The Distant Reading of Religious Texts

A “Big Data” Approach to Mind-Body Concepts in Early China

Edward Slingerland, Ryan Nichols, Kristoffer Nielbo, Carson Logan

The claim that traditional Chinese thought was characterized by mind-body holism is commonly encountered in the field, with a venerable pedigree (e.g., Granet 1934; Jullien 2007; Lévy-Bruhl 1922; Rosemont & Ames 2009; see Slingerland 2013 for a review). Scholars agree that, if there is a word for “mind” in classical Chinese, it would be *xin* 心, variously translated as “mind,” “heart” or “heart-mind.” *Xin* refers literally to the organ of the heart, but is also the locus of cognitive and emotional function. Defenders of what we term “strong mind-body holism” claim that, although the *xin* may possess its own unique functions in the early Chinese view, this is no different from the eye or ear possessing their own specific functions. This position, therefore, holds that the *xin* is not uniquely contrasted with the body, and that itwas viewed as simply one among a set of embodied organs (Geaney 2002).

Previous work (Goldin 2003, 2015, Slingerland 2013, Poli 2016) has reviewed qualitative textual and archaeological evidence that contradicts this claim, as well as cognitive science evidence that suggests that at least a “weak” form of mind-body dualism is a human cognitive universal. Unlike Cartesian, or “strong,” mind-body dualism, which postulates a razor-sharp divide between two ontological realms, cognitively natural dualism acknowledges that mind and body, although qualitatively distinct, overlap in various respects (Bloom 2004; Cohen et al. 2011). For instance, the mind is unique in being the seat of cognition, rational planning and thought, free will, and personal identity, partly because it is less material in nature than the other components of the self. The body—which, for the early Chinese, includes the organs o ther than the *xin* (“heart-mind”)—is something one can possess, or lose, but the mind/*xin* is central to one’s identity and sense of self.

In addition to this more traditional evidence against the strong mind-body holist position, several years ago the results of a methodologically novel, team-based coding project (Slingerland & Chudek 2011) bolstered the evidence typically presented in such arguments with more quantitative evidence. This study assessed evidence for and against several hypotheses about the mind in China using a corpus of pre-Qin (pre-221 BCE) classical Chinese texts. To do so, a large quantity of passages containing the keyword *xin* 心 were sampled, and human coders were asked to characterize the functions of *xin* and how *xin* -body relations are characterized. It was found that *xin* was frequently contrasted with the body, significantly more than any other organ in the body, a contrast that grew stronger over time. With regard to the function of *xin*, coders judged that, although *xin* seems to encompass both emotional and cognitive functions (and rarely refers to the actual physical organ in the body), by the Early Warring States (c. 5th c BCE), cognitive functions outnumber emotional functions by 80% to 10%, a pattern that remained stable through the Late Warring States period.

A critique of this study by Klein & Klein 2011 included charges that the study was biased by drawing a large proportion of its Pre-Warring States (before 5th century BCE) sample from a single text, the *Book of Odes*, a collection of poetry one would expect to contain an unusually high number of emotion words. Another concern voiced about the study is that a large proportion of the texts analyzed could be classed as philosophical works, which might exaggerate how much *xin* is portrayed as possessing cognitive functions. A final and more pervasive worry expressed by Klein and Klein and other critics concerns the degree to which human coders, making qualitative judgments of a given passage’s meaning, might have their judgments skewed by prior assumptions or philosophical prejudices.

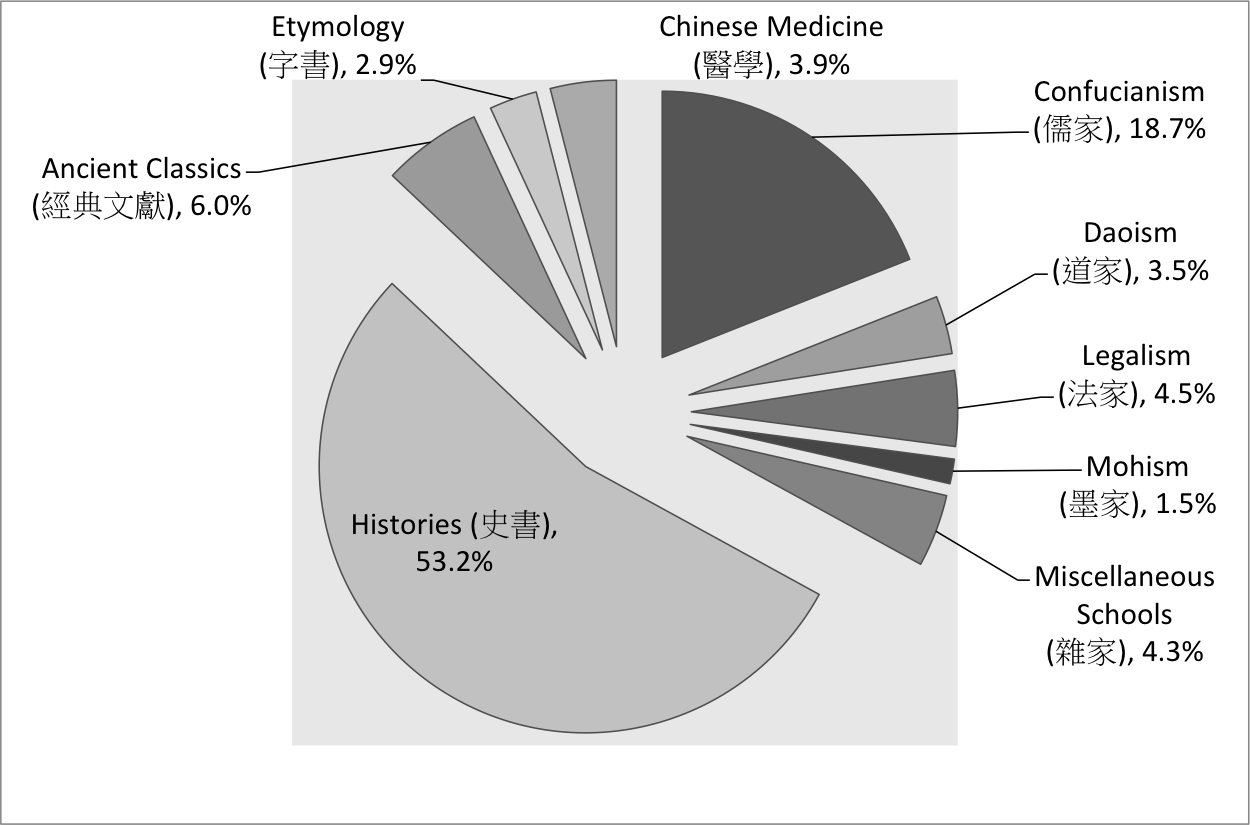
**Machine-Assisted Approaches to Religious and Philosophical Text Analysis**

Here we present a series of studies that respond to these and other critiques by applying machine-assisted, large-scale textual analysis techniques to a radically-expanded textual corpus. The immediate, and more narrow, purpose is to explore conceptions of mind and body in early China. More broadly, however, we wish to demonstrate to scholars of religion the value of supplementing our traditional close reading practices with various techniques for “distant reading” (Moretti 2013) or computer-aided analysis of texts (Rockwell & Sinclair 2016). There are a host of new methodologies for navigating massive textual corpora that have been available to scholars of religion for some time now—in some cases, a decade or two—but that remain surprisingly underutilized.

It is a sign of how conservative academic disciplines are that the manner in which scholars marshal supporting textual evidence has not changed much in the last millennium or two, despite the availability of entirely unprecedented digital tools. The Thesaurus Linguae Graecae (TLG; <http://stephanus.tlg.uci.edu/>) has been available to scholars of ancient Greece since the 1970s. For sinologists, the vast majority of the received corpus of traditional Chinese texts is available for free, on-line, in easily searchable form through a variety of sites. The Buddhist Canon, as represented in the full 85 volume *Taishō Shinshū Daizōkyō* 大正新脩大藏經 (<http://21dzk.l.u-tokyo.ac.jp/SAT/index_en.html>), is now available in searchable, on-line form, and similar resources are available for other religious and philosophical traditions around the world. Nevertheless, to date, these digital corpora have tended to be used as merely glorified concordances—as more convenient versions of tools we already had. The unprecedented and exciting analytic strategies provided by these resources have rarely been explored in Religious Studies, though other fields, especially literary studies, have taken steps toward distant reading (Moretti 2007, 2013), algorithmic criticism (Ramsay 2011), and text analysis through topic modeling (Jockers & Mimno 2013a).

The present study takes advantage of a massive textual dataset composed of 96 texts totalling 5.7m characters, compiled by Dr. Donald Sturgeon in the “Chinese Text Project” (CTP; [www.ctext.org](http://www.ctext.org)), and freely available on-line.[[1]](#footnote-1) Metadata about the period and genre of each text was included in our dataset (see Appendix 1). The vast historical sweep of our corpus means we include texts from Pre-Warring States (prior to 5th c. BCE) through the Warring States and Han Dynasty (206 BCE-220 CE), as well as a small number of post-Han texts dating up to the Song Dynasty (960-1279 CE).[[2]](#footnote-2) By increasing the number of texts we consider, we are able to meet concerns about corpus size, lack of genre diversification, and limited periodization. The great literary breadth of our corpus includes genres such as medical, military, mathematical, and historical texts (see Figure 1)[[3]](#footnote-3) and answers worries that the sample is biased in favor of poetry or philosophy. Our fully-automated information retrieval and analysis allows us to handle such a massive corpus and also respond to concerns about potential biases in human coders.

Figure 1. CTP Genre or Schools with More Than 1.5% Representation in Corpus



The textual analysis techniques we demonstrate below are obviously no substitute for traditional, close readings of texts. Indeed, most of the results are incomprehensible without the interpretative skills of experts deeply familiar with the corpus in question. As we will see, however, the sort of high altitude, broad perspective on a corpus provided by these techniques can serve as an important check on our qualitative interpretations. Moreover, there may be certain types of questions—for instance, assessing the validity of claims about general trends or prevalent themes in a given corpus—that are best addressed through machine-assisted techniques coupled with statistical analysis. As we will try to demonstrate below, the great strength of distant reading is the ability to pick up trends or patterns in large quantities of data that may be invisible to individual, human “close” readers.

**Methods, Assumptions and Goals**

In order to use a textual corpus, such as the Chinese Text Project (CTP) corpus, a certain amount of pre-processing is necessary. We applied a stop-word list to the CTP, removing overly common function words and articles.[[4]](#footnote-4) We also removed all punctuation besides sentence-ending punctuation. Inclusion of sentence-ending punctuation allowed us to use the sentence, a natural unit of semantic meaning, as our primary unit of analysis. Once the corpus was pre-processed in this manner, we explored it with a variety of analytical tools, as described below.

Strong mind-body holism attributes a broadly monist metaphysics to the authors of the classical Chinese texts in our corpus, according to which the mind and body share one and the same type of substance, and the *xin* (the most likely candidate to represent the concept of “mind”) is in no way qualitatively distinct from the other organs in the body (Ames 1993; Geaney 2002; Jullien 2007). We would expect that, if the authors of the texts in the CTP corpus *tacitly* endorsed strong mind-body holism, we would find *xin* behaving just like any other bodily organ term in proximity to the three most common words for “body” in classical Chinese (*shen* 身, *xing* 形, and *ti* 體).

It is important, in this regard, to distinguish implicit from explicit cognition. Work in various branches of the cognitive sciences has documented the “dual system” nature of human cognition (Evans 2008; Kahneman 2011). According to this mode, the explicit, “cold,” conscious aspect of the human mind rides upon a much larger, more powerful and pervasive implicit, “hot,” mostly unconscious system. People are capable of entertaining and debating any number of propositions in their explicit systems. How many angels can dance on the head of pin? If body and mind are two distinct ontological substances, how could they ever interact? As a growing literature on “theological correctness” in the cognitive science of religion has documented, however, explicit claims or endorsed beliefs do not necessarily reflect underlying beliefs and behavior. Hindus may assert in surveys that their gods are omniscient and omnipotent, whereas in narrative judgment tasks reveal that they implicitly believe them to be subject to anthropomorphic limits (Barrett 1998). Calvinists may profess to belief in predestination, but we still observe them praying on the weekends for God to favor their football team, and otherwise behaving in ways that suggest they believe their actions can change the future (Slone 2004).

When it comes to mind-body dualism in early China, it is certainly the case that one can find explicit endorsements of the strong mind-body holist position. The Confucian thinker Mencius famously remarked, in *Mencius* 6:A:7:

With regard to the mouth, all palates find the same things tasty; with regard to the ears, all find the same things pleasant to listen to; with regard to the eyes, all find the same things beautiful. Now, when it comes to the *xin*, is it somehow unique in lacking such common preferences? What is it, then, that minds share a preference for? I say that it is order and rightness. (Van Norden 2008: 151)

This passage is frequently cited by defenders of strong mind-body holism as evidence that the early Chinese saw the *xin* as equivalent to the other organs. As Jane Geaney comments, the point of the analogy set up in *Mencius* 6:A:7, as well as similar statements in other texts such as the *Xunzi*, is that “the heartmind and the senses have certain things in common—they function on the same principles regarding space and time, and they share the tendency to prefer similar things. The fact that the senses serves as analogies for the heartmind makes it unlikely that they are radically different in nature” (Geaney 2002: 101)

As Edward Slingerland has observed (2013: 19), however, to draw this conclusion from passages such as 6:A:7 is to mistake an explicit claim for a background assumption. Mencius is making an argument, which he no doubt expects to be surprising or counterintuitive, that the *xin* has a natural “taste” in the same way the other organs do. This would be a nonsensical, or, at best, a trivial, statement to make if he, and his readers, implicitly and intuitively *believed* this to be true. If we are going to claim that mind-body dualism is completely foreign to the early Chinese worldview, we need to look at background assumptions as well as explicit rhetorical claims.

