Proceedings of the Twenty-Ninth International
Florida Artificial Intelligence Research Society Conference

Semantic Parallelism in Classical Chinese Poems

John S. Y. Lee
Halliday Centre for Intelligent Applications of Language Studies
Department of Linguistics and Translation, City University of Hong Kong
jsylee@cityu.edu.hk

Abstract

We present a quantitative study on semantic parallelism in the *Complete Tang Poems*. Chinese poems consist of a sequence of couplets, which contain an identical number of characters. In a semantically "parallel" couplet, the first line in the couplet should correspond in meaning to the second. In this paper, we evaluate the extent to which a well-known semantic taxonomy (Wang, 2003) covers the phenomenon of parallelism. We then analyze which categories in this taxonomy are most frequently observed in parallel couplets.

Introduction

Classical Chinese poems consist of a sequence of couplets – pairs of adjacent lines, each with an identical number of characters, most frequently five or seven. Table 1 shows the couplet that forms the first two lines of a well-known poem from the 8th century CE. In this couplet, both lines consist of five characters.

Parallelism is a common literary device in couplet composition. In a parallel couplet, the first line in the couplet must mirror the second line syntactically and semantically; that is, the two lines must bear similar or opposite meaning, and have comparable grammatical constructions. In particular, characters occupying the same position on the two lines – these shall henceforth be referred to as *character pairs* – should have the same part-of-speech and have related meaning. For example, in Table 1, the first characters in the two lines are both colors: *bai* 'white' and *huang* 'yellow', respectively; the second character pair – *ri* 'sun' and *he* 'river' – are both objects in the natural world; the third character pair contains both verbs; and so on.

"Coupling" has been described as a universal principle behind poetic structure (Levin, 1962)[1]. It is one of the defining features of Chinese poetry, and there has been long-standing and intense intellectual interest in this feature. During the Tang Dynasty (7th to 9th century CE), rules for poem composition were formalized. Later, various semantic taxonomies were proposed to describe the phenomenon of parallelism. The book *Weiwendi shige* proposed a semantic taxonomy which classifies nouns into 8 categories. In the 19th century, a more fine-grained, 37-category semantic taxonomy was proposed in the book *Shiyunhebi*. In current scholarship, the 24-category semantic taxonomy developed by Wang (2003) is among the most well known.

Despite this long intellectual history, there has been no empirical evaluation on these taxonomies. This paper presents the first large-scale, quantitative evaluation of semantic parallelism in Chinese poems. We evaluate the extent to which Wang's (2003) taxonomy covers this phenomenon of parallelism, and then identify the semantic categories in this taxonomy that are most frequently observed in parallel couplets.

Second line	First line
黄 *huang* 'yellow'	白 *bai* 'white'
河 *he* 'river'	日 *ri* 'sun'
入 *ru* 'enter'	依 *yi* 'rest'
海 *hai* 'ocean'	山 *shan* 'mountain'
流 *liu* 'flow'	盡 *jin* 'extinguish'
Yellow River enters sea — and flows on.	White sun rests on mountains — and is gone;

Table 1. A couplet whose lines have 5 characters each[2].

Data

Textual material

The Tang Dynasty is considered the golden age of poems. The *Quantangshi* (Peng, 1960), or *Complete Tang Poems*,

[1] Aside from Chinese poetry, this principle is also exhibited in many other languages, for example in classical Hebrew poetry (Kugel, 1981).

[2] This couplet forms the beginning of the poem entitled "Climbing Crane Tower", by Wang Zhihuan. We follow the traditional format, where a line is read from top to bottom, and lines are arranged from right to left. The English translations are taken from (Cai, 2008).

originally compiled in 1705, consists of nearly 50,000 poems, totaling 210K couplets and about 2.8 million characters, by more than two thousand poets. We extracted all couplets from this body of poems and the character pairs therein.

