Describing Objects in Tang Dynasty Poetic Language: A Study Based on Word Embeddings

MARIANA ZORKINA

Abstract This article focuses on computational analysis of Tang dynasty "poems on things" (*yongwu shi* 詠物詩) and some of the most common objects described in them. Modern technology offers many possibilities for new approaches to the study of poetic language, and this article discusses some of the tools that can aid in semantic analysis of separate words or poems. These include so-called word embeddings, vector representations of word semantics, and "fingerprints" that are calculated on the basis of word embeddings to represent semantics of whole texts. Applied to classical Chinese poetry, they can show some of the paradigmatic groups of images and their distribution between concepts of happiness and sadness, loneliness and companionship. Finally, topical grouping of poems on things is discussed and explored with the help of fingerprints to look for formal principles behind the grouping of the texts.

Keywords Tang dynasty poetry, poems on things, word embeddings, distributional semantics, semantic fingerprints

Classical Chinese poetry, especially that of the Tang dynasty (618–907), is rich in figurative language, with images of objects playing an important role in the construction of a verse. Sometimes they are mere units of metaphoric language: a fan can be used to symbolize a woman; bamboo, a noble literatus. Sometimes they are just necessary elements to define the setting and mark the mood of a verse: wind, water, or cicadas indicate a sad or contemplative piece, while flowers and jewels are associated with festivities and happiness. Combined, these objects constitute an important part of the poetic language of premodern China, an

DOI 10.1215/23290048-7256989 • © 2019 by Duke University Press

image system used indirectly to tell a story, convey a message to the reader, or refer to another text.

The history of Chinese poetry reveals a genre that treated these objects as its main topic—the so-called poems on things (*yongwu shi* 詠物詩). Although the genre was recognized relatively late, conceptualized as a separate phenomenon in the fifth century at the earliest, poems dedicated to certain animals and fruit can be traced to the *Book of Odes* (*Shijing* 詩經; 11th–7th century BC) and *Songs of Chu* (*Chu ci* 楚辭; before 3rd century BC). The main topics usually include objects of material culture, animals, and plants, as well as celestial objects and geographical features. Anthologies compiled during the Qing dynasties—*Yongwushi xuan* 詠物詩選 (Selection of Poems on Things) and *Peiwen zhai yongwushi xuan* 佩文齋詠物詩選 (Collection of Poems on Things from the Peiwen Study; hereafter cited as *YWSX*)—extend the list, adding natural phenomena such as wind or rain, seasons, and different types of people.

The composition and message of *yongwu shi* could be very different: some limited themselves to the appraisal of something; others were more like riddles that encouraged the reader or listener to guess the topic. Often the object was a substitute for the author, who used it as an indirect way to convey his thoughts and feelings. However, on the surface level the vocabulary stayed close to the formal topic. The poem below, "Cicada" by Li Shangyin 李商隱 (ca. 813–58), is one of the most famous examples of the genre.

Cicada	蟬
Li Shangyin	李商隱
By high-minded nature not eating its fill,	本以高難飽
Wasted effort, voice spent in resentment.	徒勞恨費聲
Before dawn cries sparse, almost ceasing,	五更疏欲斷
A whole emerald tree doesn't care.	一樹碧無情
I, low official, a bough still adrift,	薄宦梗猶泛
Weeds already cover my garden at home.	故園蕪已平
You have kindly warned me most strongly:	煩君最相警
I and my family are pure like you.	我亦舉家清
	[*Quan Tang shi* 全唐詩 (Complete Tang Poems; hereafter cited as *QTS*) 539.6147; trans. Owen, *Late Tang*, 452]

Here, Li Shangyin uses a typical approach for *yongwu* poetry and describes the cicada as a substitute for himself: as cicadas, according to traditional assumptions, live high in the trees and feed only on dew and wind, never able to eat their fill, so is the poet with his ideals. The insect cries in vain in the first couplet, and even when it is about to vanish, the surroundings stay indifferent. In the third couplet Li Shangyin separates from the cicada and complains that he, a low-level official, has to drift from place to place restlessly, alluding to a story from *Zhanguo ce* 戰國策 (Strategies of the Warring States; 5th–3rd century BC), where a wooden figurine is warned it will one day float away with the rain waters. The next line reinforces the image of the unsettled life of the poet by describing his garden, neglected and filled with weeds, as there is no one to look after it. In the last couplet the situation is reversed: now the insect is likened to a human as it enters into a dialogue with the poet, and Li Shangyin resolves to live life as pure as the cicada. A true masterpiece of *yongwu*, this poem remains within its limits of natural, visual imagery yet conceptually leaps from one perspective to another to arrive at and communicate the both the "futility of chanting" and nonetheless the aspiration for a pure life.

Traditional analysis of poetry and its language, partially illustrated above, relies first on selecting a significant poem or a group of poems that, in the eyes of the scholar, share some common features. Then, through knowledge of other poetry, history, and culture of the era, the scholar goes about explaining the realia behind the poetic work, the usage of certain symbols, and their metaphorical meaning. This approach, however, raises a question of objectivity. One's idea about shared features may be biased, and a manual selection from the vast poetic heritage of China can hardly be comprehensive. Furthermore, taking this complex system as a whole (leaving aside the tradition of Chinese poetic criticism for the moment) and grasping its inner laws and principles are extremely challenging tasks.

Enter the digital humanities with their tools of quantitative analysis to find relations and patterns based solely on the corpus rather than preselection and interpretation of manageable texts made by humans. Far from self-sufficient, this approach is merely a means to an end that helps a researcher gain some distance from the data, to reduce its complexities and irregularities in order to see more structure behind them. Interpretation and conceptualization of the data still lie with the scholar, but the computational power of generalization brings analysis of data to a new level that may allow us to see poetry from a new perspective.

The present article has two main goals. First, I provide a short introduction to the concept of word embeddings, along with the application of neural networks to the study of word semantics. I believe many modern humanities researchers treat the digital approach with caution and distrust because of the seemingly cryptic nature of computer-based research, especially when it comes to the use of artificial intelligence. Thus, I describe the algorithms used and their potential applications in the hopes of making the principles behind this study more transparent, thereby encouraging other researchers to try these technologies and methods and explore the possibilities they bring.

The second goal is to demonstrate an application of neural networks analysis to a well-defined corpus of Chinese poetry—specifically, to analyze a selection of objects that frequently appear in Tang dynasty verse: their interrelations, metaphorical meaning, and poems about them. Some of the topics and phenomena discussed, such as the paradigmatic groupings of objects or their correlations with specific emotions in poetic language, have already been analyzed in traditional sinology. Here, I examine them through the lens of computer analysis to uncover some interesting patterns behind the presentation of objects in Tang poems and ways to examine them from the quantitative studies standpoint. Other patterns, such as the grouping of *yongwu* poems based on their inherent structures and comparison of the data with topical groupings in traditional Chinese anthologies, have not received much attention to date. With this analysis, I not only present a new angle on aspects of Chinese poetic language in the Tang but also show possible areas of continuity between qualitative and quantitative methodologies.

Word Embeddings and Mathematical Representations of Semantics

The idea of word vectors is based on the insight that a word's meaning can be deduced from its context, formalized in distributional semantics. In other words, according to the oft-cited statement of John R. Firth, "You shall know a word by the company it keeps."[1] On the other hand, if the statement is reversed, it would mean that words occurring in the same context tend to have similar meanings. As part of distributional semantics, this concept dates back to the 1950s; it is believed that Ludwig Wittgenstein was the first to articulate it,[2] but the idea was also expressed in 1954 by Zellig S. Harris in his article "Distributional Structure."[3] Attempts to measure word meanings mathematically date roughly to the same time. Against this background, the American psychologist Charles Osgood first introduced the idea of semantic vectors, now widely used and developed in distributional semantics.[4]

These vectors can be designed in different ways to fulfill a variety of tasks, but when they concern the meaning of separate words, they are based on word co-occurrence matrices, built to calculate the frequency of every word's appearance in combination with all the other words. This sequence of frequency values, corresponding to a word, can be used as a representation of its meaning and treated as a vector in calculations. Words with similar vectors and consequently their co-occurrence patterns are considered similar in meaning. This means that they have similar semantic profiles and are not necessarily strict synonyms, for example, *early* (*zao* 早) and *late* (*wan* 晚), or *day* (*ri* 日) and *night* (*ye* 夜) would be considered close in meaning since they describe similar objects and situations and are often interchangeable. However, to conform to the accepted terminology, *synonymy* is still used later in this article to describe such relations.

