大祭 食 門 婦 先 入 既 服 出 飲 相 小 哭
Mencius	0.027	10	0.35	Cognition & Planning	今 心 後 力 憂 豈 朝 死 誠 棄 觀 入 罪 古
Mencius	0.114	99	0.2	*Mencius* Stylistics	王 人 下 孟 取 相 或 士 他 好 長 舍 章 羊
Mencius	0.036	86	0.2	Benefit & Moral Excellence	文 利 學 用 大 古 賢 義 能 今 國 商 良 富
Xunzi	0.052	23	0.45	Moral-Cosmic Attunement	天 道 下 物 知 德 生 能 聖 人 得 身 言 神
Xunzi	0.058	71	0.25	Political Order vs. Disorder	人 治 事 法 世 行 功 明 主 亂 亡 得 相 用
Xunzi	0.035	17	0.21	Statecraft, Laws, Punishments & Rewards	國 法 民 兵 賞 力 利 刑 重 上 勝 官 戰 爵

Mencius and *Xunzi*. The fact that unsupervised topic modeling is able to pinpoint this scholarly insight powerfully demonstrates the value of this new research tool.

The results of Topic 5 appear to have important implications in the adjudication of an ongoing debate about the role of sacrifices and spirits in *Analects*. Consider the opinion of a key voice in Chinese intellectual history about Confucius, spirits, and sacrifices. Feng Youlan 馮友蘭 (1952–53, 1:58) uses a close-reading method to conclude that Confucius "displayed a rationalist attitude [toward spirits], making it probable that there were other superstitions of his time in which he did not believe." In contrast, Thomas

Table 7. Topic 5's text weights across texts in the corpus.

Text	*Text Weight of Topic 5*
Yílǐ 儀禮	0.175
Lǐjì 禮記	0.17
Zhōulǐ 周禮	0.052
Dàdàilǐjì 大戴禮記	0.04
Analects (Lúnyǔ) 論語	0.038
Báihǔtōngdélùn 白虎通德論	0.034
Mùtiānzǐzhuàn 穆天子傳	0.033
Kǒngzǐjiāyǔ 孔子家語	0.032
Shìmíng 釋名	0.029
Ěryǎ 爾雅	0.027

Wilson uses a close-reading method to emphasize *Analects*' advocacy of ritual, sacrificial rituals to ancestors and deities in particular. Wilson (2014, 185) reasons that "contrary to modern accounts, imperial-era commentaries on the *Analects* 論語 disclose the figure of Confucius as committed to pious sacrifices to gods and spirits." Unlike Xunzi, who explicitly reports his intentions to endorse the use of sacrifice for social-functional reasons (chap. 19, "Discourse on Ritual"; see Campany 1992), the text of the *Analects* leaves readers uncertain with regard to Confucius's intentions about sacrifice. For this reason, debate about Confucius's relation to sacrifice will not be easily settled by topic modeling or by close reading alone. Feng Youlan cites *Analects* 7.20 to argue for Confucius's pragmatic epistemology, and Wilson cites *Analects* 3.6 to demonstrate Confucius's concern with Mount Tai's sacredness and ritual importance; Feng Youlan cites 6.22 showing that Confucius keeps ghosts and spirits at a distance and prioritizes social harmony, not metaphysics, and Wilson cites 3.12 to argue for Confucius's earnestness during sacrifices to the spirits. Perhaps the process of tit-for-tat close-reading commentary will continue *ad infinitum*.

But machine-learning results from topic modeling provide two reasons to think Wilson is likely correct. First, numbers of scholars argue that belief in gods in early China had prudential, not rational, origins. Prudential concerns arose through divination and knowing the future (Overmyer et al. 1995), ancestor reverence and seeking ancestors' blessings (Eno 1990a, 1990b), and avoiding curses through shamanism (Ching 1997). To this, however, advocates of the alternative view will, as we have seen, return to the discussion with additional texts and interpretations, and the two sides will continue to trade texts in support of two mutually incompatible interpretations of *Analects* into the indefinite future. This brings us to what our model can contribute to consilience. Second, our interpretation of Topic 5 offers evidence in favor of the unique importance of practices associated with these sources of religion, especially religious ritual and sacrifice (*jì* 祭), for the compilers of the received *Analects*. If the compilers of *Analects* were as rationalist as, say, Xunzi, we would not expect to see Topic 5 so prominently, and uniquely, featured in this text. Were Feng Youlan correct, Topic 5 would probably not differentiate *Analects* from the other two texts.

Topic 82, "Rulers, Virtue and Governing the People," is a reflection of the fact that numerous passages in *Analects* discuss the importance of the charismatic virtue (*dé* 德) of rulers, dukes (*gōng* 公), and kings (*wáng* 王) as they govern (*zhèng* 政) the people (*mín* 民). Rulers are advised to employ officials with virtue (*dé* 德) and ability (*néng* 能) to serve (*shì* 事) them by bringing affairs (*shì* 事) to completion (*chéng* 成). Independent coders reported in open-ended questions that this topic concerns "history, statecraft, philosophy" and "civil-affairs, reports, officialdom." Turning to the word ranks of its keywords across the three target texts, we see governance or order (*zhèng* 政) is much more important in *Analects* (forty-three occurrences, thirty-first in rank, 5.6/1000 characters) than to authors of *Mencius* (fifty-four occurrences, fifty-seventh in rank, 3.0/1000) and *Xunzi* (ninety-five occurrences, eighty-fifth in rank, 2.3/1000). Topic 82 is the heaviest weighted topic in *Guoyu* 國語, which is a collection of historiographical and fictional anecdotes set in the pre-Qin period. Many of these anecdotes feature professional persuaders or diplomats who use their command of language to persuade rulers to "do the right thing."

Topics 5 and 82 represent themes that are unique to *Analects* and not shared with *Mencius* or *Xunzi*. The prevalence of Topic 5, "Sacrifice, Ritual, Etiquette," and Topic

82, "Rulers, Virtue and Governing the People," confirms the scholarly consensus that Confucius (as he is portrayed in *Analects*) projected social leadership emphasizing the importance of ritual and sacrifice as elements in individual self-cultivation practice and in achieving social order. The social-functional values enshrined in this tradition could be transmitted to disciples and followers through teaching, observation, and emulation. In many reported conversations with rulers, these same values were transmitted through advice on how to govern a state through virtue rather than laws and military force. These topics focus on institutionalized ritual practice and do not reveal much concern with cognition or emotion or other internal states of mind.

Mencius

Mencius was a self-proclaimed follower of the teachings of Confucius and was employed as an adviser to rulers in the middle to late fourth century BCE. *Mencius* is generally agreed to have been composed in the Warring States period. Three topics set *Mencius* apart from *Analects* and *Xunzi*: Topic 10, "Cognition and Planning," Topic 86, "Benefit and Moral Excellence," and Topic 99, "*Mencius* Stylistics." Both Topic 10 (0.027) and Topic 86 (0.036) have rather light weights in *Mencius* in comparison to Topic 99 (0.114).

Topic 10 is difficult to characterize, a point reflected in our coding results. In open-ended questions, expert coders described this topic as concerned with "loyalty, official service," "emotion, masculinity, mind," "psychology, self-cultivation, morality," and "mind, emotion, leave." Topic 10 is dominated by two sets of words. The first concerns temporality and includes present (*jīn* 今), later (*hòu* 後), and ancient (*gǔ* 古). We take this to be indicative of *Mencius*'s frequent comparison between a golden age of the past and the fallen present. The second set concerns cognition or thought. This includes heart-mind (*xīn* 心), worry (*yōu* 憂), sincere (*chéng* 誠), regard or gaze upon (*guān* 觀), and guilt or crime (*zuì* 罪). Heart-mind represents the seat of cognition and emotion and is a common term in *Analects*, *Xunzi*, and *Mencius*. As our research group has shown using similar quantitative methods, cognition and emotion terms cluster in this topic in part because *Mencius* focuses on internal reflection and mental regulation of emotion.[10] Topic 10 has heavy text weights in two texts, *Yandanzi* (0.085) and *Jian zhu ke shu* (0.084).

