

Adapting NLP and Corpus Analysis Techniques to Structured Imagery Analysis in Classical Chinese Poetry

Alex Chengyu Fang
Dept of Chinese, Translation
and Linguistics
City University of Hong Kong
Kowloon, Hong Kong SAR
acfang@cityu.edu.hk

Fengju Lo
Dept of Chinese Language
and Literature
Yuan Ze University
Taoyuan, Taiwan
gefjulo@saturn.yzu.edu.tw

Cheuk Kit Chinn
Dept of Chinese, Translation
and Linguistics
City University of Hong Kong
Kowloon, Hong Kong SAR
manigo@cityu.edu.hk

Abstract

This paper describes some pioneering work as a joint research project between City University of Hong Kong and Yuan Ze University in Taiwan to adapt language resources and technologies in order to set up a computational framework for the study of the creative language employed in classical Chinese poetry.

In particular, it will first of all describe an existing ontology of imageries found in poems written during the Tang and the Song dynasties (7th–14th century AD). It will then propose the augmentation of such imageries into primary, complex, extended and textual imageries. A rationale of such a structured approach is that while poets may use a common dichotomy of primary imageries, creative language use is to be found in the creation of complex and compound imageries. This approach will not only support analysis of inter-poets stylistic similarities and differences but will also effectively reveal intra-poet stylistic characteristics. This article will then describe a syntactic parser designed to produce parse trees that will eventually enable the automatic identification of possible imageries and their subsequent structural analysis and classification. Finally, a case study will be presented that investigated the syntactic properties found in two lyrics written by two stylistically different lyric writers in the Song Dynasty.

Keywords

Imagery, ontology, parsing, lyrics, poetry, Chinese, stylistic analysis, Su Shi, Liu Yong

1. Introduction

The language of poetry is different from that employed in other categories of writing. “Defined from a linguistic perspective, poetry represents a variant form of language, different from speech and common writing, unique in its own way as a linguistic system.” (Yuan 1989:2 [12]). The difference of poetic language from other types of writing typically exists in its intentionally polysemous readings through the creative use of imageries as part of the poet’s

artistic conception. Classical Chinese poetry, because of its formal restrictions in terms of syllables, tonal variations and rhyming patterns, commands a language system that appears to be particularly concise, finely rich, highly rhetorical, and thus linguistically complex, requiring a high degree of creativity in writing it, sophisticated interpretation in reading it and often a significant level of difficulty in understanding it.

This article addresses the issue of machine-aided analysis and understanding of classical Chinese poetry in general and attempts to establish a computational framework (cf. Figure 1) within which both inter- and intra-poet stylistic differences and similarities can be usefully investigated with firm grounding in textual analysis from a linguistic perspective. In particular, we believe that the creative language of poetry can be effectively investigated through its manipulation of imageries and through the range of linguistic devices that help to achieve the poetic articulation of the intended artistic ambience.

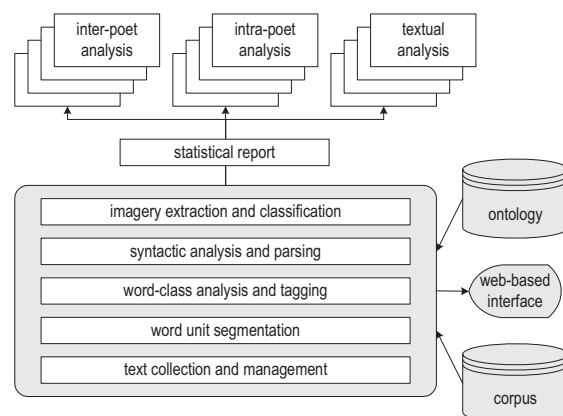


Figure 1. A computational framework for the computer-aided analysis and understanding of classical Chinese poems.