Semantic taxonomy

To evaluate the semantic closeness of automatically mined character pairs, we need a database of semantically classified characters. WordNet-like resources for Chinese (Gan, 2000; Choi, 2004; Chen et al., 2002; Xu et al., 2008) have focused on Modern Chinese rather than Classical Chinese. A recently constructed treebank (Lee & Kong, 2012) contains parts-of-speech tags and dependency relations, but no semantic information. We will utilize a semantic taxonomy proposed by Wang (2003), which has 24 categories:

- *Celestial*: heavenly bodies and other phenomena in the sky, e.g., 日 'sun', 月 'moon', 風 'wind'.
- *Seasonal*: terms for periods of time, e.g., 夜 'night', 春 'spring', 年 'year'.
- *Geographic*: geographic entities on earth, including 山 'mountain', 海 'sea', 江 'river'.
- *Architectural*: a type of building, or a component thereof, e.g., 殿 'palace', 樓 'building'.
- *Instruments*: a broad topic including tools, utensils, vehicles, household objects, weapons, etc., e.g., 劍 'sword', 琴 'violin'
- *Clothing:* e.g., 帶 'belt', 環 'ring'
- *Food*: e.g., 茶 'tea', 糕 'cake'
- *Products of civilization*: objects associated with artistic pursuits, including stationary tools (e.g., 筆 'pen', 紙 'paper') and musical instruments (e.g., 琴 'lute')
- *Literary*: Terms related to literature, e.g., 書 'book', 詩 'poem', 信 'letter'.
- *Flora*: Plants, e.g., 草 'grass', 花 'flower', 柳 'willow'.
- *Fauna*: Animals, e.g., 鳥 'bird', 魚 'fish', 龍 'dragon'.
- *Body parts*: Parts of the human body (e.g., 眼 'eye') and things produced by the body (e.g., 影 'shadow', 聲 'voice').
- *Human emotions*: Human activities (e.g., 歌 'song', 舞 'dance') and sentiments (e.g., 意 'desire', 心 'heart')
- *Human relations*: kinship, e.g., 兄, titles (e.g., 相 'minister') and professions (e.g., 兵 'soldier', 農 'farmer')
- *Pronouns*: e.g., 我 'I', 爾 'you'
- *Locations*: e.g., 北 'north', 中 'middle', 外 'outside'
- *Numbers*: e.g., 三 'three', 雙 'pair'
- *Colors*: e.g., 紅 'red'
- *Coordinates*: words used in a numbering system related to the calendar, e.g., 甲 'one'
- *Adverbs*: e.g., 亦 'also', 不 'not'
- *Conjunctions*: e.g., 和 'and' 共 'together'
- *Particles*: Usually placed at end of a phrase for emphasis, with no concrete meaning, e.g., 也 *ye*
- *Personal names*: e.g., 張 'Zhang', 王 'Wang'.
- *Place names*: proper names for locations, e.g., 楚 'Chu', 吳 'Wu'

It has been recognized that strict parallelism, if blindly followed (e.g., 'east' should always match 'west'), severely limits the space for creativity. Recognizing this, Wang (2003) also posits the concept of "related match", with 13 pairs of related categories, including, for example, 'Flora' and 'Fauna', and 'Place names' and 'Personal names'.

Based on these categories, two degrees of parallelism are differentiated. The first is "exact match", where two characters belong to the same category. The second is "related match", where two characters belong to two related categories. Our study is based on the example characters provided by Wang (2003), which account for 78% of the types in our corpus[3].

Analysis

We first evaluate the semantic taxonomy proposed by Wang (2003), and then report the semantic categories in this taxonomy that frequently participate in parallelism.

Semantic taxonomy evaluation

We will evaluate the taxonomy on the character pairs with the most statistically significant association. Mutual information is the standard metric for measuring the degree of association between two words (Church and Hanks, 1990). For any two characters *x* and *y*, their pointwise mutual information (PMI) is defined as $\log[P(x,y)/P(x)P(y)]$, where $P(x)$ and $P(y)$ are the probabilities of the characters x and y, and $P(x,y)$ is the probability of the co-occurrence of *x* and *y*.

Most previous research has examined co-occurrence as n-grams. For example, based on similarities in a window of *n* neighboring words, each word can be assigned to a class or category (Brown et al., 1992; Li & Abe, 1996). The linguistic context has also been expanded to syntactic

[3] When a character may belong to multiple categories, word sense disambiguation is needed to properly classify it. Since this task is still beyond the state-of-the-art for classical Chinese processing, we used the character's most common meaning in this study.

information (Lin, 1998). In our context, we define co-occurrence to be identical positions on two lines in a poem.