Topic 86 appears to represent "Benefit and Moral Excellence." Independent coders reported that this topic is concerned with "culture and profit," "learning, cultivation of culture," "ethics," and "study, ancient, benefit." The top term, pattern or culture (*wén* 文), is used in names (e.g., King Wen 文王, Duke Wen) in all but four of the fifty-one occurrences in *Mencius*. These four refer to "decorative pattern," "rhetoric," "(rhetorical) style," and "to refine," respectively. While *wén* does mean "high culture, civility, or civilization" in other texts, it is not used in this meaning in *Mencius* (Bergeton 2013). Topic 86 is, therefore, a case where coders may be misled by the polysemy of the word *wén*. Benefit or profit (*lì* 利) is next. *Mencius* often criticizes the pursuit of profit (*lì* 利) as inferior to what is right (*yì* 義). As indicated by the inclusion of both present (*jīn* 今) and ancient (*gǔ* 古) in Topic 10, *Mencius* often contrasts amoral behavior of the present

[10] For our team's text analytics exploration of *xīn* 心, embodiment and the metaphysics of mind in ancient China, using this same corpus, see Slingerland et al. (forthcoming).

with morally superior behavior of the ancient or Golden Age (*gǔ* 古). Study (*xué* 學) is a step on the path of self-cultivation. Study elicits innate potential to become virtuous (*xián* 賢) and good (*liáng* 良), traits needed to serve one's state (*guó* 國). Virtue terms such as these bind Topic 86 together. Topic 86 has a large text weight in only *Discourses on salt and iron*, a debate about taxation (0.159).

The contents of Topic 86 provide evidential support for an interpretation of *Mencius* as advocating internalist belief in the innate potential for goodness in human nature. This engages *Mencius*'s discussion at 3B9 in which he attacks the doctrines of Yáng Zhū 楊朱 and Mò Dí 墨翟, who advocate egoism and altruism, respectively. Mencian Confucianism repudiates these act-based ethics in favor of the cultivation of character (Csikszentmihalyi 2002). This is uncontroversial, but it leads to an ongoing interpretive problem about self-cultivation. Consider Mencius's four "sprouts" of virtues (*sì duān* 四端) in 2A6, where he writes that "if one is without the heart (*xīn* 心) of compassion, one is not human.... The feeling of compassion is the sprout of benevolence" (Van Norden 2008, 46; see also the archer analogy at 2A7). On one interpretation of these passages, the cultivation of feelings appears to be the source of moral virtue in *Mencius*, making *Mencius* representative of what is known in philosophy as an "internalist" theory of moral motivation. This allegedly contrasts with moral motivation and cultivation as found in *Analects* and *Xunzi*. These two texts are thought to advocate a greater number of, and greater roles for, externalist sources of morality like ritual (*lǐ* 禮), patterned civility (*wén*), and rectification of names (*zhèngmíng* 正名). Our evidence appears to support this interpretation of *Mencius*. We draw additional evidence for this interpretation from several sources in traditional scholarship. For example, Slingerland (2003) argues that *Mencius* is uniquely and distinctively "internalist," and Kline (2000) that *Mencius*'s ethics are "inside-out," as have others (Ihara 1991; Wong 1991). However, since Topic 86 has a high text weight in only *Discourses on salt and iron*, and not in our core Confucian texts, we must collect additional evidence for the internalist interpretation of *Mencius* before we can rest confident that it is correct.

Topic 99 appears to represent features of dialogic text and style in *Mencius*. Independent coders reported that this topic is concerned with "Mencius" and "sage, teaches, king." This bland description from coders provides some confirmation that Topic 99 stands apart from other topics that have richer semantic content and coherence. The fourth most frequently occurring word in this topic is Mencius's own name (*mèng* 孟), which obviously occurs many, many times in the text, and the third is *xia* 下, usually meaning "below" or "under." In this topic, it probably picks out the frequent use of *xia* 下 in *Mencius* chapter titles, which are classified into A (*shang* 上) and B (*xia* 下). The most frequent word is king (*wáng* 王). This reflects the fact that many of the dialogical exchanges in *Mencius* are between kings and Mencius himself. Topic 99 not only sets *Mencius* apart from *Analects* and *Xunzi*, it also sets *Mencius* apart from all other texts. Its weight in *Mencius* is 0.114, which is twice its weight in any other text. This aptly confirms our designation of this topic as reflecting "*Mencius* Stylistics."

Although it lacks thematic coherence, Topic 99 is a good example of what we have come to call "stylistic" topics. Stylistic topics often load heavily in only one or two texts in the corpus and seem to represent clusters of terms that are specific to that text. These tend to be dominated by meaningless function words not removed in our stopword list, stylistic tics, and commonly used proper names. Sometimes they also point to distinctive conceptual themes. Tied for eighth most frequent word in Topic 99, for instance, is *hào* 好, "to be fond of, like, desire," which picks up Mencius's concern with internally driven preferences and desires. Although they are perhaps less interesting philosophically or conceptually, these stylistic topics have potential use in tracing textual lineages, dating texts, classifying newly discovered texts, or picking up surprising continuities in themes between texts. After *Mencius*, the text into which Topic 99 loads with the heaviest text weight is the obscure *Yùzi* 鬻子 fragments (0.059), usually classified as "Daoist." This suggests something about stylistic or conceptual influences, or convergent thematic concerns, between these otherwise disparate texts, and marks this relationship out for further profitable study with close-reading methods by experts in the area.

Xunzi

Xunzi is a compilation of various texts, including philosophical essays, attributed to Xunzi, and exchanges between Xunzi and others. Like Mencius, Xunzi was a self-proclaimed follower of the teachings of Confucius. He was employed as a teacher and adviser to rulers in the third century BCE. Although compiled in its present form in the Han dynasty, the bulk of the material in *Xunzi* was composed in the late Warring States period. Philosophically, *Xunzi* is a third century BCE development of core ideas in *Analects* that incorporates ideas from other pre-Qin philosophies.

From a modeling perspective, what is most intriguing is the semantic scope of *Xunzi*'s heavyweight topics. Independent coders reported that Topic 71, "Political Order vs. Disorder," concerns "politics," "law, humanity, worldly," and "humanity, the world and dealing with affairs." At 0.058, this is the fourth heaviest topic in *Xunzi*. The following passage nicely illustrates how the twelve most prominent keywords in Topic 71 cluster in the *Xunzi*:

> There are men (*rén* 人) who create order (*zhì* 治); there are no rules (*fǎ* 法) creating order (法) of themselves. The rules (*fǎ* 法) of Archer Yi have not perished (*wáng* 亡), but not every age (*shì* 世) has an Archer Yi.… Thus, rules (*fǎ* 法) cannot stand alone, and categories cannot implement (*xíng* 行) themselves.… One who tries to correct the arrangements of the rules (*fǎ* 法) without understanding their meaning, even if he is broadly learned, is sure to create chaos (*luàn* 亂) when engaged in affairs (*shì* 事). And so, the enlightened (*míng* 明) ruler (*zhǔ* 主) hastens to obtain (*dé* 得) the right person (*rén* 人).… If one hastens to obtain (*dé* 得) the right person (*rén* 人),… [then] one's accomplishments (*gōng* 功) will be grand. (*Xunzi* 12.1–20, tr. adapted from Hutton 2014, 117)

As shown here, *míng* 明 (bright, clear; perspicacious, enlightened) often refers to the far-sightedness associated with sages and desired in rulers (*zhǔ* 主) in early China (Brown and Bergeton 2008). By obtaining (*dé* 得) and with the right people (*rén* 人) to assist him, the ruler can create order (*zhì* 治) and avoid chaos (*luàn* 亂), thereby achieving great accomplishments (*gōng* 功). The ruler's discernment is therefore more important than blind enforcement of rules or promulgated models (*fǎ* 法). This sounds like a classically Confucian claim, albeit with much more emphasis on institutional structures than we see in *Analects* or *Mencius*. This thematic distinctiveness is in turn reflected in the fact that Topic 71 is completely absent from *Analects* and loads lightly in *Mencius* at 0.010.

Another topic unique to *Xunzi* among the classical Confucian works is Topic 17, "Statecraft, Laws, Punishments and Rewards," which loads at 0.035 in the *Xunzi* but is absent from both *Analects* and *Mencius*. This topic similarly involves ordering the state (*guó* 國), but through what appear to be means associated with what has come to be known as Legalism. The people (*mín* 民) and officials (*guān* 官) can be managed with rewards (*shǎng* 賞), noble rank (*jué* 爵), and profit (*lì* 利). Punishments (*xíng* 刑) figure into Legalist governance, particularly heavy (*zhòng* 重) ones. This topic also includes terms related to violence and force: troops, weapons (*bīng* 兵), victory (*shèng* 勝), and war (*zhàn* 戰). These terms for warfare are arguably more frequent in *Xunzi* than in *Mencius* and *Analects* due to *Xunzi*'s Chapter 15 "Debate on Military Affairs" (*Yìbīng* 議兵), and tend to be associated with military strategists, such as Sunzi, author of *The Art of War* (*Sūnzi bīngfǎ* 孫子兵法), or Legalist thinkers such as Han Feizi, Xunzi's student. Since *The Art of War* and *Hanfeizi* are in our corpus, we can look to Topic 17's weight in them to confirm or disconfirm our reasoning. We find, indeed, that both texts appear among the heaviest texts into which Topic 17 loads, as do several other military and Legalist texts. Their presence here supports some scholars' view that *Xunzi* is more focused on institutional means of social control than is Confucius or Mencius. Independent coders reported that this topic concerned "governance" and "Legalism." (See *Hanfeizi* 53, Wáng Xiānshèn 王先慎 2006, 471–73, for a passage nicely illustrating Topic 17.)

