Exploring Chinese Poetry with Digital Assistance: Examples from Linguistic, Literary, and Historical Viewpoints

CHAO-LIN LIU, THOMAS J. MAZANEC, and JEFFREY R. THARSEN

Abstract Digital tools provide instrumental services to the study of Chinese poetry in an era of big, open data. The authors employed nine representative collections of Chinese poetry, covering the years 1046 BCE to 1644 CE, in their demonstrations. They demonstrate sophisticated software that allows researchers to extract source material that meets multiple search criteria, which may consider words, poets, collections, and time of authoring, paving the way for new explorations of Chinese poetry from linguistic, literary, artistic, and historical viewpoints. Analytic tools help researchers uncover information concealed in poetic works that are related to aesthetic expressions, personal styles, social networks, societal influences, and temporal changes in Chinese poetry. The increasing accessibility of digitized texts, along with sophisticated digital tools, such as the ones these authors developed and demonstrate here, can thereby enhance the efficiency and effectiveness for exploring and studying classical Chinese poetry.

Keywords collocations, word patterns, temporal analysis, social networks, stylometry

Studying Chinese poetry is crucial for understanding Chinese language and literature. With the rapidly increasing availability of digitized texts of Chinese poetry, digital tools can provide unprecedented flexibility for researchers to study, compare, and analyze Chinese poems from a wide variety of viewpoints. In this article, we discuss and present some digital tools to demonstrate these potentials.

The Journal of Chinese Literature and Culture • 5:2 • November 2018
DOI 10.1215/23290048-7257002 • © 2019 by Duke University Press

We can conduct basic statistical analyses of poems, offering methods for both close and distant reading. These basic analyses include the frequencies of words and positions of words in poems. For instance, we can study the roles of words describing color, weather conditions, flowers and plants, astronomical objects, rivers and mountains, and even human senses in verses. Software can also assist us in studying how words were placed in the poems to create scenes and imagery.

Digital tools can also help us find relationships among poets. We can study and compare the words, collocations, and patterns used by the poets. Although the choices of words and their combinations are often governed by rules of rhyming, it is the verses that reveal information about the background, mental state, and social status of the poets, which may offer links to other kinds of humanistic inquiry. These comparisons can be carried out at different levels of granularity, that is, person to person or corpus to corpus. Treating poets' names and other appellations as just words allows us to analyze the social network of the poets based on the titles and contents of the poems. Information about social networks of poets can be useful for studying the styles of poets and for enriching biographical databases of Chinese people, for example, the China Biographical Database (projects.iq.harvard.edu/cbdb/home; hereafter CBDB).[1]

At this moment, we have nine collections of Chinese poetry with which we can demonstrate the functions of our software. Their contents were created starting from the Zhou dynasty to the Ming dynasty, corresponding approximately to years 1046 BCE to 1644 CE. They include the *Shijing* 詩經 (The Book of Odes), *Chuci* 楚辭 (The Songs of Chu), the Han-dynasty *fu* 漢賦 found in the sixth-century anthology *Wenxuan* 文選 (Selections of Refined Literature), the collected poetry of pre-Sui China (*Xian Qin Han Wei Jin Nanbeichao shi* 先秦漢魏晉南北朝詩), *Quan Tang shi* 全唐詩 (Complete Tang-Dynasty Poems; hereafter *QTS*), *Quan Song ci* 全宋詞 (Complete Song-Dynasty Lyrics; hereafter *QSC*), *Quan Song shi* 全宋詩 (Complete Song-Dynasty Poems; hereafter *QSS*), *Yuan shi xuan* 元詩選 (Selected Yuan-Dynasty Poetry), and *Liechao shiji* 列朝詩集 (Arrayed Poetry Collections of the [Ming] Dynasty). With these collections, we can explore whether Chinese poetry changes significantly over time.

Pioneers who are interested in applying computing technologies to the study of Chinese poetry have proposed some interesting ideas.[2] Our tools help researchers explore the poems from the perspectives of words, poets, and time periods. Basically, they are designed to answer questions like who writes what, when, how, and where. Well-designed tools should facilitate any innovative research about poems that meet particular requirements for the researchers using them. Here we present the functions of the tools one by one.

The main purpose of this article is to introduce the ways in which our digital tools can assist researchers. We first give more information about our

Table 1. The corpora used in this study: Chinese poetry of 1046 BCE–1644 CE

Name (Acronym)	*Published Period*	*Name (Acronym)*	*Published Period*
Shijing 詩經 (*SJ*)	1046–476 BCE	*Chuci* 楚辭 (*CV*)	475–221 BCE
Han-dynasty *fu* in *Wenxuan* 漢賦(昭明文選) (*HF*)	202 BCE to 420 CE	*Xian Qin Han Wei Jin Nanbeichao shi* 先秦漢魏晉南北朝詩 (*PT*)	Before 589 CE
Quan Tang shi 全唐詩 (*QTS*)	618–907 CE	*Quan Song shi* 全宋詩 (*QSS*)	960–1279 CE
Quan Song ci 全宋詞 (*QSC*)	960–1279 CE	*Yuan shi xuan* 元詩選 (*YSX*)	1271–1368 CE
Liechao shiji 列朝詩集 (*LCSJ*)	1368–1644 CE		

corpora of Chinese poems, followed by an explanation of the tools constructed based on a few basic functions. We then go through several sample applications of the tools and provide a list of practical applications that demonstrate the potential of our tools. We end with a discussion of some remaining challenges for future work.

Sample Collections

We have collected nine representative corpora of Chinese poetry, each for a major period in Chinese history between 1046 BCE and 1644 CE. We list the corpora in table 1, where we assign an acronym to each corpus for easier reference to the collections. The table also shows the Chinese names of the corpora and the time periods represented by them. We do not yet have a collection for the Qing dynasty (1644–1911) because an editorial committee is still working toward this very challenging goal.[3]

We use works in *QTS* and *QSC* to illustrate functions of our tools. A piece of work in the *QTS* is called *yishou shi* 一首詩 (a poem), and a piece of work in the *QSC* is called *yique ci* 一闋詞 (a lyric). Normally, we do not refer to works in the *QSC* as poems. For the sake of simplicity, we will refer to all works in these collections as "poems" henceforth, although not all of them, strictly speaking, belong to the same indigenous literary genre.

Excluding the punctuation marks that were added by later editors, we have more than 16.5 million characters in the corpora.[4] Table 2 lists the number of items, types, and tokens in the collections. We count the number of items, rather than the number of poems, in the collections because some items appear to be just incomplete poems. For example, the titles of some items in *QTS* are *ju* 句 (lines), and their contents contain two or more lines, so it is not easy to judge whether they are complete poems. The "Types" column shows the number of distinct characters, and the "Tokens" column shows the total number of characters in the collections. We disregard obscure characters that require special fonts to show on normal computers. In practice, such characters are rare in our

Table 2. Basic statistics for the collections in table 1

Acronym	*Items*	*Types*	*Tokens*	*Acronym*	*Items*	*Types*	*Tokens*
SJ	311	2,775	29,651	*CV*	65	3,044	26,947
HF	490	5,775	73,067	*PT*	10,062	5,971	463,031
QTS	42,863	7,275	2,590,695	*QSS*	185,112	9,202	9,472,518
QSC	19,394	5,780	1,347,482	*YSX*	15,772	7,442	1,251,998
LCSJ	16,172	7,229	1,284,496				

collections, although ignoring them makes our statistics less precise than one might wish.

We cannot claim that the digitized texts we have obtained are the most complete and official possible. Although it is possible to acquire the digitized texts from online sources such as the WikiSource (zh.wikisource.org/zh-hant/), the Chinese Text Project (ctext.org/), Wenxue100 (www.wenxue100.com/), and Daizhige 殆知閣 (www.daizhige.org/), it is beyond our current capacity to make sure that the contents are perfectly authentic.[5]

We employed the collections listed in table 1 to demonstrate the functions of our software. Although sometimes we suggest possible implications of the results produced by the software, we must remind the reader that the reliability of the observations and claims depends on the quality of the collections, which could still be improved.

Basic Functions

We have built our software services based on three basic types of functions. One may use them to look for words, collocations of words, and even specific patterns of words in the poems. Locating the occurrences of words is perhaps the most fundamental function. After locating the words, we can calculate their frequencies and compute potentially useful statistics based on these frequencies to provide some distant-reading insights. It is also helpful to compare the poems between and within the collections to study how words, collocations, and patterns were shared and passed from dynasty to dynasty.

Words, Collocations, and Concordances

Looking for a specific character in the poems is perhaps the most basic service that one can imagine. The usefulness of such basic services can be supported by many printed books that provide the same type of information, such as concordances.[6]

Figure 1 lists some examples of Tang poems that use the character *bai* 白 (white). Each row in figure 1 is a partial excerpt of a Tang poem, showing at most twenty characters to the left and twenty characters to the right of *bai*. In a real

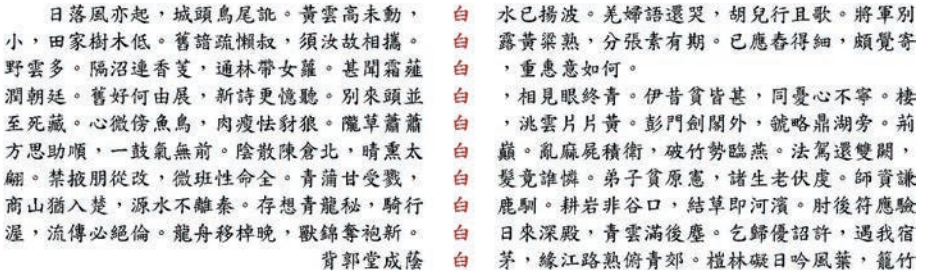

Figure 1. Samples of *bai* 白 (white) in Du Fu's poems (from top to bottom, these poems appear on the following pages: *QTS* 225.2424, *QTS* 225.2426, *QTS* 225.2426, *QTS* 225.2427, *QTS* 225.2427, *QTS* 225.2428, *QTS* 225.2428, *QTS* 225.2429, *QTS* 225.2430, *QTS* 226.2432)

system, researchers could choose to read either the complete poems or the context of the queried target—*bai* in this example. When examining only the contexts, one chooses the window size, that is, the number of characters surrounding the queried targets, that one wants to read.

When we convert the contents of printed books to digitized texts, digital tools can offer more flexible types of queries than looking for individual characters. Listing all of the Chinese words in a printed book is not practical because of the nearly infinite number of Chinese words. In contrast, digital tools can efficiently provide information about word occurrences. Figure 2 shows some examples of *bairi* 白日 (white sun) in Tang poems. Again, the target words are aligned to facilitate comparison across the contexts in which the target words appear.

In the study of Chinese poetry, researchers often look for occurrences of two words in the same poem. Two words may collocate in a poem, and two words may also form an antithetical pair consisting of words that appear at corresponding positions in a poem, with rhyming, syntactic, and semantic properties following a specific set of rules.[7] Two words that collocate may not be

Figure 2. Samples of *bairi* 白日 (white sun) in Bai Juyi's poems (from top to bottom, these poems appear on the following pages: *QTS* 432.4774, *QTS* 433.4788, *QTS* 433.4795, *QTS* 434.4801, *QTS* 434.4801, *QTS* 435.4810, *QTS* 435.4810, *QTS* 435.4823, *QTS* 435.4823, *QTS* 438.4861)

Figure 3. Samples of *bairi* 白日 (white sun) and *qingshan* 青山 (green mountains) in Tang poems (from top to bottom, these poems appear on the following pages: *QTS* 285.3267, *QTS* 306.3477, *QTS* 378.4244, *QTS* 431.4754, *QTS* 432.4771, *QTS* 515.5887, *QTS* 529.6047, *QTS* 586.6791, *QTS* 589.6833, *QTS* 711.8188)

considered an antithetical pair if the properties of the words do not meet the governing rules. Both collocation and antithetical pairs can create specific imagery in the poems, so finding the poems that contain word pairs of interest for researchers is a valuable service.

Figure 3 shows some examples of collocation and antithetical pairs for *bairi* and *qingshan* 青山 (green mountains) in the *QTS*. The poems are aligned by their instance of *bairi*. It is also possible to align them by their instance of *qingshan*, and our current tools can do so. In figure 3, *qingshan* is underlined manually to make reading easier.

Aligned characters and words in poems as shown in figures 1 and 2 are called *concordances* in linguistics. This style of listing allows researchers to study the contexts of the aligned words. To show the collocation of two words, as in figure 3, the researcher may choose which of the two words to align.

Figure 4 uses an alternate way to show the usages of *bairi*. Each row in figure 4 is a Tang poem, and each regularly consists of two pairs of two lines separated by "。."[8] The poems are organized by the locations of *bairi*, so figure 4 provides a visual impression of the distributions of positions of *bairi* in the poems, something lost in typical concordances.

Quantitative Analysis: Frequencies and Proportions

Computing the frequencies of words and their collocations offers a different view of the poems than what one perceives from simply listing the original poems for close reading. If we can find the poems that include the characters, words, and their collocations, we can record their frequencies at the same time.

As by-products of producing the listings in figures 1–3, we find that there are 8,453, 698, and 18 instances of *bai* 白 (white), *bairi* 白日 (white sun), and *bairi-qingshan* 白日-青山 (white sun–green mountains) within twenty characters in our *QTS* corpus.[9]


白日浮雲閉不開，黃沙誰問冶長猜。只憐橫笛關山月，知處愁人夜夜來。
白日蒼蠅滿飯盤，夜間蚊子又成團。每到更深人靜後，定來頭上咬楊鸞。
平流白日無人愛，橋上閑行若個知。水似晴天天似水，兩重星點碧琉璃。
紅塵白日長安路，馬足車輪不暫閑。唯有茂陵多病客，每來高處望南山。
海燕西飛白日斜，天門遙望五侯家。樓臺深鎖無人到，落盡春風第一花。
西望長安白日遙，半年無事駐蘭橈。欲將張翰秋江雨，畫作屏風寄鮑昭。
三月盡是頭白日，與春老別更依依。憑鶯為向楊花道，絆惹春風莫放歸。
手內青蛇淩白日，洞中仙果豔長春。須知物外煙霞客，不是塵中磨鏡人。
年年不見帝鄉春，白日尋思夜夢頻。上酒忽聞吹此曲，坐中惆悵更何人。
娉婷十五勝天仙，白日妲娥旱地蓮。何處閑教鸚鵡語，碧紗窗下繡床前。
春花秋月入詩篇，白日清宵是散仙。空卷珠簾不曾下，長移一榻對山眠。
漠漠輕陰晚自開，青天白日映樓臺。曲江水滿花千樹，有底忙時不肯來。
石抱龍堂蘚石幹，山遮白日寺門寒。長松瀑布饒奇狀，曾有仙人駐鶴看。
未識東西南北路，青春白日坐銷難。如何一別故園後，五度花開五處看。
李白曾歌蜀道難，長聞白日上青天。今朝夜過焦崖閣，始信星河在馬前。
軒窗縹緲起煙霞，誦訣存思白日斜。聞道昆侖有仙籍，何時青鳥送丹砂。
朱絲弦底燕泉急，燕將雲孫白日彈。嬴氏歸山陵已掘，聲聲猶帶發衝冠。
何曾解報稻粱恩，金距花冠氣遏雲。白日梟鳴無意問，唯將芥羽害同群。
慈恩雁塔參差榜，杏苑鶯花次第遊。白日有愁猶可散，青山高臥況無愁。
右翅低垂左脛傷，可憐風貌甚昂藏。亦知白日青天好，未要高飛且養瘡。
鶴老芝田雞在籠，上清那與俗塵同。既言白日升仙去，何事人間有殯宮。
群玉山頭住四年，每聞笙鶴看諸仙。何時得把浮丘袖，白日將升第九天。
十二山晴花盡開，楚宮雙闕對陽臺。細腰爭舞君沉醉，白日秦兵天下來。
曲岸蘭叢雁飛起，野客維舟碧煙裡。竿頭五兩轉天風，白日楊花滿流水。
去日家無擔石儲，汝須勤若事樵漁。古人盡向塵中遠，白日耕田夜讀書。
常言吃藥全勝飯，華岳松邊采茯神。不遣髭須一莖白，擬為白日上升人。
元和天子丙申年，三十三人同得仙。袍似爛銀文似錦，相將白日上青天。
渠水紅繁擁禦牆，風嬌小葉學娥妝。垂簾幾度青春老，堪鎖千年白日長。
文武千官歲仗兵，萬方同軌奏升平。上皇一禦含元殿，丹鳳門開白日明。
小圃初晴風露光，含桃花發滿山香。看花對酒心無事，倍覺春來白日長。