When large corpora are involved, the matrices of co-occurrences and, consequently, the vector representations of each word may contain millions of units and dimensions. This redundant information increases the complexity of the necessary calculations, so different ways of reducing the dimensionality of vectors were found to help create "dense vectors" (i.e., shorter vectors with most values that are not zero) without losing the original meanings and interrelations. When referring to mathematical representation of semantics they are called *word embeddings*. Perhaps the most popular approach at the moment is the use of predictive models based on artificial neural networks, created by a group of researchers from Google headed by Tomas Mikolov.[5] Instead of learning the actual co-occurrences in the texts, these algorithms assign random values of set length to words in the document and then go through the text multiple times, trying to adjust to more plausible values. The algorithm used in this article, CBOW, performs this task by trying to predict a word from its context and changing the value if the guess is wrong.[6] There are also numerous settings and options that considerably influence the performance of a word embedding model, for example, the choice of neural network, how many characters around the target word are treated as context, the length of the mathematical representation of the word semantics, and so on.[7] A more detailed account on these technical details can be found in the online appendix to this article.[8]

The drawback of the word embeddings produced by a neural network is that the values representing each word no longer correspond to any specific semantic features, as there is no direct connection between the frequencies of word co-occurrences and the vectors. Thus, the main operation with word embeddings is comparison: the cosine of the angle between vectors is calculated, and the smaller the angle between vectors (and the bigger its cosine), the more similar the vectors, hence the more similar the corresponding word meanings. Likewise, as the angle grows, there are fewer similarities and the cosine is smaller.

Given the technology, the questions I want to raise are what use word vectors would be outside computational linguistics, and how might they be employed to reveal something new about literary works and realia behind them. In other words, how can we detach computational methods from their original, often entirely applied purposes, and reinterpret them to align with methodologies and questions of literary studies?

One of the distinguishing features of computer analysis is that it requires data designed to be comprehensible to a machine—structured, formalized, translated into numbers. This requirement forces the same framework on the research and types of questions that can be asked. At the same time, a study of the structure of any system of signs, be it language in general, texts, or any other cultural phenomena (e.g., the "garment system" created by Roland Barthes to describe matching and nonmatching pieces of clothing in fashion),[9] often comes down to exploration of two basic types of relations: syntagmatic and paradigmatic. These concepts have been in use since Ferdinand de Saussure's *Course in General Linguistics*, redefined by Roman Jakobson,[10] and have thrived ever since, including in the fields of computational linguistics and digital humanities, heavily dependent as they are on finding and describing different types of structures.[11]

Syntagms and paradigms create two axes for the creation of meaning. Syntagmatic relations form the horizontal axis, which shows the ways units within a system can be combined into a "chain." One of the extensions of the analysis of syntagmatic relations is the study of collocations. In the field of Tang poetry studies, such an analysis was conducted, for instance, by Liu Chao-lin and colleagues.[12] Paradigmatic relations, on the other hand, function as the vertical axis and show the units that can replace each other in the same position. Whereas syntagmatic relations are the possibilities of combinations, words related paradigmatically actually have very little chance of appearing together: they are associated with each other as they all belong to a particular category, but as a general rule the choice of one excludes the choice of the other.

In the context of classical Chinese poetry and its imagery the questions about paradigmatic relations would be, simply speaking, can *white hair* be substituted with *black hair* as a legitimate poetic metaphor? Or, if a commonplace in a poem to express sadness is a tower (partly because the words for *sadness* [*chou* 愁] and *tower* [*lou* 樓] rhyme),[13] what can the *tower* be substituted with? And when such legitimate-within-the-system substitutes are found, the next question is how the choice of this specific word over the other options in the given position contributes to the meaning.

It is the paradigmatic relations that vector semantics are concerned with. Hence, although the result of computational analysis of word vectors is not as strict as structuralist descriptions of the concept of such relations, and words marked as belonging to the same paradigmatic group do not necessarily exclude each other in a sentence, the general principle stays the same: the closer the contexts in which two words appear, the higher the chance they can substitute for each other in a text.

The extraction and study of such relations have great value, as Jonathan Culler (an advocate of structuralism) described them: they "are important for what they can explain: meaningful contrasts and permitted or forbidden combinations."[14] By getting a better grip on the structures behind poetic language, we can better understand the conventions under which poetic creativity and the creation of meanings emerged.

Yet another application of computer analysis is the search for hidden patterns and structures behind the text and the language. As noted by Edward Slingerland and colleagues, the benefit of unsupervised machine learning methods, such as those described earlier, is that they make fewer assumptions about the corpus they are working with and thus cannot make mistakes based on cognitive biases. Even though some of the associations and correlations found by modern machine learning algorithms are spurious, with some caution one can use the patterns they produce to explain various phenomena or, as I show later in this article, compare them to a human-created structure.[15]

The Data Set

The performance of a vector model relies heavily on the size of the corpus chosen for training. And while modern machine learning makes use of big data and corpora comprising entries from Wikipedia and similar sources with billions of words, collections of Chinese poetry, however ample, are more modest in size, especially when one wants to stay within the limits of a reasonable time span of texts. As the focus of this study was Tang poetry, the easily accessible *Quan Tang shi* 全唐詩 (Complete Tang Poems; hereafter *QTS*) was chosen as a training corpus. This collection was commissioned in 1705 and contains more than forty thousand poems with over 2.5 million characters—an acceptable size but a small corpus in relation to what is necessary to build a word embedding model. For the same reason, no attempt to draw a distinction between different poetic forms has been made; although it is tempting to see how the distribution of words and their meanings might change from form to form, truncating the corpus would mean subverting the accuracy of the model. As a result, the vector space model created contains the generic poetic language of the Tang dynasty, which allows it to discover and describe common perceptions about the nature and qualities of objects appearing in the texts, as well as their relations to one another.

The other corpus used is the *Peiwen zhai yongwushi xuan* 佩文齋詠物詩選 (Collection of Poems on Things from the Peiwen Study; hereafter *YWSX*)—a collection compiled during the Qing dynasty (1636–1912), roughly at the same time as the *QTS*, under the main editorship of Zhang Yushu 張玉書 (1642–1711). It contains roughly fourteen thousand poems,[16] ranging from antiquity to the Ming dynasty (1368–1644), 5,751 of them marked by the editors as belonging to the Tang dynasty. The poems are grouped into topics that correspond to different "things" (*wu* 物), the traditional label for this genre. These include celestial objects, land formations, plants, animals, man-made objects like tools or buildings, foods, seasons and special dates, and people of different professions. They are also divided into groups of poetic forms. The main reason behind using the second corpus is the organization of the *YWSX*. Unlike the *QTS*, in addition to author and title it gives extra information about each poem, such as the topic and poetic form. This provides an opportunity to do more complicated types of poem analysis without time-consuming manual markup and helps avoid researcher interference as when a classification is imposed on the source text.[17]

For this study, corpora had only basic structural markup to help the algorithm distinguish poem boundaries.[18] Also, to lower the technical complexity of the task, each character was treated as a separate word; although it would be wrong to assert that classical Chinese poetry consists only of one-character words, a more precise word segmentation, that is, automatic distinction of word boundaries and proper names, would be complicated due to the problematic nature of wordhood in classical Chinese. However, studies show that in Tang poetry most words do consist of one character,[19] so this approach does not significantly change the results of the algorithm's semantic analysis, which in fact turn out to be quite robust.

Often a list of stop words is created, consisting of words used so frequently they give very little information and tend to skew the results of analysis. But considering the importance and weight of each word in a poem and the small number of function words, I decided not to use these lists, even if it contradicts the general practice. Many of the function words that would be removed in prose have relatively low frequencies in poetry and thus do not considerably skew the results; among uninformative words occurring with high frequency in the *QTS* were *no, not* (*bu* 不, *wu* 無), *have* (*you* 有), and *one* (*yi* 一), whereas other words in the top range, such as *person* (*ren* 人), *flower* (*hua* 花), and *mountain* (*shan* 山), are referential and thus important for the analysis. Of course, it might still be argued that *wu* 無 and *you* 有 also contribute greatly to the meaning of a poem. As a result of these decisions, the model with word embeddings trained on the *QTS* and discussed below contains information about the semantics of 7,125 words.[20]