Fu Peirong (2011, 872) takes the fact that Xunzi was the teacher of prominent Legalists, such as Hanfei and Li Si (the prime minister of the first emperor of the Qin dynasty), to indicate that Xunzi's theory of human nature as "bad" is opposed to the theory of a human nature with innate potential for goodness that he sees in *Analects* and *Mencius*. However, this may not be the best explanation for the prevalence of Topics 71 and 17 in the *Xunzi*. The large text weights of Topics 17 and 71 probably derive from *Xunzi*'s greater interest in discussing details of government institutions like "laws," "punishments," and "officials." Unlike Confucius and Mencius, Xunzi had more personal experience serving as an official. Hence it is only to be expected that his practical and less theoretical discussion would lead him to write about "Statecraft, Laws, Punishments and Rewards" (Topic 17) and "Political Order vs. Disorder" (Topic 71) more than *Analects* and *Mencius*.

The text weights of Topics 17, 71, and 23 show that they are all highly representative of *Xunzi*. Yet Topic 23 (0.052 in *Xunzi*), with a corpus weight of 0.453, the fifth weightiest topic in the entire corpus, has much greater overall importance to ancient and medieval Chinese literature. Topic 23 has very heavy text weights across texts in the Daoist school, including *Dao de jing* (0.434) and *Heshanggong laozi* (0.358). It constitutes only 0.007 of *Mencius*, however, and is less weighty yet in *Analects*. We dub Topic 23 "Moral-Cosmic Attunement." Its heaviest terms are heaven or sky (*tiān* 天), way (*dào* 道), and under (*xià* 下), as in the "world" (*tiānxià* 天下), literally "under heaven." These terms all tend to refer to the structure of the universe. In *Dao de jing* and *Xunzi* the "sky" or "heaven" is an impersonal force, not a moral agent as in *Mencius* and *Analects*. Prominent moral terms include way (*dào* 道), virtue (*dé* 德), and sage (*shèng* 聖). Philosophical terms include way, creature or thing (*wù* 物), knowledge (*zhī* 知), and saying or teaching (*yán* 言).

To conclude this section and summarize its results, Topic 5 suggests that *Analects* emphasizes and discusses specific ritual prescriptions more frequently than *Mencius* or *Xunzi*. Topic 82 suggests that *Analects* proposes a toolkit for moral suasion and social change differing from those of the author's close intellectual ancestors. Topic 17 reveals that *Xunzi* is more interested in detail-oriented discussions of government institutions than *Mencius* or *Analects* and Topic 71 that *Xunzi* has a more elaborate theory of the use of legal structures and military force than *Analects* and *Mencius*. Topic 23 indicates that Xunzi, despite his critique of Daoist thought, shared with Daoists an interest in moral-cosmic attunement that is absent from the *Analects* and *Mencius*. Topics 86 and 10 are not as thematically well defined as Topics 5, 17, 71, and 23 and possess large text weights in a variety of texts across different genres. The fact that Topics 86 and 10 have heavy text weights in *Mencius* is less helpful in setting this text apart from *Analects* and *Xunzi*. In spite of this, the overall match between machine-generated topics and scholarly studies of the defining characteristics of these three texts remains impressive. Overlap with existing scholarly opinion should enhance our confidence in the general method. At the same time, the specifics of our findings represent original contributions to the scholarly debate, either weighing in on one side or suggesting novel lines of attention or inquiry.

Intersecting Topics in *Analects*, *Mencius*, and *Xunzi*

We are driven to understand the semantic content and relationships between these three texts. The previous section was intended to provide an understanding of what makes these texts *different* from one another. Those discussions combine with the discussion in this section, which is about what makes these texts *similar* to one another, to preliminarily address our guiding research question about whether the contents of *Mencius* or *Xunzi* most resemble the contents of *Analects*. Using the same list of the ten most weighty topics in each of our three texts, we calculated the topic intersections between documents as a shared set of topics.[11] Now we focus on topics that load into pairwise unions of these three texts in an effort to explore their similarities. Before discussing those topics, however, we briefly report on the four topics that lie at the union of all three texts (see table 8 and figure 4).

At the intersection of *Analects*, *Mencius*, and *Xunzi* are Topics 29, "Heaven, Earth, Man and the Way"; 76, "Rulers, Ability, Knowledge"; 21, "Political and Social Order"; and 63, "Ritual, Family, and Governance." These topics possess some of the largest corpus weights of all 100 topics in our model. Topic 29 is ranked number one, Topic 76 is ranked number three, Topic 21 is ranked number four, and Topic 63 is ranked number twenty-nine. Our expert coders report that Topic 29 is concerned with "cosmology, virtue" and "cosmology, time, philosophy." They report that Topic 76 is concerned with "lordly leadership" and "politics and leadership." Topic 21 is concerned with "governance, kingdom, people" and "people, masses," and Topic 63 with "ritual, masters of ritual (*rú*)," and "ritual, rites, ceremonies."

[11] To put this point semi-formally, A ∩ B is the intersection of A with B, that is, the set of all the elements in A that are also contained in B and not contained in any other elements. We applied this definition to a 10*3 matrix, representing ten topics (for each document) in rows and three documents in columns (see figure 4).

Table 8. Formal interpretation matrix of the intersections of *Analects*, *Mencius*, and *Xunzi* with topic keywords (∩ = intersection of sets).

Document	*Topic's Weight in Text (Text Weight)*	*Topic*	*Topic's Weight in Corpus (Corpus Weight)*	*Label*	*Topic Keywords in Descending Order of Weight*
(*Mencius* ∩ *Xunzi* ∩ *Analects*)	0.04/0.06/0.03	29	0.6	Heaven, Earth, Man, & the Way	天 上 下 大 道 中 人 時 後 地 長 從 成 德
(*Mencius* ∩ *Xunzi* ∩ *Analects*)	0.1/0.05/0.04	21	0.46	Political & Social Order	民 君 行 國 治 能 得 事 政 下 食 教 官 道
(*Mencius* ∩ *Xunzi* ∩ *Analects*)	0.03/0.03/0.07	63	0.18	Ritual, Family & Governance	禮 君 人 喪 士 父 樂 母 侯 廟 親 主 命 事
Xunzi ∩ *Analects*	0.25/0.04	34	0.37	Ethical Rulership	君 人 義 禮 能 賢 莫 天 惡 安 亂 下 善 性
Xunzi ∩ *Analects*	0.08/0.07	78	0.37	Learning & Governance	人 知 言 名 用 治 能 欲 學 文 小 富 彼 盜
Mencius ∩ *Analects*	0.12/0.3	61	0.19	Language of *Analects*	孔 問 仁 言 人 禮 行 聞 道 貢 仲 學 知 路
Mencius ∩ *Analects*	0.12/0.03	33	0.08	Knowledge, Rulership, and Heaven	人 大 天 知 王 得 世 一 心 已 義 且 今 見

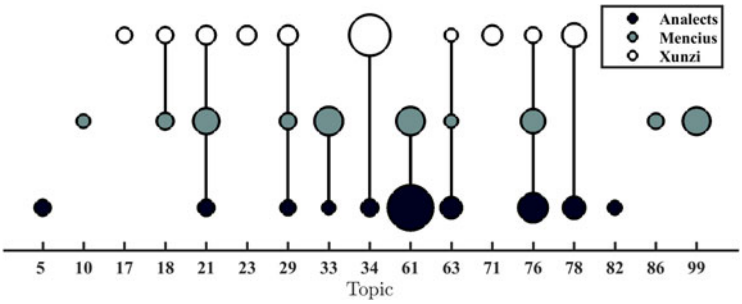

Figure 4. Topic intersections in *Analects*, *Mencius*, and *Xunzi*. Topic intersections of the ten most central topics for each document. Circles represent the presence of a topic within the ten most central topics. Circle size is proportional to the document's centrality (topic weight). Links (vertical lines) indicate an intersection.