Figure 4. Positions of *bairi* 白日 (white sun) in thirty QiJue (7_JUE) poems in *QTS* (from top to bottom, these poems appear on the following pages: *QTS* 150.1560, *QTS* 871.9876, *QTS* 477.5436, *QTS* 542.6259, *QTS* 538.6138, *QTS* 698.8031, *QTS* 446.5004, *QTS* 858.9702, *QTS* 334.3751, *QTS* 442.4947, *QTS* 804.9052, *QTS* 344.3864, *QTS* 514.5865, *QTS* 587.6812, *QTS* 699.8042, *QTS* 250.2821, *QTS* 574.6688, *QTS* 681.7810, *QTS* 711.8188, *QTS* 450.5079, *QTS* 784.8854, *QTS* 365.4123, *QTS* 477.5439, *QTS* 491.5560, *QTS* 551.6386, *QTS* 574.6683, *QTS* 784.8848, *QTS* 391.4411, *QTS* 511.5837, *QTS* 689.7914)

Text Comparison

We designed the algorithm FindCommon, shown in figure 5, to compare large sets of poems efficiently (see fig. 5). We assume that we have N collections of poems, and they are S_1, S_2, . . . , S_N. Each S_i of these N collections contains a certain number, q_i, of poems. Namely, there are q_i poems in S_i. We denote the kth

Algorithm FindCommon

Input: 1. sets of poems $S=\{S_1, S_2, \ldots, S_i, \ldots, S_N\}$, each S_i is a collection of poems (either QTS or QSC or others), i.e., $S_i = \{P_{i,1}, P_{i,2}, \ldots, P_{i,k}, \ldots, P_{i,qi}\}$, where a $P_{j,k}$ is the k-th poem in S_j
2. basic filtering conditions, F
3. output format requests, R

Output: common parts of any two poems in S

Steps:

1 Compute an indexed list of characters, V, that are used in S
2 For any two poems, P_x and P_y, do the following.
2.1 Look up the characters of P_x in V, and save the indexes for the characters in I_x. Repeat this step for P_y to create I_y.
2.2 Compare the indexes in I_x and I_y to find the characters that appear in both P_x and P_y. Record the locations of the common characters in C_x and C_y, respectively.
2.3 Emit the common words in format R, along with basic information about P_x and P_y, if the common words satisfy F.

Figure 5. Our algorithm for comparing poems. For details, see text.

poem in a collection S_i by $P_{i,k}$. Specifically, if we are handling only two sets of collections (e.g., *QTS* and *QSC*), $N=2$. If there are, respectively, 50,000 and 20,000 items in *QTS* and *QSC*, $q_1 = 50{,}000$ and $q_2 = 20{,}000$.

The main steps of FindCommon are intuitive. Given a certain number of collections, the first step identifies the set of characters used in the collections. Each character is assigned a unique identification number, which simply shows the order that FindCommon encounters the character—the identification numbers are not places of the characters in poems. Step 2.1 translates the characters in a poem into a set of integers based on the indexes of the characters of all of the collections that we obtained at the first step. Step 2.2 compares the sets of integers that encode the characters of two poems. Characters that appear in two poems can be easily revealed. Step 2.3 produces an output report about the similarity between poems based on the filtering conditions (F) and output format (R) settings.

When we compare the poems, we may find common characters, common words, or common collocations. If researchers are focusing on a special type of investigation, the filtering conditions in FindCommon allow the researchers to specify what levels of similarity they are looking for. Researchers choose the viewpoint from which the results are displayed using output format settings, a concept we explain shortly.

Let us explain the main steps of FindCommon with a running example, assuming that we have only *QTS* and *QSC* as S_1 and S_2. To further simplify the illustration, let us assume that *QTS* contains only two items and *QSC* only one.

In *QTS*, we have the following two poems by Liu Yuxi 劉禹錫 (772–842):

$P_{1,1}$:[10]

	Mountains surround the old land, situated in their midst.	山圍故國周遭在
2	Waves beat against deserted city walls, then return in silence.	潮打空城寂寞回
	East of the River Huai, the moon of old	淮水東邊舊時月
4	Returns deep in the night, passing over the ramparts.	夜深還過女牆來
		QTS 365.4117

$P_{1,2}$:[11]

	Wild flora grow beside Vermillion Bird Bridge,	朱雀橋邊野草花
2	Evening sunlight skews at the mouth of Wuyi Harbor.	烏衣巷口夕陽斜
	The swallows in front of the ancestral hall of the Wangs and Xies of old	舊時王謝堂前燕
4	Fly in, searching for a household of commoners.	飛入尋常百姓家
		QTS 365.4117

In *QSC*, we have the following item authored by Zhou Banyan 周邦彥 (1056–1121)

$P_{2,1}$:[12]

	In that land of splendor,	佳麗地
2	Who recorded the glorious deeds of the southern dynasties?	南朝盛事誰記
	The mountains surrounded the old land, encircling the clear river,	山圍故國繞清江
4	And face coiled hair.	髻鬟對起
	Raging waves in silence beat against lonely city walls,	怒濤寂寞打孤城
6	Windsails cross over the horizon afar.	風檣遙度天際
	The tree on that singular cliff seems nearly inverted.	斷崖樹猶倒倚
8	Who tied up Mo Chou's skiff?	莫愁艇子誰係
	Traces of old are thickly clustered in the emptiness,	空餘舊跡鬱蒼蒼
10	The garrisons sunk half beneath the frost.	霧沉半壘
	Deep in the night, the moon comes over the ramparts,	夜深月過女牆來

12	And with a broken heart I look east to the River Huai.	傷心東望淮水
	In what market do the alehouse banners sport and sway?	酒旗戲鼓甚處市
14	The images are faint:	想依稀
	In the Wangs' and Xies' neighborhood,	王謝鄰裏
16	The swallows don't know what year it is,	燕子不知何世
	But look for someone from a household of common lanes and alleys,	向尋常巷陌人家
18	Then face each other, as if speaking of glory and loss	相對如說興亡
	In the skewed sunlight.	斜陽裏
		QSC 2.612

First, we scan the contents of every poem in the data sets and record each different character in a list, generating V, the vocabulary of characters in all of the poems. The characters are indexed for efficient lookup operations, and this list serves as a basis for comparing the contents of individual poems. With the three poems, V is the list of characters and their indexes, that is, {山:0, 圍:1, 故:2, . . . , 月:20, 夜:21, 深:22, 還:23, 過:24, 女:25, 牆:26, 來:27, . . . }. We chose to index at the character level so that we can find all of the characters shared among poems.

The indexes show the order in which FindCommon encounters the character, not the places of the characters in the poems. A character will appear in V only once, even if the character appears in multiple poems.

At step 2.1, we convert a poem into a list of integers. We translate each character in a poem using the indexes we stored in V. In this illustration, the list of indexes for $P_{1,1}$, that is, I_{11}, will be "0, 1, 2, . . . , 27." The line 夜深月過女牆來 (Deep in the night, the moon arrives over the ramparts) in $P_{2,1}$ will be translated to "21, 22, 20, 24, 25, 26, 27" in I_{21}. (Note that number 23 is for *huan* 還, which is not in this line.)

At step 2.2, we compare the lists of indexes for the two poems under consideration (P_x and P_y) to find common characters. Comparing indexes of characters is computationally more efficient than directly comparing the characters. The character segments appearing in the first poem that contain characters used in both poems are returned (C_x), as are the character segments appearing the second poem (C_y). Because the sequences containing the characters in common may not be the same, C_x and C_y are not necessarily the same. After computing the intersection of the index lists, we can determine that *yue* 月 (moon), *yeshen* 夜深 (deep in the night), and *guo nüqiao lai* 過女牆來 (arrives over the ramparts) appear in $P_{1,1}$ and $P_{2,1}$. Note that $P_{2,1}$ does not use *huan* 還 (return), so C_{11} will read {. . . , 月， 夜深，過女牆來}. C_{11} records the character segments of $P_{1,1}$ that are formed by characters that also appear in $P_{2,1}$.

Likewise, each character in "夜深月過女牆來" of $P_{2,1}$ appeared in $P_{1,1}$, so C_{21} would read like {. . . , 夜深月過女牆來, . . . }.

At step 2.3, we can select the strings that would appear in the final report. If researchers are not interested in unigrams, like *yue* 月 (moon) in this illustration, we can remove strings that are shorter than a given threshold, and this can be requested via the filtering conditions in the input.

The example we just elaborated is a famous example of the reuse of multiple terms from diverse sources to compose a new poem. In the current case, we may report different common strings, that is, C_{11} or C_{21}, depending on various viewpoints as explained above. This can be controlled via the output format settings (R in fig. 5). Notice that the choice of viewpoint can influence the output in a variety of ways. When we compare $P_{1,2}$ and $P_{2,1}$, C_{12} and C_{21} will contain *yang xie* 陽斜 (sunlight skews, as at dusk) and *xie yang* 斜陽 (skewed sunlight), respectively. This is because of 烏衣巷口夕**陽斜** (Evening **sunlight skews** at the mouth of Wuyi harbor) and **斜陽**裏 (in the **skewed sunlight**) in $P_{1,2}$ and $P_{2,1}$, respectively.

In summary, if we compare $P_{1,1}$ and $P_{2,1}$ and report all of the common strings (including unigrams) in terms of words in $P_{2,1}$, we will find {山圍故國，寂寞打，城，空，舊，夜深月過女牆來，東， 淮水}. If we compare $P_{1,2}$ and $P_{2,1}$ and report all of the common strings in terms of words in $P_{2,1}$, we will find {舊，王謝，燕，尋常巷，家，斜陽}.

We produce the following records after we compare S_1 (*QTS*) and S_2 (*QSC*) and report all of the common strings (including unigrams) in terms of words in $P_{2,1}$. In addition to the common words, we add the names and the identifiers of the poems that are compared for each record. A record contains three fields that are separated by "|||." We put $P_{2,1}$ in the leftmost field because the common words, which are grouped in the rightmost field, are listed in the terms that appeared in $P_{2,1}$, that is, from the viewpoint of $P_{2,1}$:

Zhou-Ban-Yan_$P_{2,1}$||| **Liu-Yu-Xi_$P_{1,1}$**|||
[山圍故國，寂寞打，城，空，舊，夜深月過女牆來，東，淮水]
Zhou-Ban-Yan_$P_{2,1}$||| **Liu-Yu-Xi_$P_{1,2}$**|||
[舊，王謝，燕，家，尋常巷，斜陽]

We can offer different viewpoints for researchers to examine the words shared by the poems. Although we read 夜深月過女牆來 (Deep in the night, the moon arrives over the ramparts) in $P_{2,1}$, this string actually came from three shorter strings in $P_{1,1}$: 月 (moon), 夜深 (deep in the night), and 過女牆來 (arrives over the ramparts). Hence, a researcher can choose to see the list of common words from any of the following two viewpoints, by appropriately setting R when running FindCommon.

Liu-Yu-Xi_P$_{11}$: (QTS 365.4117)

山圍故國周遭在，潮打空城寂寞回。淮水東邊舊時月，夜深還過女牆來。

Liu-Yu-Xi_P12: (QTS 365.4117)

朱雀橋邊野草花，烏衣巷口夕陽斜。舊時王謝堂前燕，飛入尋常百姓家。

Zhou-Ban-Yan_P21: (QSC 2.612)

佳麗地，南朝盛事誰記？山圍故國繞清江，髻鬟對起。

怒濤寂寞打孤城，風檣遙度天際。斷崖樹、猶倒倚，莫愁艇子誰係？

空餘舊跡鬱蒼蒼，霧沉半壘。夜深月過女牆來，傷心東望淮水。酒旗戲鼓甚處市？

想依稀，王謝鄰裏，燕子不知何世，向尋常巷陌人家。相對如說興亡，斜陽裏。

Figure 6. A sample result of running FindCommon. For colors and underlining, see text.

Zhou-Ban-Yan_P$_{2,1}$||| Liu-Yu-Xi_P$_{1,1}$|||
[山圍故國，寂寞打，城，空，舊，夜深月過女牆來， 東，淮水]
Liu-Yu-Xi_P$_{1,1}$ ||| Zhou-Ban-Yan_P$_{2,1}$|||
[山圍故國，打空城寂寞，淮水東，舊，月，夜深，過女牆來]

The example that we just elaborated is a famous example of using several terms from multiple sources in a new poem.[13] If we were to give a more complete account, Zhou Banyan also used a poem by Xie Tiao 謝朓 (464–99) and an anonymous *yuefu* (music bureau) poem 樂府詩 in P$_{2,1}$.[14] We do not discuss these additional poems because they are not part of *QTS* or *QSC*.

It is relatively difficult to read the output of FindCommon directly in text forms. Showing the common characters in colors may make the reading easier. Figure 6 shows one possible way of doing this: characters in P$_{1,1}$ or P$_{1,2}$ that also appear in P$_{2,1}$ are shown in red, characters that appear in both P$_{2,1}$ and P$_{1,1}$ are shown in blue, characters that appear in both P$_{2,1}$ and P$_{1,2}$ are shown in orange, and characters that appear in all the poems are shown in green. To facilitate reading in black-and-white printouts, the words shown in colors are also underlined in figure 6.

Applications

With the basic functions explained in the previous section, we can create tools to help researchers explore Chinese poetry from a variety of perspectives, including from the perspectives of words, poets, poems, time, and their combinations. We examine the outcomes of these explorations in this section.

Basic Statistics

With digitized poetry corpora, we can conduct basic search and comparison operations as described in the section on *Basic Functions*. In addition to looking for the contexts and frequencies of particular characters, words, and collocations

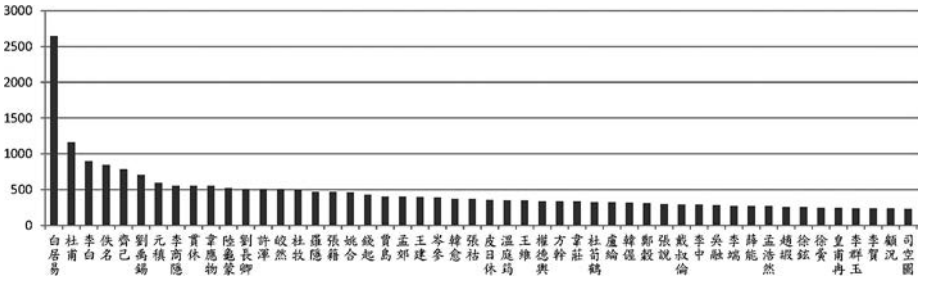

Figure 7. Numbers of items produced by the top fifty poets in *QTS*

as explained in the subsection on *Quantatitive Analysis*, we can search, calculate, and compare the frequencies of a particular type of information. For example, we found that there are about 2,500 poets in *QTS* and about 1,300 poets in *QSC*. Figures 7 and 8 show the statistics about the top fifty poets in the *QTS* and *QSC*, respectively.

In addition to searching for poets' names, we can search for the names of the tunes to which the lyrics were written (*cipai* 詞牌) for *QSC*. Figure 9 shows the number of items using the top fifty tunes that appear in our *QSC*. Again, the numbers vary widely. Another interesting result is to see word frequency in the poems. Figures 10 and 11 show the fifty most frequent bigrams in *QTS* and *QSC*, respectively.

By comparing the most frequent words listed in figures 10 and 11, one may observe and wonder about the differences in *feng* 風 (wind) in *QTS* and *QSC*. *Chunfeng* 春風 (spring wind) and *qiufeng* 秋風 (autumn wind) are among the fifty most frequent bigrams in *QTS*, but *dongfeng* 東風 (east wind) and *chunfeng* and *xifeng* 西風 (west wind) are among the fifty most frequent in *QSC*. The increased rankings of *dongfeng* and *xifeng* from *QTS* to *QSC* are quite remarkable, and *dongfeng* is even the most frequent bigram in *QSC*. With the help of digital tools, it is easy for us to look into the words that are related to *feng* in *QTS* and *QSC*.