1	南北 'north, south'	6	雨風 'rain, wind'
2	山水 'hill, water'	7	風月 'wind, moon'
3	雲月 'cloud, moon'	8	水雲 'water, cloud'
4	玉金 'jade, gold'	9	西東 'west, east'
5	青白 'green,white'	10	東北 'east, north'

Table 2. The ten character pairs with the highest pointwise mutual information.

Table 2 shows the character pairs with the highest PMI. As shown in Table 3, among the top 25 character pairs, 76% exhibit "exact match" according to Wang's (2003) taxonomy. The figure drops below 50% when the top 500 are considered. The figures for "Related" match, as expected, are considerably higher. It reaches 88% for the top 25 character pairs, and is still close to 60% for the top 500.

A number of unrelated character pairs have notably high PMI. Often, these pairs may not seem related on the surface by their literal meaning, but are evocative of certain events or situations. The highest pair is 酒 *jiu* 'wine' and 詩 *shi* 'poem', where 'wine' is classed as 'food' and 'poem' as 'civilization'. This pair in fact reflects a tradition of Chinese poets. Wine, then as now, is seen as a catalyst for inspirations to write and interpret poems, and is a central theme for many prominent poets. At banquets in Ancient China, poets would drink while playing the lute and compose poems or articles. It is no surprise that two other unrelated pairs, *jiu* and 琴 *qin* 'lute' as well as *jiu* and 書 *shu* 'book', also have high PMI.

Several unrelated pairs with high PMI belong to the categories 'celestial' and 'flora', such as 樹 *shu* 'tree' and 雲 *yun* 'cloud'. Others represent a kind of meronymy, such as the pairs 樹 *shu* 'tree' and 山 *shan* 'mountain', and 鳥 *niao* 'bird' and 雲 yun 'cloud'. In both of these cases, a natural object ('hill' and 'cloud') is paired with a typical object in its realm ('tree' and 'bird').

Top-N pairs	Exact match	Related match
25	76%	88%
50	64%	80%
100	69%	74%
500	46%	58%

Table 3. The percentage of the top character pairs, as determined by pointwise mutual information, that exhibit exact or related match

Frequent semantic categories

Having evaluated the extent to which Wang's taxonomy (2003) covers the phenomenon of parallelism, we now identify those categories in the taxonomy that most frequently participate.

Exact match

Table 4 shows the number of character pairs that belong to each semantic category in exact match. Three of the categories, 'Place names', 'Personal names', 'Literary references', do not occur. It is perhaps not unexpected that categories involving proper names tend to be the most difficult to match.

By absolute counts, 'Geographic' is the most frequent category in the corpus. The literary device called *xing* (興) is widely recognized in Classical Chinese poems. With this device, poets paint an atmosphere and mood to trigger the subsequent expression of feelings and emotions, often through descriptions of time and environment (Yuan, 1990). 'Geographic', together with six other more frequent categories, ('Celestial', 'Color', 'Locations', 'Fauna', 'Seasonal', 'Flora'), likely reflect the use of this device.

The absolute counts, however, do not tell how likely a noun in a particular category participates in exact match. To normalize these counts, we divide them by the number of character pairs in which at least one character belong to the category. The categories 'Numbers', 'Colors' and 'Locations' have the highest percentage (Table 4). In general, categories with a narrower range of meaning, such as 'Numbers' and 'Colors', are more likely to yield exact matches. Conversely, these categories may be expected to be less likely to participate in related matches, since it is more difficult to balance a character from such a category with one from outside the category. This is indeed the case for 'Numbers', 'Colors' and 'Locations', all of which have relatively low rate of related matches (Table 5).

Category	Count	Percentage
Numbers	10980	40.0%
Colors	9226	36.4%
Locations	8077	30.4%
Fauna	4235	22.3%
Geographic	12926	18.5%
Seasonal	8875	18.4%
Celestial	10002	16.5%
Flora	5063	16.3%
Body parts	3821	14.3%
Human emotions	5287	13.5%
Architectural	2594	11.1%
Instruments	2845	10.4%
Human relationships	2224	9.6%
Products of civilization	915	6.1%
Clothing	370	5.6%
Food	137	2.5%
Coordinates	89	1.6%

Table 4. Frequency of exact match[4]

[4] 'Percentage' is the frequency count divided by the number of character pairs in which at least one character belong to that semantic category.