This information supports two major inferences about shared semantic content in these texts. First, these heavy corpus weights provide strong evidence that topics at the heart of early Confucianism are exceptionally well seeded throughout ancient and medieval Chinese literature. When we compare topics weighty in these texts with topics weighty in what are traditionally referred to as "Daoist," "Legalist," and "Mohist" texts, we find that those loading into Confucian texts are much, much more likely to be represented with heavy corpus weights.[12] Second, consider that Mencius and Xunzi both self-identified as masters of ritual (*rú* 儒), who followed in Confucius's *rú* footsteps. The fact that *Analects*, *Mencius*, and *Xunzi* have a shared interest in Topics 29 ("Heaven, Earth, Man and the Way"), 76 ("Rulers, Ability, Knowledge"), 21 ("Political and Social Order"), and 63 ("Ritual, Family and Governance") indicates that the pre-Qin concept of *rú* referred to a set of philosophies characterized by a high degree of internal coherence. Earlier we noted that Topic 86 separated *Mencius* from other texts by virtue of its internalist perspective about moral motivation and normativity. We contrasted internal moral motivation with external motivation, which we associated with ritual and law. Here, however, we see that *Mencius* (0.027) and *Xunzi* (0.025)

[12] The Chinese Text Project's genre classification emerges from traditional Chinese content-based and form-based library taxonomies as well as recent categories. The five genre categories "Confucianism," "Mohism," "Daoism," "Legalism," and "School of Names" are English translations of a classification system that can be traced back to the Western Han scholar and historian Sima Tan (c. 165–100 BCE; see Csikszentmihalyi 2002; see also Goldin 2011 on "Legalism"). The "School of the Military" and the "Miscellaneous Schools" categories can be traced back to Ban Gu's (32–92) classification of books in the *Han shu*. Knowledge of the representation of genre in our corpus is helpful for the sake of interpreting topics. For example, the biggest genre is history (53 percent), followed by Confucianism (16 percent) and ancient classics (6 percent). However, due to several shortcomings of the Chinese Text Project's classification of genre, we do not use genre for analyses. Consider CTP's "Excavated texts" category. The fact that it includes documents that CTP calls "*Mawangdui*" and "*Guodian*" mistakenly leads users to infer that these documents represent the Mawangdui and Guodian manuscripts. In fact, at the time of writing, these CTP documents only represent different versions of the *Dao de jing*. Further, at present the "ancient classics" genre includes Song dynasty material. Thus we exercise extreme caution in using some of the CTP genres.

share in Topic 63, "Ritual, Family and Governance," to the same degree. This topic is fronted by ritual or rites (*lǐ* 禮, word weight = 0.042), which is why it is strong in *Analects* (0.069). This apparent conflict in our interpretation can perhaps be resolved by keeping in mind that Mencius's internalist stance naturally made him less interested in discussing the external tradition embodied in the rituals and rites, but this does not mean that he dismissed them altogether.

To appreciate how results at the intersection of all three texts might inform current debate, let us continue examination of Topic 63 in light of some secondary literature. Topic 63 does not prominently feature moral terms. The difference between Topic 34 and Topic 63 allows us to grasp a subtle but important difference in the scope of the shared social ideals across the three books. Topic 34 sees the virtue of duty or right action (*yì* 義, word weight = 0.031) traveling together with terms connoting high social status like lord or nobleman (*jūn* 君, word weight = 0.037) and rituals (*lǐ* 禮, word weight = 0.022). This informs our understanding of what Brindley (2009) has called the "sociology of the *junzi*." Specifically, considerations about how the distribution of Topics 63 and 34 differs between the three books can add considerable subtlety to this scholarly debate. In contrast to *Mencius*'s appeal to internal states (see discussion of Topic 86 above), *Xunzi* appears to link rites that accord high social status and certain moral virtues. *Yì* 義, with its connotations with animal sacrifices to gods, not benevolence, has a particularly heavy word weight in Topic 34. Topic 34 is represented in *Xunzi* at five times the level and *Analects* at twice the level it is represented in *Mencius*.

This suggests an interpretive hypothesis worth exploring through traditional scholarship. Some scholars, including Ivanhoe (2008, 5), argue that the Confucian "ethical ideal" is "something anyone can achieve and a way of being human that can be manifested in a wide range of social roles." Others concur (Hsu 1977, 162; Wills 2012, 25). However, Brindley (2009), echoing Hall and Ames (1987, 188), argues with some force that achievement of the status of gentleman or nobleman (*jūnzi* 君子) is restricted to high-status males, or males who are entitled to perform certain rites.

While we concur with Brindley on this matter, our topic-modeling results suggest value in further research on two pleasingly concrete questions: First, is the *jūnzi* ideal preferentially associated with ritual and the virtue of duty or righteousness, rather than other virtues like benevolence (*rén* 仁)? An affirmative answer is suggested by an analysis of Topic 63 and the word weights of its keywords. Second, consider that Brindley (2009) states in the title of her paper that her thesis is restricted to *Analects*. When that restriction is accompanied by claims about "Confucian" morality, it can sow confusion. This leads to another research question: Might there be significant cross-textual variety in early Confucian texts' association of the *jūnzi* ideal with high social status? Topic 63's text weights in *Analects*, *Mencius*, and *Xunzi* place it in each text's top ten, but its distribution in *Analects* is twice that of its distribution in *Mencius* and *Xunzi* (see table 5). This raises the probability that Brindley and Ivanhoe are talking past one another, which is easy to do using exclusively close-reading methods. Could it be that Brindley emphasizes what is true but only of *Analects*, while Ivanhoe emphasizes what is true of *Mencius* and *Xunzi* but not *Analects*? On the strength of our results, we surmise this is likely to be true. More important, our interpretation of the results suggests value in pursuing a concrete research question with close-reading methods to narrow in on an answer.

We now move from our brief review of topics at the intersection of all three texts to discussion of those topics that intersect only in pairwise fashion. The topics at the intersection of *Analects* and *Xunzi* include Topics 34, "Ethical Rulership," and 78, "Learning and Governance." The corpus weights of Topics 34 and 78 rank number eight and number nine of 100 total topics, suggesting their wide influence. Independent coders report that Topic 34 is concerned with "subject-ruler relations" and "virtue, politics," and that Topic 78 is concerned with "governance, learning, talent" and "human, knowledge, culture."

Topic 34 has a heavy text weight in *Xunzi* (0.256) and a moderate weight in *Analects* (0.043) but a very low weight in *Mencius* (0.023). Why might Topic 34, on "Ethical Rulership," load lightly in *Mencius*? Moral leadership is a common theme in that text, after all. Investigating heavyweight characters in Topic 34, *Analects* and *Xunzi* represent rites (*lǐ* 禮) at very similar rates (9.8/1000 and 8.5/1000 respectively). But the rate of *lǐ* in *Mencius* falls far below this (3.8/1000). Since in *Analects* and *Xunzi* human nature (*xìng* 性) is initially ignorant of normative values, these texts recommend use of the rituals and rites (*lǐ*) to build morally refined gentlemen (*jūnzǐ* 君子) who possess traits such as right action (*yì* 義), worthiness (*xián* 賢), and goodness (*shàn* 善). With rites and rituals, the Confucianism of *Analects* and *Xunzi* says that the nobleman orders himself and leads a state that is neither chaotic (*luàn* 亂) nor characterized by widespread badness (*è* 惡), but rather at peace (*ān* 安). While not unimportant, rites play a much less significant role in the philosophy of *Mencius*, since the Mencian strategy of self-cultivation is directed at motivating normative behavior through appeal to and training of the innate sprouts of virtue (Ivanhoe [1993] 2000). Topic 34's emphasis on rites in *Analects* and *Xunzi* is what we expect to observe given the "internalist" morality represented in Topic 86, which had a heavy topic weight in *Mencius*.