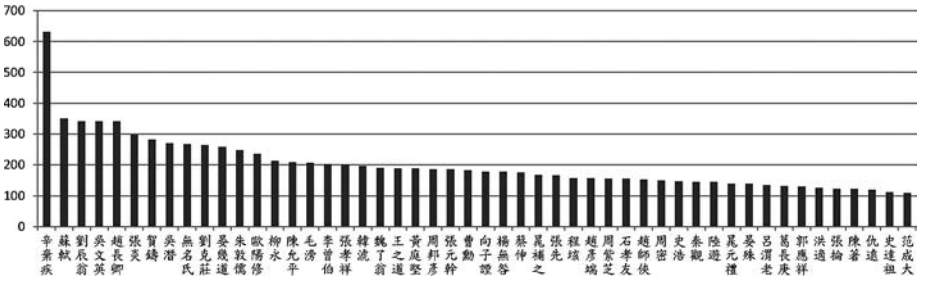

Figure 8. Numbers of items produced by the top fifty poets in *QSC*

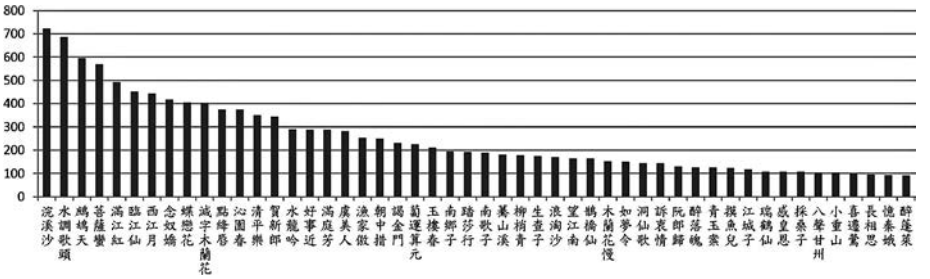

Figure 9. Numbers of items using the the top fifty named tunes in *QSC*

Table 3 shows the statistics for ten bigrams that include *feng*. The second and the fourth rows show the word frequencies in *QTS* and *QSC*, respectively. *Chunfeng* and *dongfeng* 東風 appear 1,128 and 1,360 times, respectively, and are leaders in *QTS* and *QSC*. We can divide the frequencies by the total number of items in a collection (cf. table 2) to obtain the percentages of poems that use the words, as shown in table 3. The word *chunfeng* appears in 2.63 percent (i.e., 1,128 of 42,863) of works in *QTS*, and *dongfeng* appears in 7.01 percent (1,360 of 19,394) of works in *QSC*.

Comparing the changes of percentages from *QTS* to *QSC*, *dongfeng* 東風, *xifeng*, and *chunfeng* became more popular in *QSC*. Although *qiufeng* seemed to be less popular in *QSC* due to its decreased frequency, the chance that we read an item that uses *qiufeng* remains approximately the same for both *QTS* and *QSC*. Notice that, although there are fewer lyrics in *QSC* that use *chunfeng* than there are poems that use *chunfeng* in *QTS*, the percentage of lyrics that use *chunfeng* in *QSC* is actually higher than that of the poems that use *chunfeng* in *QTS*. *Xiafeng* 夏風 (summer wind) and *dongfeng* 冬風 (winter wind) are rare in both *QTS* and *QSC*.

Statistics in table 3 also suggest that we may attempt to make the argument that *nuanfeng* 暖風 (warm wind) and *hanfeng* 寒風 (cold wind) were preferred choices for *xiafeng* 夏風 and *dongfeng* 冬風 in poems like 暖風花繞樹，秋雨草

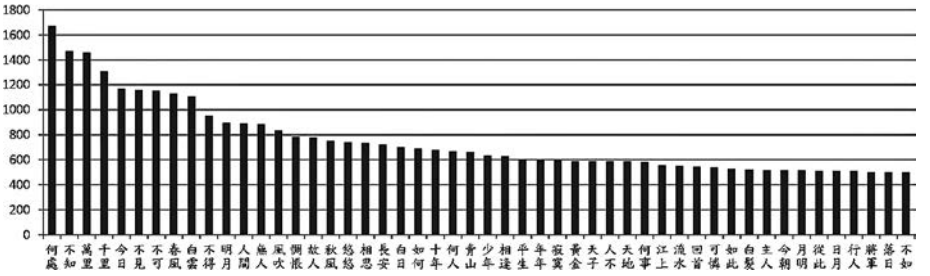

Figure 10. Frequencies of the fifty most frequent bigrams in *QTS*

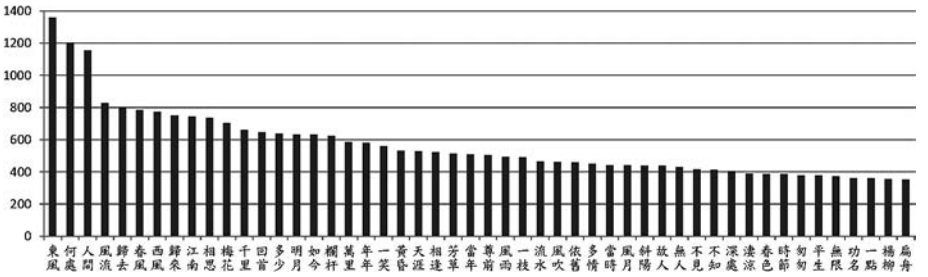

Figure 11. Frequencies of the fifty most frequent bigrams in *QSC*

沿城 (flowers surrounding the trees in warm wind, grass lining around the city in autumn rains).[15] *Xiafeng* and *dongfeng* are extremely rare, and *nuanfeng* and *hanfeng* were far more common. Further investigation would determine if *xiafeng* and *dongfeng* would make sense in the context in which *nuanfeng* and *hanfeng* appear.

Positions of Words

Much like a sequence of screenshots in movies, the positions of words may affect imagery created by the words in poems. We can quantify information about word positions like those shown in figure 4 in different ways. We may record that there are two and four instances in which *bairi* 白日 (white sun) appears at the first and tenth characters (not counting the punctuation). We may record that, in figure 4, there are three instances in which *bairi* starts at the fifth position in the last sentences, as in the line 堪鎖千年白日長 (To be locked up for a thousand years—white-sun everlasting).[16]

Table 4 provides a summary of the positions of *bairi* in four types of Tang poems in yet another way. In terms of the order of lines in poems, the first row in table 4 indicates the positions of *bairi* in poems. We use *WuJue*, *WuLu*, *QiJue*, and *QiLu* to denote, respectively, pentametric quatrains (*wuyan jueju* 五言絕句), regulated pentametric octaves (*wuyan lüshi* 五言律詩), heptametric quatrains (*qiyan jueju* 七言絕句), and regulated heptametric octaves (*qiyan lüshi* 七

Table 3. Frequencies and percentages of ten bigrams that use *feng* 風 (wind) in *QTS* and *QSC*

Corpus	Dongfeng 東風 *(east wind)*	Xifeng 西風 *(west wind)*	Nanfeng 南風 *(south wind)*	*Beifeng* 北風 *(north wind)*	Chunfeng 春風 *(spring wind)*	Xiafeng 夏風 *(summer wind)*	Qiufeng 秋風 *(autumn wind)*	Dongfeng 冬風 *(winter wind)*	Nuanfeng 暖風 *(warm wind)*	Hanfeng 寒風 *(cold wind)*
QTS	444	239	135	204	1128	3	749	3	35	64
	1.04%	0.56%	0.31%	0.48%	2.63%	0.01%	1.75%	0.01%	0.08%	0.15%
QSC	1360	772	50	21	784	0	241	0	49	26
	7.01%	3.98%	0.26%	0.11%	4.04%	0.00%	1.24%	0.00%	0.25%	0.13%

Table 4. Positions of *bairi* 白日 (white sun) in four specific types of poems in *QTS*

	Position								
Category	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Total*
Counts									
5_JUE	5	4	2	2	0	0	0	0	13
5_LU	7	10	30	27	13	14	10	5	116
7_JUE	8	9	4	9	0	0	0	0	30
7_LU	9	12	15	13	11	12	8	8	88
Percentages									
5_JUE	38.46%	30.77%	15.38%	15.38%	0	0	0	0	100%
5_LU	6.03%	8.62%	25.86%	23.28%	11.21%	12.07%	8.62%	4.31%	100%
7_JUE	26.67%	30.00%	13.33%	30.00%	0	0	0	0	100%
7_LU	10.23%	13.64%	17.05%	14.77%	12.50%	13.64%	9.09%	9.09%	100%

5_JUE, WuJue (pentametric quatrains); 5_LU, WuLu (regulated pentametric octaves); 7_JUE, QiJue (heptametric quatrains); 7_LU, QiLu (regulated heptametric octaves)

言律詩). "Position" indicates the line in a poem (e.g., 2 = second line), and the rows for WuJue and QiJue poems contain data for four lines in the columns.

The numbers in the upper part of table 4 show the frequencies of the appearances of *bairi* in WuJue, WuLu, QiJue, and QiLu. The numbers in the QiJue row are based on the QiJue poems listed in figure 4; for example, *bairi* appears in the second sentence in nine instances of QiJue poems in *QTS*. The statistics in other rows of the table were gathered based on other poems in *QTS*.

The numbers in the lower part of the table show the percentages of the appearances of *bairi* in WuJue, WuLu, WuJue, and QiLu, calculated based on the data in the upper part. We divide the count for a position by the total in the count row to obtain the percentage; for example, for WuJue, position 1, 5/13 = 38.46 percent. Sums of the percentages may not equal 100 percent because of rounding.

The statistics in table 4 indicate that *bairi* usually appears in the beginning halves of the poems, that is, the first two lines in WuJue and QiJue poems and the first four lines in WuLu and QiLu poems. From this viewpoint, *bairi* appears in the beginning halves of WuJue, WuLu, QiJue, and QiLu poems 69.3 percent, 63.8 percent, 56.7 percent, and 55.7 percent of the time.[17] More specifically, in WuLu and QiLu poems, it is more common for *bairi* to appear in the second couplet, that is, the third and the fourth lines, than in the other pairs.[18]

Colors and Scenes

Colors for poems are like soundtracks for movies: they are essential to setting the mood and creating imagery in their respective works. When we checked the most frequent unigrams in *QTS*, we found that *bai* 白 (white) is the most

frequent color word.[19] With this clue, we started to investigate the reasons for this phenomenon.[20]

We can find words starting with *bai* and calculate the percentage of a poet's poems in *QTS* that used these words. The statistics collected for thirteen renowned poets are listed in table 5, which contains two main parts. The "Frequencies" section of table 5 shows the total frequencies of the *bai*-words that appeared more than ten times in the works of thirteen poets (listed in the second column). The percentages indicate how often an individual poet used this *bai*-word; the thick boxes indicate the three (and ties) most frequent words used by the poets.[21]

For the ratios in table 5, ratio A shows the total percentage of the poet's poems in *QTS* that used the *bai*-words listed. For the column of Li Bai 李白 (701–62), it is sum of all of the percentage values listed under "Frequencies": 46.65 = 6.92 + 2.34 + · · · + 0.89. Our data show that Li Bai liked to use *bai* much more than other poets. Ratio B shows the total percentage of poets' poems that used the terms shown in boldface in the "Word" column: *baifa* 白髮 (white hair), *baitou* 白頭 (white head), *baishou* 白首 (white head), *baixu* 白鬚 (white beard), *baigu* 白骨 (white bones), and *baizi* 白髭 (white mustache). For the column of Li Bai, it is the sum of 2.34, 0.67, 1.56, 0.11, 1.23, and 0.00 (the values in the shaded rows). These six terms typically appeared in works of a pessimistic emotional tenor.

It is thus possible to glimpse some differences between the main themes of poets' works with ratio B. The B ratios of Meng Haoran 孟浩然 (689–740), Li Shangyin 李商隱 (812?–58), and Wen Tingyun 溫庭筠 (812–970) are less than 2 percent. In sharp contrast, the B ratios of Du Fu 杜甫 (712–70) and Bai Juyi 白居易 (772–846) exceed 7 percent. Traditionally, Meng has been considered part of a leisurely pastoral tradition (*tianyuan shipai* 田園詩派), and Li and Wen are considered to use expressions that lead to "a beautiful and gorgeous conception" (「唯美穠麗的意境」).[22] Both Du Fu and Bai Juyi, on the other hand, are considered social poets who expressed deep concern about society at large.

We also found that *hong* 紅 (red) is the most frequent color word used in *QSC*. Words that included *hong*, such as *hongchen* 紅塵 (red dust), *canhong* 殘紅 (scattered red; i.e., flower petals), *hongzhuang* 紅妝 (red makeup; i.e., rouge), and *hongxiu* 紅袖 (red sleeves) were used as indirect means of referring to difficult goals or transient things. Hence, the popularity of *hong* in *QSC* can be partially understood via the social status of the Song poets.[23] *Canhong* (scattered red) in the following *ci* by Wang Anguo 王安國 (1028–74) is related to the passage of time:[24]

Table 5. Words that include *bai* 白 (white) in poems of thirteen poets in *QTS*

Metric	*Word*	孟浩然	孟郊	李商隱	李白	李賀	杜牧	杜甫	溫庭筠	王維	白居易	許渾	賈島	韓愈
Ratios														
Ratio A		8.96	18.41	9.73	46.65	23.83	12.55	26.94	15.67	18.80	17.37	15.19	16.30	10.48
Ratio B		1.87	5.72	1.80	5.92	2.13	4.66	7.94	1.99	2.28	7.19	4.54	3.70	3.23
Frequencies														
217	白日	**0.75**	**4.73**	**1.62**	**6.92**	**2.98**	**1.01**	**2.42**	0.00	**1.14**	**2.04**	**1.18**	**2.22**	**3.23**
164	**白髮**	**1.12**	**3.73**	**0.54**	**2.34**	**1.28**	**1.62**	**1.99**	0.00	**0.85**	**2.50**	**1.58**	**2.22**	**0.54**
158	白雲	**2.99**	**1.99**	**0.54**	**3.79**	0.85	**1.42**	0.86	0.28	**7.41**	0.95	**2.96**	**4.44**	0.27
149	**白頭**	0.00	0.75	**0.72**	0.67	0.43	**2.23**	**3.37**	**1.14**	0.57	**2.23**	**2.37**	0.49	**1.61**
86	**白首**	**0.75**	**1.00**	0.18	1.56	0.43	0.20	**1.99**	**0.85**	**0.85**	**1.02**	0.59	0.25	**0.81**
74	白玉	0.00	0.50	**2.34**	**3.01**	0.85	**0.81**	0.60	0.00	**0.85**	0.53	0.20	0.00	0.27
74	白馬	0.37	0.50	0.00	**2.34**	**4.68**	0.00	1.38	**2.85**	**0.85**	0.30	0.39	0.00	0.00
63	白雪	0.37	0.25	0.36	**2.34**	0.00	0.40	1.04	0.28	0.00	0.68	**0.79**	0.00	0.27
59	白帝	0.00	0.00	0.18	1.00	0.43	0.00	**3.54**	0.28	0.00	0.08	0.39	0.00	**0.54**
58	白露	0.00	0.50	0.18	1.56	0.43	0.00	0.86	0.28	0.28	0.79	0.39	**0.99**	0.27
54	白石	0.00	**1.00**	**0.90**	1.12	0.43	0.00	0.26	0.57	0.57	0.68	0.20	**1.23**	**0.81**
38	白蘋	0.37	0.75	0.18	0.22	**1.28**	0.20	0.52	**2.85**	0.00	0.30	0.20	0.00	0.54
32	白水	0.00	0.25	0.00	0.89	**1.28**	0.00	1.12	0.00	0.57	0.08	0.39	0.00	0.27
31	**白鬚**	0.00	0.00	0.18	0.11	0.00	0.20	0.00	0.00	0.00	**0.98**	0.00	0.49	0.00
30	白鷺	0.00	0.25	0.00	1.79	0.00	0.40	0.26	0.00	0.85	0.15	0.00	0.25	0.00
25	白璧	0.37	0.25	0.18	1.79	0.85	0.40	0.00	0.28	0.00	0.00	0.00	0.25	0.00
23	白楊	0.00	0.00	0.36	1.12	0.00	0.00	0.26	0.00	0.00	0.26	0.20	0.00	0.00
22	白蓮	0.00	0.00	0.00	0.00	0.00	0.40	0.00	0.57	0.00	0.61	0.39	0.00	0.00
21	白羽	0.37	0.25	0.00	0.67	0.00	0.20	0.52	0.28	0.85	0.08	0.00	0.00	0.00
21	**白骨**	0.00	0.25	0.18	1.23	0.00	0.00	0.60	0.00	0.00	0.00	0.00	0.00	0.27
19	白鷗	0.00	0.00	0.00	0.78	0.00	0.20	0.69	0.00	0.00	0.11	0.00	0.00	0.00
19	白屋	0.00	0.25	0.18	0.00	0.43	0.20	0.86	0.00	0.00	0.11	0.20	0.25	0.00
19	白鶴	0.37	0.25	0.00	0.33	0.00	0.00	0.43	0.00	0.57	0.15	0.39	0.00	0.27
19	素琴	0.00	0.00	0.18	0.78	0.00	0.00	0.00	0.57	0.28	0.19	0.39	0.25	0.00
19	素手	0.00	0.00	0.00	1.56	0.00	0.00	0.00	0.85	0.00	0.04	0.20	0.00	0.00
18	白浪	0.37	0.00	0.00	0.56	0.00	0.00	0.17	0.00	0.00	0.23	0.20	0.74	0.00
17	白衣	0.00	0.00	0.18	0.22	0.00	0.00	0.17	0.00	0.57	0.19	0.20	**0.99**	0.00
16	白鹿	0.00	0.00	0.00	0.89	1.28	0.20	0.09	0.00	0.00	0.11	0.00	0.00	0.00
15	白波	0.00	0.25	0.00	0.78	0.43	0.00	0.09	0.00	0.00	0.08	0.59	0.00	0.00
15	**白髭**	0.00	0.00	0.00	0.00	0.00	0.40	0.00	0.00	0.00	0.45	0.00	0.25	0.00
15	皓齒	0.00	0.00	0.00	0.78	0.85	0.00	0.26	**0.85**	0.00	0.00	0.00	0.00	0.00
14	白沙	0.00	0.00	0.00	0.33	0.00	0.20	0.43	0.00	0.57	0.11	0.00	0.00	0.00
14	白鳥	0.00	0.00	0.18	0.00	0.00	0.61	0.35	0.28	0.28	0.08	0.00	0.49	0.00
14	白花	0.00	0.00	0.00	0.22	0.00	0.40	0.26	0.57	0.00	0.15	0.00	0.00	0.27
12	白社	0.75	0.00	0.18	0.00	0.00	0.20	0.00	0.57	0.28	0.08	0.39	0.25	0.00
12	白龍	0.00	0.25	0.00	0.89	0.00	0.00	0.00	0.28	0.00	0.08	0.00	0.00	0.00
12	白紵	0.00	0.00	0.18	0.89	0.00	0.00	0.00	0.28	0.28	0.04	0.00	0.00	0.00
12	白晝	0.00	0.00	0.00	0.11	1.70	0.20	0.09	0.00	0.00	0.15	0.20	0.00	0.00
11	白如	0.00	0.00	0.00	0.33	0.43	0.20	0.00	0.57	0.00	0.11	0.00	0.00	0.27