The fundamental hypothesis we put to the test in the studies reported below is that, whatever early Chinese thinkers might explicitly *say* about *xin*-body relations, an analysis of the overall patterns of language use would reveal that the *xin* was at least implicitly understood as being unlike any other organ, with a unique relationship to the physical body. Our basic assumption is that large-scale patterns of language use will tell us something about implicit cognition. Even if Mencius claims that the *xin* is the same as the other organs, if we find that he, and other early Chinese writers, habitually mention *xin*, and only *xin*, as opposed to the other organ terms, in relation to the body, this suggests that *xin* occupies a distinct cognitive space in the early Chinese mindset.

To evaluate this hypothesis, we performed two collocation analysis studies, Studies 1 and 2, where we analyze the patterns of co-occurrence between various key terms in our corpus. In Study 1, we analyze semantically uncontroversial pairs of terms in order to see if we can establish some semantic benchmarks—that is, determine whether or not particular collocation patterns correspond to particular semantic relationships such as contrastive pairs (“many”::“few”) or part-whole (“wheel”::“cart”). In Study 2 we then present collocation data describing *xin*’s relationship with the three body terms, as well as similar data describing other bodily organs’ relationships with the body terms and some simple statistics that compare the two sets of data. The goal is not only to note differences between *xin* and the other organs, but also to see if any of the observed collocation patterns match the semantic benchmarks established in Study 1.

In Studies 3 and 4, we turn to “unsupervised” methods of textual analysis, which involve machine-learning techniques for processing and analyzing patterns in textual corpora (Dan Jurafsky & Martin 2009, Miner et al. 2012). Specifically, our paper uses two methods, hierarchical clustering and topic modeling. Hierarchical clustering is a form of unsupervised information retrieval used to extract structured information from unstructured data (C. D. Manning, Raghavan, & Schütze 2008). Topic models are generative probabilistic models that seek out sets of words (“topics”) that reliably travel together throughout either a document or a corpus of documents (D. Blei, Ng, & Jordan 2003).

A great advantage of unsupervised methods is that they make no assumptions about target terms of interest or size of word window for collocations, let alone about experimenters’ potential hypotheses about the texts in question. They therefore provide a relatively objective measure of relationships between lexical items in the corpus. Although they are methodologically quite distinct from collocation analysis, with both of these methods, hierarchical clustering and topic modeling analysis, we make the same prediction: if the early Chinese authors were strong mind-body holists, *xin* should behave like the other organs of the body in their writings. *Xin* and other organ terms should all appear with equal frequencies in significant topics, and should not differ with regard to how they cluster vis-a-vis the body terms, or other terms for cognition or emotion.

**Introduction to Studies 1 and 2: Collocation Analysis**

Collocation analyses involve measurement of how frequently, and how closely, terms of interest occur with regard to one another in a textual corpus, and inference from those measures to, typically, syntactic features of words and parts of speech. In the field of corpus linguistics, this technique has long been used to track patterns and changes in idiom usage, bigrams, etc. Most humanities scholars, however, would be more interested in the semanticimplications of word collocation. There have been some advances on this front (S. T. Gries 2013; Daniel Jurafsky & Martin 2015; Rohde, Gonnerman, & Plaut 2005), as well as practical demonstrations of how “dumb” collocation pattern extractors can “learn” something about natural language semantics. Collocation patterns have been used, for instance, to train a machine-learning algorithm to perform reasonably well on the multiple-choice synonym questions found in the Tests of English as a Foreign Language (TOEFL) exam (Landauer & Dumais 1997). Indeed, as Bullinaria & Levy 2007 observe, semantic inferences generated from experienced collocation patterns in word use probably play a central role in how infants acquire language (2007: 510). A literature deriving inferences from textual collocation patterns in a variety of genres (diaries, prose, sermons, emails, survey responses and more) to patterns of thought and emotion in their authors has also been growing (see Teubert & Čermáková 2007; Sampson & McCarthy 2005; C. Manning & Schütze 1999). With regard to classical Chinese, a study by Lee & Wong 2012 analyzed collocation patterns in the *Complete Tang Poems* to show affinities between, for instance, particular seasons and distinctive semantic classes of words.

Corpus linguistics techniques often begin with collocational studies focusing on a particular target word, or word set, in an effort to explore associations between the target word and other terms in a large corpus. Of studies that would be of interest to religious studies scholars, we can look at three focused on portrayals of Islam in contemporary sources that help to illustrate the breadth and potential of these techniques. Using a large corpus of contemporary British newspapers, Baker, Gabrielatos and McEnery pose the following research question: “What does a collocational analysis of the word *Muslim* reveal about the construction of this group?” (Baker, Gabrielatos, & McEnery 2013: 260). To conduct their study, they used a part-of-speech tagging instrument that assigns to each token word a code indicating whether it is a noun, an adjective, etc. The most frequent noun collocates of the adjective *Muslim* found in this study were “community” (1st), “world” (2nd), and “woman” (3rd), though “cleric” (6th) and “extremist” (10th) appeared in the top 10 (2013: 261). After manually coding collocates of *Muslim*, the authors conclude that the term’s use in the British press is rarely of a specifically religious nature, and more frequently serves as either a reference to national or ethnic identity, what they call a “shortcut to typical, or differentiating, attributes” (265).

In a similar study using two corpora, Al-Hejin tracks the contrastive representations of Muslim women in Arab online news media and BBC online news media (Al-Hejin 2015). Al-Hejin identifies that *Muslim women* and cognate terms were primarily found in a semantic macrostructure concerning war and violence. Muslim women were also found to be portrayed primarily as victims (2015: 39-40). Al-Hejin puts that study’s collocational data in conversation with surveys of public opinion, both in Britain and the Arab world, to offer a fresh perspective on the portrayal of Muslim women in international media beyond traditional forms of commentary based upon one, or a handful, of written works.

Finally, in an examination of 250 statements promoting Islamic extremist violence, Prentice, Rayson, & Taylor 2012) use mixed methods, including collocation, to identify patterns and infer motives and ends. 160 of the 250 documents they studied were authored by well-known groups, including Al-Qa’ida, Hamas and Hezbollah, and all promoted Islamic terrorism. Also using a part-of-speech tagger and a semantic tagger, researchers assigned words to semantic fields (2012: 266). The authors then used collocation analyses to identify commonly occurring words within each semantic field. They then repeated the same set of procedures with two comparison corpora (one written, one spoken) comprised of neutral content, which served as ‘control conditions’ in their study. Core terms from Islam play a large role in marking the unique features of their extremist corpus. Its most important keywords are *Allah*, *Muslim*, *jihad*, *Mujahideen*, *Islam* and *against* (2012: 269-70). Collocational research into the extremist corpus revealed fascinating patterns in which authors polarize their readers by using semantic pairings emphasizing sharp contrasts. Among the most important were darkness/light, success/failure, serenity/violence, selfish/unselfish, and religious/non-religious (2012: 273). Prentice and co-authors also generate collocational data within specific semantic fields, allowing them to identify, for example, collocates of terms that fall into the ‘Places’ field. This feature of their study revealed noticeable rates of strategic thinking in the form of the collocation of place names with terms such as *enter, stationing,* and *fortified* (2012: 277). These studies demonstrate the value of collocational analyses to reveal hidden features of semantic content in discursive practices related to religious identity.

*Methods for Statistical Analysis*

These three studies have in common an interest in religion, the use of collocational techniques for textual analysis, and the broader goal of revealing semantic relationships often difficult to make out when studying a handful of texts at a time. One difference, however, is concerning. Each of the three studies mentioned above employs a different statistical test when evaluating results of corpus collocations. In addition to working with word counts and raw frequencies of collocations, Baker, Gabrielatos and McEnery use logdice; Al-Hejin uses loglikelihood; Prentice, Rayson, & Taylor 2012. This variation might be confusing to humanities scholars unfamiliar with statistics, and also provoke suspicions that these studies’ authors were picking the measure most likely to give them the answers they wanted. For this reason, before proceeding to a discussion of our studies, it is worthwhile to first discuss the strengths and weaknesses of the various statistical measures employed in large-scale textual analysis.

First, it is important to see that different collocational statistics are suitable for different research questions. For example, though nearly all measures reflect the association of word1 and word2 to each other, most do not distinguish between whether word1 predicts word2, or vice versa. This will not matter in collocational studies where the directionality of relationships is unimportant. Nonetheless, this problem is frequently overlooked by researchers who assume a collocational statistic describes reciprocal association (Ellis & Ferreira-Junior, 2009). In a demonstration of this problem, Stefan Greis calculates common collocational statistics including Mutual Information, T-score, Dice, and others for word1 (‘of’) and word2 (‘course’) using the British National Corpus (S. Gries 2013: 145). These statistics all reveal an extremely strong relationship between the pair of terms, but these values are almost entirely composed by the one-way association of ‘course’ predicting ‘of’.

Some corpus linguists (e.g. Michelbacher, Evert & Schütze, 2007) argue that this casts serious doubt on using traditional collocational statistics, and such concerns have led to the development of collocational tests, such as ΔP (Gries, 2013), that take into account the directionality of collocation. Unfortunately, these techniques are often only of use in highly restricted environments. For example, ΔP can only be calculated for terms adjacent to one another. This makes it useful for analyzing linguistic compounds, but renders it unavailable for use by researchers like us interested in looking at the co-occurrence of keywords within a broader textual window, often referred to as a “KWIC” (keywords-in-context) window. Because we are not concerned about radically asymmetric collocational relationships, the directionality issue found in English between ‘of’ and ‘course’ does not affect our study. Furthermore, it is worth keeping in mind that, in Gries’s studies, mutual information, T-score, logdice, minimal sensitivity and others were all roughly consistent.

Of the statistical tests developed by pioneers in the field of corpus linguistics, a handful are particularly relevant for our purposes. We are interested in the collocation of *xin*, other organ terms, and physical body terms within sentences. The most widespread test statistic in corpus linguistics is mutual information (MI), an extended measure of collocational strength drawn from information theory. MI is calculated by dividing the observed frequency of a co-occurring word within a specific KWIC window by the expected frequency of a co-occurring word in that window, and taking the logarithm to the base 2 of the result (Biber & Jones 2009: 1287). In other words, it measures the association strength between two words by comparing their probability of appearing together, while also considering their individual distributions (K. W. Church & Hanks 1990). The result of this calculation offers a score that indicates the strength of the relationship between two terms, or between two sets of terms. Since MI compares the observed co-occurrence of two words to what would be expected at chance level—that is, if the words were independent—interpretation of MI is straightforward. A positive score indicates that the words are more strongly associated than chance would predict, a negative score indicates that they are anti-associated, and a score of zero indicates that the association is at chance level (Oakes 1998, Paperno et al. 2014).

MI is good for tracking associations between words for which the joint probability distribution is similar to the words’ individual probabilities distributions—that is, where the words in question are more or less equally common in the corpus. If one of the associated words occurs very infrequently in the overall corpus and the other frequently, however, the association is likely to be a product of chance. “MI score highlights lexical items that are relatively infrequent by themselves but have a higher-than-random probability of co-occurring with the node word” (Mautner 2007: 55), but it “gives too much weight to rare events” (Oakes 1998: 171). That is, MI scores for collocations involving rare words might be artificially high. When it comes to comparing *xin* to other organ terms, this is a particular worry. *Xin* is quite common in the corpus, whereas several of the organ terms we compare it with appear only rarely, especially outside of the medical genre.

We employ several approaches proposed to remedy this problem (Paperno et al. 2014, Dunning 1993). First, Oakes recommends MI3, a measure that corrects for rare terms by cubing the normal MI measure—thereby making a small term that much smaller (Oakes 1998: 171-2). We, therefore, report MI3 in our supplementary materials. In addition, we also report two measures that, in our view, do a better job of dealing with the potential distortions caused by rare words: the t-score and conditional probability.

Instead of measuring the association strength, t-score estimates the confidence in two words being associated (K. Church et al. 1991). T-scores are calculated by subtracting the expected frequency of a term (given its overall frequency in the corpus) in a given window relative to a target term from its actual frequency. Then the result is divided by the amount of variance in the frequency of the term in question relative to the average frequency of terms in the corpus. T-score solves the shortcoming of MI by adjusting the joint probability distribution according to the weight of individual terms within the corpus, in effect lowering the t-score for rare terms as compared to more common terms. In terms of interpretation, t-score is similar to MI in that a positive score indicates an above chance level word association, a negative score indicates a negative association, and zero indicates independence between the words.

We also report conditional probability, which calculates the probability of the appearance of one word given the occurrence of another word. Specifically, conditional probability is calculated by multiplying the probability of the occurrence of word 1 given the occurrence of word 2 by the probability of the occurrence of word 2. Unlike other test statistics for use in corpus linguistics, conditional probability is presented as a pair of values. This is because it is sensitive to asymmetries in word frequencies between, for example, the probability that the occurrence of the word *of* predicts the occurrence of the word *course* (a low probability, because of the ubiquity of ‘of’) versus the probability that the occurrence of the word *course* predicts the occurrence of the word *of* (a high probability, because of the infrequency of ‘course’).