Topic 78, "Learning and Governance," contains a few concepts, including *wén* 文 (pattern; patterned civility, high culture), emphasized in *Analects* and *Xunzi* but not *Mencius*. Learning (*xué* 學) and knowledge (*zhī* 知) have higher saturation in *Analects* and *Xunzi*. Their presence in Topic 78 suggests that keys to rulership involve cognitive preparation of the mind for rule or management (*zhì* 治). This contrasts with the model of rulership discussed in *Mencius* Books 1 and 2, in which kings are challenged to deeper levels of empathy and emotion. The contents of Topic 78 raise the probability that *Analects* and *Xunzi* are semantically linked by virtue of their advocacy of a set of normative values deriving from learning (*xué*) and an external, refined (*wén*) tradition. The sizeable text weights of Topic 78 in *Analects* (0.074) and *Xunzi* (0.084) provide counter-evidence to the claim that differences between *Mencius* and *Xunzi* on the contents of human nature are merely a matter of emphasis rather than the result of different views of the moral resources located within the individual.[13]

In terms of differentiating the influence of *Analects* on *Mencius* and on *Xunzi*, evidence weighs in favor of greater discursive overlap between *Analects* and *Xunzi*. This appears to reduce the probability that the traditional theory about *Mencius*'s closer relation to *Analects* is correct. But one might suspect that consideration of topics at the intersection of *Analects* and *Mencius* will *increase* the justification of a closer relation between

[13] For a representation of this view, see Lau ([1970] 2005, xix–xxii).

Analects and *Mencius*. At this intersection we have Topics 61, "*Analects* Stylistics," and 33, "Knowledge, Rulership and Heaven." Both Topics 61 and 33 occur in the top ten largest loading topics in *Analects* (0.307 and 0.026) and *Mencius* (0.121 and 0.122). Yet note they also both fall within the top thirteen topics of *Xunzi* (0.020 and 0.021). From this we draw the important inference that no heavy loading topics falling at the intersection of *Analects* and *Mencius* successfully differentiate their semantic contents from *Xunzi*'s contents. This alone increases the probability that *Xunzi* tracks the semantic contents of *Analects* more closely than does *Mencius*, all things considered.

Ranked sixty-fifth, with a corpus weight of 0.077, Topic 33 is not commonly represented in the corpus. We call this topic "Knowledge, Rulership and Heaven." Independent coders report that Topic 33 is concerned with "heaven, knowledge," "ruling," and "kingship." Three of three independent coders in forced-choice questions report that Topic 33 concerns leadership, though kingship and statecraft were also regarded as important. The dominant keywords in this topic are people (*rén* 人), big or great (*dà* 大), heaven (*tiān* 天), know (*zhī* 知), rulership (*wáng* 王), thinking (*xīn* 心), and right action (*yì* 義). Examining character frequencies, we find that terms from this topic often appear at similar rates in *Xunzi* and *Mencius*, and dissimilar rates in *Analects*. For example, 心 heart-mind is the fourteenth most frequent term in *Mencius* (126 occurrences, 7.1/1000 characters) and thirty-second in *Xunzi* (168 occurrences, 4.1/1000) but only the 255th in *Analects* (six occurrences, 0.78/1000). This is so despite the fact that Topic 33 sits at the intersection of *Mencius* and *Analects* but not all three texts. Right action (義) is the thirteenth most frequent term in *Xunzi* (315 total, 7.8/1000 characters) and the twenty-fifth in *Mencius* (107 occurrences, 6.0/1000) but sixty-third in *Analects* (twenty-four occurrences, 3.1/1000). The effect of these character level data is to raise doubts about Topic 33's ability to pull *Analects* and *Mencius* together and away from *Xunzi*.

Topic 61 ranks twenty-ninth in the corpus with a corpus weight of 0.188. Topic 61, "*Analects* Stylistics," contains keywords including Confucius's name (*kǒng* 孔), as well as three other characters used in names of followers of Confucius (*zhòng* 仲, *lù* 路, and *gòng* 貢). We infer that Topic 61 represents linguistic features of *Analects*' and *Mencius*'s literary style, particularly dialogic prose. This explains why it is prominent neither in *Xunzi* nor in the corpus as a whole. While *Analects* and *Mencius* consist mostly of reported dialogues, *Xunzi* contains lengthy essays. In sum, although Topics 33 and 61 feature among the top ten topics in *Analects* and *Mencius* (Topic 61 is number thirteen and Topic 33 is number twelve in *Xunzi*), *Xunzi* also contains high word frequencies of keywords found in Topics 33 and 61. Examination of topics at the intersection of our three texts, and at the pairwise intersection of two of three of our texts, appears to serve as evidence to shift the burden of proof onto those traditionalists who argue that Mencius is the inheritor of Confucius's mantle.

Conclusion

Topic modeling is an extremely powerful tool for the study of the intellectual tradition embodied in the extant corpus of early Chinese texts, and it is made more powerful when its machine-learning methods are combined with close-reading methods and experimental text analysis. As illustrated here, topic-modeling algorithms produced a set of topics that accurately reflects insights by scholars who use close-reading methods. Our algorithm did this without being fed prior knowledge of ancient and medieval Chinese thought and literature. We interpreted the topics with the help of expert volunteers in a process familiar from experimental text analysis; the algorithm does no interpretation. The ability to replicate such scholarly consensus is quite remarkable and underlines the robustness of topic-modeling data. More importantly, this "unsupervised" technique can uncover new or unexpected connections invisible to the individual scholar reading through the enormous early Chinese corpus on his or her own.

Textual scholars will benefit greatly from new methodologies such as topic modeling, as well as other automated, machine learning means for the "distant reading" of texts, in the near future. The widespread availability of textual corpora in digital form has yet to substantially alter the manner in which we approach our material. To date, they have been used primarily as glorified concordances. Techniques such as topic modeling represent entirely new ways to analyze and explore texts that can generate novel insights and allow us to grapple with prodigious amounts of textual material. In the end, however, the true usefulness of topic modeling lies in how it can be brought to bear on controversial questions that divide scholarly opinions. In this article, we have attempted to show how topic modeling can provide a fresh source of input that may help resolve age-old scholarly debates concerning the intellectual relationships of *Analects*, *Mencius*, and *Xunzi*.

To be sure, our topic-modeling approach has a number of limitations. First, in inexpert hands, far too many topics might be dismissed as uninterpretable "junk topics." Second, extensive polysemy in classical Chinese presents interpretive challenges for us and those who follow, for example, as Topic 86 loads into *Mencius,* 文 might be interpreted as "culture" without expert knowledge of its use in names and in other meanings (decoration, etc.). This is why topic modeling Chinese corpora requires teamwork between expert Asian studies scholars with deep familiarity with the text, humanities programmers, and statisticians. Close collaboration is essential. This will no doubt require traditional scholars to challenge themselves to overcome aversions to the use of machine learning. One goal of the research projects that have funded the project culminating in this article is not only to raise awareness among humanities scholars of the existence of such techniques, but also to demystify them and make them, and their results, more easily accessible to the scholarly community. Most importantly, it should always be emphasized that distant-reading techniques can never be a substitute for qualitative, close reading. Besides the obvious ways in which close reading is necessary for any genuine understanding of a text, the actual significance of automated results can never be assessed without use of such understanding.

To conclude, our study of the thematic relationships between *Analects*, *Mencius*, and *Xunzi* has been presented in the spirit of advancing new threads in old conversations. We are confident that the coming wave of like-minded machine-learning research, to be conducted by a new generation of researchers in philosophy, religion, and Asian studies, will lead to groundbreaking changes to our knowledge about early Confucianism—albeit only if traditional scholars are ready to receive them. We see this potential for a couple reasons.

First, while machine-learning efforts like ours are subject to several forms of bias, such biases are less than those associated with traditional close-reading methods, where scholars implicitly and explicitly loyal to their professors, to a privileged theory, and/or to Confucius (as opposed to Legalists) sometimes fall into unproductive patterns of textual commentary reduplicated across generations. Some researchers find it likely that contemporary literature about early Confucian moral philosophy contains such undesirable features in part due to its extreme culture of authority when compared to other traditions within Chinese history and across the world. An argument that one of us recently advanced for this conclusion contends that large groups of scholars use ongoing close-reading interpretations to conclude that early Confucian moral thought is best represented by one of each of a half a dozen different Western normative ethical theories (pragmatism, Aristotelian virtue theory, sentimentalism, care ethics, etc.). These theories are logically inconsistent with one another (namely, if one is true, the others must be false). This situation appears to many out-group members as a crisis. Not only does this subfield show little sign of worry, its in-group members treat the confusions about what moral theory Confucianism represents as an opportunity for more growth and publications. This is remarkable in the face of the fact that the underlying state of affairs deductively implies that the majority of these interpretations must be false (Nichols 2015). Machine-learning efforts are likely to be especially fruitful in contexts like this in which interpretive stalemates, proof-texting, allegiance to one's intellectual ancestors, or cognitive biases threaten to dominate discussion in secondary literatures.

Second, our small contribution confirms a number of scholarly opinions on several shared themes across these three documents. This is important since it suggests that our methods are sound. For example, our findings from Topic 34 support a theory that *Analects* and *Xunzi* share an "externalist" theory of human nature and moral self-cultivation, while findings from Topic 86 support attribution to *Mencius* of an "internalist" moral philosophy, confirming widely disseminated interpretations in the secondary literature.