Table 5. continued

Metric	*Word*	孟浩然	孟郊	李商隱	李白	李賀	杜牧	杜甫	溫庭筠	王維	白居易	許渾	賈島	韓愈
11	素書	0.00	0.00	0.00	0.33	0.43	0.00	0.26	0.00	0.00	0.08	0.20	0.25	0.00
10	素月	0.00	0.25	0.00	0.56	0.43	0.00	0.26	0.00	0.00	0.00	0.00	0.00	0.00
10	白魚	0.00	0.00	0.00	0.00	1.28	0.00	0.52	0.00	0.00	0.04	0.00	0.00	0.00
10	白刃	0.00	0.25	0.00	0.56	0.00	0.00	0.26	0.00	0.00	0.04	0.00	0.00	0.00
10	素絲	0.00	0.00	0.00	0.33	0.43	0.00	0.17	0.28	0.28	0.08	0.00	0.00	0.00
10	白家	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.38	0.00	0.00	0.00
10	白猿	0.00	0.00	0.00	0.89	0.00	0.20	0.00	0.00	0.00	0.04	0.00	0.00	0.00

For details, see text

	The lingering spring does not stay,	留春不住
2	The orioles' words are fully exhausted.	費盡鶯兒語
	The ground is filled with scattered red, the palace brocade dirty:	滿地殘紅宮錦汙
4	Last night, there was wind and rain in the southern garden.	昨夜南園風雨
	As Feng Xiaolian approached the pipa,[25]	小憐初上琵琶
6	Her thoughts encircled the horizon at dawn.	曉來思繞天涯
	I cannot bear the embroidered hall's vermillion door;	不肯畫堂朱戶
8	The spring wind is naturally in the willow catkins.	春風自在楊花
		QSC 1.216

In addition to discovering and studying the most frequent colors of collections of poems,[26] reading pairs of colors that appear in an antithesis is also interesting. Bai Juyi used *bai* 白 (white) and *hong* 紅 (red) in the following two samples to create colorful scenes (table 6 lists more examples of other poets):[27]

	Stretching out my arms, I grab red cherries,	引手攀紅櫻
2	And red cherries fall like sleet.	紅櫻落似霰
	Lifting my head, I see the white sun,	仰首看白日
4	And the white sun races off like an arrow.	白日走如箭
		QTS 434.4801

	Do you not see the southern mountains far, far off, full of white clouds?	君不見南山悠悠多白雲
18	And do you not see the western capital surging, surging, with only red dust?	又不見西京浩浩唯紅塵
		QTS 453.5123

Table 6. Paired colors in antitheses (shown in boldface) in *QTS*

Poet	*Verse*	*QTS*
李白 (701–62)	朝弄**紫**沂海，夕披**丹**霞裳	161.1670
	清切**紫**霄迥，優遊**丹**禁通	164.1704
	雲臥留**丹**壑，天書降**紫**泥	168.1740
杜甫 (712–70)	內分金帶**赤**，恩與荔枝**青**	224.2400
	南望**青**松架短壑，安得**赤**腳蹋層冰	225.2415
	赤日石林氣，**青**天江海流	227.2459
白居易 (772–846)	日欲沒時**紅**浪沸，月初生處**白**煙開	439.4892
	白藕新花照水開，**紅**窗小舫信風回	450.5077
	白首林園在，**紅**塵車馬回	450.5087
杜牧 (803–53)	杉樹**碧**為幢，花騅**紅**作堵	520.5947
	一嶺桃花**紅**錦甃，半溪山水**碧**羅新	524.5993
	別夜酒餘**紅**燭短，映山帆滿**碧**霞殘	524.6007
李商隱 (813–58)	露氣暗連**青**桂苑，風聲偏獵**紫**蘭叢	539.6160
	昨日**紫**姑神去也，今朝**青**鳥使來賒	540.6203
	青門弄煙柳，**紫**閣舞雲松	540.6211
溫庭筠 (812–70)	水極晴搖泛灩**紅**，草平春染煙綿**綠**	576.6701
	綠楊陰裡千家月，**紅**藕香中萬點珠	582.6750
	舞衫萱草**綠**，春鬢杏花**紅**	583.6763

Social Networks Analysis

Poets often mentioned the names of their friends or other people in the titles and contents of their poems, so we can study the social network of poets in selected collections by connecting their names.[28] Information about poets' social network can be interesting itself, and the information is also useful for enriching the contents of databases like the CBDB. One may also wonder whether poets who refer to each other may share words or show similar styles in their poems.

In *QTS*, Li Bai mentioned himself in his own poems: "I, Li Bai, climb aboard a skiff, wanting to move, / And suddenly I hear a stomping song on the shore" 李白乘舟將欲行，忽聞岸上踏歌聲[29] and "Although you are Li Bai's bride, / How you marvel at the wife of the Chamberlain" 雖為李白婦，何異太常妻.[30] At least eight poets mentioned Li Bai in fifteen works, among which Du Fu contributed seven. Similarly, Luo Yin 羅隱 (833–910) mentioned Du Fu in his writings, as in the lines, "The flowers of Weiqu are in Du Fu's poetry, / To this day, they are heartless flirts, revered in wealthy households" 杜甫詩中韋曲花，至今無賴尚豪家.[31]

Of course, mentioning a person's name may not imply a direct friendship. The title "Passing by Jia Yi's Home in Changsha" 長沙過賈誼宅 does not mean that Liu Changqing 劉長卿 (d. 790?), the author, personally knew Jia Yi 賈誼, which is impossible as Jia passed away some time prior to 168 BCE, and Liu was born in the early eighth century.[32] Similarly, the fact that Luo Yin mentioned Jia

Yi in his poem does not mean that Luo Yin really knew Jia Yi: "Jia Yi met with misfortune in Luoyang, / And Du Fu had literary skill in Shaoling" 洛陽賈誼自無命，少陵杜甫兼有文.[33] Instead, we can conclude from such evidence that Luo Yin must have visited a memorial site or a reconstruction of Jia Yi's home. Their connection is a commemorative or literary one, not a social one.

We can employ biographical information about poets in CBDB to extend our search for names. Both style names (*zi* 字) and pen names (*hao* 號) are useful. Information about these alternative names helped us find that Pi Rixiu 皮日休 (834?–83?) mentioned Lu Guimeng 陸龜蒙 (d. 881?) by his alternative names, in this case Luwang 魯望, in two titles of poems by Pi: "On Events in Early Summer, Sent to Luwang" 初夏即事寄魯望[34] and "Respectfully Matching Luwang's 'Poem on a White Gull'" 奉和魯望白鷗詩.[35] Bai Juyi often referred to Yuan Zhen 元稹 (779–831) as Yuan the Ninth (Yuan jiu 元九) in his poems (e.g., "In the Morning I Heard Yuan the Ninth Chant His Poems" 早聞元九詠君詩) and titles (e.g., "Seeing Yuan the Ninth's Poetry at Indigo Bridge Station" 藍橋驛見元九詩).[36]

It is easy to establish a relationship of "mentioning the name of" in poems, but it takes more discretion to judge direct friendships. Li Bai—whose name literally means "Plum White"—was not mentioned in the line "**plums white** and peaches red fill the city walls" 李白桃紅滿城郭, nor was Yuan Zhen mentioned in the line "In the **ninth** year of the Kai**yuan** era, the Duke of Yan said" 開元九年燕公說.[37] An alternate name of the poet-warlord Gao Pian 高駢 (821–87) is Qianli 千里 (literally, "a thousand leagues"), a term that occurs very frequently in poetry. Certainly not all occurrences of *qianli* refer to Gao Pian.

We can apply heuristics to filter the strings that are not names, for instance, recognizing the alternate names only when a poet mentioned the full name of another poet at least once. Such filtering procedures are doable,[38] but it is also possible for the filters to ignore strings that actually refer to poets. The quality of the filters affects the precision and recall rates of the outcome.

We can apply visualization techniques to show the social network of poets in a specific period, as shown in figure 12, where the arrows go from the poet who mentioned others, and the thickness of the arrows reflects the number of mentions. In the CBDB project, historians manually verified the contents of their source, *Tang Wudai ren jiaowangshi suoyin* 唐五代人交往詩索引 (Indexes to the Exchange Poems of Tang and Five Dynasties) before information about the relationship between poets was entered into the database.[39]

Comparisons among Poets

Studying poets' preferences and innovations in using words in poems is popular in stylometry. Jiang Shaoyu has studied the poems of Li Bai and Du Fu and compared words that are related to *feng* 風 (wind) and *yue* 月 (moon, month)

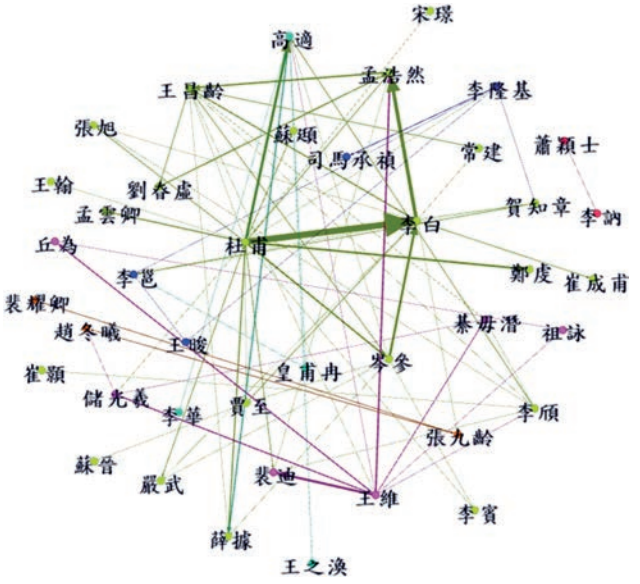

Figure 12. Poets' social network for High Tang (Luo, "*Quan Tangshi* de chubu fenxi")

in their poems.[40] Jiang then used this as a prompt to investigate the contents of individual poems to point out the differences in Li Bai's and Du Fu's use of *feng* and *yue*.

Identifying the words that are related to *feng* and *yue* and calculating their frequencies in different poets' poems offer an alternative way to observe Li Bai's and Du Fu's differences.[41] Table 7 lists the most frequent words of Li Bai and Du Fu that use *feng*. When we compare the numbers in the tables, we should recall that we have more poems of Du Fu than of Li Bai in *QTS* (cf. figure 7). The data indicate that Li Bai used *feng* more often than Du Fu did. The most frequent five words in table 7 also suggest that there are subtle differences in their uses of *feng*.

Table 8 lists the most frequent words related to *yue* in Li Bai's and Du Fu's poetic corpora. *Yue* is an ambiguous character, as it can represent the moon or a month. It is easy to notice that Du Fu used the names of the months much more often than Li Bai. In contrast, the table shows why Li Bai is famous for moon imagery in his poems: he used *yue* in many different and interesting ways.

The idea of comparing Li Bai's and Du Fu's uses of *feng* and *yue* can be extended. Table 9 lists the frequencies of frequent bigrams that appeared in the poems of four poets: Li Shangyin, Li Bai, Du Mu 杜牧 (803–52), and Du Fu. These bigrams are special in that they are formed by concatenating either *chun* 春 (spring) or *qiu* 秋 (autumn) with another character, and they represent something related to these two seasons.[42] The numbers "14;2" in the row of 春風; 秋風 (spring wind; autumn wind) under Li Shangyin (LSY) indicate that we

Table 7. Li Bai's and Du Fu's use of *feng* 風 (wind) in *QTS*

Li Bai: *Bigram*	*Freq.*	*Bigram*	*Freq.*	*Bigram*	*Freq.*	*Bigram*	*Freq.*	*Bigram*	*Freq.*
春風	72	松風	17	南風	8	悲風	6	高風	4
清風	28	隨風	14	北風	8	飄風	5	西風	4
秋風	26	香風	11	涼風	8	胡風	5	扶風	4
東風	24	天風	10	狂風	7	從風	5	屏風	4
長風	22	英風	8	雄風	6	嚴風	5	動風	4
Du Fu: *Bigram*	*Freq.*	*Bigram*	*Freq.*	*Bigram*	*Freq.*	*Bigram*	*Freq.*	*Bigram*	*Freq.*
秋風	30	朔風	8	高風	6	江風	4	南風	4
春風	19	微風	8	清風	6	驚風	4	涼風	4
北風	14	隨風	7	天風	6	山風	4	東風	4
悲風	10	回風	7	長風	5	多風	4		
裡風	8	臨風	7	陰風	4	含風	4		

observed 14 and 2 instances of *spring wind* and *autumn wind*, respectively, in Li Shangyin's poems.