This leaves unanswered specific questions about what measures of collocation can tell us about semantic relationships. For instance, what is the most useful collocational window size for capturing significant semantic relationships? Are particular types of semantic relationships characterized by distinctive collocation scores? Although there has been important preliminary work exploring these and related questions (see especially Rhode et al. 2005, Bullinaria & Levy 2007, and a recent review in Daniel Jurafsky & Martin 2015), the answers are likely to vary from language to language, and possibly genre to genre. For this reason we begin with an effort to discover latent collocational patterns in early Chinese texts by letting the corpus talk to us rather than by attempting to test any hypotheses. In this semantic benchmarking exercise, our Study 1, we sought to gain knowledge about applications of corpus linguistics techniques to our historical Chinese corpus by analyzing classical Chinese word pairs with known, and fairly uncontroversial, semantic relationships to one another.

**Study 1: Using Collocation Measures to Semantically Benchmark Classical Chinese Word Pairs**

For Study 1, we chose a series of semantic relationships most likely to prove useful in our subsequent analysis of mind-body relations: various types of contrastive pairs, pairs linked by endemic function, and pairs characterized by a part-whole relationship (see **Table 1 below**). To these semantic pairs we added two control conditions, where the second character in the pair was replaced by 1) another word semantically-related to the target character, but not with the same semantic relationship as the originally-paired word; and 2) a word semantically unrelated to the target character. In both control conditions, we picked substitution characters that matched as closely as possible the word frequency of the original character. Exploring these known semantic pairs, raw collocation counts were recorded for the sentence level and at the 10L10R, 5L5R, 2L2R, and 1L1R windows. T-scores, conditional probability, MI and MI3 were calculated for each pair at these various windows.

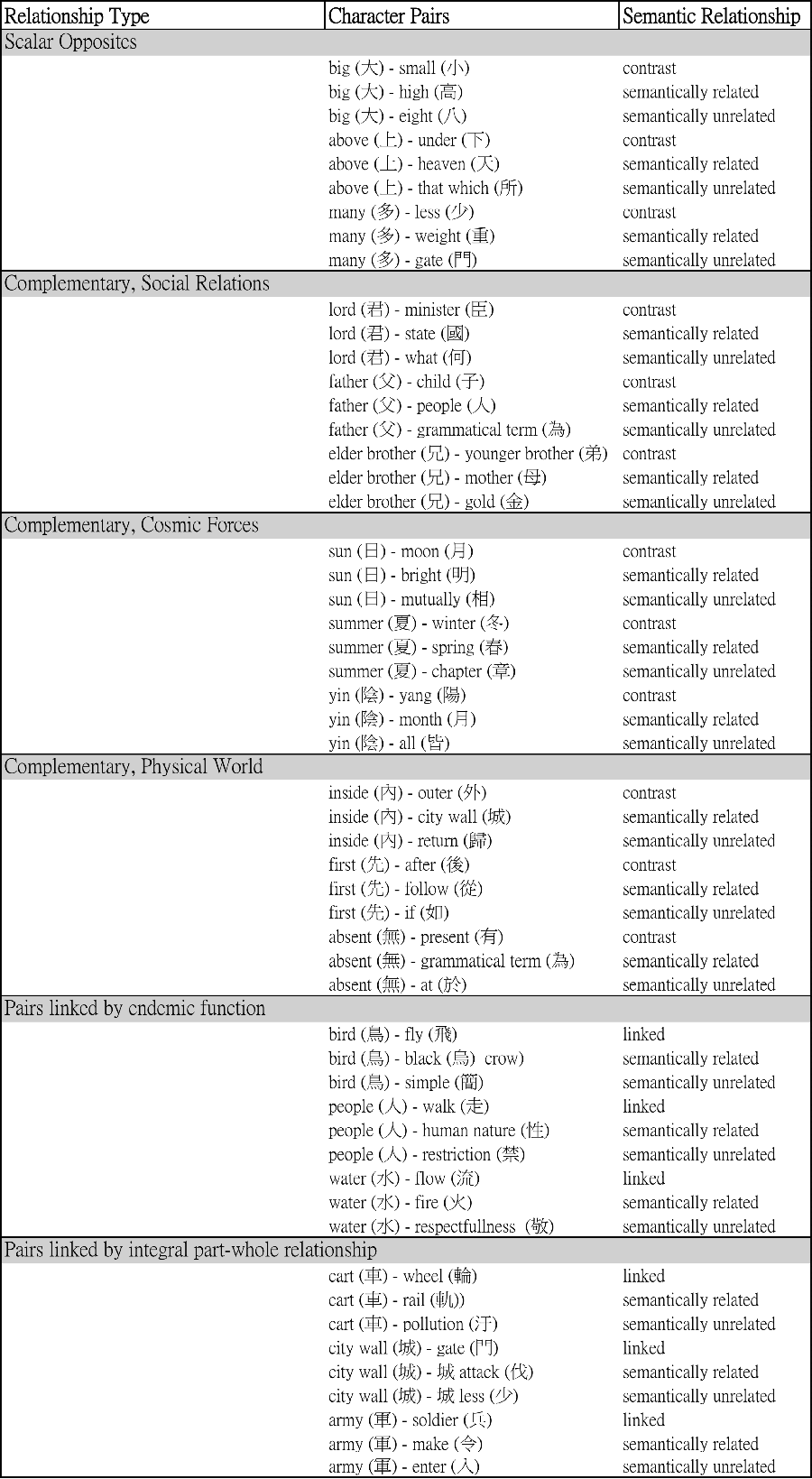


Table 1. Semantic Relationship Benchmarking Pairs

Our full results are available in our on-line supplementary materials (**URL**). When it came to our contrastive, functionally-related and part-whole pairs, we found that, rather than clustering nearby the target character, we see a relatively even distribution of collocations from the sentence and 10L10R level down to the 1L1R level. This suggests that there is no obvious “sweet spot” for capturing semantically-significant collocations. That said, the differences between the target pair and the semantically-related and -unrelated pairs become somewhat starker at smaller windows, which may mean that windows such as 2L2R or 1L1R do a better job of capturing semantic relations, or may simply reflect the fact that our target pairs often appear together as set pairs in the early Chinese corpus. The sentence, uniquely among our KWIC windows, reflects authorial or editorial decisions about semantic relatedness. It, therefore, has a kind of organic validity to it, and we accordingly have chosen to focus on this measure in our discussion here, concluding that we have the most confidence in drawing inferences about the psychology of authors when looking at co-occurrences of two terms within the same sentence, as opposed to within the same 100 word or 10 word string. When it comes to statistical measures, with the exception of the MI score (which was not surprising, for reasons discussed above), t-score, conditional probability, and MI3 measures all provided broadly similar results. We will therefore focus on the t-score at the sentence level, although all of the measures, at all of the various KWIC windows we employed, are available in our on-line supplementary materials.

The basic results are striking, and are presented here in **Table 2**.

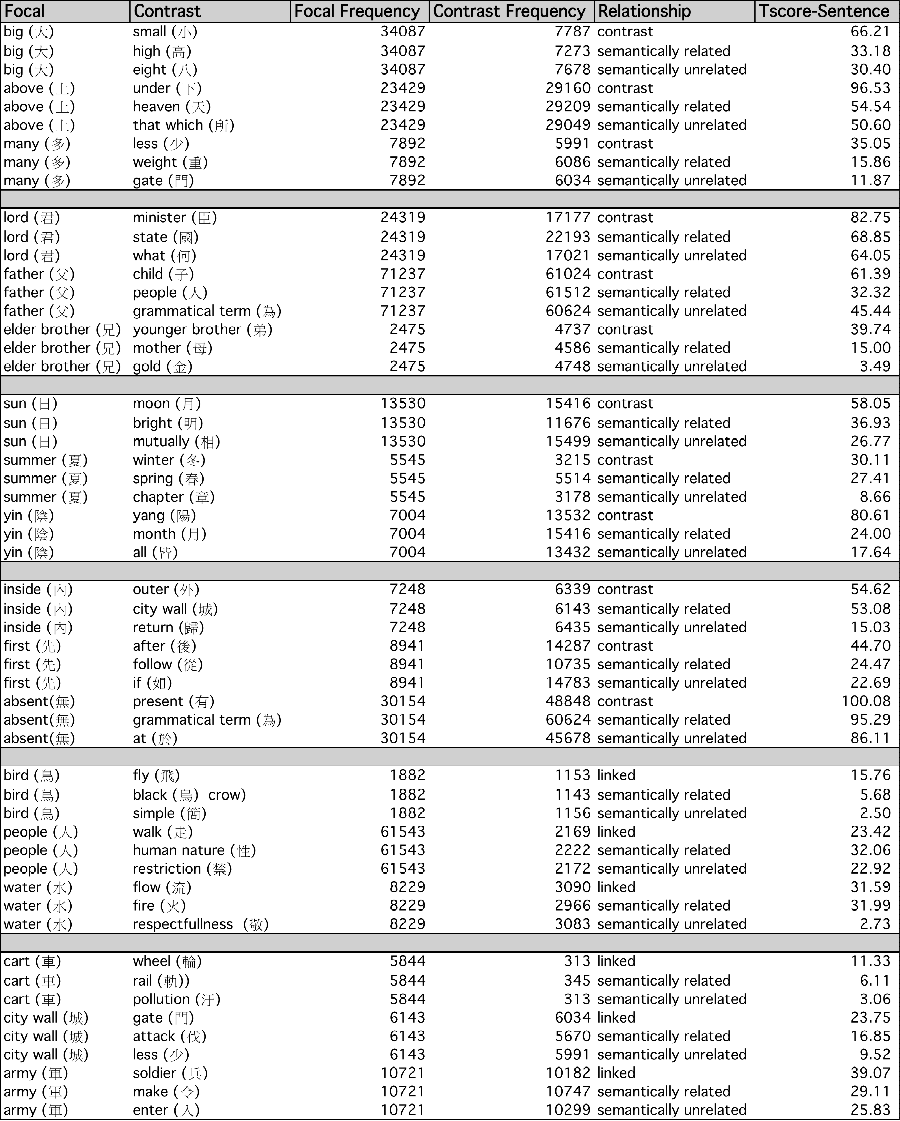


Table 2. Selected Results from Benchmarking Study

To begin with, our contrastive pairs *all* show higher t-scores than the control pairs. In most cases, we see a clear pattern where the contrastive pairs have very high t-scores, with the scores falling off considerably for the semantically-related pairs and then further still for the semantically-unrelated pairs. This validates our prediction that semantically-related terms will have strong collocation patterns, with contrastive semantic relationships being the most powerful of all. The only two cases where we do not get a sharp drop-off as we move from contrastive pair to semantically-related pairs (although we do still get a drop) are in the case of moving from summer::winter to summer::spring, and inside::outside to inside::city wall. The first case may be a result of the fact that the four seasons frequently appear in lists together, artificially increasing the collocation measure for what was intended as a merely semantically-related word. The latter case is more ambiguous. It is possible that “city wall,” which our classical Chinese expert chose as a term likely to be related to “inside” (and conceptions of containment generally), was a poorly-chosen semantically-related control. In any case, this pair remains an outlier to our observed pattern.

Interestingly, the part-whole pairs also showed the same pattern as the contrastive pairs: a consistent fall-off moving from the target pair to the merely semantically-related and then the unrelated controls. This suggests that the strong signal we saw with the contrastive pairs may not be linked to that narrow semantic relationship, but might rather serve as a general signal of a specific, and well-defined, semantic relationship between two terms. In other words, we can imagine vaguely semantically related terms hovering around each other in the logical space of the text, kicking off the sort of middling t-scores that we see in the semantically-related control pairs. Pairs of terms with specific, and well-defined, semantic relationships to one another exert a stronger attraction on one another, and therefore display dramatically higher t-scores. It is possible, then, that collocation measures, such as t-scores, can only tell us about the intensity or specificity of the semantic link between two terms, rather than the exact nature of that link. We will return to this topic again below when we look at collocation patterns for *xin* and the other organs.

It is worth noting that, when it comes to endemic function, the pattern of sharp fall-off as we lose semantic specificity is a bit muddied. Our first set of terms, involving ‘bird,’ shows the pattern very clearly. The pattern breaks down, however, with our other two pairs, people::walk and water::flows. This may be because endemic functions are just not as clearly signalled as contrastive or part-whole pairs. It is, in our opinion, more likely the result of methodological errors.

Water::flow is a well-chosen endemic function pair, but it was probably a mistake to pick “fire” (*huo* 火) as the semantically-related control, since it also commonly appears in conjunction with water in the list of the Five Phases (*wuxing* 五行). In other words, as in the case of summer::spring, we made a mistake in choosing a semantically-related term that also commonly appears with the target in compounds. It probably would have been better to choose a term related to ‘wetness’ or ‘moisture’ instead. In the case of ‘people/humans,’ *ren*人 is an extremely high-frequency term that also appears in many compounds. As noted in the topic modeling study below, it is common enough to have been included in our “stop list” of highly pervasive terms with limited semantic use. Moreover, as we note in the hierarchical clustering study, many species besides people “walk/run” (*zou* 走)*,* so we should have chosen a species-verb pair that was more specific, as in bird-fly. Finally, one of the compounds *ren* appears in is *renxing* 人性; we, therefore, erred in picking our semantically-related term in this case as well, since the chosen term is not just semantically-related, but also linked with the target term in a common compound or set phrase, as with water-fire and the seasons. These are all issues that can, and will, be untangled in follow-up studies. We choose to report our original results not only in the interest of transparency, to also to show how challenging this sort of work can be, and why it requires experts with a deep knowledge of the corpus under study.