The last decision concerned the choice of words referring to objects included in the study. The point of departure for the selection was the table of contents of *YWSX* listing all the objects and phenomena that, according to the editors, could be things in the poems on things. To reduce complexity, only topics consisting of or reducible to one character (e.g., *juhua* 菊花 [chrysanthemum] can be reduced to *ju* 菊 without loss of meaning) were chosen. Then, only topics that were connected to more than ten Tang poems were selected. These numbers do not simply reflect the popularity of these things described as poetic objects during the Tang; it is more a mixture of what was available to the compilers of the *YWSX*, how many Tang poems related to a certain thing were judged worthy of inclusion, and which topics were considered interesting during the Qing period, when the *YWSX* was created. But ideally a logical connection is still at work—the more popular the object as a poetic topic in Tang, the more poems would be written and survive and the greater chance more of them would be included. In the end, ninety-three words for objects were selected for the analysis.[21]

Synonyms in the System of Poetic Language

There are several ways to test the quality of word vectors and to use them.[22] The main and easiest type of task consists of extracting a word's synonyms and presenting them in a table for an intuitive estimation by a human reader. The general problem of intuitively assessing the performance of models trained on texts rich in metaphors but limited in word usage and vocabulary is that even an ideal system will not be able to produce perfect synonyms for a given word from the point of view of the language as a whole: a poetic language is to a large extent a closed system that chooses to include some types of words and relationships and excludes others or prioritizes some types of semantic relations over the rest. The relations the model records are relations between the items of such a system, and they can show strong paradigmatic connections not because they are actually synonymous but because they belong to the same paradigmatic group as poetic images: expressing the same mood, emotion, or situation. I believe the more implications and the bigger the metaphorical role of the word, the stronger its connections to other important images, although in reality these words would hardly be synonymous. From this point of view, the trained model provides plausible information. Table 1 shows such the results of a synonym search.

The synonyms produced depend on many factors. First, they depend on the number of similar words in the model, as in such obvious examples as *clothes* (*yi* 衣), *mountain* (*shan* 山), and *swallow* (*yan* 燕), all producing a large number of closely related words. When the number of synonyms in a given vocabulary is limited, the model produces concepts that are associated with the source word. This happens, for instance, in the case of *wind* (*feng* 風): the list contains different synonyms meaning "cold," as well as some wind instruments, connected to the target word by the concept of air movement.

As expected, when it comes to strong poetic images, the picture gets more interesting, as in the case of *moon* (*yue* 月), *mirror* (*jing* 鏡), *star(s)* (*xing* 星), and *fisherman* (*yu* 漁): the connection between the moon, the spiritual world, and souls is revealed, as well as an association between the moon and a mirror. The stars, though not associated with the moon in a straightforward way, are

Table 1. Synonyms produced by the vector model

Buddhist monk (seng 僧)	*Mountain* (shan 山)	*Clothes* (yi 衣)	*Swallow* (yan 燕)	*Moon* (yue 月)
chan 禪 (Chan buddhism)	*feng* 峯 (peak)	*pao* 袍 (robe)	*que* 鵲 (magpie)	*jing* 景 (landscape)
zhai 齋 (fasting)	*yan* 巖 (cliff)	*jin* 巾 (soft hat)	*die* 蝶 (butterfly)	*ri* 日 (sun)
cha 茶 (tea)	*cen* 岑 (high hill)	*shan* 衫 (gown)	*hu* 鵠 (crane)	*ying* 影 (shadow)
song 松 (pine)	*xi* 谿 (gorge)	*qiu* 裘 (fur coat)	*zhi* 雉 (pheasant)	*jing* 鏡 (mirror)
ren 人 (person)	*cun* 村 (village)	*ni* 霓 (variegated)	*fu* 鳧 (wild duck)	*zhu* 燭 (candle)
shi 師 (master)	*xi* 溪 (creek)	*ru* 襦 (jacket)	*yuan* 鳶 (hawk)	*ye* 夜 (night)
qi 棋 (chess)	*ling* 嶺 (mountain ridge)	*chang* 裳 (skirt)	*lei* 鸓 (flying squirrel)	*xue* 雪 (snow)
weng 翁 (old man)	*lou* 樓 (tower)	*qun* 裙 (skirt)	*sun* 隼 (falcon)	*che* 徹 (penetrate)
yuan 猿 (ape)	*zhang* 嶂 (high cliff)	*qin* 衾 (bed quilt)	*qin* 禽 (birds)	*xi* 夕 (evening)
reng 仍 (still)	*yuan* 原 (source)	*ying* 纓 (ribbon)	*zhuo* 啅 (chirp)	*po* 魄 (soul)
Star (xing 星)	*Sword* (jian 劍)	*Wind* (feng 風)	*Fishing* (yu 漁)	*Mirror* (jing 鏡)
ying 螢 (firefly)	*ji* 戟 (halberd)	*xu* 噓 (sigh, hiss)	*bian* 扁 (small, often boat)	*jian* 鑑 (mirror)
chan 蟾 (toad)	*jian* 箭 (arrow)	*biao* 飆 (whirlwind)	*qiao* 樵 (firewood)	*xia* 匣 (sheath)
yin 銀 (silver)	*xia* 匣 (sheath)	*qi* 淒 (cold, cloudy)	*ji* 楫 (oar)	*yan* 眼 (eye)
chan 躔 (course of stars)	*mao* 矛 (spear)	*liang* 涼 (cold)	*diao* 釣 (to fish)	*di* 的 (bright)
shao 筲 (bucket)	*pei* 轡 (reins)	*jia* 笳 (reed flute)	*meng* 艋 (small boat)	*lian* 臉 (face)
mao 昴 (name of a constellation)	*bu* 簿 (book)	*huai* 淮 (Huaihe river)	*nong* 儂 (here *nong* 農, peasant)	*bang* 蚌 (oyster)
diao 刁 (sly)	*tuo* 橐 (sack)	*lin* 霖 (continued rain)	*ji* 磯 (a rock projecting over the water)	*yue* 月 (moon)
dou 斗 (dipper)	*tie* 鐵 (metal)	*sa* 颯 (sound of wind)	*bo* 泊 (berth)	*jiao* 皎 (bright)
hong 虹 (rainbow)	*zhen* 鎮 (restrain, garrison)	*xiao* 簫 (bamboo flute)	*ze* 舴 (small boat)	*dian* 殿 (palace)
deng 燈 (lamp)	*bi* 筆 (brush)	*fei* 霏 (falling of snow and rain)	*chuan* 船 (boat)	*mian* 面 (face)

connected to one of its symbols, a toad. A fisherman or the act of fishing has a strong connection to another symbol of a simple life, gathering firewood. The examples here are rather simple, to illustrate the credibility of the algorithm, but this method of defining paradigmatic groups may prove useful in finding lesser-known associations.

Another type of task is performing simple arithmetical operations on vectors, and the model has shown it can answer straightforward ones correctly.

One possible operation involves subtraction, that is, asking a model what a given concept would become if we deprive it of one of its important features. For example, a *married woman* (*fu* 婦) without a *husband* (*fu* 夫) (vector of a *married woman* minus vector of a *husband*) produces an *unmarried woman* (*nü* 女), *small* (*xiao* 小), and *courtesan* (*ji* 妓) in the top ten results; a tree (*mu* 木) without leaves (*ye* 葉) produces *to wither* (*fei* 腓).

Word embeddings on their own can only help assess word semantics and give no information about the surrounding text, but with information provided by the algorithm one discovers that in the *QTS* the words *to wither* and *leaves* exclude each other, so a tree can have only one state in a piece, with leaves or without, as in the following lines by Sun Ti 孫逖 (696–761) in the poem "Pool at the Mountain Geshan" 葛山潭 (Geshan tan):

	It is so cold that grass and trees wither,	凉哉草木腓
6	White dew soaks people's clothes.	白露沾人衣
		[*QTS* 118.1187]

The only poem in *QTS* that combines both words is "Worries on the Borders" 邊愁 (Bian chou) by Cui Shi 崔湜 (671–713), who simply used *leaves* instead of the more conventional combination of grass and trees at the position:

	On the ninth month the fleabane stems break,	九月蓬根斷
2	On the borders grass and leaves wither.	三邊草葉腓
	Wind and dust change the horse color,	風塵馬變色
4	Frost and snow dress the sword in new clothes.	霜雪劍生衣
		[*QTS* 54.663]

It is also possible to perform addition, to add features to a selected word. For example, if one combines *person* (*ren* 人) with *virtue* (*de* 德), the result of the calculations would be *righteousness* (*yi* 義); in other words, if a person is virtuous, he or she can additionally or alternatively be called righteous.