The statistics in table 9 shed light on differences in word preferences among these poets. Note that the samples in table 9 are limited, and close reading is necessary to reach a decisive conclusion. Despite these limitations, we still can explore comparisons from multiple perspectives. *Spring wind* and

Table 8. Li Bai's and Du Fu's use of *yue* 月 (moon, month) in *QTS*

Li Bai: *Bigram*	*Freq.*	*Bigram*	*Freq.*	*Bigram*	*Freq.*	*Bigram*	*Freq.*	*Bigram*	*Freq.*
明月	57	溪月	9	有月	5	湖月	3	夜月	3
秋月	40	八月	9	轉月	4	漢月	3	夕月	3
五月	28	雲月	9	曉月	4	樓月	3	喘月	3
日月	23	花月	8	孤月	4	新月	3	向月	3
海月	14	見月	7	台月	4	待月	3	古月	3
上月	13	江月	6	落月	3	弄月	3	十月	3
三月	13	蘿月	5	片月	3	如月	3	二月	3
山月	10	素月	5	滿月	3	好月	3	乘月	3
Du Fu: *Bigram*	*Freq.*	*Bigram*	*Freq.*	*Bigram*	*Freq.*	*Bigram*	*Freq.*	*Bigram*	*Freq.*
日月	20	明月	7	落月	4	正月	3	從月	3
歲月	14	江月	6	秋月	4	星月	3	九月	3
十月	10	五月	6	漢月	4	新月	3		
三月	9	夜月	5	門月	3	四月	3		
八月	8	二月	5	素月	3	六月	3		

Table 9. Word frequencies in poems of four poets in *QTS*

	Poet					*Poet*			
Word Pair	*LSY*	*LB*	*DM*	*DF*	*Word Pair*	*LSY*	*LB*	*DM*	*DF*
春風;秋風	14; 2	72; 26	18; 11	19; 30	春草;秋草	0; 0	15; 12	2; 1	13; 5
春水;秋水	2; 3	3; 10	4; 5	8; 12	春色;秋色	0; 1	9; 11	3; 6	20; 7
春月;秋月	0; 0	0; 40	0; 0	0; 4	春來;秋來	4; 1	0; 3	2; 6	8; 6
春日;秋日	2; 2	2; 1	3; 1	13; 5	春光;秋光	2; 1	6; 0	2; 3	9; 1
春山;秋山	2; 0	2; 6	0; 4	2; 5	春天;秋天	1; 0	2; 2	0; 0	5; 11
春雨;秋雨	0; 2	0; 2	3; 3	4; 4	春江;秋江	0; 1	1; 2	0; 2	6; 2

Li Shangyin (LSY) Li Bai (LB), Du Mu (DM), and Du Fu (DF)

autumn wind were the most common choices among all rows.[43] In contrast, none of these poets used *spring moon* (*chunyue* 春月), and only Li Bai and Du Fu used *autumn moon* (*qiuyue* 秋月). In terms of personal preference, *spring wind* appeared in Li Bai's poems three times more often than *autumn wind*. Li Shangyin is similar to Li Bai, but Du Fu seems to prefer *autumn wind* instead.[44]

The entries that have zeros can be linked to strong personal preferences. For instance, Li Bai would not use *chunyu* 春雨 (spring rain) or *chun lai* 春來 (spring comes), though he used *qiuyu* 秋雨 (autumn rain) and *qiu lai* 秋來 (autumn comes). Du Mu is special in that he did not use *chuntian* 春天 (spring sky) or *qiutian* 秋天 (autumn sky).

Comparisons among Poems

Applying the FindCommon algorithm explained in the section on *Text Comparison* to compare poems, we can find many types of connections among poems.[45] Sometimes, poets would directly reuse the same sentences that had been used in other poems.[46] In *QSC*, He Zhu 賀鑄 (1052–1125) reused two lines from a poem of Du Mu in *QTS* (shown in boldface):

Du Mu:[47]

	These untroubled times have appeal, but I lack ability.	**清時有味是無能**
2	**Idle, I love the solitary cloud; serene, I love monks.**	**閒愛孤雲靜愛僧**
	About to take a pennon in hand to go off to the rivers and lakes,	欲把一麾江海去
4	Out upon Leyou Plain I gaze at Zhaoling	樂游原上望昭陵
		QTS 521.5961

He Zhu:[48]

	Idle, I love the solitary cloud; serene, I love monks	**閒愛孤雲靜愛僧**
2	and find good friends.	得良朋
	These untroubled times have appeal, but I lack ability.	**清時有味是無能**
4	Rectify this Deaf Cheng.[49]	矯聾丞
	And then, in my early years, I trod haughty tracks,	況復早年豪縱過
6	And am a sickly child still:	病嬰仍
	Even now I'm dull and stupid as a winter fly,	如今痴鈍似寒蠅
		QSC 1.505

In another example, He Zhu reorganized a few terms of Li Shangyin in his own poem:

Li Shangyin:[50]

	Because **cloud screens are endlessly charming**,	為有**雲屏無限嬌**
2	Winter's end in the capital means fear of **spring nights**.	鳳城寒盡怕**春宵**
	For no reason she was married to a **groom of golden tortoises**,[51]	無端嫁得**金龜婿**
4	Unworthy to bear her fragrant quilt or serve in the **morning court**.[52]	辜負香衾事**早朝**
		QTS 539.6168

He Zhu:[53]

	A **groom of golden tortoises** seeking pleasure in Zhangtai district,	章台遊冶**金龜婿**
2	Still reeling, reeling from drunkenness when he comes home.	歸來猶帶醺醺醉
	The floral drip shrinks from **spring nights**:	花漏怯**春宵**
4	**Cloud screens are endlessly charming.**	**雲屏無限嬌**
	Her back is set against the shadows of the scarlet-gauze lantern—	絳紗燈影背
6	The sound of jade pillow and hairpin shattering.	玉枕釵聲碎
	Not waiting for his hangover to clear,	不待宿醒銷
8	His horse neighs, speeding on to **morning court**.	馬嘶催**早朝**
		QSC 1.520

The follow example shows that He Zhu shared wording with three poets (also in *QTS*): Zhang Ji 張籍 (766?–830?), Xu Hun 許渾 (788?–858), and Cui Tu 崔塗 (fl. c. 888) in a single poem. Whether such shared wording is the result of conscious borrowing or drawing from a common poetic language is difficult to determine in this preliminary investigation, but in either case, it provides material for literary scholars to investigate more carefully from other angles.

Zhang Ji:[54]

	The **green mountains are arrayed**, the river stretches out far;	**青山**歷歷水悠悠
2	Today we run into one another, and tomorrow it's autumn.	今日相逢明日秋
	I tie my horse to the **willow tree** beside the city walls	系馬城邊**楊柳**樹
4	And buy you a drink to detain you a moment longer.	為君沽酒暫淹留
		QTS 386.4354

Xu Hun:[55]

	Red blossoms are **half-fallen**, the swallows are in flight;	紅花**半落**燕於飛
2	Though we were travelers in **Chang'an** together, today you return alone.	同客**長安**今獨歸
	A letter home on a piece of paper will report to my brothers;	一紙鄉書報兄弟
4	Once you've returned, I, abashed, will still wear the robes from our parting.	還家羞著別時衣
		QTS 538.6137

Cui Tu:[56]

	A three-year traveler beneath apple blossoms	海棠花底三年客
2	**Does not see** that the blossoms of the apple have fully flowered,	**不見**海棠花盛開
	Then, facing **Jiangnan**, sees a depiction	卻向**江南**看圖畫
4	And starts to feel ashamed for having come to Shu's city in vain.	始慚虛到蜀城來
		QTS 679.7784

He Zhu:[57]

	Preparing, stringing up lanterns: spring matters come early	排辦張燈春事早

2	Within the twenty city gates.	十二都門
	Phenomena and forms befit the new dawn.	物色宜新曉
4	The golden calf's cart is light, the jade horse small—	金犢車輕玉驄小
	Brushing its head, in the **willows** we traverse its path.	拂頭**楊柳**穿馳道
6	Water-mallow broth and minced perch are not what I like.	蓴羹鱸鱠非吾好
	Leaving the land, I sing	去國謳吟
8	The tune of **half-fallen [blossoms] in Jiangnan.**	**半落江南**調
	The **green mountains** filling my eyes, I resent the western light:	滿眼**青山**恨西照
10	**Not seeing Chang'an** makes a man old.	**長安不見**令人老
		QSC 1.506

He Zhu is not the only poet who shared phrases with poems in *QTS.* Find-Common also shows that Xin Qiji 辛棄疾 (1140–1207) and Wen Bing 文丙 (Tang, dates unknown) shared some words in their poems.

Wen Bing:[58]

	[This tree is] as endearing as the hundred flora	可憐同百草
2	And bears a **countenance of frost and snow**.	況負**雪霜姿**
	The ground does not esteem song and dance,	歌舞地不尚
4	But people naturally move it in **winter season.**	**歲寒**人自移
	On the stairs it adds **cool and simplicity**,	階除添**冷淡**
6	With brush-tip in enters one's thoughts.	毫末入思惟
	Clouds and grottoes emerge at the end of the way:	盡道生雲洞
8	**Who understands** that **the road** is craggy and perilous?	**誰知路**嶮巇
		QTS 887.10028

Xin Qiji:[59]

	In the dark, a fragrance stretches across **the road**, and **snow** falls down.	暗香橫**路雪**垂垂
2	The night wind blows,	晚風吹
	The night wind blows.	曉風吹
4	Flowers' intentions compete with the spring:	花意爭春
	To put forth their **winter-season** branches early.	先出**歲寒**枝
6	After all, when it comes to completing spring's business,	畢竟一年春事了
	Because they're too early,	緣太早

8	They should slow down.	卻成遲
	They don't need to have a complete **countenance of frost and snow**	未應全是**雪霜姿**
10	When they're about to open	欲開時
	And when they've yet to open:	未開時
12	Powdered face and red lips,	粉面朱唇
	Halfway dotted with rouge.	一半點胭脂
14	The flowers don't resent my drunken slander of the flowers:	醉裡謗花花莫恨
	Mixing **cool and simplicity**,	渾**冷淡**
16	**Who understands** it?	有**誰知**
		QSC 3.1957

It is also possible for us to compare poems within *QTS.* Doing so reveals some of the limitations of *QTS* as a source for studying Tang poetry. Our algorithm makes it easy to find cases of the same poem, with minor variants, listed under different names. These demonstrate some of the problems that faced the *QTS* editors, many of which remain unresolved.[60] For example, in *QTS*, the following two items are listed under the names Lu Lun 盧綸 (d. 799?) and Lu Shangshu 盧尚書 (Minister Lu) and given very different titles. Despite these differences, the poems are extremely similar and differ only in two characters (shown in boldface):

Lu Lun:[61]

	Dusk shines on the **overlooking** window, dark dust rises;	夕照**臨**窗起暗塵
2	The green pines surround the palace, unaware of spring.	青松繞殿不知春
	Take a look at the white-**haired** scripture reciters:	君看白**髮**誦經者
4	Half of them were singers and dancers in the palace.	半是宮中歌舞人
		QTS 279.3169

Lu Shangshu:[62]

	Dusk shines on the **embroidered** window, dark dust rises;	夕照**紗**窗起暗塵
2	The green pines surround the palace, unaware of spring.	青松繞殿不知春
	Take a look at the white-**headed** scripture reciters:	君看白**首**誦經者
4	Half of them were singers and dancers in the palace.	半是宮中歌舞人
		QTS 783.8843

According to the biographical information of Lu Lun, he once served at the head of the Ministry of Revenue (Hu bu 戶部) as its minister (*shangshu* 尚書). From this perspective, it is possible that this Lu Shangshu is Lu Lun and the editors of *QTS* repeated attribution of this poem to both his specific name and his generic title out of carelessness. It is also possible that when the poem first appeared in early sources, such as the late eleventh-century *Tang yulin* 唐語林 (Garden of Stories of the Tang), it was attributed to the generic "Lu Shangshu" and editors assigned it to the famous poet Lu Lun only in later collections.[63] Our digital tools are good at highlighting such problems that would require further investigation from literary scholars to be resolved.

Below are two more pairs of poems in *QTS* whose authors might be the same. In the following pair, the poems are similar but their titles ("Parting with a Beauty" 別佳人 vs. "Parting with My Wife" 別妻) are related yet different. The names of their authors, Cui Ying 崔膺 (fl. c. 788) and Cui Ya 崔涯 (early 9th century), are different, but their pronunciations are very similar. The variant characters, *long* 壠 (ridge) versus *long* 隴 (hillock), are also very similar in pronunciation and appearance.

Cui Ying:

	The spring flowing atop the ridges splits beneath the ridges	**壠**上流泉**壠**下分
2	My insides breaking, I wail and moan, can't bear to hear it.	斷腸嗚咽不堪聞
	Once Chang E had left for the moon	嫦娥一入月中去
4	There were white clouds in the sky over Wuxia for a thousand autumns.	巫峽千秋空白雲
		QTS 275.3119

Cui Ya:

1	The spring flowing atop the hillocks splits beneath the hillocks	**隴**上泉流**隴**下分
		QTS 505.5741

The following poems of Luo Yin and Lu Yin 盧殷 (746–810) differ in just one character. They have the same title, "Running into a Frontier Emissary" 遇邊使, and the pronunciations of the names of their authors are very similar.

Lu Yin:

	With no real news for many years,	累年無的信

2	He looks out every night from the frontier ramparts.	每夜望邊城
	His sleeves hold a thousand streams of tears,	袖掩千行淚
4	And in his letters he sends his square-inch heart's **feelings**.	書封一尺**情**
		QTS 470.5342

Luo Yin:

4	And in his letters he sends his square-inch heart's **gold**.	書封一尺**金**
		QTS 665.7622

The correct attribution of this poem is to Lu Yin, but we most move away from digital methods to discover this. If we look through historical sources, we will find that this poem was included in the anthology *Yulan shi* 御覽詩 (Poems for Imperial Perusal), compiled in the early ninth century by Linghu Chu 令狐楚 (766–834), at least thirteen years before Luo Yin was born. Therefore, it could not have been written by Luo Yin. Thus, we can conclude the misattribution to Luo Yin must have been a mistake introduced by the editors of *QTS* or one of the collections they drew from.[64]

The following two poems in *QTS* also differ in only one character. The poets are Zhang Bayuan 章八元 (fl. 770s) and Zhu Fang 朱放 (d. 788?), two very different persons, and the titles of the poems are the same, "Sent to Administrator Yan upon Returning to My Old Residence at Tonglu" 歸桐廬舊居寄嚴長史. Looking at the pages of a hard copy of the *QTS*, we can verify that there are indeed different characters, that is, *fu* 夫 and *tian* 天, which led to the different bisyllabic words "master" (*fuzi* 夫子) and "emperor" (*tianzi* 天子). Hence, we have identified another type of authorship problem.

Zhang Bayuan:

	I took leave of you yesterday, **master**, to paddle a boat home.	昨辭**夫**子棹歸舟
2	My home being in Tonglu, I had been thinking of my old hills.	家在桐廬憶舊丘
	In the third month, when we were close, the flowers vied to bloom;	三月暖時花競發
4	At the two streams, where we parted, the water fought as it flowed.	兩溪分處水爭流
	When near, I had heard an old riverman tell news of my village;	近聞江老傳鄉語

6	Now far, I see the mountains of my home dispel my worries of travel.	遙見家山滅旅愁
	Perhaps sometime when drunk on a snowy night,	或在醉中逢夜雪
8	I'll think of you and how I should travel to the rivers of Shan.	懷賢應向剡川遊
		QTS 281.3193

Zhu Fang:

1	I took leave of you yesterday, **emperor**, to paddle a boat home.	昨辭**天**子棹歸舟
		QTS 315.3540

Again, we can figure out the correct attribution only if we return to traditional methods and compare evidence from historical sources. The "Administrator Yan" referred to in the poem's title likely refers to Yan Wei 嚴維 (mid-8th century), with whom only Zhang was acquainted, and Zhang, not Zhu, grew up in Tonglu, meaning it makes more sense for Zhang to refer to his "old hills" in this poem. In addition, the poem is attributed to Zhang in the early sources *Wenyuan yinghua* 文苑英華 (Finest Flowers of the Garden of Literature) and *Tangshi jishi* 唐詩紀事 (Recorded Contexts of Tang Poems).[65] Therefore, it is much more likely that the poem should be attributed to Zhang Bayuan. Our digital tools can highlight this problem, but philological investigation must solve it.

The following poems in *QTS* show yet another type of challenge. The poets are Dai Shulun 戴叔倫 (732–89), Qingjiang 清江 (late 8th century), and Kezhi 可止 (860–934). Dai was the eldest, and Kezhi was born at least fifty years after Qingjiang deceased. The titles and contents of Qingjiang's and Kezhi's poems are exactly the same: "Encountering Rain in a Meditation Hovel" 精舍遇雨. The title of Dai's poem is different: "Facing the Rain in a Meditation Hovel" 精舍對雨. Moreover, both Qingjiang's and Kezhi's poems differ from Dai's in just one character.[66] However, because the title and content of the poem provide very few clues to its authorship, it is difficult to resolve this problem with any certainty.