Another problem that we recognize in this pilot study is that we are cherry-picking our target and control pairs. An ideal assessment of the link between collocation and semantics would obtain collocation measures for *all* term pairs in the entire corpus, rank them in strength, and then turn to experts familiar with the corpus to see if the trends are robust. The problem with this approach is that it is computationally intractable, given the vast number of relationships that would have to be tested (i.e., 215,696). It is possible to mitigate this challenge by vigorously pruning the corpus—that is, removing rare words, grammatical terms, overly common words, etc.—but this may not be enough. The most productive route forward would be a random sampling of word pairs followed by expert semantic evaluation, a project that our research team is currently planning.

**Study 2: Applying Semantic Benchmarks to *Xin* - Body Relations**

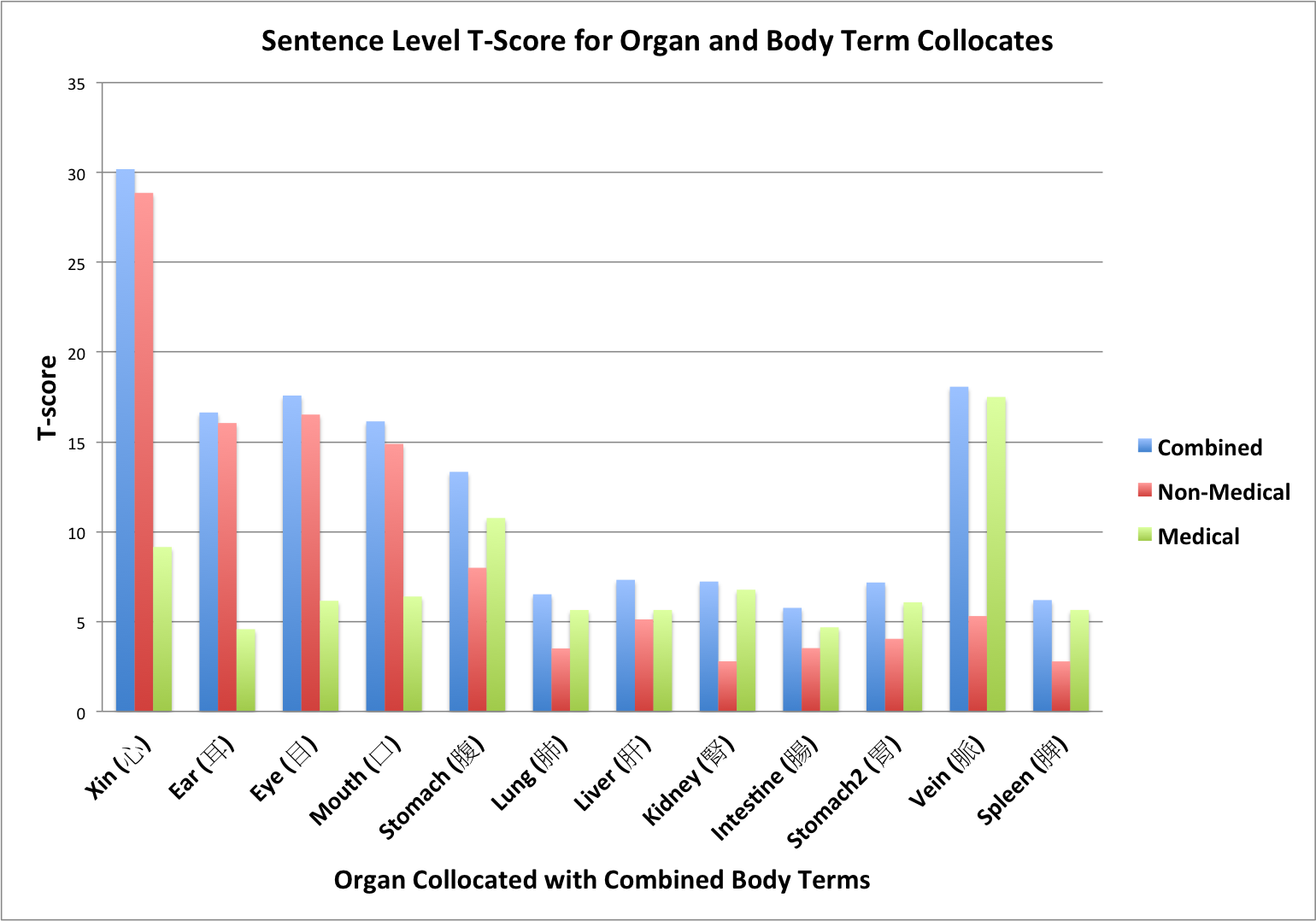
In Study 2, we determined collocation measures and patterns for *xin* and other organs in the body in relationship to the three standard terms for “body” in classical Chinese (*shen* 身, *ti* 體, and *xing* 形). Unlike the semantic pairs above, we also gathered this data separately for three different genre groupings within our corpus: all texts, all texts except for medical texts, and medical texts alone.[[5]](#footnote-5) This is because, whatever early Chinese views about mind-body dualism, we expected there might be a genre effect on the degree to which the *xin* was portrayed as a physical organ in the body, with this being more likely in medical texts, given their technical nature and physiological goals. Moreover, the frequency of certain organ terms—most notably, “vein/artery/pulse/meridian/pulse” (*mai* 脈)*—*is much higher in medical texts than in the corpus as a whole,[[6]](#footnote-6) which could skew the results.

The results of our analysis, broken down to isolate the potential effect of the medical text genre, are reported in Table 3 and Figure 3 below:

Table 3. Xin and other organs :: the body, split by genre groupings

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Term 1 | Term 2 | Sentence Level T-Score | | |
| Combined | Non-Medical | Medical |
| Xin (心) | Body Terms (身,形,體) | 30.18 | 28.86 | 9.14 |
| Vein (脈) | Body Terms (身,形,體) | 18.07 | 5.3 | 17.5 |
| Eye （目） | Body Terms (身,形,體) | 17.58 | 16.52 | 6.15 |
| Ear (耳) | Body Terms (身,形,體) | 16.63 | 16.05 | 4.57 |
| Mouth （口） | Body Terms (身,形,體) | 16.14 | 14.88 | 6.39 |
| Stomach （腹) | Body Terms (身,形,體) | 13.33 | 7.99 | 10.76 |
| Liver (肝 ) | Body Terms (身,形,體) | 7.32 | 5.12 | 5.64 |
| Kidney (腎 ) | Body Terms (身,形,體) | 7.22 | 2.78 | 6.77 |
| Stomach2 (胃) | Body Terms (身,形,體) | 7.17 | 4.04 | 6.07 |
| Lung (肺) | Body Terms (身,形,體) | 6.51 | 3.5 | 5.64 |
| Spleen (脾) | Body Terms (身,形,體) | 6.19 | 2.78 | 5.64 |
| Intestine (腸) | Body Terms (身,形,體) | 5.76 | 3.52 | 4.68 |

Figure 3. Visualization of Xin and other organs :: the body, split by genre groupings



*Xin* looks very different from the other organs. In both the Combined corpus and the Non-Medical corpus, its t-score is almost double that of the next highest organ.

The only place this is not true is in the Medical Texts, where the term *mai* 脈 takes *xin*’s place as the odd-organ out, having almost double the t-score of *xin* or stomach/belly (*fu* 腹). Typically translated as “vein, artery, meridian, pulse,” *mai* is a central term in traditional Chinese medicine, referring to the channels through which vital energy (*qi* 氣) flows in the body.

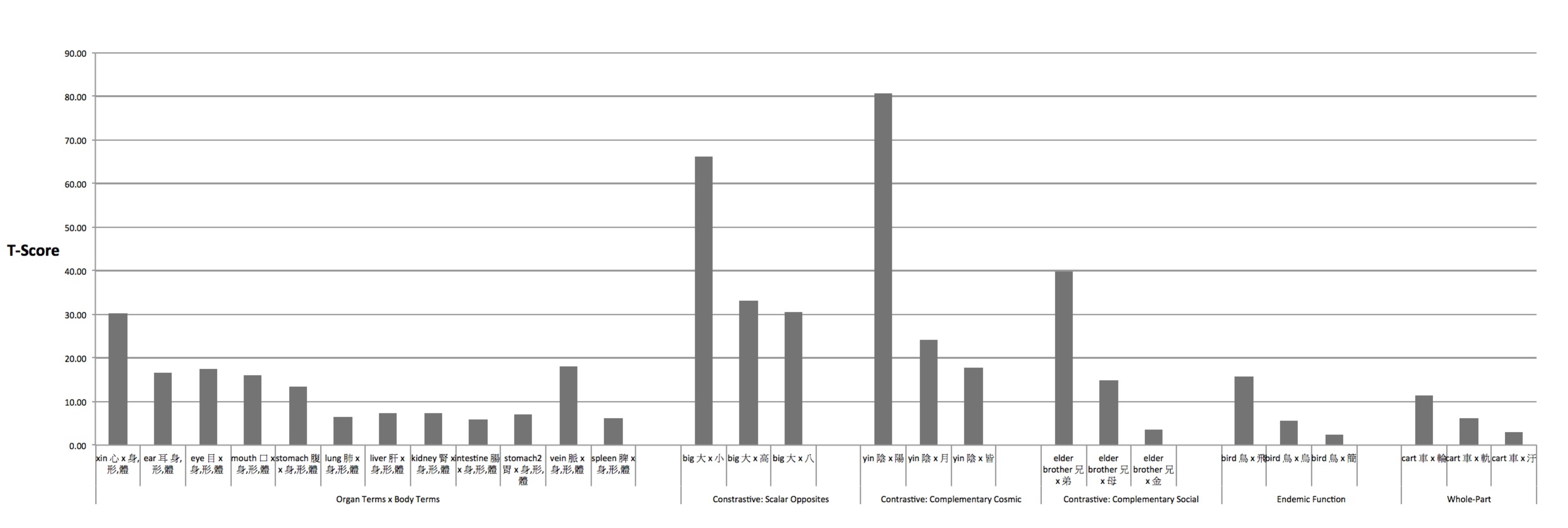
Tunnelling back down in the actual passages behind these collocation results, we can see that, in the medical texts, *mai* 脈 does often appear in conjunction with the body terms, sometimes in ways that suggest a complementary or contrastive relationship. However, we may also be seeing the effect of the occasional occurrence of *mai* in the compound *maixing* 脈形, which refers to the “shape of the pulse” that is used to diagnose ailments. For example, in the Han medical text *Treatise on Cold Injury* (*Shang Han Lun* 傷寒論), a student asks, “When a person is sick from intense fear, how does their pulse present itself?” (*maihezhuang*脈何狀). The Teacher replies, “The shape of the pulse (*maixing*脈形) is like following a thread as it winds around and around, and their face is white and devoid of color.”[[7]](#footnote-7) *Xing* 形 in this case has its basic meaning of “shape” rather than “body,” so collocations between the terms in such cases would be a false signal. Overall, we think the pair’s observable collocation patterns are best attributed to the unique focus of this genre, which is to manipulate the channels of vital energy in a patient to restore them to health. It is worth noting that a special role for *mai* entirely disappears once the Medical texts are excluded, and is greatly diminished in the overall corpus as a whole.

Turning back to *xin*, we can compare its pattern of collocation with body terms in the ex-Medical corpus, as opposed to the other organ terms, with our semantic benchmarking efforts in Study 1. As Table 4 and Figure 4 below indicate, the *xin*-body relationship looks more like a strong, well-defined semantic relationship (Contrast or Part-Whole) than any of the other organ-body collocations patterns, which look more like the generally semantically-related control pairs.

Table 4. Xin and other organs :: the body, Combined corpus, with semantic benchmarking pairs

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Texts** | **Pinyin Organ Term** | **Chinese Organ Term** | **English Body Term** | **Chinese Body Terms (combined)** | **Sentence Level T-Score** |
| Combined | xin | 心 | body | 身,形,體 | 30.18 |
| ear | 耳 | body | 身,形,體 | 16.64 |
| eye | 目 | body | 身,形,體 | 17.58 |
| mouth | 口 | body | 身,形,體 | 16.14 |
| stomach | 腹 | body | 身,形,體 | 13.33 |
| lung | 肺 | body | 身,形,體 | 6.52 |
| liver | 肝 | body | 身,形,體 | 7.32 |
| kidney | 腎 | body | 身,形,體 | 7.23 |
| intestine | 腸 | body | 身,形,體 | 5.76 |
| stomach2 | 胃 | body | 身,形,體 | 7.17 |
| vein | 脈 | body | 身,形,體 | 18.07 |
| spleen | 脾 | body | 身,形,體 | 6.20 |
|  | | | | | |
| Contrastive (scalar opposites) | big | 大 | small | 小 | 66.21 |
| big | 大 | high | 高 | 33.18 |
| big | 大 | eight | 八 | 30.40 |
|  |  |  |  |  |  |
| Contrastive (complementary cosmic) | yin | 陰 | yang | 陽 | 80.61 |
| yin | 陰 | moon | 月 | 24.00 |
| yin | 陰 | all | 皆 | 17.64 |
|  | | | | | |
| Contrastive (complementary social) | elder brother | 兄 | younger brother | 弟 | 39.74 |
| elder brother | 兄 | mother | 母 | 15.00 |
| elder brother | 兄 | gold | 金 | 3.49 |
|  | | | | | |
| Endemic Function | bird | 鳥 | fly | 飛 | 15.76 |
| bird | 鳥 | black crow | 烏 | 5.68 |
| bird | 鳥 | simple, bamboo strip | 簡 | 2.50 |
|  | | | | | |
| Whole-Part | cart | 車 | wheel | 輪 | 11.33 |
| cart | 車 | rail or track | 軌 | 6.11 |
| cart | 車 | pollution | 汙 | 3.06 |

Figure 4. Xin and other organs :: the body, Combined corpus, with semantic benchmarking pairs



*[Note: the above figure will have to be oriented in landscape mode to be visible in article]*

The overall pattern gives the strong sense that *xin* and the body share a special relationship, although all of the organ terms share the same broad semantic space with the body terms.