The more complex the task, the more mistakes the model produces—this is when the lack of more detailed tagging of the corpus starts to show. Thus, the model solves problems of excluding a nonmatching word relatively well in successions like *tiger* (*hu* 虎), *deer* (*lu* 鹿), *dog* (*quan* 犬), and *flower* (*hua* 花) (*flower* being excluded), as well as more subtle ones: out of *willow* (*liu* 柳), *part* (*bie* 別), *sad* (*chou* 愁), and *happy* (*xi* 喜) the last one is marked as non-matching. However, a complex logical problem typical for machine learning tests—looking for a connection "a to b is like c to __" ("Paris is to France as London is to __")—produces uneven results.[23] Although the model is able to recognize a number of associations for gender relations and finds analogues among the "five relations" (*wulun* 五倫) of Confucianism (i.e., father and son, husband and wife, ruler and subject, elder and younger brother, older and younger friend), this type of analysis often produces unreliable results and is not used further in the article.

I hope it has become clear that the model performs well, especially when carrying out lower-level tasks like finding synonyms. On the other hand, finding more complex logical correspondences between different concepts is the hardest task and usually requires very specific and clear questions. When strong connections are searched for, the model of Tang poetic language produces plausible results. This capability of defining different types of relations between words and, as explained further below, whole texts is the basis of many text analysis tasks and determines the clustering of poems into semantic groups later in this article.

Relations between Objects in Tang Poetic Language

Another way to explore the objects in Tang poetic language is to see their relationships with abstract concepts, popular moods, and topics. Classical examples of such associations include autumn, often connected to contemplations about the past and sad thoughts; ape cries reminding the poet of his loneliness; and willow trees, often a symbol of someone's will to stay.[24] The question is, how do these images correlate with each other? Which things would be associated for Tang poets with sadness and grief as powerfully as autumn? Which drink is more common to enjoy with company—tea or wine? Which plants are considered the most fragrant?

Such questions can be answered with word embeddings, which allow us to visualize semantic relations of a group of words in a two-dimensional space with their "maps of images." For this purpose, pairs of antonymous (at least to some degree) attributes, such as *happiness* (*xi* 喜) and *worries* (*chou* 愁), *alone* (*du* 獨) and *accompanying* (*pei* 陪), *fragrant* (*xiang* 香) and *(bad) smell* (*chou* 臭), have been chosen and set up to represent the x- and y-axes of the plots. Then, for each word from the list, its cosine similarities to both attributes have been calculated to represent the coordinates on corresponding axes. The resulting plot shows how the chosen concepts correlate in a space between two chosen extremes. It is important that this method is prone to producing false correlations: one could use unrelated words to represent the axes and, in the spirit of Jorge Luis Borges and the fictional Chinese taxonomy created by him, explore how animals distribute between being fabulous and embalmed.[25] The choice of concepts to compare against lies completely with the researcher, as does the choice of material for any quantitative analysis, and without control can lead to misinterpretations. In this study the intent was to use pairs of antonyms that are known to be associated with poetic images, are relatively frequent in the corpus, and produced no errors in tests for synonymy.

Since the question of smell—as opposed to feelings or mood—is easier to verify against widespread conceptions of what is pleasant and not, it is used here as proof of concept. Figure 1 shows the distribution of plants: *bamboo* (*zhu* 竹), *willow* (*liu* 柳), *pine tree* (*song* 松), *lotus (flower)* (*he* 荷), *peach (blossom)* (*tao* 桃), *chrysanthemum* (*ju* 菊), *plum (blossom)* (*mei* 梅), *grass* (*cao* 草), *osmanthus flower* (*gui* 桂), *apricot (blossom)* (*xing* 杏), *pomegranate (flower)* (*liu* 榴), *vine (flower)* (*teng* 藤), *leaves* (*ye* 葉), *tangerines* (*ju* 橘), *grain* (*gu* 穀), and *orchids* (*lan* 蘭). The model produces plausible data: willow, bamboo, grass, and leaves have no distinct features and are little associated with smells, while pine trees; plum, apricot, and peach blossoms; orchids; and (somewhat less) osmanthus flowers are shown as pleasant to smell. Grains seem to be the least pleasant in the group: this is not a straight association, as they never actually occur with the word *(bad) smell*. But, unlike other plants in the selection, they have strong connections to food, which on many occasions in the *QTS*, especially when it comes to meat, is described as bad smelling, as in the lines of "Going from the Capital to Fengxian County, Singing My Feelings in Five Hundred Characters" 自京赴奉先縣詠懷五百字 (Zi jing fu Fengxian xian yonghuai wubai zi) by Du Fu 杜甫 (712–70):

	Behind the vermilion gates rot wine and meat,	朱門酒肉臭
68	On the road lie the bones of those frozen to death.	路有凍死骨
		[*QTS* 216.2265]

This illustrates how, in contrast with flowers and trees, food has a stronger connotation with rotting, hence with waste.

After ascertaining that the model handles this type of problem well, one can make more general observations. While reading the visualizations, however, it is important to note the x- and y-axes are not always symmetrical. Depending on the point coordinates on the plot, both can be skewed more to the negative or positive side, so a point that visually appears in the middle is not necessarily equally related to both attributes. As for the coordinates (and as a general principle for word embeddings), the cosine similarities carry no special meaning in themselves and are considered only in relation to the others. However, the general rule is that a number close to 1 means the compared words have very strong synonymy (1 is the highest possible number when the directions of the two word vectors are identical and the angle between them equals 0°), and outputs close to 0 (meaning the angle between the vectors is close to 90°) and below 0 (when the vector has an opposite direction) generally signal the absence of connection or even negative correlation.

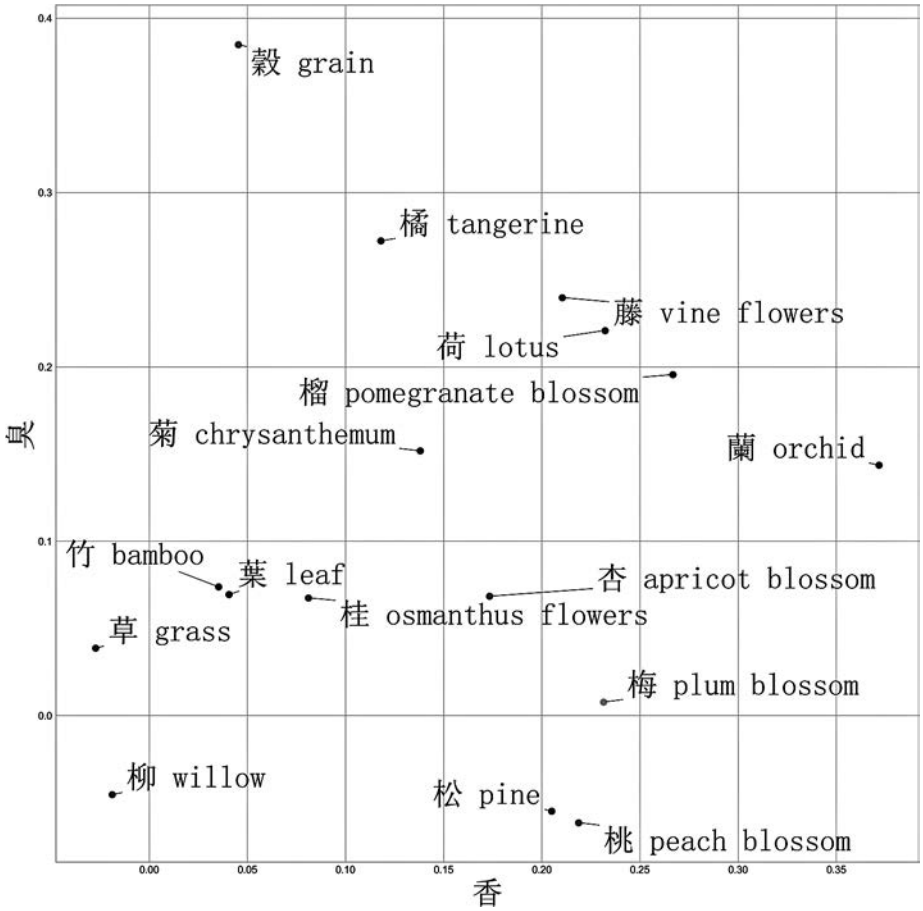

Figure 1. Plants in relation to unpleasant smell (*chou* 臭) and fragrance (*xiang* 香)

Some of the most interesting observations are those suggesting how various objects relate to emotions. Figures 2 and 3 show how animate objects and natural phenomena from the list of objects relate to *happiness* (*xi* 喜) and *worries* (*chou* 愁) (chosen over *sorrow/sadness* [*bei* 悲] because it is much more frequent in the corpus and for the most part shows stronger correlations). In both cases, there is a clear division between groups of images associated with emotions.