Dai Shulun:

	The empty gate is silent, silent, and **calm** is my person.	空門寂寂**澹**吾身
2	Rain by the creek is faint, faint, washing this traveler of dust.	溪雨微微洗客塵
	Lying down, I look at the white clouds, not fully sunlit	臥向白雲晴未盡

4	Let other yellow birds be intoxicated with the fragrant spring.	任他黃鳥醉芳春
		QTS 274.3111

Qingjiang and Kezhi:

1	The empty gate is silent, silent, and **tranquil** is my person.	空門寂寂**淡**吾身
		QTS 812.9147; *QTS* 825.9292

We can search and uncover more connections among the poems if we compare all of the poems in *QTS* and the lyrics in *QSC*. Recall that we have 42,863 items in *QTS* and 19,394 items in *QSC* (cf. table 2). An exhaustive comparison procedure that considers two pairs of poems from two perspectives would conduct more than 1.9 billion comparisons in FindCommon.[67] The computational load is quite high, although it is still possible to complete the procedure on a personal computer. Further technical details were discussed in another article by the first author.[68]

Temporal Analysis: Zipf's Law and Beyond

Many researchers have studied the Zipfian distributions of literary works in both Chinese and English, such as Chin-Kun Hu and Wei-Cheng Kuo's work.[69] Since we have collections of Chinese poems that were produced over a period of 2,000 years, it is interesting to examine the Zipfian distributions of our collections and compare the changes of the curves.[70] We create charts that are based on the following typical form of the Zipf's law:[71]

$$\log\left(\frac{f(w)}{N}\right) = k - \alpha \log\left(r(w)\right), \tag{1}$$

where w, $f(w)$, and $r(w)$ denote a word, its frequency, and rank in a corpus, respectively. The rank of the most frequent word in a corpus is 1. N is the size of the corpus, and k and α are constants.

For each of the collections listed in table 1, we can compute the frequency of every Chinese character in the collection. We rank the characters according to their frequencies and assign 1 to the most frequent character, 2 to the second most frequent, and so on. With the rank and frequency for each individual character in a collection, we can plot its Zipfian distribution.

Figure 13 shows nine curves for the collections listed in table 1. For nine collections of poetry that were produced over more than 2,000 years, the curves are extremely similar overall. Purely based on visual impression, we can tell that

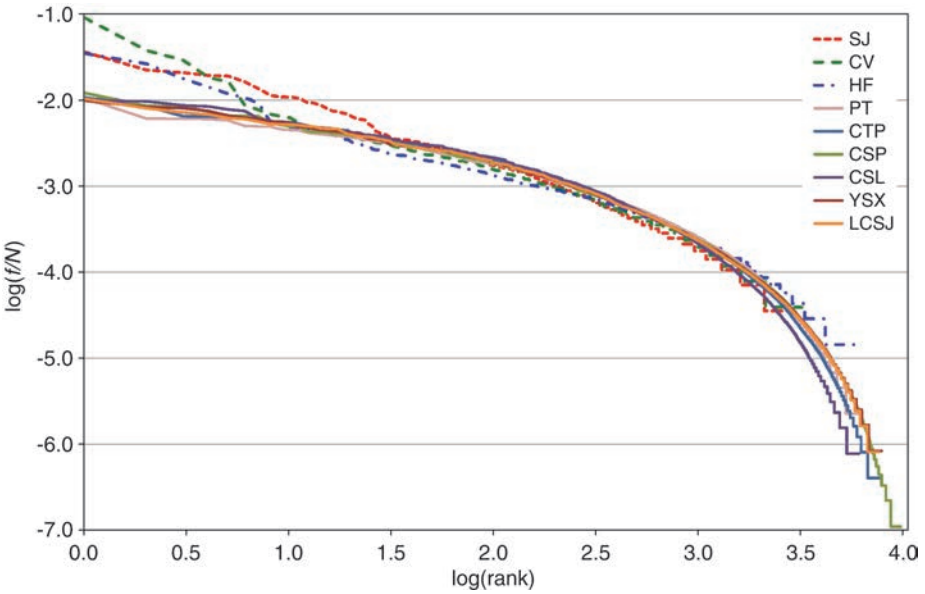

Figure 13. Zipf curves for ancient poetic works (*SJ*, *CV*, and *HF*; see table 1) do not coincide with those of later poems. See table 1 for abbreviations.

the curves for *SJ*, *CV*, and *HF* are different from the other six curves, which are very close to one another. We can calculate the average of the squares of the differences between the values of log(*f*/*N*) of the curves or the correlation coefficients between the curves to reach the same conclusions suggested by our visual impressions.[72]

Although a poem may be included in a collection more than once, as described above, it is expected that such repeated inclusions are not significantly common enough to influence the trends of the curves. Consider *QTS*, *QSC*, and *QSS*, for example. The frequency of the 1,000th most frequent characters in *QTS*, *QSC*, and *QSS* are more than 500, 2,000, and 250, respectively. Unless all of the repeatedly included poems used a common character such that the frequency of the common character was boosted by hundreds, the repeatedly included poems will not influence the rank of a character significantly. Even if this extreme situation did occur, the influence should affect only a few characters, so the trends of the distributions of the frequency of the characters will not be significantly influenced by repeated inclusions. Similar phenomena about the trends have been observed by other researchers.[73]

Table 10 lists the ten most frequent characters of the corpora whose curves are plotted in figure 13. These lists are very similar, and although sixty different characters could potentially be included in the table, only sixteen distinct characters are listed.[74] In fact, we can compare the most frequent characters of any

Table 10. Ten most frequent characters in poems and lyrics remain stable over time

	Rank									
Corpus	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
PT	不	無	風	有	人	雲	之	何	日	我
QTS	不	人	山	無	風	一	日	雲	有	何
QSS	不	人	一	無	山	有	風	來	天	日
QSC	人	風	花	一	不	春	無	雲	來	天
YSX	不	人	山	風	一	雲	天	日	有	無
LCSJ	不	人	風	山	一	花	日	雲	有	無

two corpora, for example, *QTS* and *QSS*, to further investigate their similarity.[75] We found that the most frequent 1,700 characters in *QTS* and *QSS* are the same set of characters, although the ranks of these characters are not exactly the same.

Table 11 lists the ten most frequent bigrams in each of the four types of poems in *QTS* and *QSS*. This table could potentially list eighty bigrams, but only twenty-nine distinct bigrams actually occur. Three bigrams, *buzhi* 不知 (don't know), *hechu* 何處 (where), and *chunfeng* 春風 (spring wind), appear in all of the eight categories.

Temporal Analysis: Word History

The availability of poetic writings across different dynasties allows us to study the history of words.[76] We can observe the appearance and disappearance of words in poems, and we can investigate the contexts in which these words appear to study whether their lexical meanings changed over time.

We can compute the occurring portion of a bigram, β, in a collection, *C*, with the following formula:[77]

Table 11. Ten most frequent bigrams in poems and lyrics remain stable over time

	Rank									
Category	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
QTS.5_JUE	何處	不知	不見	春風	明月	千里	秋風	青山	萬里	無人
QSS.5_JUE	不知	何處	不可	如何	春風	不見	明月	天地	無人	萬里
QTS.5_LU	何處	萬里	白雲	千里	故人	不可	不知	秋風	春風	相思
QSS.5_LU	平生	千里	何處	春風	不可	萬里	不知	故人	秋風	風雨
QTS.7_JUE	不知	春風	何處	今日	人間	無人	萬里	惆悵	不得	何事
QSS.7_JUE	不知	春風	人間	梅花	無人	東風	何處	今日	風吹	不是
QTS.7_LU	何處	今日	不知	萬里	春風	千里	人間	惆悵	白雲	十年
QSS.7_LU	千里	人間	萬里	春風	不知	十年	平生	歸來	何處	故人

5_JUE, WuJue (pentametric quatrains); 5_LU, WuLu (regulated pentametric octaves); 7_JUE, QiJue (heptametric quatrains); 7_LU, QiLu (regulated heptametric octaves)

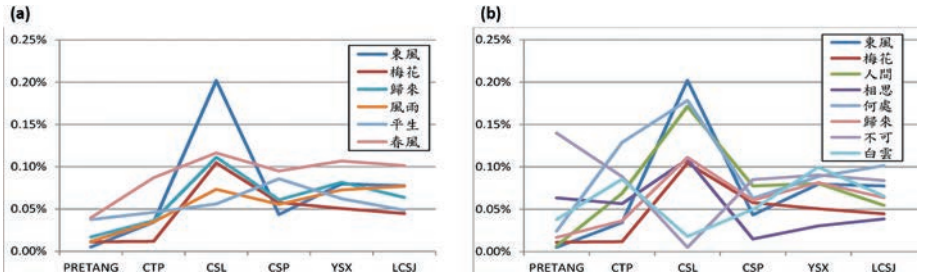

Figure 14. Percentages of selected words in different collections

$$\text{portion}(\beta, C) = \frac{2 \times \text{frequency}(\beta)}{\text{total number of characters in } C} \tag{2}$$

In the numerator, we multiply the β frequency by 2 because each occurrence of β contributes two characters. The two charts in figure 14 show how the portions of two lists of words change over time. On the horizontal axis, we have ordered the collections by the time period in which their contents were produced (both *QSC* and *QSS* were from the Song dynasty). On the vertical axis, we show the proportions of a selected list of words in table 11. The proportions of the words in chart (a) increase gradually from pre-Tang to *QTS* and from *QTS* to *QSC* and *QSS* and remain higher than the proportions in *QTS* afterward. Chart (b) suggests that *QSC* is special in its use of some words. The proportions of six of the selected words in *QSC* are higher than their proportions in *QTS* and *QSS*, while the proportions of two other words, *bu ke* 不可 (cannot) and *bai yun* 白雲 (white cloud), are very lower in *QSC*.

Using the birth and death years of the poets recorded in CBDB, we can draw a chart like figure 15 to show which poets used these words. The horizontal axis of figure 15 shows the years of Tang and Song dynasties, and the widths of the rectangles that contain the poets' names indicate the poets' life span.[78] We used *QTS* for the Tang dynasty and *QSC* and *QSS* for the Song dynasty when drawing figure 15. The figure does not show poets whose life spans are unknown, so it is not complete. The figure is divided into four parts, from top to bottom, for *hongyan* 紅顏 (red face), *xuanfa* 玄髮 (dark hair), *kongmen* 空門 (empty gate), and *xingsong* 惺忪 (sleepy), each showing the poets who used these four words.

An interface like figure 15 can provide useful information that a traditional lexicon may not easily achieve. First, the chart offers a distant reading of the history of the words' occurrences, along with the poets who used the words. Although there were fewer poets in *QTS* than in *QSC* and *QSS*, more *QTS* poets used *kongmen* (and *xuanfa*), which could prompt further research. We can easily see that *xingsong* might have been an invented by Yan Shu 晏殊 (991–1055) in

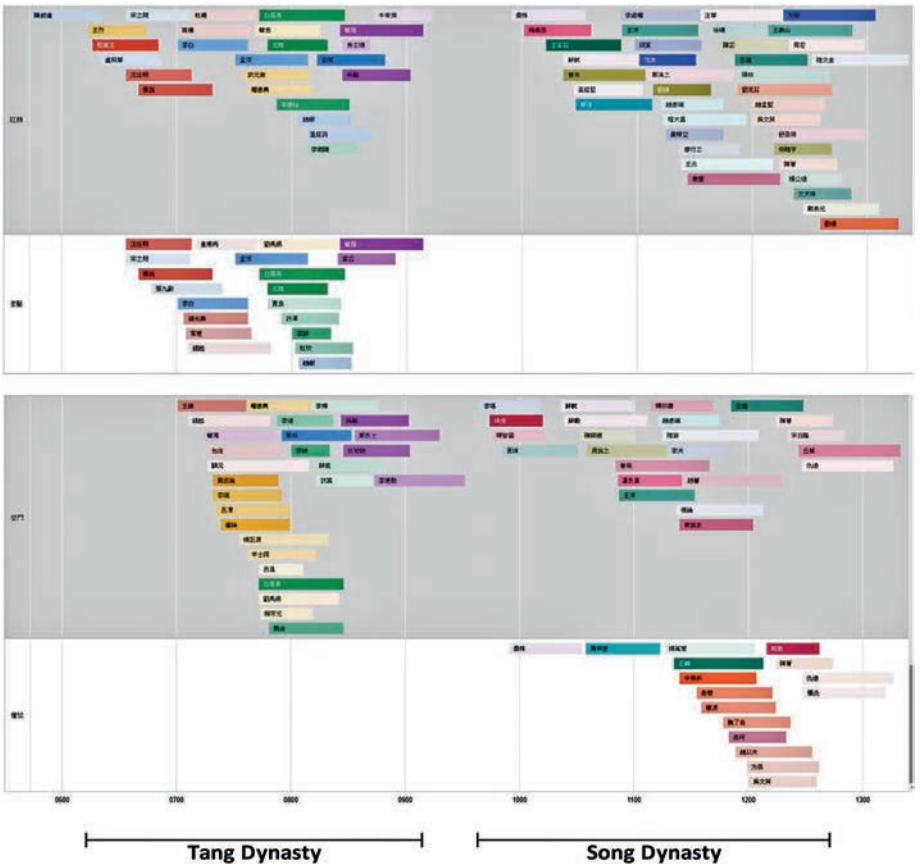

Figure 15. Four examples of word occurrences in *QTS* (Tang) and *QSC* + *QSS* (Song). For details, see text. Produced with the support of Google Charts: developers.google.com/chart/

the Song dynasty.[79] Extending the exploration to *CV*, we will find a clue for the introduction of *baopu* 抱璞 (clasping the jade block; i.e., internal excellence not noticed by others).[80]

Second, software developers can add more functions to the charts for close reading, style comparisons, and other applications. Researchers can click on the poets' names to read the poems that actually used the specific words, for example, *hongyan* (red face), for further investigation. Given the time stamps on the horizontal axis, one may study how poets used *hongyan* in a specific time period, for example, the High Tang or Southern Song period. A potentially more fruitful application would be to automatically extract the poems that used a specific word to study whether the meanings carried by the word changed over time, using advanced techniques like word embedding.[81] Additionally, this tool can be useful to language learners, for whom the poems can serve as a source of uses of the selected words.

Concluding Remarks and Discussion

There are many kinds of research into Chinese poetry, not all of them confined to literature and linguistics.[82] It is not difficult to find real-world studies in literature that are related to the applications presented in this article. Liu Dianjue and colleagues compiled concordances for individual characters in both *Chuci* and Xie Tiao's works.[83] Yi-Hsin Lai produced a detailed account about words, including those about colors, in Li Shangyin's WuLu poems.[84] It is certainly possible to study words about birds, animals, tea, or plants in poems.[85] Yao Gui studied the exchange poems between two Tang poets, so social network analysis can provide some leads for further research.[86] Wang Wei-Yung spent many years comparing the contents of Tang poems and Song lyrics.[87] It is very interesting to study repetitions in English and Chinese poetry.[88] Hu and Kuo's study of the Zipfian distributions of Chinese literary works is just one example of this kind.[89] If studying the roles of words about birds in literary works from the Han to the Tang dynasty was possible in the past,[90] conducting *longue durée* research on Chinese poetry in an era of digital humanities should require relatively less time to gather source materials and allow more time for studying them.