Of course, the particular strong semantic relationship between *xin* and the body might be a part-whole rather than contrastive, which might be seen as corroboration for the mind-body holist position. We think this a poor inference from our results for a variety of reasons. First of all, if the semantic relationship being picked up by the t-scores were of a part-whole nature, it would make sense for it to be shared by *all* of the organs equally, which is not what we see. A basic tenet of the mind-body holist position is that the *xin* is merely one organ among the others. However we interpret the precise nature of the semantic link between *xin* and body, it is—in contrast to the holist prediction—clearly of a qualitatively different order than any other organ.

The possibility that the *xin*-body relationship is one of Part-Whole is further weakened by a follow-up study we performed comparing *xin* to two other common organs, the eye and the ear, on a series of collocation measures: the body terms, an endemic function term, a semantically-related term and a semantically-unrelated term. The results are presented in Table 5 and Figure 5 below.

Table 5. T-scores of xin, ear and eye and selected characters

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Focal Term** | **Chinese** | **English** | **Comparison Term** | **Relation** | **T-Score** |
|
| xin | 心 | body | 身,形, 體 |  | 30.18 |
| xin | 心 | think, reflect | 思 | function | 17.98 |
| xin | 心 | intention | 意 | related | 22.26 |
| xin | 心 | end, completion | 終 | unrelated | 12.36 |
| ear | 耳 | body | 身,形, 體 |  | 16.64 |
| ear | 耳 | hear, listen | 聽 | function | 16.94 |
| ear | 耳 | announce, command | 告 | related | 6.51 |
| ear | 耳 | end, completion | 終 | unrelated | 12.65 |
| eye | 目 | body | 身,形, 體 |  | 17.58 |
| eye | 目 | see, perceive | 見 | function | 20.90 |
| eye | 目 | bright, clear | 明 | related | 17.78 |
| eye | 目 | life, living | 生 | unrelated | 12.46 |

Figure 5. T-scores of xin, ear and eye and selected comparison characters

As we can see, the organs show a fairly flat set of t-scores across the semantic categories, falling off as expected for the semantically-unrelated pairs. “Ear” is the one exception, with an unusually low collocation score with its semantically-related control, “to announce/to tell” (*gao* 告). This is likely because of the fact that “announce” is not terribly closely related to “ear,” but was a compromise choice because we were unable to find a good candidate that came anywhere near to matching “to hear/listen/obey” (*ting* 聽) in terms of word frequency. (We would have liked something along the lines of ‘noise’ or ‘sound,’ but these terms were either extremely common or extremely uncommon).

The relevant pattern to notice, however, is that *xin*—again, alone among the organs—has a collocation score with the body terms almost twice as high as any of the other collocations for any of the organs. This would make no sense if, like the other organs, it was simply one part of the embodied organism, with its own particular function but no qualitatively special status.

In sum, the ability to identify precise semantic relationships in classical Chinese from patterns of collocation scores alone remains elusive, although a much larger-scale exploration of the corpus may bring progress on this front. What we believe we *have* been able to demonstrate here, though, is that specific and strong semantic relations yield higher t-scores than vague ones, and that *xin*, alone among the organs, is characterized by just such a collocation profile vis-a-vis the physical body terms. This is difficult to reconcile with the claim that *xin* is simply one organ among many, as the strong mind-holist position would assert. Below we turn to other methods of automated textual analysis that strongly corroborate these results.

**Study 3: Hierarchical Clustering Analysis**

A complementary, and in some ways even more exciting, approach to word co-occurrences involves employing unsupervised machine learning to extract significant patterns (Jurafsky & Martin 2015; C. D. Manning, Raghavan, & Schütze 2008; Plasse et al. 2007). The first of these methods that we applied to the issue of mind-body dualism in early China is called hierarchical clustering analysis. There are a variety of methods for performing such analyses, but the most common (and the one we employed) is referred to as bottom up, agglomerative hierarchical clustering.[[8]](#footnote-8) In this method, the corpus is first converted into a “vector space,” which, in our case, is essentially an enormous multi-dimensional table, with each row representing an individual document and each column an individual term.[[9]](#footnote-9) An algorithm runs through the space, measuring the geometric distances between individual terms. It then begins clustering them together in an iterative manner: the two terms with the shortest distances are “agglomerated” into a group, which then becomes a unit for the next stage of agglomeration, until the algorithm has built up a set of hierarchical clusters (clusters within clusters). The results are typically represented in a tree form or “dendrogram.” Figure 6 below, from a hierarchical clustering analysis of a large contemporary English corpus targeting various classes of nouns, performed by Rohde, Gonnerman, & Plaut 2005, shows how HCA can do an impressive job of tracing how individual target terms in a corpus are related to one another semantically.

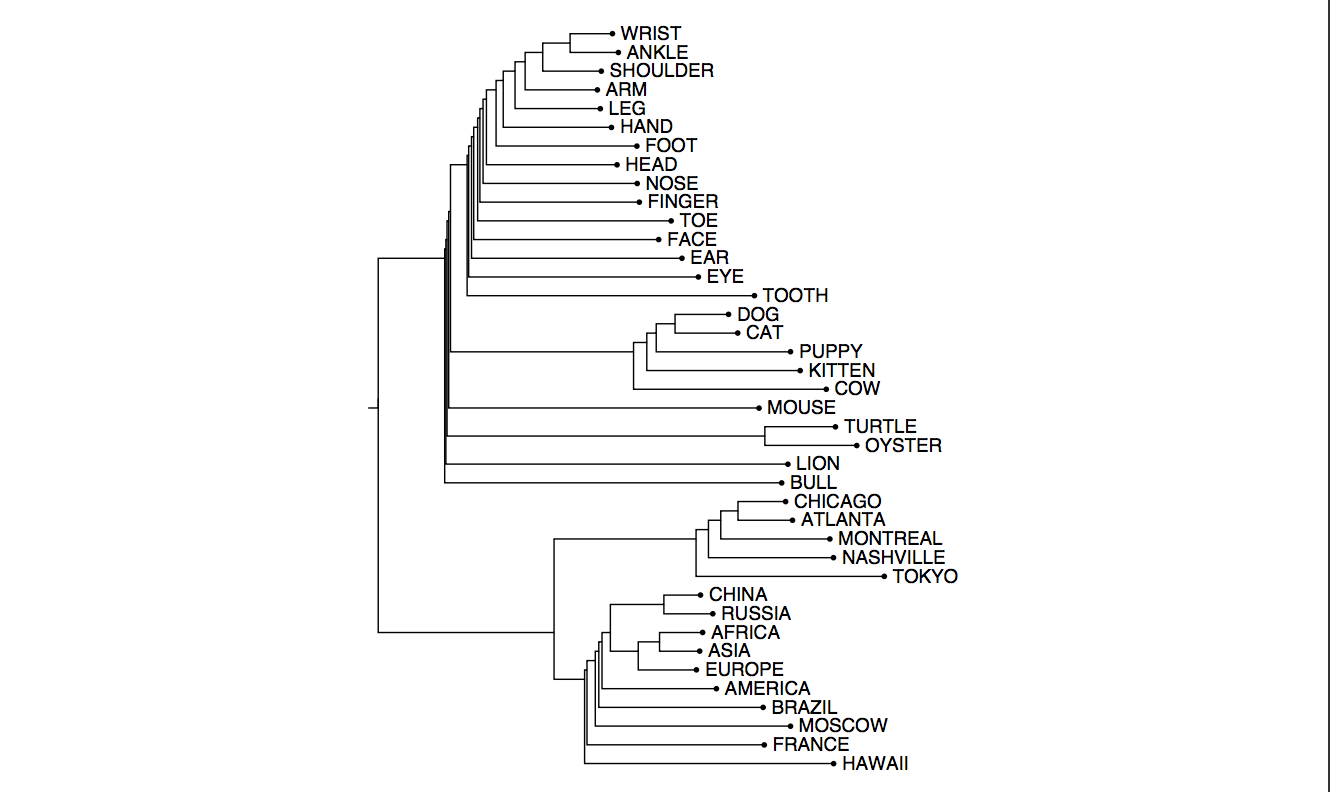


Figure 6. Dendrogram of Noun Classes Based on Vector Distances in an English Corpus (from Rohde et al. 2005: 20, Figure 9).

Approaches such as hierarchical clustering are called “unsupervised” because they involve a type of machine learning in which algorithms explore a set of completely unlabelled or unclassified data and attempt to identify statistically-significant clusters based on a single, or small set, of parameters, such as corpus distance. The great advantage of unsupervised approaches is that they are as objective as one could desire. Although various assumptions are built into the processing of the document, the parameter or parameters selected, and the specifics of the algorithm, these can be easily varied and the resulting patterns compared. The running of the program involves no human input, and, in our case, was performed by a colleague with no knowledge of classical Chinese. This greatly reduces the potential for interpretative bias.

For Study 3, we began by producing a dendrogram representing relations between some of the control terms[[10]](#footnote-10) from Study 1, with the results visualized in Figure 7:

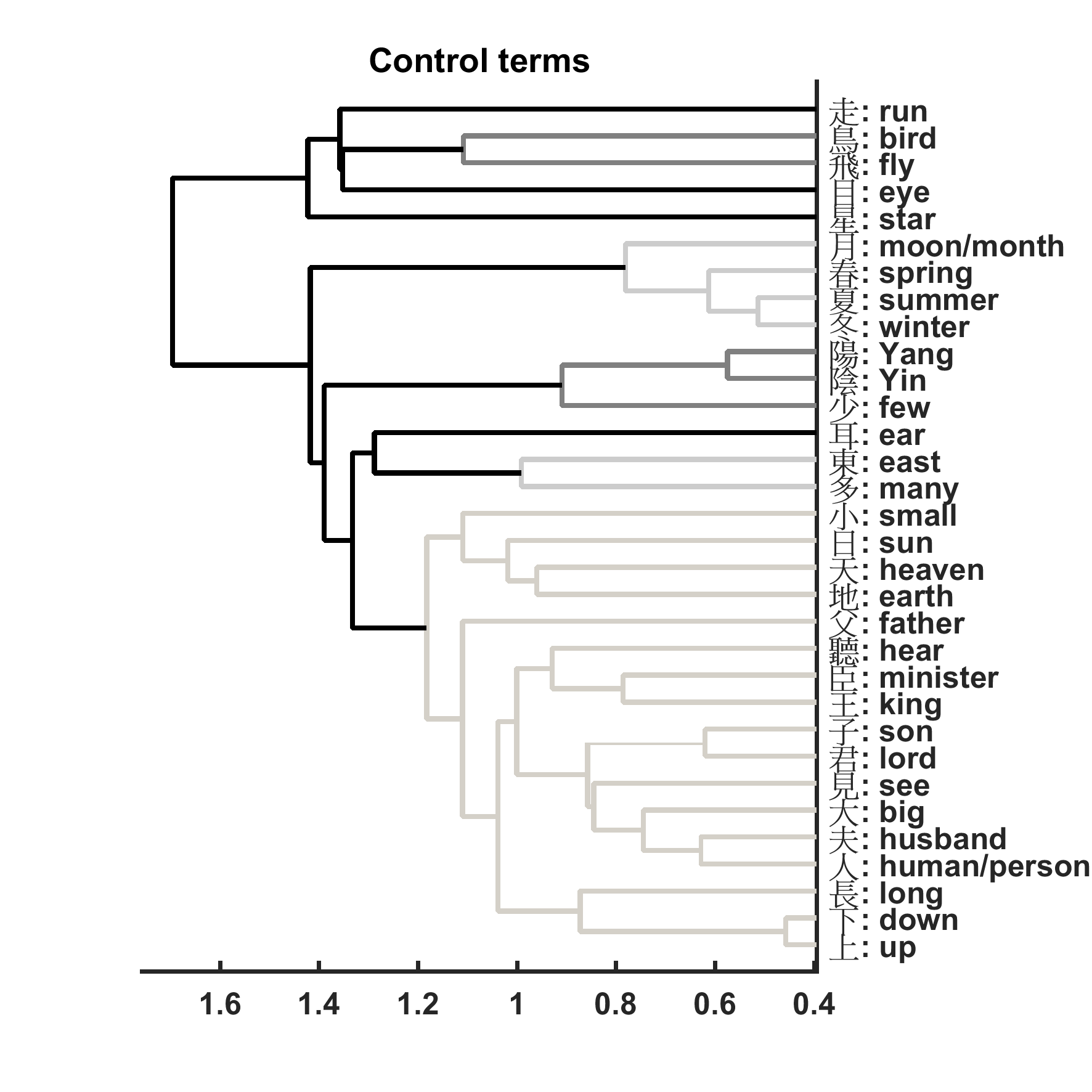


Figure 7. Dendrogram of Control Terms

The results are striking. With only a few exceptions, the tree relations are precisely what we expect given the well-understood semantic relations between these terms. For instance, we see the seasons clustering together (pictured in lighter grey near the top of the diagram), with the two classic opposites (summer and winter) sharing the most basic node, but then joining with spring the next node up, and finally *yue* 月, which refers to both the “moon” and “month.” We also see most of the scalar opposites and complementary pairs (*yinyang*陰陽, Heaven/earth *tiandi*天地, king/minister *wangchen*王臣, above/below *shangxia*上下) clustering tightly together.