Many of the world's literary traditions are available in digital, fully searchable form, the result of enormous effort. This new format affords exciting possibilities for supplementing, confirming, or challenging our traditional qualitative techniques with entirely new quantitative methods capable of perceiving patterns invisible to human minds. Our results call for attention to a handful of explicit issues in ancient and medieval Chinese textual studies. More broadly, we hope that our preliminary distant reading of *Analects*, *Mencius*, and *Xunzi* here gives a sense of the power, scope, and possibility of these new tools—not as replacements for our traditional modes of analyzing texts, but as sources of potential new discoveries, interventions in ongoing interpretive cruxes, and catalysts for new conversations.

Acknowledgments

This project received financial support from a Canadian Social Sciences and Humanities Research Council Partnership Grant to the University of British Columbia's Centre for Human Evolution, Cognition and Culture and its Cultural Evolution of Religion Consortium. Many thanks to "The Chinese Challenge," a Templeton World Charities Foundation project, for financial resources, and to Justin Barrett for the vision to support this new research methodology. We also thank the College of Humanities and Social Sciences, California State University, Fullerton; the University of North Carolina at

Chapel Hill; and the Interacting Minds Centre, Aarhus University. Part of this research was performed while co-author Kristoffer Nielbo was visiting the Institute for Pure and Applied Mathematics, which is supported by the National Science Foundation; part of the research was performed while co-author Edward Slingerland was an Andrew W. Mellon Foundation Fellow at the Center for Advanced Study in the Behavioral Sciences, Stanford University; and part of the research was performed while co-author Ryan Nichols was a fellow at the University of British Columbia's Centre for Human Evolution, Cognition and Culture. Techniques for generating Chinese topic clouds were developed as part of the Lexomics project funded by the National Endowment for the Humanities. We thank the many experts in ancient Chinese thought and language who replied to our questionnaire on our topics, and to Steve Angle at Warp, Weft and Way for allowing us to post our request for expert coders on that blog. We give special thanks to Robban Toleno and Brenton Sullivan for help, and to Donald Sturgeon for permission to use his CTP online textual corpus.

List of References

Ames, Roger T. 2001. "New Confucianism: A Native Response to Western Philosophy." In *Chinese Political Culture: 1989–2000*, ed. Shiping Hua, 70–99. Armonk, N.Y.: M.E. Sharpe.

Barbrook, Adrian C., Christopher J. Howe, Norman Blake, and Peter Robinson. 1998. "The Phylogeny of *The Canterbury Tales*." *Nature* 394(6696):839–40.

Bergeton, Uffe. 2013. "From Pattern to 'Culture'?: Emergence and Transformations of Metacultural Wén." PhD diss., University of Michigan.

Blei, David M. 2012a. "Probabilistic Topic Models." *Communications of the ACM* 55(4): 77–84. doi:10.1145/2133806.2133826.

——. 2012b. "Topic Modeling and Digital Humanities." *Journal of Digital Humanities* 2(1). http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/ (accessed September 24, 2017).

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3:993–1022.

Boltz, William G. 2007. "The Composite Nature of Early Chinese Texts." In *Text and Ritual in Early China*, ed. Martin Kern, 50–78. Seattle: University of Washington Press.

Brindley, Erica. 2009. "'Why Use an Ox-Cleaver to Carve a Chicken?' The Sociology of the *Junzi* Ideal in the *Lunyu*." *Philosophy East and West* 59(1):47–70. doi:10.1353/pew.0.0033.

Brooks, E. Bruce, and A. Taeko Brooks, trans. 1998. *The Original Analects: Sayings of Confucius and His Successors*. New York: Columbia University Press.

Brown, Miranda, and Uffe Bergeton. 2008. "'Seeing' Like a Sage: Three Takes on Identity and Perception in Early China." *Journal of Chinese Philosophy* 35:641–62.

Campany, Robert. 1992. "Xunzi and Durkheim as Theorists of Ritual Practice." In *Discourse and Practice*, eds. Frank Reynolds and David Tracy, 197–231. Albany: State University of New York Press.

Chen, Jack W., Zoe Borovsky, Yoh Kawano, and Ryan Chen. 2014. "The *Shishuo Xinyu* as Data Visualization." *Early Medieval China* 20:23–59. doi:10.1179/1529910414Z.00000000013.

Cheng, Anne. 1993. "Lun yǔ" 論語 [Analects]. In *Early Chinese Texts: A Bibliographical Guide*, ed. Michael Loewe, 313–23. Berkeley: Society for the Study of Early China, Institute of East Asian Studies, University of California.

Ching, Julia. 1997. "Son of Heaven: Sacral Kingship in Ancient China." *T'oung Pao* 83 (1–2):3–41.

Clark, Kelly James, and Justin T. Winslett. 2011. "The Evolutionary Psychology of Chinese Religion: Pre-Qin High Gods as Punishers and Rewarders." *Journal of the American Academy of Religion* 79(4):928–60. doi:10.1093/jaarel/lfr018.

Csikszentmihalyi, Mark. 2002. "Traditional Taxonomies and Revealed Texts in the Han." In *Daoist Identity: History, Lineage, and Ritual*, eds. Livia Kohn and Harold David Roth, 81–101. Honolulu: University of Hawai'i Press.

Dietrich, Eric. 2011. "There Is No Progress in Philosophy." *Essays in Philosophy* 12(2): 329–44.

Draper, Paul, and Ryan Nichols. 2013. "Diagnosing Bias in Philosophy of Religion." *Monist* 96(3):420–46.

Eno, Robert. 1990a. *The Confucian Creation of Heaven: Philosophy and the Defense of Ritual Mastery*. Albany: State University of New York Press.

——. 1990b. "Was There a High God *Ti* in Shang Religion?" *Early China* 15:1–26.

Feng Youlan (Fung Yu-Lan). 1952–53. *A History of Chinese Philosophy*. Princeton, N.J.: Princeton University Press.

Fù Pèiróng 傅佩榮. 2011. *Lúnyǔ sānbǎi jiǎng* 論語三百講 [The Analects: Three hundred lectures]. Taipei: Liánjīng chūbǎn gōngsī.

Goldin, Paul R. 2011. "Persistent Misconceptions about Chinese 'Legalism.'" *Journal of Chinese Philosophy* 38(1):88–104.

Goldstone, Andrew, and Ted Underwood. 2014. "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us." *New Literary History* 45(3):359–84. doi:10.1353/nlh.2014.0025.

Hall, David L., and Roger T. Ames. 1987. *Thinking through Confucius*. Albany: State University of New York Press.

Hou, Yufang, and Anette Frank. 2015. "Analyzing Sentiment in Classical Chinese Poetry." In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 15–24. Beijing: Association for Computational Linguistics and the Asian Federation of Natural Language Processing. http://www.aclweb.org/anthology/W15-3703 (accessed July 15, 2016).

Hsu [Xu], Zhuoyun. 1977. *Ancient China in Transition: An Analysis of Social Mobility, 722–222 B.C*. Stanford, Calif.: Stanford University Press.

Hunter, Michael. 2014. "Did Mencius Know the Analects?" *T'oung Pao* 100(1–3): 33–79. doi:10.1163/15685322-10013p02.

Hutton, Eric L., trans. 2014. *Xunzi: The Complete Text*. Princeton, N.J.: Princeton University Press.

Ihara, Craig K. 1991. "David Wong on Emotions in Mencius." *Philosophy East and West* 41(1):45–53.

Ivanhoe, Philip J. [1993] 2000. *Confucian Moral Self Cultivation*. 2nd ed. Indianapolis: Hackett.

——. 2008. "The Shade of Confucius: Social Roles, Ethical Theory, and the Self." In *Polishing the Chinese Mirror: Essays in Honor of Henry Rosemont, Jr.*, eds. Ronnie Littlejohn and Marthe Chandler, 34–49. New York: Global Scholarly Publications.

Jockers, Matthew L. 2013. *Macroanalysis: Digital Methods and Literary History*. Urbana: University of Illinois Press.

Kern, Martin. 2015. "Speaking of Poetry: Pattern and Argument in the 'Kongzi Shilun.'" In *Literary Forms of Argument in Early China*, eds. Joachim Gentz and Dirk Meyer, 175–200. Leiden: Brill.

Kline, T. C. 2000. "Moral Agency and Motivation in the *Xunzi*." In *Virtue, Nature and Moral Agency in the Xunzi*, eds. T. C. Kline and Philip J. Invahoe, 155–75. Indianapolis: Hackett.