The main claims of "big data" can be useful to the study of Chinese poetry. Having more digitized texts of Chinese poems is essential. If we can link data that are related to the poets or poems distributed in different databases, then digital tools will broaden the outlook for our research. Information about friendships among Tang poets may be available in the *Quan Tang wen* 全唐文 (Complete Tang Prose). Information about poets' lives should help us better understand and appreciate poets' poems, so obtaining poets' biographical information from databases like the CBDB can be useful. Specialized types of poems like monk poetry and Buddhist poetry are certainly important for studying Chinese literature and culture.[91]

Further improvements are needed. Developing digital tools is just one step toward a complete research project. We need to have reliable sources of data, which is missing in this work. We have good reasons to believe that the data we have gathered are good enough for our demonstrations, and our observations show some interesting directions for computer-assisted research. However, not all of the reported statistics are precise. In addition to reliable data, we also need to expand the availability of digitized texts of Chinese poetry to achieve large-scale studies. The need is clear; for example, although the *QTS* is a representative collection of Tang poems, it actually does not include all Tang poems. To know Du Fu more, we need to gather many more poems by Du Fu from *Quan Tang shi bubian*.[92]

Collecting more original data is a key step, but raw data alone are not really helpful. In the applications we report here, we work on characters and frequent bigrams in the poems. The texts in our corpora were not segmented into words,

Figure 16. Positions of "春風" and "秋風" in Du Fu's 5_Lu and 7_Lu poems in *QTS* (from top to bottom, the poems can be found on pages *QTS* 234.2581, *QTS* 228.2475, *QTS* 229.2493, *QTS* 225.2422, *QTS* 227.2469, *QTS* 225.2423 *QTS* 224.2396, *QTS* 225.2409, *QTS* 228.2473, *QTS* 225.2426, *QTS* 225.2419, *QTS* 225.2422, *QTS* 229.2489, *QTS* 230.2520, *QTS* 231.2549, *QTS* 230.2523, *QTS* 228.2483, *QTS* 222.2369, *QTS* 227.2467, *QTS* 227.2458)

limiting the scope of research that software can help with. Although we can use close reading to alleviate part of the resulting problems, one should demand that the corpora be segmented.[93] If possible, texts with phonological, part-of-speech, or grammatical annotations will lead to more research opportunities.[94]

It is best to create tools that are sufficiently flexible for researchers to make effective use of the services provided by the software.[95] Each row in figure 16 lists one of twenty Du Fu's WuLu (regulated pentametric octaves) and QiLu (heptametric regulated octaves) that use *chunfeng* 春風 (spring wind) and *qiufeng* 秋風 (autumn wind) in *QTS*. The figure orders these poems according to the positions of *chunfeng* and *qiufeng* in the poems while considering the types of the poems. The positions of *chunfeng* and *qiufeng* in these poems suggest that *qiufeng* is more likely than *chunfeng* to appear in the first part of Du Fu's Lu poems. Does this positional difference shed light on Du Fu's feelings about spring and autumn winds? To create the presentation in figure 16, we combined multiple functions that we discussed individually in this article.

CHAO-LIN LIU 劉昭麟
National Chengchi University, Taiwan
chaolin@nccu.edu.tw

THOMAS J. MAZANEC 余泰明
University of California, Santa Barbara, USA
mazanec@ucsb.edu

JEFFREY R. THARSEN 康森傑
University of Chicago, USA
tharsen@uchicago.edu

Acknowledgment

The authors benefited from discussions with Wen-Huei Cheng 鄭文惠, Wei-Yun Chiu 邱偉雲, Hongsu Wang 王宏甦, Shuhua Zhang 張淑華, Yuanli Geng 耿元驪, and Ching-Chun Hsieh 謝清俊. Communications with graduate students, including Kuo-Feng Luo 羅國峯, Chih-Kun Huang 黃植琨, and Chu-Ting Hsu 許筑婷, sometimes led to interesting hunches. This work was supported in part by research contract 104-2221-E-004-005-MY3 from the Ministry of Science and Technology of Taiwan and by internal contract 107H9999 of National Chengchi University. A large part of the reported observations took place during the author's visit to Harvard University, which was supported by the grant USA-HAR-105-V02 from the Top University Strategic Alliance and by the 2016–17 Fulbright Scholar Program from the Fulbright Foundation.

Notes

1. Fuller, *China Biographical Database User's Guide.*
2. Hu and Yu, "Computer Aided Research Work of Chinese Ancient Poems"; Lo, "Shilun yinyong zixun keji."
3. Zhu, "Quan Qing shi bianzuan."
4. It is important to briefly mention the differences between Chinese characters and words for readers who are not familiar with the Chinese written language. Characters are basic units for Chinese words. A Chinese word can be formed by one or more characters. For instance, *shui* 水 and *guo* 果 are two characters. They can be used individually to represent *water* and *results*, respectively. A word consisting of *n* Chinese characters can be called an n-gram in linguistics; for example, *shuiguo* 水果 is a bigram that represents *fruit*. In vernacular Chinese, the majority of words are bigrams and trigrams, but in classical Chinese, the proportion of unigrams is very large. There is no dictionary that can exhaustively list all understandable Chinese words because fluent speakers can create and understand words on the fly.
5. The authors thank Professor Geng Yuanli 耿元驪 of Liaoning University for obtaining texts from Daizhige.
6. E.g., Liu, *Chuci zhuzi suoyin*; Liu, Chen, and He, *Xie Tiao ji zhuzi suoyin*.
7. We use *antithetical* as a translation for *duizhang* 對仗, although the term has synonymic ("parallel") as well as contrastive connotations.
8. These are heptametric quatrains (*qiyan jueju* 七言絕句).
9. For the sake of simplicity, we use XXX-YYY to denote the collocation of two words, XXX and YYY.
10. The titles of $P_{1,1}$ is "Shitou cheng" 石頭城 (*QTS* [Peng et al., *Quan Tang shi*], 365.4117).
11. The titles of $P_{1,2}$ is "Wuyi gang" 烏衣巷 (*QTS* 365.4117).
12. The titles of $P_{2,1}$ is "West river; Dashi jinling" 西河大石金陵 (*QSC* [Tang, *Quan Song ci*], 2.612).
13. Chen and Wang, *Song ci qingshang*.
14. Xie Tiao's poem is "Drum and Pipe Songs of the Prince of Sui (4 of 10): Entering the Court" 隋王鼓吹曲十首 （其四）：入朝曲, and the *yuefu* is "Music of No Worries (1 of 2)" 莫愁樂二曲（其一）.
15. Li Bai, "Seeing Off Governor Yuan to Assume His Post at Changsha" 送袁明府任長沙 (*QTS* 185.1890).
16. *QTS* 391.4411.

17. Here, e.g., 69.3 percent is the sum of 38.5 percent and 30.8 percent, and 63.8 percent is the sum of 6.0 percent, 8.6 percent, 25.9 percent, and 23.3 percent.
18. The eight sentences in WuLu and QiLu poems are grouped into four couplets. Each couplet consists of two consecutive sentences. The second pair is known as the "chin couplet" (*hanlian* 頷聯) in traditional poetics.
19. Liu et al., "Color Aesthetics and Social Networks."
20. Cheng et al., "Qinggan xianxiangxue yu secai zhengzhixue."
21. The degree of being often is defined as the frequency of a word divided by the number of items of the poet in *QTS*.
22. Lee, "Wan Tang 'Wen-Li' zuopin."
23. Sun, "Tang Song ci benti tezheng de biaoxian xingshi."
24. "Clear and Level Music (Spring Evening)" 清平乐（春晚）, *QSC* 1.216.
25. Feng Xiaolian 馮小憐: concubine of Gao Wei 高緯 (557–77), penultimate ruler of the northern Qi dynasty, was known for her skill in dance and pipa playing.
26. Liu et al., "Color Aesthetics and Social Networks."
27. From "Beneath the Flowers with Alcohol: 2 of 2" 花下對酒二首（其二）, *QTS* 434.4801; and from "Waking Up Late in the Snow, Singing What I Feel and Showing It to Attendant in Ordinary Zhang, Mentor Wei, and Director Huangfu" 雪中晏起偶咏所怀兼呈张常侍韦庶子皇甫郎中, *QTS* 453.5123.
28. Liu and Luo, "Tracking Words in Chinese Poetry"; Liu et al., "Color Aesthetics and Social Networks."
29. "Given to Wang Lun" 贈汪倫, *QTS* 171.1765.
30. "Given to My Wife" 贈內, *QTS* 184.1884.
31. "Sent to Recluse Wei, South of the City" 寄南城韋逸人 (*QTS* 657.7550; Li, *Luo Yin ji xinian jiaojian*, 3.141). Luo Yin's line alludes to the opening of Du Fu's poem, "Respectfully Accompanying Escort Zheng in Weiqu: 1 of 2" 奉陪鄭駙馬韋曲二首（其一）: "In Weiqu the flowers are heartless flirts, / At every home they drive people crazy" 韋曲花無賴，家家惱殺人 (*QTS* 225.2413; Xiao, *Du Fu quanji jiaozhu*, 1064; Owen, *Poetry of Du Fu*, sec. 3.1). A variant in Luo Yin's poem gives "every household" (*jiajia* 家家) for "wealthy households" (*haojia* 豪家), which would more closely imitate Du Fu's line.
32. *QTS* 151.1566.
33. *QTS* 656.7543.
34. *QTS* 609.7027.
35. *QTS* 614.7082.
36. *QTS* 459.5226; *QTS* 438.4870.
37. Yang Shi'e 羊士諤 (762–862?), "Hearing a Bamboo Flute in a Mountain Gallery" 山閣聞笛, *QTS* 332.3696; Gu Kuang 顧況 (727?–816?), "Song on the Fifth Day of the Eighth Month" 八月五日歌, *QTS* 265.2944.
38. Luo, "*Quan Tangshi* de chubu fenxi."
39. Wu, *Tang Wudai ren jiaowangshi suoyin*. This task was completed by Shuhua Zhang 張淑華 of Northwestern Polytechnical University, China.
40. Jiang, "Li Bai Du Fu shi zhong de 'yue' he 'feng.'"
41. Liu et al., "Color Aesthetics and Social Networks."
42. When used alone, *chun* 春 and *qiu* 秋 typically represent *spring* and *autumn*, respectively.
43. They appeared 192 times, i.e., 16 + 98 + 29 + 49.

44. The ratio of 春風 to 秋風 (*spring wind* to *autumn wind*) is 72:26 for Li Bai, 14:2 for Li Shangyin, and 19:30 for Du Fu.
45. Liu and Luo, "Tracking Words in Chinese Poetry."
46. For instance, in the genre of *jijushi* 集句詩 (poems of gathered lines), the lines of other poems are borrowed and rearranged to form new poems.
47. "A Quatrain upon Ascending Leyou Plain upon Going to Wuxing" 將赴吳興登樂游原一絕, *QTS* 521.5961; translation adapted from Owen, *Late Tang*, 303–4.
48. "A Time of Great Peace, 6 of 7: I Love the Solitary Clouds" 太平時七首（其六）：愛孤雲, *QSC* 1.505.
49. Deaf Cheng 聾丞: provincial official, referring here to the speaker; alludes to Xu Cheng 許丞, who remained an upright and loyal official despite growing deaf with age. See Ban, *Han shu*, 89.3631.
50. "Because" 為有, *QTS* 539.6168.
51. Groom of golden tortoises: groom from a wealthy household.
52. Morning court: a meeting between a ruler and his trusted councilors convened early in the day.
53. "Pusa Man, #2" 菩薩蠻其二, *QSC* 1.520.
54. "Parting with a Traveler" 別客, *QTS* 386.4354.
55. "Seeing Off Yang Fa to the East" 送楊發東歸, *QTS* 538.6137.
56. "Depiction of an Asiatic Apple" 海棠圖, *QTS* 679.7784. Asiatic apple: *Malus spectabilis*. Jiangnan: the Lower Yangzi region. Shu's city: Chengdu.
57. "Phoenix Perching in a Paulownia, 3 of 3: Gazing at Chang'an" 鳳棲梧三首 （其三）：望長安, *QSC* 1.506.
58. "Newly Planted Pine Tree" 新栽松, *QTS* 887.10028.
59. "River God, #1: On a Plum, Sent to Yu Shuliang" 江神子（其一）：賦梅寄余叔良, *QSC* 3.1957.
60. For an introduction in English to the circumstances of *QTS*'s compilation, see Kroll, "*Ch'üan T'ang shih*."
61. "Stopping by Princess Yuzhen's Palace" 過玉真公主影殿, *QTS* 279.3169.
62. "Inscribed on Anguo Abbey" 題安國觀, *QTS* 783.8843.
63. See Wang, *Tang yulin jiaozheng*, 7.881–82.
64. See Li, *Luo Yin ji xinian jiaojian*, 991–92.
65. See Li, *Wenyuan yinghua*, *juan* 254; Ji, *Tangshi jishi jiaojian*, 26.702–04; and Xin, *Tang caizi zhuan jiaojian*, 4.109–14.
66. *QTS* 812.9147 notes that Kezhi might be the poet of this poem.
67. To compare any two items in a pool of 62,257 (42,863 + 19,394) items, we conduct 62,257 × 62,256 ÷ 2 = 1,937,935,896 comparisons.
68. Liu and Luo, "Tracking Words in Chinese Poetry."
69. Hu and Kuo, "Universality and Scaling in the Statistical Data of Literary Works."
70. Liu et al., "Character Distributions of Classical Chinese Literary Texts."
71. Zipf, *Selected Studies of the Principle of Relative Frequency in Language*; Zipf, *Human Behavior and the Principle of Least Effort*.
72. Since the numbers of distinct characters (*types* in linguistics) in the collections are different, we need to choose a specific number of items from each curve to compare. *SJ* has the fewest types in our data, so our choice must be no more than the number of types in *SJ*. If we use the data for the 1,000 most frequent characters for each curve when we compute the average of the squares of the differences of $\log(f/N)$ between the curves, the values

between *QTS* and *PT*, *QSC*, *QSS*, *YSX*, and *LCSJ* are smaller than 0.0006, whereas the values between *QTS* and *SJ*, *CV*, and *HF* are 0.0129, 0.0080, and 0.0071, respectively. If we use the number of types of *SJ* for each curve when we compute the correlation coefficients between the curves, the values between *QTS* and *PT*, *QSC*, *QSS*, *YSX*, and *LCSJ* are larger than 0.999, whereas the values between *QTS* and *SJ*, *CV*, and *HF* are 0.989, 0.991, and 0.995, respectively.

73. Chen, Guo, and Liu, "Statistical Study on Chinese Word and Character Usage"; Hu and Kuo, "Universality and Scaling in the Statistical Data of Literary Works."
74. Some of these sixteen characters are homonyms, for which only one pronunciation is provided: *bu* 不 (not), *wu* 無 (no, there is not), *feng* 風 (wind), *ren* 人 (person), *you* 有 (have, there is), *yun* 雲 (clouds), *ri* 日 (sun, day), *yi* 一 (one), *shan* 山 (mountain), *tian* 天 (sky), *he* 何 (what, how), *hua* 花 (flower blossoms), *lai* 來 (come), *zhi* 之 (various meanings, most often a particle indicating noun-phrase modification), *wo* 我 (first-person pronoun), and *chun* 春 (spring).
75. Liu et al., "Character Distributions of Classical Chinese Literary Texts."
76. Liu, "Quantitative Analyses of Chinese poetry of Tang and Song dynasties"; Liu, "Flexible Computing Services"; Liu and Luo, "Tracking Words in Chinese Poetry."
77. Note that we use *C* here to stand for a collection, whereas in figure 5 we used *C* to stand for the locations of common characters.
78. All Chinese characters within the boxes in figure 15 are poets' names, and we do not provide their Hanyu Pinyin here.
79. See "Butterfly Enamored of Flowers" 蝶戀花: "A mat of sapphire and a cuisine of sand, / Sleeping to noon in a haze, / The oriole's warble is sprightly, as if it understands" 碧簟紗廚，晌午朦朧睡。鶯舌惺忪如會意 (*QSC* 1.104). This verse has also been mistakenly attributed to Su Shi 蘇軾 (1037–1101). See Zou and Wang, *Su Shi ci biannian jiaozhu*, 936.
80. "Bian He clasps his block of jade and weeps tears of blood: / Where can he find a craftsman good enough to shape it?" 和抱璞而泣血兮，安得良工而剖之. From "Reckless Remonstrance" 謬諫 in *CV*; see Hong, *Chuci buzhu*, 13.254; and Hawkes, *Songs of the South*, 257.
81. Mikolov et al., "Distributed Representations of Words and Phrases."
82. See, e.g., Jack Chen's work, esp. Chen, *Poetics of Sovereignty*.
83. Liu, *Chuci zhuzi suoyin*; Liu, Chen, and He, *Xie Tiao ji zhuzi suoyin*.
84. Lai, "Li Shangyin wuyan lüshi cihui fengge zhi yanjiu."
85. Gao, *Quan Tang shi zhong qinniao rushi zhi yanjiu*; Hsu, *Zhong Tang tungwu yuyen shi yanjiu*; Lin, *Tang dai chashi yanjiu*; Pan, *Caomu yuanqing*.
86. Yao, *Pi Rixiu Lu Guimeng changhe shi yanjiu*.
87. Wang, *Songci yu Tangshi zhi duiying yanjiu*.
88. Sun, *Poetics of Repetition in English and Chinese Lyric Poetry*.
89. Hu and Kuo, "Universality and Scaling in the Statistical Data of Literary Works." See also Chen, Guo, and Liu, "Statistical Study on Chinese Word and Character Usage."
90. Wu, *Yongwu yu xushi*.
91. Luo, *Liuchao senglü shi yanjiu*; Mazanec, "Invention of Chinese Buddhist Poetry."
92. Chen, *Quan Tang shi bubian*; Owen, *Poetry of Du Fu*.
93. Lo, "Design and Applications of Systems for Word Segmentation."
94. Chu, "Ting Tangshi de jiaoxiang"; Tharsen, "Chinese Euphonics"; Lee, "Classical Chinese Corpus"; Lee, Kong, and Luo, "Syntactic Patterns in Classical Chinese Poems." See, e.g., a talk by Jack W. Chen on "The Quan Tang Shi and Topic Modeling: An Experiment in

Macroscopic Literary Analysis" : ceas.yale.edu/events/quan-tang-shi-and-topic-modeling-experiment-macroscopic-literary-analysis.