The exceptions to this pattern are easily explainable as a result of either common compounds or semantic ambiguity. For instance, “son” (*zi*子) clusters with “lord” (*jun* 君), rather than with “father” (*fu* 父), as one might expect. This, however, reflects the high frequency of the compound term *junzi* 君子(“gentleman, lit. son of a lord”) in the corpus. “Big” (*da* 大) has been pulled out of its expected tight fit with “small” (*xiao* 小) because of its common appearance in the compound *dafu* 大夫 (the name of a government rank), with “human/person” (*ren* 人) being similarly pulled into the orbit of “husband” (*fu*夫) by the common compound form of “husband,” *furen* 夫人.

In some cases, as in Studies 1 and 2, we could have done a better job of picking specifically-defined and semantically unambiguous pairs. For instance, “run/walk” (*zou* 走), which we had initially envisioned as an endemic function pair for “human/person” (*ren* 人), appears quite far away in the tree. This is almost certainly because 1) *many* species besides humans *zou* 走 ‘walk, run,’ and 2) *ren* 人 actually only rarely refers to ‘humans’ as a distinct species, commonly referring to “people” in general. Our much more narrowly defined and unambiguous endemic function pair, “bird-fly” (鳥飛), appears, as expected, tightly linked at the first node. *Ting* 聽 ‘to hear,’ which we had expected to pair with the appropriate organ (*er* 耳 ‘ear’), instead clusters with minister and king. This seems odd until one considers that a common alternative meaning of *ting* is “to obey,” the central imperative of minister-king relations. A few surprising tight pairings (e.g., “east” *dong* 東 and “many” *duo* 多) or great divides (e.g., “eye” *mu* 目 and “see” *jian*見) aside, our hierarchical clustering algorithm appears to give an accurate model of expected co-occurrences within the early Chinese corpus, which, in turn, presents us with a coherent conceptual map of semantic relationships.

Having established a benchmark, let us now turn to the dendrogram representing our controversial terms of interest, the *xin* and the other organs in relation to the three primary body terms. This tree is represented in Figure 8 below

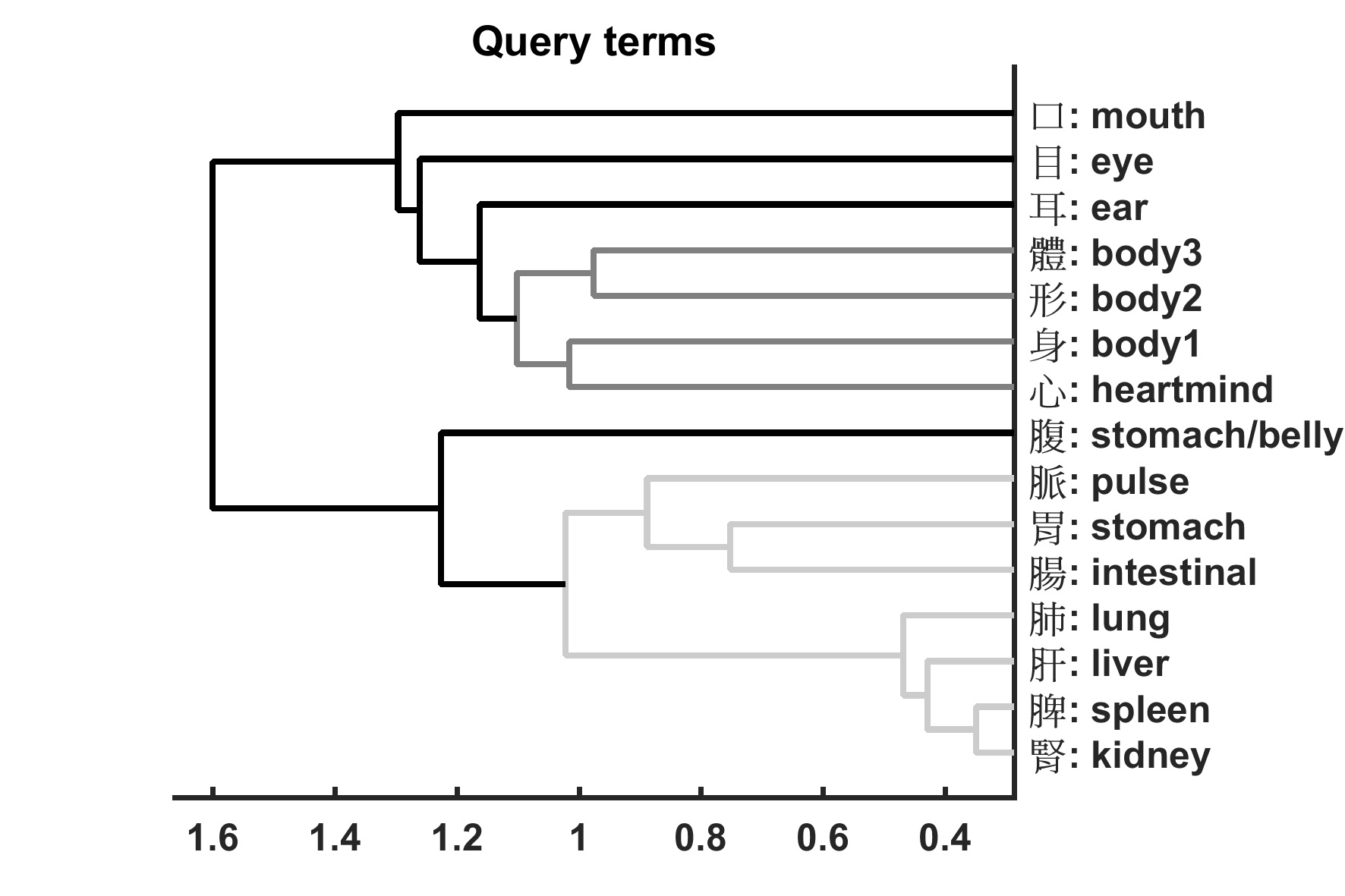


Figure 8. Dendrogram of Query Terms

It is difficult to imagine a clearer representation of mind-body dualism than Figure 8. The nodes pictured in the middle-shaded grey show the *xin* as being uniquely paired with the first of the body terms, 身, and then, in the next node out, also uniquely paired with the *other* two body terms, 形 and 體. This mind-body nexus then clusters with the three organs most associated with perception and communication, the mouth, eye and ear. In other words, these are the three organs that most directly serve the *xin* in its role as the center of cognition and perception. Finally, what we might think of as the more “physiological” organs all cluster together in an entirely separate tree. The “stomach/belly” *fu* 腹 is something of an outlier in Figure 8. (Note that this is the same pattern we found in our t-score results above in Study 2, which provides an important confirmation that our mixed quantitative methods are converging on similar results.) This is possibly because of its occasional figurative usage as a metonym for basic desires or simple needs, as in *Daodejing* 12: “the sage is for the belly, not the eye” (*shengren wei fu, bu wei mu* 聖人為腹，不為目).[[11]](#footnote-11)

It is important to observe that the hierarchical clustering algorithm (which was run on the entire corpus, without stop words) reproduces almost exactly the word collocation results reported in Study 2 for the entire corpus (Table 3 and Figure 3 above). The dendrogram in Figure 8 serves as a nearly perfect visual representation of our earlier t-score results relative to the body terms. There, the ear, eye, and mouth cluster closely together with t-scores in the neighborhood of 17, but are distinguished from *xin* with its t-score of 30. The “stomach/belly” (*fu* 腹) then appears at one remove with a t-score of 13, with all of the other organs representing a distinct, and internally tightly integrated, cluster with t-scores in the 5-7 range. The one exception is the *mai* 脈, which does appear as an outlier in the diagram vis-a-vis the “physiological” organs, but which we would expect, given its t-score of 18 above, to be joining the ear, ear, and mouth in the upper part of figure. The difference here between the dendrogram and t-score results most likely reflects the fact that, in the textual vector space of the corpus, *mai* 脈is rarely encountered but densely-clustered where it does appear. It is highly represented well in the corpus with 2362 appearances, but these are almost all concentrated in a small number of medical texts (2135 appearances). One methodologically significant conclusion that we can draw from the contrast between the t-score and hierarchical clustering results is that hierarchical clustering seems to do a better job of putting words in their proper place, as it were. T-scores alone fail to communicate the sometimes extremely lumpy distribution of certain key terms, and may, therefore, distort their relation to other terms of interest. Another equally important conclusion, however, is that the overall tight fit between the hierarchical clustering results and the t-score results should increase our confidence in t-scores as a reliable measure of collocation, at least in this classical Chinese corpus.

Our hierarchical clustering algorithm, and the dendrogram it produced, represents patterns of geometrical distances between terms within the Chinese Text Project’s early Chinese corpus. Alternative semantic interpretations of this data are possible, but frankly difficult for us to imagine. Study 3’s results appear to represent a confirmation of the special status of the *xin*, its unique relationship to the body, and its special connection to perception and communication. In other words, like our earlier studies, Study 3 strongly confirms the view that, at least in terms of implicit, background assumptions, the authors of this corpus of early Chinese texts were mind-body dualists.

**6 Topic Modeling**

Topic modeling is another unsupervised method that uses a complex form of statistics, Bayesian probability, to discern latent patterns of regularly co-occurring terms in a textual corpus.[[12]](#footnote-12) These patterns are called “topics.” Topic modeling begins with the assumption that the surface structure of the texts in a given corpus can be viewed as the product of latent, or hidden, themes, and its task is identifying these themes (topics), as well as identifying the individual words that belong to these topics. Each topic thus produced consists of a list of words ordered by “weight” in the topic, with words at the top of the list contributing much more to the formation of the topic than words lower on the list.

The use of topic modeling in the humanities is still in its infancy, and to date has been used primarily in literary studies and political science.[[13]](#footnote-13) Within Religious Studies, a related analytic technique, principal component analysis (PCA), was used by a group of scholars in Taiwan to resolve a controversy concerning the authorship of various translations of Indian Buddhist texts into Chinese (Hung, Bingenheimer, & Wiles 2010). Topic modeling has also been applied to classical Chinese corpora to extract topics related to positive and negative emotions in Tang poetry (Hou & Frank 2015). With regard to the CTP corpus, some of our team has also been experimenting with using topic modeling to explore dating and authenticity controversies surrounding early Chinese texts, such as the *Shu Jing* or *Zhuangzi* (Nichols et al. Under Submission).

In a 100-topic model of the CTP corpus that we created, *xin* appears in 6 of the topics, and is the primary component of one the “heaviest loading,” or most common, topic, in the entire corpus, Topic #97). The three conceptual topics[[14]](#footnote-14) in which *xin* appears, in order of their overall weight in the corpus, and with the top 10 loading words (ranked by importance), are listed below in Table 6.

Table 6. Conceptual Topics in Which *Xin* Appears

|  |  |  |  |
| --- | --- | --- | --- |
| Topic # | Weight | Name | Top Ten Words (in order, left column first) |
| 97 | 0.47514 | Cognition/Perception/Cosmic Fortune | *xin*心 peace/balance平  see/perceive見 yang 陽  bright/intelligent/clear明 intention意  accord/harmonize合 spirit神  lose/miss失 fortune/luck 福 |
| 10 | 0.34877 | Temporal Cognition and Planning | now今 interrogative豈  *xin* 心 morning/court朝  after後 death死  strength/effort力 sincerity/integrity誠  worry 憂 abandon棄 |
| 33 | 0.07733 | Human, Heaven and Political Order | person人 get/obtain得  big/great 大 world/era/generation世  Heaven/sky天 one/unified ─  know/knowledge知 *xin*心  king王 stop/already已 |

Topic #97 is the most important for *xin*, since it is the second most heavily weighted in the entire corpus and *xin* constitutes its most important term. We have characterized Topic #97 as “Cognition/Perception/Cosmic Fortune,” since the most heavily weighted words in the topic are the first three: *xin*, “to see/perceive” (*jian* 見), and “bright, intelligent, clear” (*ming*明). The word cloud in Figure 9 visualizes the relative weights or contributions of each term to the topic in an intuitive manner:



Figure 9. Word cloud for Topic #97

The main focus of the topic seems to be cognition and perception, with—significantly—no mention of emotion. We should also note that no other organs are mentioned, not even the ones most closely associated with perception, such as the eye (*mu* 目) or ear (*er* 耳). The secondary references to according or harmonizing with things (*he*合), missing an opportunity or making a mistake (*shi*失), peace (*ping* 平), good fortune (*fu* 福), and intention (*yi* 意) suggest a connection with planning or navigating the world, whereas the mention of spirits (*shen* 神) and yin-yang (*yinyang* 陰陽)[[15]](#footnote-15) suggests that cosmic forces are some of the variables to be considered. In any case, the most important topic in which *xin* forms a major component seems to reflect a worldview that sees it as the sole seat of perception, cognition, planning, and personal responsibility, which in turn fits with a mind-body dualist account.