The connections the algorithm is able to find depend not just on the meaning of the images. Tiger, for example, is one of the few animals in the poetic language model with a strong correlation with happiness and none with sorrow. This contradicts known depictions of the animal in Tang poetry. Here, for example, is one of the poems on tigers by Du Fu, selected by the *YWSX* editors to represent the topic:

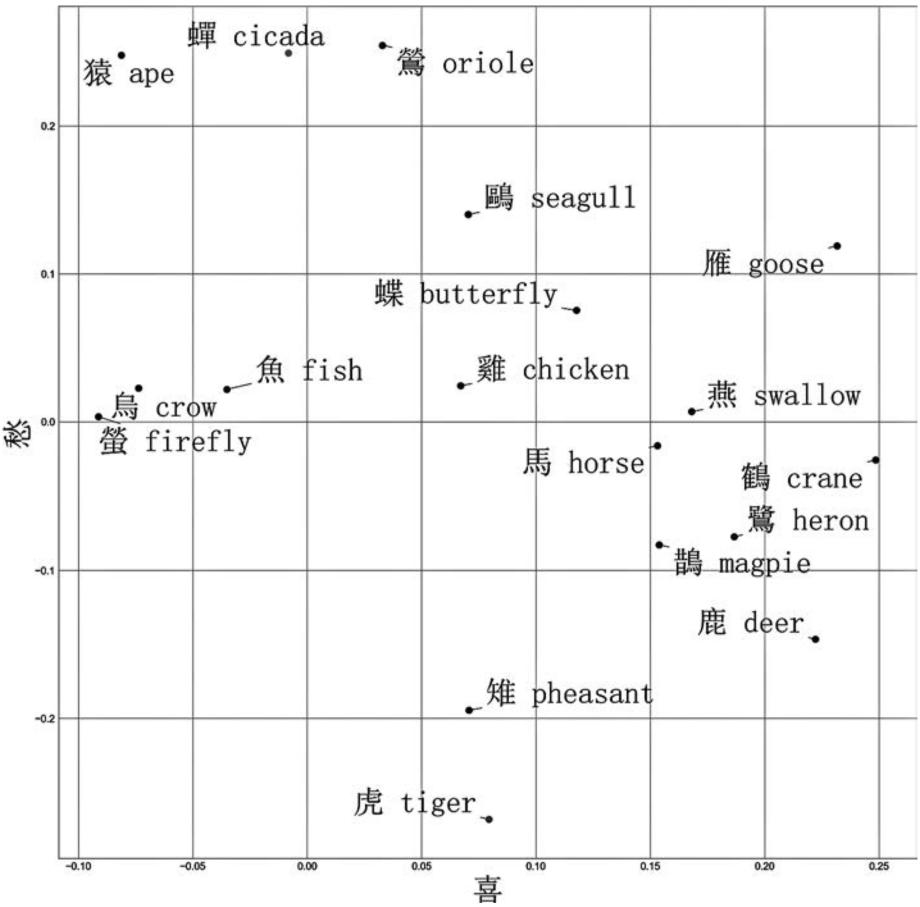

Figure 2. Animate objects in relation to happiness (*xi* 喜) and worries (*chou* 愁)

	Getting Out What Stirred Me	遣興
	Du Fu	杜甫
	The fierce tiger depends upon threatening power	猛虎憑其威
2	and always gets roped up tight.	往往遭急縛
	With a thunderous howl it roars in vain,	雷吼徒咆哮
4	the wooden braces are already on its paws.	枝撐已在腳
	All at once you look at its pelt spread as bedding,	忽看皮寢處
6	the flash of its pupils is no more.	無復睛閃爍
	For people it is worse than this,	人有甚於斯
8	which is enough to warn the most evil men.	足以勸元惡
		[*QTS* 218.2291; trans. Owen, *Poetry of Du Fu*, 2:131]

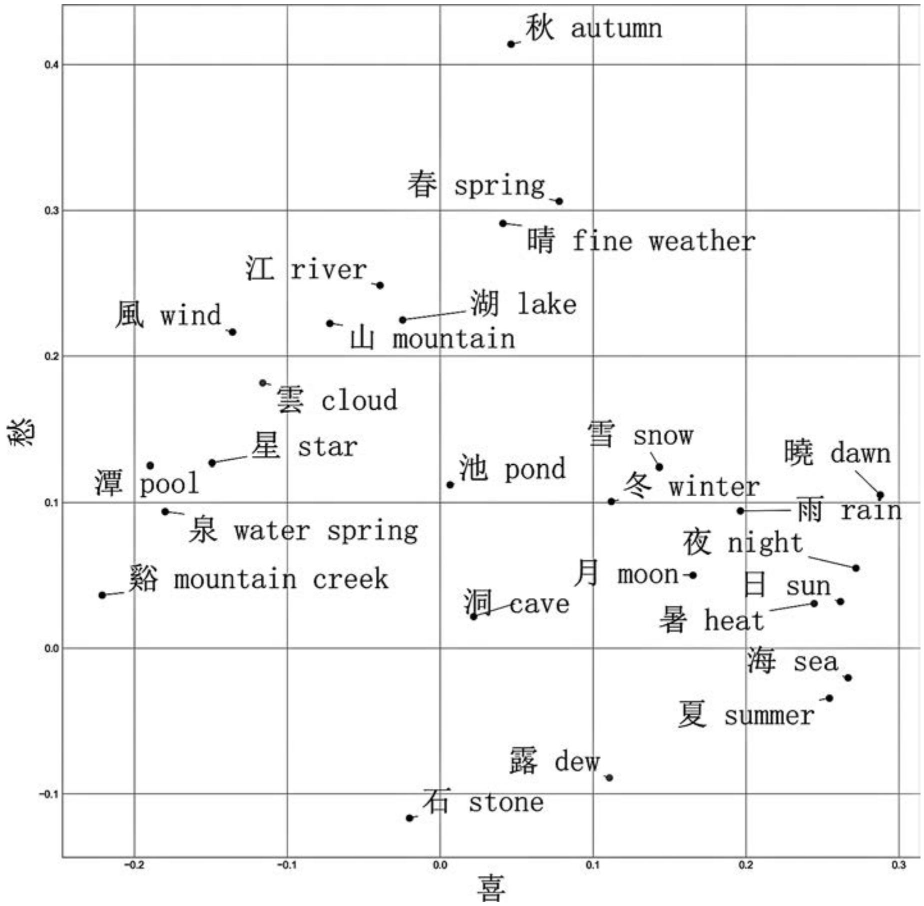

Figure 3. Nature phenomena related to happiness and worries

Here, although the tiger ultimately loses to men, he is described as a mighty foe, fierce and powerful, with a roar like thunder and finally compared to criminals and evil men.

Numerous poems on tigers follow the same mood, with the exception of those where the animal symbolizes celestial phenomena, such as the Orion constellation or the third of the Twelve Earthly Branches, part of the sexagenary ordering system that could be also associated with zodiac. The fixed epithet for the animal, often appearing even in titles, is *fierce* (*meng* 猛). A particularly vivid image is created by Dugu Ji 獨孤及 (725–77) in "A Song on a Painting of Shooting a Tiger Made with Li Shangshu" 和李尚書畫射虎圖歌 (He Li Shangshu hua she hu tu ge):

	A hungry tiger, mouth wide open, stands blocking the road,	飢虎呀呀立當路
2	Ten thousand men shake in fear, one hundred animals are in rage.	萬夫震恐百獸怒

[*QTS* 247.2770]

Given the examples, one can see that tigers did not bring joy, so the correlation with happiness is either a mistake of the model or a result of relationships more complex than simple word co-occurrences. The most probable explanation for this phenomenon is the different linguistic modes governing poetic expression, namely, according to the terminology of Tsu-lin Mei and Yu-kung Kao, imagistic and propositional language.[26] These two modes utilize different lexicons and grammar, as imagistic language is more concerned with a static picture and uses many noun phrases, whereas propositional language makes assertions. While Mei and Kao were describing language differences within individual poems, the same distinction can be applied more widely to distinguish between lyrical and narrative poetry.[27]

It may be argued that this distinction is the reason behind the strange correlation: poetry about sorrow is predominantly lyrical and describes static images, while happiness is more associated with action—just like tigers generally appear in narratives with vivid scenes and many events. On the other hand, fish, apes, and cicadas appear in contemplative scenes that are full of thought, in relatively static settings. Deer, cranes, and magpies are common signs of happiness, and egrets were admired for their appearance and could be used as a metaphor for courtiers,[28] so their closeness to happiness on the plot raises no questions.