Lau, D. C., trans. [1970] 2005. *Mencius*. New York: Penguin.

Loewe, Michael, ed. 1993. *Early Chinese Texts: A Bibliographical Guide*. Berkeley: Society for the Study of Early China, Institute of East Asian Studies, University of California.

Makeham, John. 1996. "The Formation of *Lunyu* as a Book." *Monumenta Serica* 44(1): 1–24.

Marshall, Emily A. 2013. "Defining Population Problems: Using Topic Models for Cross-National Comparison of Disciplinary Development." *Poetics* 41(6):701–24. doi:10.1016/j.poetic.2013.08.001.

McCallum, Andrew Kachites. 2002. "MAchine Learning for LanguagE Toolkit (MALLET)." http://mallet.cs.umass.edu/ (accessed July 15, 2016).

Mohr, John W., and Petko Bogdanov. 2013. "Introduction—Topic Models: What They Are and Why They Matter." *Poetics* 41(6):545–69. doi:10.1016/j.poetic.2013.10.001.

Moretti, Franco. 2000. "Conjectures on World Literature." *New Left Review* 1:54–68.

Nelson, Robert K. 2015. "Mining the *Dispatch*." http://dsl.richmond.edu/dispatch/pages/home (accessed July 15, 2016).

Nichols, Ryan. 2015. "Early Confucianism Is a System for Social-Functional Influence and Probably Does Not Represent a Normative Ethical Theory." *Dao* 14(4):499–520. doi:10.1007/s11712-015-9464-8.

Overmyer, Daniel L., David N. Keightley, Edward L. Shaughnessy, Constance A. Cook, and Donald Harper. 1995. "Chinese Religions—The State of the Field Part I: Early Religious Traditions: The Neolithic Period through the Han Dynasty, ca. 4000 B.C.E. to 220 C.E." *Journal of Asian Studies* 54(1):124–60.

Qu Wanli. 1983. *Shījīng quánshì* 詩經詮釋 [Explanatory notes to the *Book of odes*]. Taipei: Guójiā túshūguǎn chūbǎnshè.

Rhody, Lisa M. 2012. "Topic Modeling and Figurative Language." *Journal of Digital Humanities* 2(1). http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/ (accessed July 15, 2016).

Schmidt, Benjamin M. 2012. "Words Alone: Dismantling Topic Models in the Humanities." *Journal of Digital Humanities* 2(1). http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/ (accessed July 15, 2016).

Slingerland, Edward. 2000. "Why Philosophy Is Not 'Extra' in Understanding the *Analects*." Review of *The Original Analects* by E. Bruce Brooks and A. Taeko Brooks. *Philosophy East and West* 50(1):137–41.

——. 2003. *Effortless Action: Wu-wei as Conceptual Metaphor and Spiritual Ideal in Early China*. New York: Oxford University Press.

Slingerland, Edward, and Maciej Chudek. 2011. "The Prevalence of Mind-Body Dualism in Early China." *Cognitive Science* 35(5):997–1007. doi:10.1111/j.1551-6709.2011.01186.x.

Slingerland, Edward, Ryan Nichols, Kristoffer Nielbo, and Carson Logan. Forthcoming. "The Distant Reading of Religious Texts A 'Big Data' Approach to Mind-Body Concepts in Early China." *Journal of the American Academy of Religion*.

Twitchett, Denis Crispin, and Michael Lowe. 1986. *The Cambridge History of China Volume 1: The Ch'in and Han Empires, 221 BC–AD 220*. Cambridge: Cambridge University Press.

Underwood, Ted. 2012a. "Topic Modeling Made Just Simple Enough." *The Stone and the Shell*, April 7. http://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/ (accessed July 15, 2016).

——. 2012b. "What Kinds of 'Topics' Does Topic Modeling Actually Produce?" *The Stone and the Shell*, April 1. http://tedunderwood.com/2012/04/01/what-kinds-of-topics-does-topic-modeling-actually-produce/ (accessed July 15, 2016).

Van Norden, Bryan W. 1992. "Mengzi and Xunzi: Two Views of Human Agency." *International Philosophical Quarterly* 32(2):161–84. doi:10.5840/ipq199232212.

——. trans. 2008. *Mengzi: With Selections from Traditional Commentaries*. Indianapolis: Hackett.

Wáng Xiānshèn 王先慎, ed. 2006. *Hánfēizi jíjiě* 韓非子集解 [Collected commentaries on the *Hanfeizi*]. Beijing: Zhōnghuá shūjú.

Weingart, Scott. 2012. *Topic Modeling for Humanists: A Guided Tour*. http://www.scottbot.net/HIAL/?p=19113 (accessed July 15, 2016).

Wills, John Elliot. 2012. *Mountain of Fame: Portraits in Chinese History*. Princeton, N.J: Princeton University Press.

Wilson, Thomas. 2014. "Spirits and the Soul in Confucian Ritual Discourse." *Journal of Chinese Religions* 42(2):185–212. doi:10.1179/0737769X14Z.00000000013.

Wong, David B. 1991. "Is There a Distinction between Reason and Emotion in Mencius?" *Philosophy East and West* 41(1):31–44. doi:10.2307/1399716.

Xú Fuguan 徐復觀. 1969. *Zhongguo renxing lun shi: Xian Qin pian* [History of Chinese views on human nature: The pre-Qin period]. Taipei: Commercial Press.

Zhang, Longxi. 2012. *The Concept of Humanity in an Age of Globalization*. Göttingen: V&R Unipress.

Zipf, George Kingsley. 1949. *Human Behaviour and the Principle of Least Effort*. Reading, Mass.: Addison-Wesley.

Appendix 1. Texts, Genres, and Dates.

Text	*Genre*	*Era*
Analects (論語)	Confucianism (儒家)	WS
Mengzi (孟子)	Confucianism (儒家)	WS
Liji (禮記)	Confucianism (儒家)	WS
Xunzi (荀子)	Confucianism (儒家)	WS
Xiao Jing (孝經)	Confucianism (儒家)	WS
Shuo Yuan (說苑)	Confucianism (儒家)	Han
Chun Qiu Fan Lu (春秋繁露)	Confucianism (儒家)	Han
Han Shi Wai Zhuan (韓詩外傳)	Confucianism (儒家)	Han
Da Dai Li Ji (大戴禮記)	Confucianism (儒家)	Han
Baihutong (白虎通)	Confucianism (儒家)	Han
Xin Shu (新書)	Confucianism (儒家)	Han
Xin Xu (新序)	Confucianism (儒家)	Han
Yangzi Fayan (揚子法言)	Confucianism (儒家)	Han
Zhong Lun (中論)	Confucianism (儒家)	Han
Kongzi Jiayu (孔子家語)	Confucianism (儒家)	Han
Qian Fu Lun (潛夫論)	Confucianism (儒家)	Han
Lunheng (論衡)	Confucianism (儒家)	Han
Tai Xuan Jing (太玄經)	Confucianism (儒家)	Han
Fengsu Tongyi (風俗通義)	Confucianism (儒家)	Han
Kongcongzi (孔叢子)[14]	Confucianism (儒家)	Han
Shen Jian (申鑒)	Confucianism (儒家)	Han
Zhuangzi (莊子)	Daoism (道家)	WS
Dao De Jing (道德經)	Daoism (道家)	WS
Liezi (列子)	Daoism (道家)	Post-Han
He Guan Zi (鶡冠子)	Daoism (道家)	Han
Wenzi (文子)	Daoism (道家)	Han
Wen Shi Zhen Jing (文始真經)	Daoism (道家)	Post-Han
Lie Xian Zhuan (列仙傳)	Daoism (道家)	Post-Han
Yuzi (鬻子)	Daoism (道家)	WS
Heshanggong (河上公)	Daoism (道家)	Han
Hanfeizi (韓非子)	Legalism (法家)	WS
Shang Jun Shu (商君書)	Legalism (法家)	WS
Shen Bu Hai (申不害)	Legalism (法家)	WS
Shenzi (慎子)	Legalism (法家)	WS
Jian Zhu Ke Shu (諫逐客書)	Legalism (法家)	WS
Guanzi (管子)	Legalism (法家)	WS
Mozi (墨子)	Mohism (墨家)	WS
Mo Bian Zhu Xu (墨辯注敘)	Mohism (墨家)	Post-Han
Gongsunlongzi (公孫龍子)	School of Names (名家)	Post-Han
The Art of War (孫子兵法)	School of the Military (兵家)	WS
Wu Zi (吳子)	School of the Military (兵家)	WS

Continued

[14] As observed by Kern (2015, 189), "traditionally dated to the late third century BCE but most likely composed only in Eastern Han times, even if including earlier material."