95. Liu, "Flexible Computing Services."

References

Ban Gu 班固 (32–92). *Han shu* 漢書 (History of the Han Dynasty). Beijing: Zhonghua shuju, 1962.

Chen, Jack W. *The Poetics of Sovereignty: On Emperor Taizong of the Tang Dynasty*. Cambridge, MA: Harvard University Asia Center, 2010.

Chen, Qinghua, Jinzhong Guo, and Yufan Liu. "A Statistical Study on Chinese Word and Character Usage in Literatures from the Tang Dynasty to the Present." *Journal of Quantitative Linguistics* 19, no. 3 (2012): 232–48.

Chen Shangjun 陳尚君. *Quan Tang shi bubian* 全唐詩補編 (A Supplement to *Quan Tang Shi*). Beijing: Zhonghua shuju 中華書局, 1992.

Chen You-Bing 陳友冰 and Wang De-Shou 王德壽. *Song ci qingshang: Bei Song pian* 宋詞清賞、北宋篇 (Selected Appreciations of Song Lyrics: Northern Song), 138–39. Taipei: Cheng Chung shuju 正中書局, 2001.

Cheng Wen-Huei 鄭文惠, Chao-Lin Liu 劉昭麟, Chiu Wei-Yun 邱偉雲, and Hsu Chu-Ting 許築婷. "Qinggan xianxiangxue yu secai zhengzhixue: Zhong Tang shige baise shuqing xipu de shuwei renwen yanjiu" 情感現象學與色彩政治學：中唐詩歌白色抒情系譜的數位人文研究 (Phenomenology of Emotion Politics of Color: Digital Humanities Research on the Lyrical Genealogy of "White" in the Poetry of Mid-Tang Dynasty). In *Digital Humanities: Between Past, Present, and Future*, edited by J. Hsiang, 207–57. Taipei: National Taiwan University Press, 2016.

Chu Chia-Ning 竺家寧. "Ting Tangshi de jiaoxiang: You shengyun fenxi shige de yinyuexing" 聽唐詩的交響－由聲韻分析詩歌的音樂性 (A Linguistic Analysis on the Rhythm of Tang Poetry). *Chinese Phonology* 16 (2009): 25–45.

Fuller, Michael A. *The China Biographical Database User's Guide*. Harvard University, 2015. https://projects.iq.harvard.edu/files/cbdb/files/cbdb_users_guide.pdf.

Gao Yi-Lu 高旖璐. "Quan Tang shi zhong qinniao rushi zhi yanjiu" 《全唐詩》中「禽鳥入詩」之研究 (A Study on Birds in Poems in the *Complete Tang Poems*). PhD diss., National Changhua University of Education, 2009.

Hawkes, David, trans. *Songs of the South: An Anthology of Ancient Chinese Poems by Qu Yuan and Other Poets*. 3rd ed. Harmondsworth, UK: Penguin. 1985.

Hong Xingzu 洪興祖, ed. *Chuci buzhu* 楚辭補注 (The *Songs of Chu*, with Supplementary Commentary). Beijing: Zhonghua shuju, 1983.

Hsu Jing-Yi 許靜宜. "Zhong Tang tungwu yuyen shi yanjiu" 中唐動物寓言詩研究 (A Study on the Animals in Fable Poems in Mid-Tang Period). Master's thesis, National Taiwan Normal University, 2008.

Hu, Chin-Kun, and Wei-Cheng Kuo. "Universality and Scaling in the Statistical Data of Literary Works." In *POLA Forever: Festschrift in Honor of Professor William S.-Y. Wang on His Seventieth Birthday*, edited by Dah-an Ho and Ovid J. L. Tzeng, 115–39. Taipei: Academia Sinica, 2005.

Hu Junfeng 胡俊峰 and Yu Shiwen 俞士汶. "The Computer Aided Research Work of Chinese Ancient Poems." *ACTA Scientiarum Naturalium Universitatis Pekinensis*, 37, no. 5 (2001): 725–33.

Ji Yougong 計有功. *Tangshi jishi jiaojian* 唐詩紀事校箋 (Recorded Contexts of Tang Poems, Collated and Annotated). Annotated by Wang Zhongyong 王仲鏞. Chengdu: Ba-Shu shushe, 1989.

Jiang Shaoyu 蔣紹愚. "Li Bai Du Fu shi zhong de 'yue' he 'feng'—jisuanji ruhe yong yu gudian shici jianshang" 李白杜甫詩中的"月"和"風"——電腦如何用於古典詩詞鑒賞 ("Moon" and "Wind" in Li Bai's and Du Fu's Poems—Using Computers for Studying Classical Poems). In *Proceedings of the First International Conference on Literature and Information Technologies*, 2003. http://cls.lib.ntu.edu.tw/LIT/papers/summary2_c.doc.

Kroll, Paul. "*Ch'üan T'ang shih* 全唐詩." In *The Indiana Companion to Traditional Chinese Literature*, edited by William Nienhauser, 1:364–65. Bloomington: Indiana University Press, 1986.

Lai, Yi-Hsin 賴宜欣. "Li Shangyin wuyan lüshi cihui fengge zhi yanjiu" 李商隱五言律詩詞彙風格之研究 (A Study on the Word Style of Li Shang-yin Five-Character Verses Lexical Style Research). Master's thesis, National Taichung University of Education, 2012.

Lee, John. "A Classical Chinese Corpus with Nested Part-of-Speech Tags." In *Proceedings of the Sixth EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 75–84. 2012. http://www.aclweb.org/anthology/W12-1011.

Lee, John, Yin Hei Kong, and Mengqi Luo. "Syntactic Patterns in Classical Chinese Poems: A Quantitative Study." *Digital Scholarship in the Humanities* 33, no. 1 (2018): 82–95.

Lee, Wei-Chih 李瑋質. "Wan Tang 'Wen-Li' zuopin dui Nanchao gongtishi zhi chengchuan yu chuangbian" 晚唐「溫李」作品對南朝宮體詩之承傳與創變 (Wen Ting-Yun and Li Shan-Yin's Works in the Late Tang Receive to the Gong-Ti Poetry of the Southern Dynasties). Master's thesis, National Central University, Taiwan, 2009.

Li Dingguang 李定廣, ed. *Luo Yin ji xinian jiaojian* 羅隱集繫年校箋 (The Works of Luo Yin, Chronologically Arranged, Collated, and Annotated). Beijing: Renmin wenxue chubanshe, 2013.

Lin Chen-Ying 林珍瑩. *Tang dai chashi yanjiu* 唐代茶詩研究 (A Study on Tea Poems in Tang Dynasty). Taipei: Huamulan wenhua chubanshe, 2007.

Liu, Chao-Lin. "Flexible Computing Services for Comparisons and Analyses of Classical Chinese Poetry." In *Proceedings of the 2017 International Conference on Digital Humanities*, 505–7. 2017. https://dh2017.adho.org/abstracts/612/612.pdf.

———. "Quantitative Analyses of Chinese Poetry of Tang and Song Dynasties: Using Changing Colors and Innovative Terms as Examples." In *Proceedings of the 2016 International Conference on Digital Humanities*, 260–62. 2016. https://arxiv.org/abs/1608.07852.

Liu, Chao-Lin, and Luo Kuo-Feng 羅國峯. "Tracking Words in Chinese Poetry of Tang and Song Dynasties with the China Biographical Database." In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities, the Twenty-Sixth International Conference on Computational Linguistics*, 172–80. Osaka: COLING 2016 Organizing Committee, 2016. https://aclanthology.coli.uni-saarland.de/volumes/proceedings-of-the-workshop-on-language-technology-resources-and-tools-for-digital-humanities-lt4dh.

Liu, Chao-Lin, Wang Hongsu 王宏甦, Hsu Chu-Ting 許筑婷, Cheng Wen-Huei 鄭文惠, and Chiu Wei-Yun 邱偉雲. "Color Aesthetics and Social Networks in Complete Tang Poems: Explorations and Discoveries." In *Proceedings of the Twenty-Ninth Pacific Asia Conference on Language, Information and Computation*, 132–41. 2015. http://aclweb.org/anthology/Y/Y15/Y15-2016.pdf.

Liu, Chao-Lin, Zhang Shuhua 張淑華, Geng Yuanli 耿元驪, Lai Huei-ling 賴惠玲, and Wang Hongsu 王宏甦. "Character Distributions of Classical Chinese Literary Texts: Zipf's Law, Genres, and Epochs." In *Proceedings of the 2017 International Conference on Digital Humanities*, 507–11. 2017. https://dh2017.adho.org/abstracts/080/080.pdf.

Liu Dianjue 劉殿爵, ed. *Chuci zhuzi suoyin* 楚辭逐字索引 (A Concordance to the *Chuci*). Hong Kong: Commercial Press, 2000.

Liu Dianjue 劉殿爵, Chen Fangzheng 陳方正, and He Zhihua 何志華, eds. *Xie Tiao ji zhuzi suoyin* 謝朓集逐字索引 (A Concordance to the Works of Xie Tiao). Hong Kong: Chinese University Press, 2000.

Lo, Fengju 羅鳳珠. "Design and Applications of Systems for Word Segmentation and Sense Classification for Chinese Poems." In *Proceedings of the Fourth Conference on Technologies for Digital Archives*. 2005. http://datf.iis.sinica.edu.tw/Papers/2005datfpapers/B-4.pdf.

———. "Shilun yinyong zixun keji zuowei shixue yanjiu buzhu gongju de fazhan fangxiang yu jiegou fangfa" 試論引用資訊科技作為詩學研究輔助工具的發展方向與建構方法 (On the Development and Construction of Research Tools for Studies of Poetry with Computing Technologies). In *Language, Literature, and Information*, edited by F. Lo, 319–63. Hsin-Chu: National Tsing-Hua University Press, 2004.

Luo Kuo-Feng 羅國峯. "*Quan Tangshi* de chubu fenxi: Banben bidui, shige duiying yu shehui wangluo" 《全唐詩》的初步分析：版本比對、詩歌對應與社會網路 (Some Studies of the *Complete Tang Poems*: Version Comparison, Word Alignment, and Social Network Analysis). Master's thesis, National Chengchi University, Taiwan, 2016.

Luo Wen-Lin 羅文伶. "Liuchao senglü shi yanjiu" 六朝僧侶詩研究 (A Study of Monk Poetry of the Six Dynasties). PhD diss., Tunghai University, 2002.

Mazanec, Thomas J. "The Invention of Chinese Buddhist Poetry: Poet-Monks in Late Medieval China." PhD diss., Princeton University, 2017.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. "Distributed Representations of Words and Phrases and Their Compositionality." In *Proceedings of the Twenty-Sixth International Conference on Neural Information Processing Systems – Volume 2*, 3111–19, 2013. https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

Owen, Stephen. *The Late Tang: Chinese Poetry of the Mid-Ninth Century (827–860).* Cambridge, MA: Harvard University Asia Center, 2006.

———, trans. *The Poetry of Du Fu*. 6 vols. Berlin: De Gruyter, 2016.

Pan Fu-Jun 潘富俊. *Caomu yuanqing: Zhongguo gudian wenxue zhong de zhiwu shijie* 草木緣情：中國古典文學中的植物世界 (Plants in Classical Chinese Literature). Taipei: Shangwu yinshuguan, 2015.

Peng Dingqiu 彭定求 (1645–1719), Shen Sanceng 沈三曾, Yang Zhongne 楊中訥, Wang Shihong 汪士鋐, Wang Yi 汪繹, Yu Mei 俞梅, Xu Shuben 徐樹本, Che Dingjin 車鼎晉, Pan Conglu 潘從律, and Cha Sili 查嗣瑮, eds. *Quan Tang shi* 全唐詩 (The Complete Tang Poems). Beijing: Zhonghua shuju, 2003.

Sun, Ceclie Chu-Chin. *The Poetics of Repetition in English and Chinese Lyric Poetry.* Chicago: University of Chicago Press, 2011.

Sun Yanhong 孫豔紅. "Tang Song ci benti tezheng de biaoxian xingshi" 唐宋詞本體特徵的表現形式 (Expressive Styles of the Tang and Song Lyrics). *Zhongguo shehuikexue wang: Zhongguo shehuekexue bao* 中國社會科學網-中國社會科學報 (Chinese Social Sciences Today), July 5, 2016.

Tang Guizhang 唐圭璋 (1901–90), ed. *Quan Song ci* 全宋詞 (The Complete Song Lyrics). Beijing: Zhonghua shuju. Reprint. Taipei: Hongshi chubanshe, 1981.

Tharsen, Jeffrey R. "Chinese Euphonics: Phonetic Patterns, Phonorhetoric and Literary Artistry in Early Chinese Narrative Texts." PhD diss., University of Chicago, 2015.

Wang Dang 王讜. *Tang yulin jiaozheng* 唐語林校證 (Grove of Stories of the Tang, Collated and Investigated). Edited by Zhou Xunchu 周勛初. Beijing: Zhonghua shuju, 1987.

Wang Wei-Yung 王偉勇. *Songci yu Tangshi zhi duiying yanjiu* 宋詞與唐詩之對應研究 (A Study on the Correspondences between Song Lyrics and Tang Poems). Taipei: Wenshizhe chubanshe, 2003.

Wu Ruyu 吳汝煜, ed. *Tang Wudai ren jiaowangshi suoyin* 唐五代人交往詩索引 (Indexes to the Exchange Poems of Tang and Five Dynasties). Shanghai: Shanghai guji chubanshe, 1993.

Wu Yi-feng 吳儀鳳. *Yongwu yu xushi—Han Tang qinniao fu yanjiu* 詠物與敘事—漢唐禽鳥賦研究 (A Study of Fu on Birds from Han to Tang Dynasty: Description and Narration). Taipei: Huamulan wenhua chubanshe, 2007.

Xiao Difei 蕭滌非, ed. *Du Fu quanji jiaozhu* 杜甫全集校注 (Complete Poetry of Du Fu, Collated and Annotated). Beijing: Renmin wenxue chubanshe, 2014.

Xin Wenfang 辛文房. *Tang caizi zhuan jiaojian* 唐才子傳校箋 (Biographies of Tang Talents, Collated and Annotated). Annotated by Fu Xuancong 傅璿琮. Beijing: Zhonghua shuju, 1995.

Yao Gui 姚垚. "Pi Rixiu Lu Guimeng changhe shi yanjiu" 皮日休、陸龜蒙唱和詩研究 (A Study on the Exchange Poems between Pi Rixiu and Lu Guimeng). Master's thesis, National Taiwan University, 1980.

Zhu Zejie 朱則傑. "Quan Qing shi bianzuan choubei weiyuan hui chengli" 《全清詩》編纂籌備委員會成立 (Establishing the Editorial Board for the *Complete Qing Poems*). *Qingshi yanjiu* 清史研究 (Studies in Qing History) 0, no. 3 (1994): 96.

Zipf, George K. *Human Behavior and the Principle of Least Effort: An Introduction of Human Ecology.* Boston, MA: Addison-Wesley Press, 1949.

———. *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge, MA: Harvard University Press, 1932.

Zou Tongqing 鄒同慶 and Wang Zongtang 王宗堂, eds. *Su Shi ci biannian jiaozhu* 蘇軾詞編年校注 (The Lyrics of Su Shi, Chronologically Arranged, Collated with Commentary). Beijing: Zhonghua shuju, 2007.