The plot with natural phenomena shows that most water sources, with the exception of the sea, were typically places for melancholy feelings. This confirms Stephen Owen's observation on a group of poems presenting a "meditation on the past," where a popular rhyme containing the words *autumn* (*qiu* 秋), *to worry* (*chou* 愁), *seagull* (*ou* 鷗), *boat* (*zhou* 舟), *tower* (*lou* 樓), and *stream* (*liu* 流) sets the stage for such meditations. The images are not bound to appear only at the lines' end,[29] as we see, for example, in the following poem by Liu Changqing 劉長卿 (?–ca. 786):

Berthing Late at the Xiangjiang River and Missing Close Friends	晚泊湘江懷故人
Liu Changqing	劉長卿
Far in the sky clouds are floating away,	天涯片雲去

2 Pointing to the imperial city they stir memories. 遙指帝鄉憶
Melancholy feelings grow with the old age, 惆悵增暮情
4 Xiaoxiang river reflects the autumn colors. 瀟湘復秋色
At what place did the small boat moor? 扁舟宿何處
6 The setting sun longs to recover its wings. 落日羨歸翼
For ten thousand miles around there are no close friends, 萬里無故人
8 The river gulls do not recognize each other 江鷗不相識
[*QTS* 148.1522]

This poem is a meditation not, strictly speaking, on the grand past but, rather, on the past of the poet himself. However, it contains all the imagery required for the sad scene, and it reflects the results on the graph: the season, the gulls, and the water on which one can travel by boat are present and connected to sadness. This confirms that it is indeed a common setting.

Another interesting observation that can be made is the balance between being in solitude versus being in companionship. Writing verse was often a social activity: poems would be composed to commemorate an event, to thank a host for an invitation, to give a compliment, and for many other reasons. But despite the social function of much of classical Chinese poetry and the abundance of pieces that might even indicate their role quite straightforwardly by explaining the occasion in the title, it seems that the contents often have a very different mood. One of the ways to see just how social the verses are is to analyze which paradigms the words belong to more often, solitude or companionship. Figure 4 shows how all of the selected objects are distributed between these concepts. Taking into account the asymmetry between the axes, one discovers that most of the things that appear in Tang poetry are associated with being alone; that is, they are used in the context of being lonely. It doesn't necessarily reflect the number of poems with such mood in the corpus, but it does show that solitude, with more words related to the paradigm, could be expressed in more ways and settings.

Tang Dynasty Poems on Things in the *YWSX* and Their Interrelations

The previous sections dealt with objects as they appear in poems and their relation to different moods. The next step is to move beyond the objects themselves and consider the poetry about them. Although the poems in the *YWSX* are topically grouped according to the object to which each work is dedicated, the assignment of a poem to a topic seems to obey no formal criteria and depend much more on editor judgment. Thus, for example, "Reading Classic of *Mountains and Waterways*" 讀山海經 (Du *Shanghai jing*) by Tao Yuanming

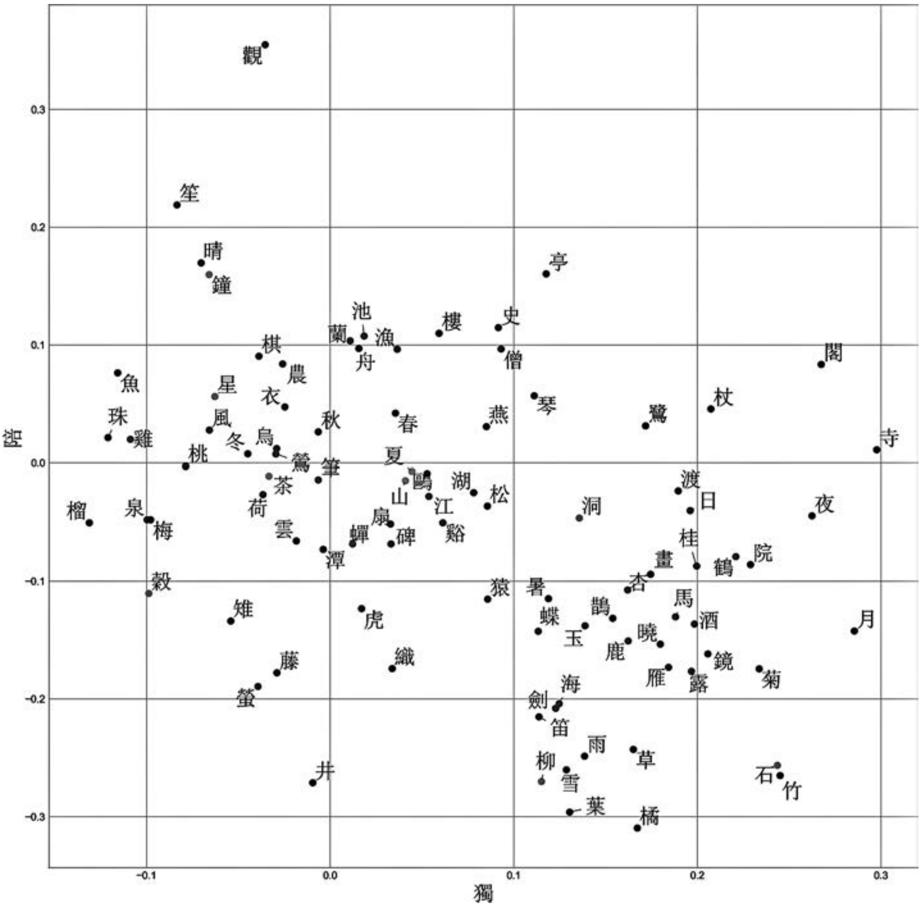

Figure 4. All objects related to being lonely (*du* 獨) and with company (*pei* 陪)

陶淵明 (365?–427) is split among different chapters: the first of the thirteen stanzas is included in the “reading books category” (*dushu lei* 讀書類), the second in the “eight immortals category” (*baxian lei* 八仙類), and the fourth in the “all kinds of trees category” (*zongshu lei* 總樹類), while other stanzas are omitted.

In many cases a poem could theoretically fall under one topical category but was assigned to another, as with “Climbing Qingshan on the Ninth Day” 九日登青山 (Jiuri deng Qingshan) by Zhu Wan 朱灣 (8th century): the *YWSX* has chapters for both mountains and *jiuri* 九日, the “ninth day of the ninth month,” when the “Double Nine Festival” (*chongyang jie* 重陽節) was celebrated. In this case, the poem was classified in the “mountains category” (*zhongshan lei* 衆山類), but since traveling to the mountains was a popular part of the festival, many poems with similar titles appear in the section dedicated to the festival itself. Among them, for example, "Accompanying Runzhou Governor Shao on the Ninth Day to Climb the Beigushan Mountain" 九日陪潤州邵使君登北固山 (Jiuri pei Runzhou Shaoshijun deng Beigu shan) by Zhang Zirong 張子容 (8th century) or "Accompanying Director Cui Dong Feasting on the Beishan Mountain" 九日陪崔峒郎中北山宴 (Jiuri pei Cui Dong langzhong Beishan an) by Yan Wei 嚴維 (?–780).

Another illustrative example are the poems titled "Spring Frost" (*chunhan* 春寒). There are seven poems from different dynasties with this title in the *YWSX*, and although all of them describe cold spells in spring, two of them (by poets Wen Tong 文同 [1018–79] and Wang Yue 王越 [1426–99])[30] were attributed to the group "frost category" (*han lei* 寒類), and two (by Xu Ning 徐凝 [9th century] and Huang Geng 黃庚 [13th century]) to spring (*chun lei* 春類); a poem by the Ming poet Tan Shao 郯韶 (14th century) is found in the category "fragrance" (*xiang lei* 香類), and two pieces by the Ming poet Zhang Yuchu 張宇初 (?–1410) and the Yuan poet Fang Hui 方囘 (1227–1307) are placed in the "dove" (*jiu lei* 鳩類) and "geese" (*e lei* 鵞類) categories, respectively.