Appendix 1 (*contd.*)

Text	*Genre*	*Era*
Liu Tao (六韜)	School of the Military (兵家)	WS
Si Ma Fa (司馬法)	School of the Military (兵家)	WS
Wei Liao Zi (尉繚子)	School of the Military (兵家)	Han
Three Strategies (三略)	School of the Military (兵家)	Han
Hai Dao Suan Jing (海島算經)	Mathematics (算書)	Han
The Nine Chapters (九章算術)	Mathematics (算書)	Han
Sunzi Suan Jing (孫子算經)	Mathematics (算書)	Post-Han
Zhou Bi Suan Jing (周髀算經)	Mathematics (算書)	Han
Huainanzi (淮南子)	Miscellaneous Schools (雜家)	Han
Lü Shi Chun Qiu (呂氏春秋)	Miscellaneous Schools (雜家)	WS
Gui Gu Zi (鬼谷子)	Miscellaneous Schools (雜家)	Han
Yin Wen Zi (尹文子)	Miscellaneous Schools (雜家)	WS
Deng Xi Zi (鄧析子)	Miscellaneous Schools (雜家)	WS
Shiji (史記)	Histories (史書)	Han
Chun Qiu Zuo Zhuan (春秋左傳)	Histories (史書)	WS
Lost Book of Zhou (逸周書)	Histories (史書)	WS
Guo Yu (國語)	Histories (史書)	WS
Yanzi Chun Qiu (晏子春秋)	Histories (史書)	WS
Wu Yue Chun Qiu (吳越春秋)	Histories (史書)	Han
Yue Jue Shu (越絕書)	Histories (史書)	Han
Zhan Guo Ce (戰國策)	Histories (史書)	WS
Yan Tie Lun (鹽鐵論)	Histories (史書)	Han
Lie Nü Zhuan (列女傳)	Histories (史書)	Han
Guliang Zhuan (穀梁傳)	Histories (史書)	Han
Gongyang Zhuan (公羊傳)	Histories (史書)	Han
Han Shu (漢書)	Histories (史書)	Han
[Qian] Han Ji ([前]漢紀)	Histories (史書)	Han
Dong Guan Han Ji (東觀漢記)	Histories (史書)	Han
Hou Han Shu (後漢書)	Histories (史書)	Post-Han
Zhushu Jinian (竹書紀年)	Histories (史書)	Han
Mutianzi Zhuan (穆天子傳)	Histories (史書)	WS/Han[15]
Gu San Fen (古三墳)	Histories (史書)	Post-Han
Yandanzi (燕丹子)	Histories (史書)	Post-Han
Xijing Zaji (西京雜記)	Histories (史書)	Post-Han
Book of Poetry (詩經)	Ancient Classics (經典文獻)	Pre-WS
Shang Shu (尚書)	Ancient Classics (經典文獻)	Han
Book of Changes (周易)	Ancient Classics (經典文獻)	WS
The Rites of Zhou (周禮)	Ancient Classics (經典文獻)	WS
Chu Ci (楚辭)	Ancient Classics (經典文獻)	WS
Yili (儀禮)	Ancient Classics (經典文獻)	WS
Shan Hai Jing (山海經)	Ancient Classics (經典文獻)	Han
Jiaoshi Yilin (焦氏易林)	Ancient Classics (經典文獻)	Han
Jingshi Yizhuan (京氏易傳)	Ancient Classics (經典文獻)	Song (forgery)[16]

Continued

[15] Text in six parts. Parts 1–4 are authentic 350 BCE texts. Part 5 is a post-Han addition. Part 6 is a compilation of WS texts.

[16] Probably a Song forgery. Twitchett and Lowe (1986, 692) state that the *Jingshi yizhuan* is not authentic, but was written during the Song dynasty.

Appendix 1 (*contd.*)

Text	*Genre*	*Era*
Shi Shuo (詩說)	Ancient Classics (經典文獻)	Post-Han
Shuo Wen Jie Zi (說文解字)	Etymology (字書)	Han
Er Ya (爾雅)	Etymology (字書)	WS
Shi Ming (釋名)	Etymology (字書)	Han
Fang Yan (方言)	Etymology (字書)	Han
Ji Jiu Pian (急救篇)	Etymology (字書)	Han
Huangdi Neijing (黃帝內經)	Chinese Medicine (醫學)	Han
Nan Jing (難經)	Chinese Medicine (醫學)	Han
Shang Han Lun (傷寒論)	Chinese Medicine (醫學)	Han
Jinkui Yaolue (金匱要略)	Chinese Medicine (醫學)	Han
Guodian(郭店)	Excavated texts (出土文獻)	WS
Mawangdui (馬王堆)	Excavated texts (出土文獻)	Han

Appendix 2. Stopwords.

之	是	于	元	后	哉	還	甚	求	氏	焉
不	與	在	正	作	難	絕	本	說	外	我
也	夫	非	多	因	稱	往	止	左	同	復
以	可	六	西	雖	屬	己	興	起	受	千
而	五	諸	足	始	宜	邪	耳	會	反	亦
其	將	必	又	里	聽	固	廣	定	少	九
為	使	然	高	請	終	首	益	通	常	七
曰	何	若	內	女	遠	由	應	對	過	方
者	至	及	當	右	盡	共	十	所	此	乃
子	四	未	去	敢	異	徒	則	故	太	百
有	矣	萬	北	前	進	任	無	三	謂	皆
於	自	吾	來	易	初	更	一	二	如	乎

Appendix 3. Survey Given Independent Coders.

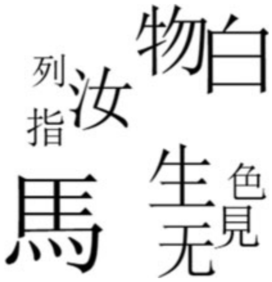

Figure 5. Word cloud for Topic 27.

Survey Text

1. Suppose you had to guess what is the theme of this word cloud. What are one to three English words you would use to describe this theme?

2. Please indicate how confident you are about your answer to the previous question by using the slider bar below.

0 = Completely Uncertain 7 = Completely Certain
0 1 2 3 4 5 6 7

3. Consider the categories below. Please select ALL categories into which you believe the content of this word cloud belongs.

Virtue or Morality Philosophy Religion Military Uncategorizable
Knowledge Fortune or Luck Mysticism Mind Leadership
Politics Body Cosmos

4. Since you selected "Mind" among the previous answers, please select ALL concepts below that represent the content of the word cloud.

Cognition Emotion Belief Rationality Feelings Perception
Judgement Skill Soul Memory

5. Since you selected "Military" among the previous answers, please select ALL concepts below that represent the content of the word cloud.

Victory Government Weaponry State Peace Violence Order War

6. Since you selected "Politics" among the previous answers, please select ALL concepts below that represent the content of the word cloud.

Lord Emperor Statecraft Minister Sage King Law Official

7. Since you selected "Fortune" among the previous answers, please select ALL concepts below that represent the content of the word cloud.

Weather Dates Fate Calendar Law

8. Since you selected "Cosmos" among the previous answers, please select ALL concepts below that represent the content of the word cloud.

Sagehood Ability Seasons Human Benefit World

9. Since you selected "Knowledge" among the previous answers, please select ALL concepts below that represent the content of the word cloud.

Earth Reflection Dao World Human Culture

10. Since you selected “Virtue” among the previous answers, please select ALL concepts below that represent the content of the word cloud.

Speak   Order   Life   Desire   Wisdom   Goodness   Worthy
Peace   Respect

11. Since you selected “Leadership” among the previous answers, please select ALL concepts below that represent the content of the word cloud.

Royalty   Statecraft   King   Education   Family

12. Since you selected “Philosophy” among the previous answers, please select ALL concepts below that represent the content of the word cloud.

Language   Emperor   Ruism   Sage King   Qi   Mencius
Confucius   Logic

13. Since you selected “Body” among the previous answers, please select ALL concepts below that represent the content of the word cloud.

Medicine   Health   Bodily   organs   Yin   Medicine   Qi   Biology
Yang

14. Since you selected “Religion” among the previous answers, please select ALL concepts below that represent the content of the word cloud.

Gods   Spirit   Sacrifice   Religion   Heaven   Deities

15. Since you selected “Mysticism” among the previous answers, please select ALL concepts below that represent the content of the word cloud.

Divination   Ritual   Sacrifice   Spirit   Deities   Mourning
Qi   Gods