Interestingly, the second most important topic for *xin*, Topic #10, also seems to center on similar themes. We have termed this “Temporal Cognition and Planning.” The topic is heavily dominated by “now/today” (*jin*今), followed by *xin*, “after/future” (*hou* 後), “strength/effort” (*li*力), and “worry/concern” (*you*憂) (see word cloud, Figure 10).

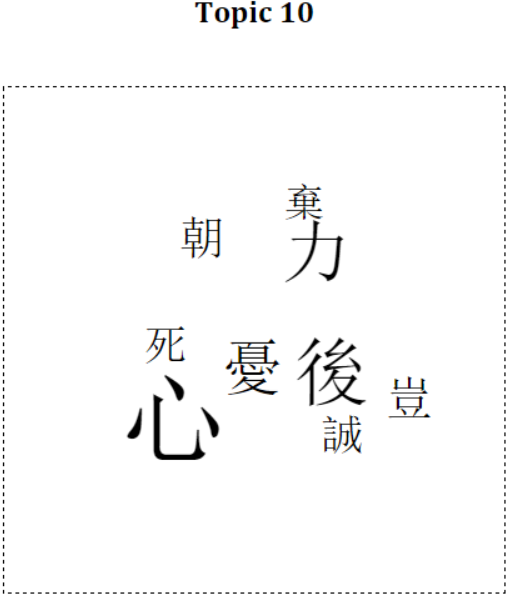


Figure 10. Word cloud for Topic #10

Like Topic #97, the focus seems to be on the *xin*’s role in thinking about the future, planning, and exerting effort. A commonly used interrogative (*qi*豈) and mention of “sincerity” (*cheng*誠) suggests interior thought and resolve, whereas “death” (*si*死) hints at potential dire consequences.[[16]](#footnote-16) “Morning” (*chao*朝; also “court,” as in royal court) adds to the sense of cognition stretched over time. It is worth noting that both #97 and #10 are distributed quite widely across the corpus.

*Xin* plays a relatively minor role in Topic #33, appearing ninth and having a weight of only 0.017, as opposed to .052 for the first term, “person/human” (*ren* 人). To the extent it is conceptually coherent, this topic seems to concern humans, Heaven and political order, with knowledge (*zhi*知) also playing a role. This topic also serves to underscore *xin*’s primarily cognitive nature.

The only other organ that appears in any of our 100 topics is *mai*脈 (“vein/artery/pulse”), which shows up, as one might expect, in two wonderfully coherent “Traditional Chinese Medicine” topics (#73 and #84), that consist almost entirely of technical, medical terminology and load almost exclusively in the Medical text portion of the corpus.[[17]](#footnote-17) In the Ex-Medical corpus, then, it is significant that *xin* is the *only* bodily organ[[18]](#footnote-18) to appear among the most important characters in any of our 100 topic models. In other words, it is the only organ conceptually, or stylistically, salient enough to appear in distinctive thematic clusters. This is yet another example of the qualitative uniqueness of *xin*, and further evidence against strong mind-body holist claims about early China.

**7. Conclusion**

Some of the studies reported above would have benefited from a more careful consideration of candidate control words, or a better-formulated stop-word list. Overall, however, our results are quite robust, especially because they come to similar conclusions by means of very different methodologies. Whether we are looking at word collocation measures, hierarchical cluster analyses or topic models, *xin* stands out as entirely, qualitatively unique among the organs. The locus of the most important of human capacities—thought, planning, and decision-making—it shows a distinctive and highly-salient relationship to the physical body, one that makes little sense unless the authors of the texts in which *xin* appears were operating against a background assumption of at least “weak” folk mind-body dualism.

There are no doubt those among our colleagues in Religious Studies who view talk of collocation measures, KWIC windows and hierarchical clustering algorithms as a sinister encroachment of the sciences upon the humanities, and further evidence that the twilight of the humanities is truly upon us. Even some early practitioners and advocates of digital humanities have, more recently, begun portraying the movement as part of a “neoliberal” conspiracy to undermine the core mission of the university and transform humanistic scholars into disposal tech flunkies (Allington, Brouillette, & Golumbia 2016). We believe, on the contrary, that the judicious adoption of digital humanities techniques is simply a way for humanities scholars to employ the best techniques and theories available to answer questions that matter to them. “Distant” reading, employed merely as a supplement to qualitative analysis, can help us to situate our close reading, give us new perspectives on our corpora, and, perhaps, decisively tip the balance of evidence in hermeneutic disputes. Back in the 1960s, the linguist Margaret Masterman described the potential for computers to analyze texts from a new perspective as a “telescope of the mind,” a powerful new tool whose true potential had yet to be scratched (Masterman 1962: 38).

Over fifty years later, this potential remains underexploited, despite major gaps in our analytic ability that computer-assisted analysis can help to fill. In the field of Religious Studies, as well as in the humanities, more generally, there is, in our opinion, a need for new methods for settling hermeneutic disputes, or at least for narrowing the scope of reasonable views. For instance, in a well-known work on early Chinese thought, David Hall and Roger Ames argue that their claims about broad, “cultural determinants” in early China—for instance, strong mind-body holism—should not be subject to what they call “the Fallacy of the Counterexample” (Hall & Ames 1995: xv). That is, legitimate generalizations about trends in the corpus of early Chinese texts cannot be invalidated by isolated, unrepresentative counterexamples, and we should not allow ourselves to “become lost in the details” (xv) to the point that we lose sight of general trends. The problem is that they present no clear criteria for determining what constitutes a genuine trend and what counts as a distracting, irrelevant counterexample. We argue that the sort of large-scale textual analysis techniques described here could be useful in this regard. They give us a way to pan out from the intricacies of individual passages and texts to gain a panoramic view of an entire corpus at once. Perhaps more importantly, they also allow us to support our generalizations about a given corpus with relatively objective evidence rather than mere assertion or argument from authority.[[19]](#footnote-19)

In the 1970s and 80s, John B. Smith created one of the first tools for conducting such analyses, the Archive Retrieval and Analysis System (ARRAS), that originally ran on a mainframe computer accessed via remote terminal. Although he was a computer scientist, Smith was profoundly sensitive to the humanistic enterprise and saw his platform as a tool to help humanities scholars do their work better, not as a replacement for humanistic expertise. “Humanists have always been explorers,” he wrote. “They sail not the seas of water but on seas of color, sound, and, most especially, words” (Smith 1984: 20). To extend Smith’s analogy, contemporary sailors still rely on such venerable tools as the compass and anemometer, and base the bulk of their decision-making on their qualitative feel for the ocean, wind, waves, and ship. Nevertheless, no one seriously concerned with sailing effectively, especially on a long journey or far from shore, would turn up their nose at GPS, radar, and advanced satellite-based weather forecasting.

We feel the analogy is apt for humanities scholars. Computer-assisted textual interpretation can help us to gain our bearings as we travel through massive textual corpora, allow us to evaluate our qualitative intuitions in light of quantitative data, reveal hidden “topics” or themes invisible to individual human readers, and help us to more rigorously distinguish between unrepresentative counterexamples and instances of broader trends. Like GPS or a marine weather forecast, they can be misused, especially if they allow unqualified novices to set out to sea under the illusion that they know what they are doing, with possibly disastrous consequences. Knowledgeable experts are required to evaluate whether or not the new tools are useful in a given context, or to answer a particular question. They are, also, needed to determine when the results of content-blind algorithms or abstract statistical measures are best ignored because of the complexity or make-up of the material to be analyzed. In the right hands, however, new tools—whether navigational or scholarly—are unqualified gifts. Besides the light that our studies have shed on debates concerning mind-body dualism in early China, we hope that we have succeeded in showing the potential for large-scale analytic techniques to augment our ability to understand and map meanings in religious or philosophical textual corpora, and to weigh in decisively on otherwise intractable scholarly debates.

**Acknowledgements**

Most of this work was funded by a Social Sciences and Humanities Research Council (SSHRC) of Canada Partnership Grant awarded to ES, with some performed while KN was visiting the Institute for Pure and Applied Mathematics (IPAM), which is supported by the National Science Foundation, and ES was an Andrew W. Mellon Foundation Fellow, Center for Advanced Study in the Behavioral Sciences, Stanford University.

**References**

Al-Hejin, B. (2015). Covering Muslim women: Semantic macrostructures in BBC News. *Discourse and Communication, 9*(1), 19-46.

Allington, Daniel, Brouillette, Sarah, & Golumbia, David. (2016, May 1, 2016). Neoliberal Tools (and Archives): A Political History of Digital Humanities. *LA Review of Books*.

Ames, Roger. (1993). The Meaning of the Body in Classical Chinese Philosophy. In Thomas Kasulis, Roger Ames & Wimal Dissanayake (Eds.), *Self as Body in Asian Theory and Practice* (pp. 157-177). Albany, NY: State University of New York Press.

Andrews, Nicholas O., & Fox, Edward A. (2007). Recent developments in document clustering.

Baker, P., Gabrielatos, C., & McEnery, T. . (2013). Sketching Muslims: A Corpus Driven Analysis of Representations Around the Word “Muslim” in the British Press 1998-2009. *Applied Linguistics, 34*(3), 255-278.

Barrett, Justin L. (1998). Cognitive constraints on Hindu concepts of the divine. *Journal for the Scientific Study of Religion, 37*, 608-619.

Biber, Douglas, & Jones, James. (2009). Quantitative methods in corpus linguistics. In A Lüdeling & M. Kytö (Eds.), *Corpus linguistics: an international handbook* (pp. 1286-1304). Berlin: De Gruyter.

Blei, David M. (2012). Topic Modeling and Digital Humanities. *Journal of Digital Humanities, 2*.

Blei, David, Ng, Andrew Y., & Jordan, Michael I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning, 3*, 993-1022.

Bloom, Paul. (2004). *Descartes' baby: How the science of child development explains what makes us human*. New York: Basic Books.

Brett, Megan. (2012). Topic modeling: A basic introduction. *Journal of the Digital Humanities, 2*(1).

Bullinaria, John, & Levy, Joseph. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavioral Research Methods, 39*(3), 510-526.

Church, Kenneth, Gale, William, Hanks, Patrick, & Kindle, Donald. (1991). Using statistics in lexical analysis. *Lexical acquisition: exploiting on-line resources to build a lexicon*, 115.

Church, Kenneth Ward, & Hanks, Patrick. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics, 16*, 22–29.

Cohen, Emma, Burdett, Emily, Knight, Nicola, & Barrett, Justin. (2011). Cross-Cultural Similarities and Differences in Person-Body Reasoning: Experimental Evidence From the United Kingdom and Brazilian Amazon. *Cognitive Science, 35*, 1282-1304.

Dunning, Ted. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics, 19*, 61–74.

Evans, Jonathan. (2008). Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annual Review of Psychology, 59*, 255–278.

Geaney, Jane. (2002). *On the Epistemology of the Senses in Early Chinese Thought*. Honolulu, HI: University of Hawaii Press.

Goldin, Paul. (2003). A Mind-Body Problem in the Zhuangzi? In Scott Cook (Ed.), *Hiding the World in the World: Uneven Discourses on the Zhuangzi* (pp. 226-247). Albany: State University of New York Press.

Goldin, Paul. (2015). The Consciousness of the Dead as a Philosophical Problem in Ancient China. In R.A.H. King (Ed.), *The Good Life and Conceptions of Life in Early China and Greek Antiquity* (pp. 59-92). Berlin: De Gruyter.

Granet, Marcel. (1934). *La Pensée chinoise*. Paris: La Renaissance du livre.

Gries, Stefan. (2013). 50-something years of work on collocations: What is or should be next …. *International Journal of Corpus Linguistics, 18*, 137-166. doi: 10.1075/ijcl.18.1.09gri

Gries, Stefan Th. (2013). 50-something years of work on collocations: What is or should be next.. *International Journal of Corpus Lingistics, 18*(1), 137-165.

Hall, David, & Ames, Roger. (1995). *Anticipating China: Thinking through the Narratives of Chinese and Western Culture*. Albany, NY: SUNY Press.

Hou, Yufang, & Frank, Anette. (2015). *Analyzing Sentiment in Classical Chinese Poetry.* Paper presented at the Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Beijing.

Hung, Jen-Jou, Bingenheimer, Marcus, & Wiles, Simon. (2010). Quantitative evidence for a hypothesis regarding the attribution of early Buddhist translations. *Literary and Linguistic Computing, 25*(1), 119-134.

Jockers, Matthew L., & Mimno, David. (2013a). Significant themes in 19th-century literature. *Poetics, 41*, 750-769. doi: 10.1016/j.poetic.2013.08.005

Jockers, Matthew L., & Mimno, David. (2013b). Significant themes in 19th-century literature. *Poetics, 41*(6), 750-769. doi: 10.1016/j.poetic.2013.08.005

Jullien, François. (2007). *Vital nourishment: Departing from happiness* (Arthur Goldhammer, Trans.). New York: Zone Books.

Jurafsky, Dan, & Martin, James H. (2009). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed ed.). Upper Saddle River, N.J: Pearson Prentice Hall.