The vocabulary of the poem itself is also not constitutive in defining the topic. Consider, for instance, several Tang poems in pentasyllabic regulated verse (*wuyan lü* 五言律) that mention the moon. Although the topics seem to be clear in these selections, there is a lot of overlap in imagery:

Spring Mountains in the Moonlight 春山夜月
Yu Liangshi (8th century) 于良史

Spring mountains are full of beauty and joy, 春山多勝事
2 Delighted I forget to return for the night. 賞翫夜忘歸
I scoop the water and the moon is in my hand, 掬水月在手
4 I play with flowers and aroma fills my clothes. 弄花香滿衣
When the inspiration comes, there is no near or far, 興来無遠近
6 About to leave, I cherish the fragrant verdure. 欲去惜芳菲
On looking south, to where a bell tolls, 南望鳴鐘處
8 The tower is deep in blue mist. 樓臺深翠微
[*QTS* 275.3118]

Moon Round 月圓
Du Fu 杜甫

Lonely moon, full, facing the upper story, 孤月當樓滿
2 the cold river stirs on my door by night. 寒江動夜扉
Cast into waves, golden light unsettled, 委波金不定

4 shining on the mat, figured work even more faint. 照席綺逾依
Not yet waning, empty mountains serene, 未缺空山靜
6 suspended on high, the constellations sparse. 高懸列宿稀
In the gardens of home pine and cassia flourish, 故園松桂發
8 I share its clear glow with them, thousands of miles away. 萬里共清輝

[*QTS* 230.2525; trans. Owen, *Poetry of Du Fu*, 4:330]

New Moon 初月
Du Fu 杜甫

Its light so thin, how could it be half-full? — 光細弦豈上
2 Rays oblique, the orb not yet steady. 影斜輪未安
Faintly ascending beyond ancient passes, 微升古塞外
4 Already hidden by twilight clouds' edge. 已隱暮雲端
The Star River does not change its color, 河漢不改色
6 The barrier mountains are cold on their own. 關山空自寒
There is white dew in the front yard, 庭前有白露
8 In darkness filling the chrysanthemums. 暗滿菊花團

[*QTS* 225.2421; trans. Owen, *Poetry of Du Fu*, 2:168]

The first of the poems belongs to the category "mountains"; the other two fall into the "moon" category. Yet the vocabulary of all three is very similar: mention of the moon brings a whole cluster of further associations, determined by the rules of poetic creation and cultural concepts. The first word to appear, quite logically, is *night*. Then, it becomes likely that some kind of water will be mentioned—be it water in general or a particular river. The rules of parallelism favor the introduction of mountains with the water. Then, a popular cliché would include the autumn moon, cold weather with accompanying snow reflecting the white moon in the sky. Flowers can also appear in such poems as a contrast to the moon and, if chrysanthemums, as a symbol of autumn. Taking into account that the space in many forms of classical poems is very limited and that there are rules partially dictating composition and plot development within a verse, connecting these images into a coherent text can occupy most of the poem, such that the main difference between verses often becomes the way these images are connected to each other.

It is clear, then, that the poems are not sorted into categories according to a formal principle predicted by the vector model; the same vocabulary and imagery can be used to depict different situations, and poems describing very similar situations can end up in different groups, depending on the editors' judgment as to what is most prominent and important in each piece. With this in mind, it is interesting to see if the word embedding model, which uses only the contents of the texts, is able to distinguish between poems that belong to one topic and those that do not. This should indicate whether the selection of Tang poems from the *YWSX* lends itself to a meaningful sorting.

To perform such sorting, it is necessary to go beyond the semantics of separate words and create mathematical representations of whole poems based on their contents. For this task, so-called semantic fingerprints were used.[31] There are many different understandings and usages of the term *semantic fingerprints*, but in this article I use it to refer to a representation of the semantics of a document based on word embeddings. There are two common ways to calculate a document's fingerprint. The first is to combine the vectors of all the words in it. This will create a vector representation that takes into account not only word meanings but also the length of the document. The second way is to calculate an average of the vectors of all the words in the document. This approach was chosen over the first, as it concentrates more on semantics and disregards structural differences between texts. The method of creating the fingerprints, despite its seeming simplicity, outperforms other, more complicated methods of defining groups.[32]

For the purposes of this article, a fingerprint for each piece was calculated as an average of all its word vectors; thus, the length of a poem did not influence its semantic representation.[33] When looking at separate poems, the fingerprint comparison seems to some extent to align with the decisions of the *YWSX* editors. Taking the example of the poems mentioned above, those placed in the same category show higher cosine similarities: 0.735 between "Accompanying Runzhou Governor Shao on the Ninth Day Climb the Beigushan Mountain" and "Accompanying Director Cui Dong Feast on the Beishan Mountain," 0.677 for the latter compared with "Climbing Qingshan on the Ninth Day," and 0.592 between "Climbing Qingshan on the Ninth Day" and "Accompanying Runzhou Governor Shao on the Ninth Day Climb the Beigushan Mountain." Similar results were achieved when comparing poems containing the moon and mountains, with the highest similarity score (0.798) appearing between the two poems by Du Fu. However, when asked to present the works most similar to a given piece, the model more often puts poems from other topics in first place. This means that, although the system recognizes the semantic similarity between poems assigned to the same topics by the editors, it often decides that poems from other topics are closer in meaning to the target one. Although the instrument used has limited power, this phenomenon may indicate that there is no direct and formal connection between the contents of a piece and its topic and substantiates the assumption that the grouping could have been the result of the editors' wish to cover a wider variety of topics, not strictly the contents of the poems.

Conclusion

This article has shown a new way of exploring the imagery of Tang poetry using "poems on things" and the "objects" that were popular topics for such poems. Through the employment of distributional semantics and neural networks, one can study semantics of separate words, their groups, or whole poems, as well as their relations and the groups these entities can form. From the point of view of traditional methodologies, this may be considered a study of paradigms: groups formed by separate images or verses have been discussed, as have the similarities and differences between groupings made by humans versus those created by an algorithm. The article has shown that, although the objects of Tang poetry often have clear correlations with different types of states and emotions, such as sadness and happiness or solitude and companionship, the semantic fingerprints of the poems about them do not fall into clear and straightforward groups. Instead, they show that poems with similar contents and imagery were often assigned to different topics with seemingly no formal principles behind the editors' choices.

The results are not conclusive, as many technical details, such as part-of-speech tagging or increasing of the corpus size, could improve the overall accuracy. However, it contributes to the discussion of Chinese classical poetic language and global patterns or cross-topical tones in poetry and ideally sheds some light on the possibilities to come with the implementation of computational methods.

MARIANA ZORKINA 左曉艷
University of Zurich
mariana.zorkina@aoi.uzh.ch

Acknowledgment

The research was funded by Swiss National Science Foundation grant 100015_163032, "The Imagery of Tang 'Poems on Things' (*yǒngwùshī* 詠物詩): A Semiotic Approach."

Notes

[1] Firth, *Synopsis of Linguistic Theory*, 11.

[2] Wittgenstein, *Philosophical Investigations*, 2–3.

[3] Harris, *Distributional Structure.*

[4] Osgood, Suci, and Tannenbaum, *Measurement of Meaning*, 6–7.

[5] Mikolov et al., “Distributed Representations”; Mikolov et al., “Efficient Estimation.”

[6] Mikolov et al., “Efficient Estimation,” 4.

[7] The settings in this study were CBOW, dimensionality 500, window 5, minimum word frequency 2, negative sampling, 5 iterations.

[8] See Zorkina, “Appendix: Technical Details,” www.chinesepoetryforum.org/?page_id=1512.

[9] Barthes, *Fashion System.*

[10] Jakobson, “Two Aspects of Language.”

[11] The applicability of Jakobson’s theory to classical Chinese poetry was discussed in Kao and Mei, “Meaning, Metaphor and Allusion.”

[12] Liu et al., “‘*Quantangshi*’ de fenxi”; Liu et al., “Color Aesthetics.”

[13] Owen, *Late Tang*, 192.

[14] Culler, *Structuralist Poetics*, 14.

[15] For explaining various phenomena, see, e.g., Slingerland et al., “Distant Reading of Religious Texts,” 7.

[16] An article on *yongwu shi* states that there are exactly 14,590 poems (Knechtges and Chang, *Ancient and Early Medieval Chinese Literature*, 3:1959). According to automatic counts of my corpus, there are 13,812 to 15,871 poems, depending on whether pieces in poem cycles are counted as separate or one entity.

[17] Both texts were obtained from the Chinese Text Project website: ctext.org.

[18] For a more detailed description of the corpora and the training process of the algorithm, see Zorkina, “Appendix: Technical Details.”

[19] Lee and Wong, “Glimpses of Ancient China,” 623.

[20] The model can be accessed at projector.tensorflow.org/?config=https://dl.dropboxusercontent.com/s/i5xyyq99anyiogn/QTS_embeddings.json?dl=0.

[21] The full list can be accessed at Zorkina, “Selection of Objects,” uofi.app.box.com/s/pz3wvsssja2s7hkk4seqg2ww6prn15ay.

[22] Mikolov et al., “Efficient Estimation,” 5–8.

[23] Mikolov, Yih, and Zweig, “Linguistic Regularities,” 747.