Related match

Use of the *xing* device is manifested not only in exact matches, but also in related matches. Among the six most frequent semantic categories, all but 'Body parts – Human Emotions' are bound up with time and environment (Table 5). Again, two category-pairs involving proper names, 'Human relationships - Personal names' and 'Personal names - Place names', do not occur.

The most frequent character pairs for some related match deserve some comments. In the 'Body parts - Human emotions' related match, the most frequent pair 'heart/thought' is consistent with the ancient Chinese belief that the heart, rather than the brain, was the organ for thinking and feeling. As for the 'Instruments - Products of civilization' related match, one would expect the instrument to be one for writing literary work, such as a writing brush or an ink slab; instead, 'sword' comes out on top, perhaps reflecting the respect for martial arts among the elites.

Related match	Count	Most frequent pair
Celestial - Geographic	11553	雲/水 'cloud/water'
Celestial - Seasonal	10306	雪/春 'snow/spring'
Geographic – Architectural	6147	路/門 'road/door'
Body parts - Human emotions	4756	心/思 'heart/thought'
Architectural - Instruments	2171	寺/船 'temple/boat'
Flora- Fauna	1868	鳥/花 'bird/flower'
Locations - Numbers	1576	中/一 'middle/one'
Instruments - Products of civilization	1512	劍/書 'sword/book'
Numbers - Colors	1472	一/青 'one/green'
Instruments - Clothing	1148	劍/衣 'sword/clothes'
Clothing - Food	178	酒/衣 'wine/clothes'

Table 5. Frequency of related match

Conclusion

We have presented a quantitative study on semantic parallelism in the *Complete Tang Poems*. First, we retrieved the character pairs that are most statistically significantly associated. An evaluation shows that a well-known semantic taxonomy (Wang, 2003) covers most of these character pairs. Nonetheless, we identified how a number of frequent pairs that are deemed "unparallel" to be related in other aspects. Finally, we investigated how often the individual categories in this taxonomy exhibit parallelism. Our statistics suggest that those with narrow semantic ranges tend to have higher rates for exact match, but lower rates for related match.

Acknowledgment

The work described in this paper was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 11606515).

References

Brown, P., deSouza, P., Mercer, R., Della Pietra, V., and Lai, J. 1992. Class-based n-gram models of natural language. *Computational Linguistics* 18(4).

Cai, Z.-Q. 2008. *How to Read Chinese Poetry*. New York: Columbia University Press.

Chen, H.-H., Lin, C.-C., and Lin, W.-C.. 2002. Building a Chinese-English WordNet for Translingual Applications. *ACM Transactions on Asian Language Information Processing* 1(2):103—122.

Choi, K.-S., Bae, H.-S., Kang, W., Lee, J., Kim, E., Kim, H., Song, Y., and Shin, H. 2004. Korean-Chinese-Japanese Multilingual WordNet with Shared Semantic Hierarchy. *Proc*. *LREC*.

Church, K. and Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1).

Gan, K. W., and Wong, P. W. 2000. Annotating Information Structures in Chinese Texts using HowNet. *Proc*. *2nd Workshop on Chinese Language Processing*.

Li, H. and Abe., N. 1996. Clustering words with the MDL principle. *Proc*. *COLING*.

Lin, D. 1998. Automatic retrieval and clustering of similar words. *Proc*. *ACL-COLING*.

Kugel, J. L. 1981. *The Idea of Biblical Poetry: Parallelism and Its History*. New Haven and London: Yale University Press.

Levin, S. R. 1962. *Linguistic Structures in Poetry*. The Hague: Mouton.

Lee, J. and Kong, Y. H. 2012. A Dependency Treebank of Classical Chinese Poems. *Proc*. *NAACL*.

Peng, D. 1960. *Quantangshi* **全唐詩** [The complete Shi poetry of the Tang]. Beijing: Zhonghua shuju.

Xu, R., Gao, Z., Pan, Y., Qu, Y., and Huang, Z. 2008. An Integrated Approach for Automatic Construction of Bilingual Chinese-English WordNet. *Proc*. *3rd Asian Semantic Web Conference*.

Wang, L. 2003. *Hanyu shiluxue* 漢語詩律學 [The metric of Chinese poems]. Hong Kong: Zhonghua shuju.

Yuan, X. 袁行霈. 1990. Zhongguo wenxue gailun 中國文學概論 [The introduction of Chinese literature]. Beijing: Higher Education Press.