Jurafsky, Daniel, & Martin, James. (2015). Vector Semantics. In Daniel Jurafsky & James Martin (Eds.), *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (3rd edition draft)*.

Kahneman, Daniel. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus, Giroux.

Klein, Esther, & Klein, Colin. (2011). Did the Chinese have a change of heart? *Cognitive Science, 36*, 179–182.

Landauer, Thomas, & Dumais, Susan. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review, 104*(2), 211-240.

Lee, John, & Wong, Tak-sum. (2012). *Glimpses of Ancient China from Classical Chinese Poems.* Paper presented at the Proceedings of COLING 2012: Posters, Mumbai, India.

Lévy-Bruhl, Lucien. (1922). *La mentalité primitive* Paris: Alcan.

Manning, Christopher D., Raghavan, Prabhakar, & Schütze, Hinrich. (2008). *Introduction to information retrieval*. New York: Cambridge University Press.

Manning, Christopher, & Schütze, Hinrich. (1999). *Foundations of Statistical Natural Language Processing* (1 edition ed.). Cambridge, Mass: The MIT Press.

Masterman, Margaret. (1962). The Intellect’s New Eye. In D.J. Foskett (Ed.), *Freeing the Mind: Articles and Letters from The Times Literary Supplement during March-June, 1962* (pp. 38-44). London: Times.

Mautner, Gerlinde. (2007). Mining large corpora for social information: The case of elderly. *Language in Society, 36*(01), 51-72.

Miner, Gary, Elder, John, Hill, Thomas, Nisbet, Robert, Delen, Dursun, & Fast, Andrew. (2012). *Practical text mining and statistical analysis for non-structured text data applications* (1st ed ed.). Waltham, MA: Academic Press.

Mohr, John W., & Bogdanov, Petko. (2013). Introduction—Topic models: What they are and why they matter. *Poetics, 41*(6), 545-569. doi: 10.1016/j.poetic.2013.10.001

Moretti, Franco. (2007). *Graphs, maps, trees: abstract models for literary history*. London; New York: Verso.

Moretti, Franco. (2013). *Distant Reading*. London: Verso.

Nichols, Ryan, Nielbo, Kristoffer, Slingerland, Edward, Bergeton, Uffe, Logan, Carson, & Kleinman, Scott. (Under Submission). Topic Modeling Ancient Chinese Texts: Knowledge Discovery in Databases for Asianists. *Journal of Asian Studies*.

Oakes, M. (1998). *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Paperno, Denis, Marelli, Marco, Tentori, Katya, & Baroni, Marco. (2014). Corpus-based estimates of word association predict biases in judgment of word co-occurrence likelihood. *Cognitive Psychology, 74*, 66-83. doi: 10.1016/j.cogpsych.2014.07.001

Plasse, Marie, Niang, Ndeye, Saporta, Gilbert, Villeminot, Alexandre, & Leblond, Laurent. (2007). Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. *Computational Statistics & Data Analysis, 52*, 596-613. doi: 10.1016/j.csda.2007.02.020

Poli, Maddalena. (2016). *Me, Myself and I: The Notion of Self in the Zhuangzi.* (MA), Università Ca'Foscari, Venezia, Italy.

Prentice, S., Rayson, P., & Taylor, P.J. (2012). The language of Islamic extremism: Towards an automated identification of beliefs, motivations and justifications. *International Journal of Corpus Linguistics, 17*(2), 259-286.

Ramsay, Stephen. (2011). *Reading machines: toward an algorithmic criticism*. Urbana: University of Illinois Press.

Rockwell, Geoffrey, & Sinclair, Stéfan. (2016). *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. Cambridge, MA: MIT Press.

Rohde, Douglas, Gonnerman, Laura, & Plaut, David C. (2005). An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence.

Rosemont, Henry, Jr., & Ames, Roger. (2009). *The Chinese Classic of Family Reverence*. Honolulu: University of Hawai'i Press.

Sampson, G, & McCarthy, D. (Eds.). (2005). *Corpus linguistics: readings in a widening discipline*. New York: Continuum.

Slingerland, Edward. (2013). Body and Mind in Early China: An Integrated Humanities-Science Approach. *Journal of the American Academy of Religion, 81*(1), 6-55.

Slingerland, Edward, & Chudek, Maciej. (2011). The prevalence of mind-body dualism in early China. *Cognitive Science, 35*, 997-1007.

Slone, D. Jason. (2004). *Theological incorrectness : why religious people believe what they shouldn't*. Oxford ; New York: Oxford University Press.

Smith, John B. (1984). A new environment for literary analysis. *Perspectives in computing, 4*(2/3), 20-31.

Tangherlini, Timothy R., & Leonard, Peter. (2013). Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research. *Poetics, 41*(6), 725-749. doi: 10.1016/j.poetic.2013.08.002

Teubert, W., & Čermáková, A. . (2007). *Corpus linguistics: a short introduction*. New York: Continuum.

Torjet, Andrew, & Christensen, Jon. (2012). Building new windows into digitized newspapers. *Journal of Digital Humanities, 1*(3).

Van Norden, Bryan. (2008). *Mengzi: With Selections from Traditional Commentaries*. Cambridge, MA: Hackett Publishing Company.

1. We are grateful to Dr. Sturgeon, Postdoctoral Fellow in Chinese Digital Humanities and Social Sciences at the Fairbank Center for Chinese Studies, Harvard University, for permission to download the corpus in its entirety for analysis purposes. The CTP also has some built-in analysis tools that can be extremely useful for scholars, and can be subscribed to for full download access. [↑](#footnote-ref-1)
2. See our on-line materials (URL), Appendix 1 for a complete list of texts, with their Era and Genre tags. Because of controversies concerning precise dating of early Chinese texts, and well as concerns about some of the genre labels employed in CTP, none of this meta-data was employed in studies reported here except for the separating out of Medical texts from all other genres in Study 2. [↑](#footnote-ref-2)
3. We recognize many objections that historians could make to the CTP genre and school categories, but believe this figure gives a sense of the breadth of the corpus. [↑](#footnote-ref-3)
4. Our stop-word list is presented in Appendix 2 (URL). In retrospect, the stop list used for the current study was overly aggressive in certain respects (removing, for instance, words such as “woman” (*nu* 女) or “listen/obey” (*ting* 聽) that would have been useful to include) and failing to eliminate common words such as *ren* 人 (person, human) that may have distorted some of the topic modeling or other analyses presented below. The creation of properly calibrated stop-word lists is obviously an area where careful thought and expertise are required, and what constitutes a well-chosen list very much depends on the purpose of the analysis. For instance, although pronouns and grammatical particles are typically included in stop lists, a scholar interesting in problems of textual dating may very well want to include them. [↑](#footnote-ref-4)
5. Medical texts contain a total of 226,071 characters, or 3.94% of our corpus’s 5.74m characters, and include the *Huangdi Neijing* (黃帝內經, *The* *Yellow Emperor’s Inner Canon*), the *Huangdi Bashiyi Nanjing*(黃帝八十一難經, *The Yellow Emperor‘s Canon of Eighty-one Difficult Issues*), the *Shang Han Lun* (傷寒論, *Treatise on Cold Injury*), and the *Jinkui Yaolue* (金匱要略, *Essential Medical Treasures of the Golden Chamber*). [↑](#footnote-ref-5)
6. *Mai*脈appears 2135 times in the four medical texts listed in footnote 4, but only 227 times anywhere else in the entire CTP corpus. [↑](#footnote-ref-6)
7. 問曰：人病恐怖者，其脈何狀？師曰：脈形如循絲纍纍然，其面白脫色也. From the “Methods for Determining Standard Pulses” (*pingmaifa*平脈法) chapter, 7.1. [↑](#footnote-ref-7)
8. Specifically, Unweighted Pair Group Method with Arithmetic Mean (UPGMA). For technical discussions of this and related techniques, see Andrews & Fox 2007. [↑](#footnote-ref-8)
9. Although not strictly necessary, the preprocessed corpus was normalized by slicing it into 1000 characters slices. To verify association robustness, additional trials were run with non-normalized texts and slice lengths of 100 and 500 characters. All trials produced almost identical clusterings of the query terms, suggesting that our results are extremely robust. [↑](#footnote-ref-9)
10. The hierarchical clustering algorithm was run before we had finalized the details of Study 1, and so the control terms differ somewhat from what we have reported above. Since the point is merely to validate the methodology’s ability to identify semantically coherent clusters, we decided not to rerun the study, given the enormous time and computational power required. [↑](#footnote-ref-10)
11. Reiterated runnings of the hierarchical clustering algorithm continued to produce the same tree, with the exception of *fu* 腹, which sometimes flipped up to the tree containing *xin*, the sensory and communication organs, and the body. This makes sense considering that 腹 is the outlier in the bottom of the tree as presented in Figure 8. [↑](#footnote-ref-11)
12. The most commonly-used method for topic modeling in the humanities (and the method that we employed) is called Latent Dirichlet Allocation (LDA). A reasonably non-technical (at least in the beginning!) introduction to LDA from one of its creators, David Blei, can be found in D. M. Blei 2012; perhaps more useful for most humanists is Brett 2012 and a blog post by one of the pioneers in digital humanities, Ted Underwood (<https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/>

    ). Helpful special journal issues include Volume 2, Number 1 (Winter 2012) of the *Journal of Digital Humanities*, which includes some application pieces in addition to Brett 2012 and Blei 2012. Also see the contributions to the special issue of *Poetics* 41.6 (December 2013), especially the introduction to topic modeling by Mohr & Bogdanov 2013. [↑](#footnote-ref-12)
13. See, for instance, an analysis of a Texas newspaper article archive (Torjet & Christensen 2012), an 18th-century midwife’s diary (<http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/>), and literary themes in 19th century literature (Jockers & Mimno 2013b). Topic modeling has also been used as a more effective methods than simple keyword searches for turning up themes of interest in massive, relatively unknown corpora (Tangherlini & Leonard 2013). [↑](#footnote-ref-13)
14. The other three topics in which *xin* appears are “stylistic” topics, picking up unique clusters of specialized terminology or grammatical particles that are distinctive of a particular text. They are described and discussed in Appendix 4 (URL). [↑](#footnote-ref-14)
15. Although only *yang* 陽 appears in the top ten characters, *yin* 陰is not far behind at fourteenth place. [↑](#footnote-ref-15)
16. “Crime/guilt” (*zui*罪) also appears thirteenth in the topic. [↑](#footnote-ref-16)
17. Topic #73 loads at high levels into two of our four medical texts, 60% in the *Huangdi Neijing* (黃帝內經, *The* *Yellow Emperor’s Inner Canon*) and 70% in the *Huangdi Bashiyi Nanjing*(黃帝八十一難經, *The Yellow Emperor‘s Canon of Eighty-one Difficult Issues*), with some minor loading in the other two texts *Shang Han Lun* (傷寒論, *Treatise on Cold Injury*) (9%) and *Jinkui Yaolue* (金匱要略, *Essential Medical Treasures of the Golden Chamber*) (6%). Topic #84, on the other hand, loads at 80% into the *Shang Han Lun* and the *Jinkui Yaolue*, although only 4% into the *Nanjing* and 1% in the *Shang Han Lun*. The former texts are generally attributed to the late Warring States or earlier portion of the Han Dynasty (206 BCE – 220 CE) and the latter texts to the second half, so our topic model might here be corroborating these attributions by picking up an interesting chronological shift in medical terminology, one that could then be used to date less well-attested texts. [↑](#footnote-ref-17)
18. Topics #73 and #84 are shown in Appendix 4 (URL), along with one very minor topic, Topic #11. Topic #11 is included because, at first glance, “stomach” (*wei* 胃) seems to appear in it. Topic #11 loads almost exclusively into the two archaeological versions of the *Daodejing*—the Mawangdui (20%) and Guodian (6%)—in the CTP corpus. Digging into the passages themselves, however, it is clear that *wei*胃 here is simply the graphic variant for *wei* 謂(“to be called”) that is employed in the Mawangdui *Daodejing*. Topic #11 thus seems to be another stylistic topic, picking out terms distinctive to the archaeological versions of the text. These include the use of *bang*邦in place of *guo* 國 to convey “state,” a feature of texts written before the tabooing of the personal name of the first Han Emperor, Liu Bang 劉邦, upon his accession in 206 BCE. This topic also loads in the two lexicons in the corpus (the *Shuowen* and *Erya*, both at 2%), which no doubt reflects their inclusion of archaic word forms. It also shows up in the poetry collection, the *Songs of Chu* (*chuci* 楚辭), at 1%, which might reflect stylistic features common to texts from the ancient state of Chu, where the Mawangdui and Guodian tombs were located. Topic #11 thus gives us another good example of how topic modeling can pick up clusters of distinctive grammar or terminology that could be useful in dating texts or adjudicating debates about origins or transmission of received texts. [↑](#footnote-ref-18)
19. Cf. the comment by Rockwell and Sinclair that literary scholars frequently employ “semi-quantitative words” such as “more” or “less” in their arguments, as in “There is a lot more discussion in *Frankenstein* about technology than other novels.” “Whether or not [claims like this] are right,” they observe, “we are making a claim that can be investigated by using quantitative tools to count words. That is why we should be beware of hard distinctions such as that between hermeneutical and quantitative methods” (Rockwell & Sinclair 2016: 41). [↑](#footnote-ref-19)