[24] Stephen Owen (*Late Tang*, 461) claims that the reason behind the association is the similar pronunciation of the words *to stay* (*liu* 留) and *willow* (*liu* 柳). This may, however, be questioned, as the characters have different tones (Pulleyblank, *Lexicon of Reconstructed Pronunciation*, 197) and thus belong to different rhyme groups.

[25] Borges, *Other Inquisitions*, 103.

[26] Kao and Mei, “Syntax, Diction, and Imagery,” 58. This topic was studied from the quantitative point of view in Lee, Kong, and Luo, “Syntactic Patterns.”

[27] Levy, *Chinese Narrative Poetry*, 26.

[28] Cai, *How to Read Chinese Poetry*, 30.

[29] Owen, *The Late Tang*, 192.

[30] The *YWSX* attributes Wang Yue to the Song dynasty (960–1279).

[31] Semantic fingerprints as presented in Kutuzov et al., “Clustering Comparable Corpora,” 1. For a short overview and an example of different uses, see Han et al., “Semantic Fingerprints-Based Author Name Disambiguation,” 1883.

[32] Kutuzov et al., “Clustering Comparable Corpora,” 5.

[33] The model the fingerprints of Tang poems in *YWSX* can be accessed at: projector.tensorflow.org/?config=https://dl.dropboxusercontent.com/s/g6boq5qaifckend/YWSX_fingerprints.json?dl=0.

References

Allen, Colin, Hongliang Luo, Jaimie Murdock, Jianghuai Pu, Xiaohong Wang, Yanjie Zhai, and Kun Zhao. "Topic Modeling the Hàn diǎn Ancient Classics (汉典古籍)." *CA: Journal of Cultural Analytics* (2017). doi:10.22148/16.016.

Barthes, Roland. *The Fashion System*. Translated by Matthew Ward and Richard Howard. New York: Hill and Wang, 1983.

Borges, Jorge Luis. *Other Inquisitions, 1937–1952*. Translated by Ruth L. C. Simms. Austin: University of Texas Press, 1964.

Cai, Zong-qi, ed. *How to Read Chinese Poetry: A Guided Anthology.* New York: Columbia University Press, 2007.

Culler, Jonathan. *Structuralist Poetics: Structuralism, Linguistics, and the Study of Literature*. London: Routledge and Kegan Paul, 1975.

Firth, John R. "A Synopsis of Linguistic Theory 1930–1955." In *Studies in Linguistic Analysis*, edited by John R. Firth, 1–32. Oxford: Blackwell, 1957.

Han, Hongqi, Changqing Yao, Yuan Fu, Yongsheng Yu, Yunliang Zhang, and Shuo Xu. "Semantic Fingerprints-Based Author Name Disambiguation in Chinese Documents." *Scientometrics* 111, no. 3 (2017): 1879–96.

Harris, Zellig S. "Distributional Structure." *Word* 10, no. 23 (1954): 146–62.

Jakobson, Roman. "Two Aspects of Language and Two Types of Aphasic Disturbances." In Jakobson and Halle, *Fundamentals of Language*, 69–96. The Hague: Mouton, 1980.

Kao Yu-kung and Mei Tsu-lin. "Meaning, Metaphor, and Allusion in T'ang Poetry." *Harvard Journal of Asiatic Studies* 38, no. 2 (1978): 281–356.

———. "Syntax, Diction, and Imagery in T'ang Poetry." *Harvard Journal of Asiatic Studies* 31 (1971): 49–136.

Knechtges, David R., and Taiping Chang, eds. *Ancient and Early Medieval Chinese Literature: A Reference Guide*. 4 vols. Leiden: Brill, 2010.

Kutuzov, Andrey, Mikhail Kopotev, Tatyana Sviridenko, and Lyubov Ivanova. "Clustering Comparable Corpora of Russian and Ukrainian Academic Texts: Word Embeddings and Semantic Fingerprints." Preprint. arXiv:1604.05372v1 [cs.CL], 2016.

Lee, John, and Wong Tak-sum. "Glimpses of Ancient China from Classical Chinese Poems." Poster in *Proceedings of the 24th International Conference on Computational Linguistics*, edited by Kay, Martin, and Christian Boitet, 621–32. Mumbai, 2012. http://www.aclweb.org/anthology/C12-2061

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. doi:10.1093/llc/fqw059.

Levy, Dore. *Chinese Narrative Poetry: The Late Han through T'ang Dynasties.* Durham, NC: Duke University Press, 1988.

Levy, Omer, Yoav Goldberg, and Ido Dagan. "Improving Distributional Similarity with Lessons Learned from Word Embeddings." *Transactions of the Association for Computational Linguistics*, no. 3 (2015): 211–25.

Liu, Chao-Lin 劉昭麟, Chun-Ning Chang 張淳甯, Chu-Ting Hsu 許筑婷, Wen-Huei Cheng 鄭文惠, Hongsu Wang 王宏甦, and Wei-Yuan Chiu 邱偉雲. "'*Quantangshi*' de fenxi, tankan yu yingyong—fengge, duizhang, shehuiwanglu yu duilian" 《全唐詩》的分析、探勘與應用－風格、對仗、社會網路與對聯 (Textual Analysis of Complete Tang Poems for Discoveries and Applications—Style, Antitheses, Social Networks, and Couplets). *ROCLING XXVII*, edited by Sin-Horng Chen, Hsin-Min Wang, Jen-Tzung Chien, Hung-

Yu Kao, Wen-Whei, Chang, Yih-Ru Wang, and Shih-Hung Wu, 43–57. Hsinchu: Association for Computational Linguistics and Chinese Language Processing (ACLCLP) 2015. http://aclweb.org/anthology/O15-1.

Liu, Chao-Lin, Hongsu Wang, Wen-Huei Cheng, Chu-Ting Hsu, and Wei-Yun Chiu. "Color Aesthetics and Social Networks in *Complete Tang Poems*: Explorations and Discoveries." In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, edited by Hai Zhao, 132–41. Shanghai, 2015. http://aclweb.org/anthology/Y15-2

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient Estimation of Word Representations in Vector Space." Preprint. arXiv:1301.3781v3 [cs.CL], 2013.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. "Distributed Representations of Words and Phrases and Their Compositionality." *Advances in Neural Information Processing Systems* 26 (2013): 3111–19.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. "Linguistic Regularities in Continuous Space Word Representations." In *Proceedings of NAACL-HLT*, edited by Lucy Vanderwende, Hal Daumé III, Katrin Kirchhoff, 746–75. Atlanta: Association for Computational Linguistics, 2013. http://aclweb.org/anthology/N13-1

Osgood, Charles E., George J. Suci, and Percy H. Tannenbaum. *The Measurement of Meaning*. Urbana: University of Illinois Press, 1967.

Owen, Stephen. *The Late Tang: Chinese Poetry of the Mid-Ninth Century (827–860)*. Cambridge, MA: Harvard University Press, 2009.

———. trans. *The Poetry of Du Fu*. 6 vols. Berlin: De Gruyter, 2015.

Peiwen zhai yongwushi xuan 佩文齋詠物詩選 (Collection of Poems on Things from the Peiwen Study). In *Wenyuange Siku quanshu* 文淵閣四庫全書 (Literary Profundity Pavilion Edition of the Complete Books of the Four Treasuries), edited by Yang Ne 楊訥 and Li Xiaoming 李曉明. Taipei: Shangwu yinshuguan, 1983–86.

Pulleyblank, Edwin G. *Lexicon of Reconstructed Pronunciation in Early Middle Chinese, Late Middle Chinese, and Early Mandarin*. Vancouver: UBC Press, 1991.

Quan Tang shi 全唐詩 (Complete Poems of the Tang). Compiled by Peng Dingqiu 彭定求. Beijing: Zhonghua shuju, 1960.

Řehůřek, Radim, and Petr Sojka. "Software Framework for Topic Modelling with Large Corpora." In *Proceedings of the LREC 2010 workshop New Challenges for NLP Frameworks*, edited by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, 46–50. Valletta: University of Malta, 2010.

Slingerland, Edward, Ryan Nichols, Kristoffer Neilbo, and Carson Logan. "The Distant Reading of Religious Texts: A 'Big Data' Approach to Mind-Body Concepts in Early China." *Journal of the American Academy of Religion* 85, no. 4 (2017): 985–1016. doi:10.1093/jaarel/lfw090.

Sturgeon, Donald. "Unsupervised Identification of Text Reuse in Early Chinese Literature." *Digital Scholarship in the Humanities* 33, no. 3 (2017): 670–84. doi:10.1093/llc/fqx024.

Wittgenstein, Ludwig. *Philosophical Investigations*. Translated by G. E. M. Anscombe. Oxford: Blackwell, 1953.