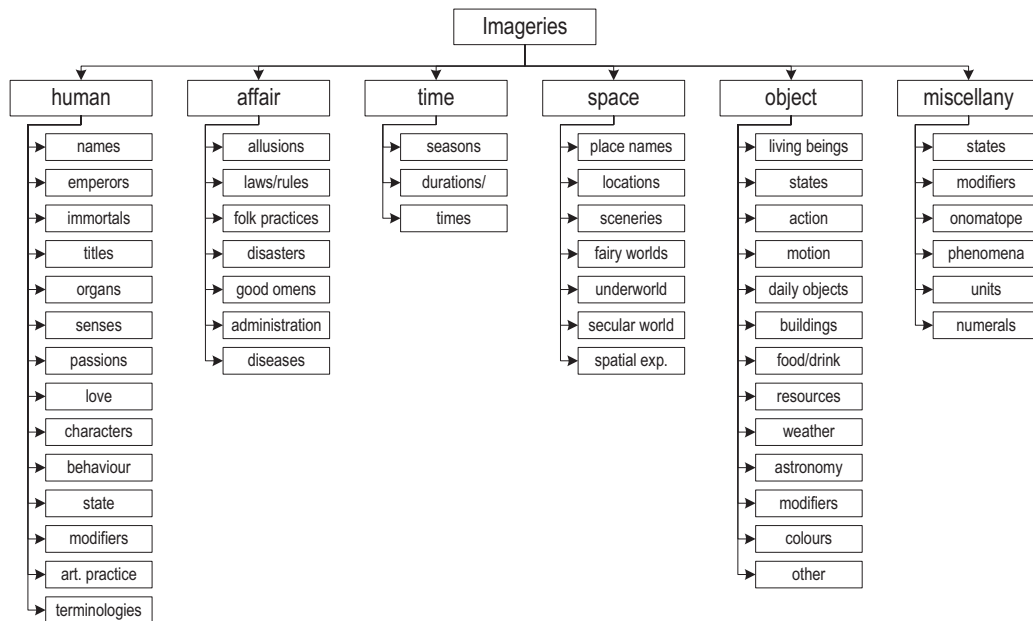


Figure 2. An ontology of imageries in classical Chinese poems

This article thus will first of all describe an ontology of imageries that has been created for the poems written during the Tang and the Song dynasties, ranging from the seventh century AD to the thirteenth century AD. It will then propose a new, structured approach towards the extraction and classification of imageries, according to which poetic imageries can be categorised into primary, complex, extended and textual sub-types. Since the automatic processing of imageries in this fashion requires syntactic analysis, we shall then move on to the description of a syntactic parser that provides a structured description of the syntactic constituents for classical Chinese poetry. We finally present a preliminary syntactic analysis of two contemporary poets from the Song dynasty in support of a syntax-based approach to the processing and understanding of classical Chinese poetry.

2. An Ontology of Imageries

Lo (2008 [8]) describes an ontology of imageries designed for the study of classical Chinese poetry. The complete collection of the poems from the Tang dynasty was processed at the lexical level. The collection comprises 51,170 poems by 2,821 poets totaling 4,767,979 characters. Individual characters were segmented into meaningful word units (WUs) before WUs were indexed according to their semantic class. A synset was created for synonymous WUs and each synset was described by a keyword. For example, “一年” (one year) is the key word for the following five variant synonymous WUs forming the synset:

一年	一載、一歲、一春、一年、一秋
----	----------------

Six classes are constructed: human, affair, time, space, object, and miscellany. See Figure 2. Each item in Figure 2 is categorised further into subcategories which will eventually include the actual words and expressions found in a poem. Human/characters, for example, is subdivided into positive moral characters and negative moral characters. As another example, object/astronomy is subdivided into sun, moon, star, sky, etc. Thus, in addition to the six pandects, the system notes 54 subclasses with a further 372 subdivisions before reaching the terminal classes comprising the actual WUs indexed from the complete collection of poems. The system is now available from <http://cls.hs.yzu.edu.tw/tang/Database/index.html>.

As an example, a search for imageries involving the use of the concept “冬” (winter) in the seasons category would yield 221 records, representing 214 lines from 193 poems by 107 poets. They contain 31 different WUs. Table 1 shows a list of such expressions sorted according to frequency in descending order.

Table 1. Poetic expressions found in the Complete Collection of Tang Poems involving the imagery *winter*.

Frequency	Chinese	English
32	三冬	three winters
22	冬春	winter spring
22	嚴冬	harsh winter
20	窮冬	impoverished winter
20	經冬	enduring winter
13	無冬	no winter
6	冬夏	winter summer
6	冬雪	winter snow
6	清冬	clear winter
5	冬天	winter day
5	冬盡	winter end
5	冬衣	winter clothes
5	去冬	previous winter
4	一冬	one winter
4	冬來	winter arriving
4	冬寒	winter cold
4	冬景	winter scene
4	凌冬	cold winter
4	過冬	spend winter
3	冬令	winter season
3	冬秋	winter autumn
3	寒冬	cold winter
2	入冬	entering winter
2	冬去	winter departing
2	冬深	winter deep
2	初冬	early winter
2	見冬	showing winter
1	冬初	winter start
1	早冬	early winter
1	殘冬	remnant winter
1	逢冬	meeting winter

3. A Structured Approach to the Analysis of Imageries

In this article, we propose a more structured approach to imageries than what was described in the previous section. While the ontology remains more or less sufficient for descriptions of classical Chinese poems, the imageries themselves need to be reprocessed to reveal their inner structures. For instance, the imageries involving “冬” (winter) in Table 1 can be analysed into the following according to the role of *winter* within the phrase structure. Consider Table 2.

Table 2. A structured analysis of winter imageries.

$Function_{winter}$	$Function_{collocate}$	Collocate	F
head	modifier	嚴, 窮, 清, 去, 凌, 寒, 初, 殘, 早	64
head	determiner	三, 無, 一	42
complement	verbal	經, 過, 見, 深, 去, 入, 逢	33
modifier	head	雪, 衣, 天, 景, 寒, 令, 初	28
head	coordination	春, 秋	25
subject	verbal	盡, 來	9

Table 2 has four columns. $Function_{winter}$ shows the phrase internal function of *winter*, the imagery in question. $Function_{collocate}$ indicates the phrase internal function of the collocates co-occurring with *winter*. *Collocate* lists the actual collocates and F the number of occurrence of $Function_{winter}$. The rows are arranged according to F in descending order. As can be seen, of the 216 poetic expressions in the Complete Collection of Tang Poems involving the imagery winter, there are seven types of structural analysis, of which modifier+headwinter is the most frequent, occurring 64 times. This structure also has nine different types of instantiations of the modifier as collocates with winter, namely, “嚴” (harsh), “窮” (impoverished), “清” (clear), “去” (previous), “凌” (cold), “寒” (cold), “初” (early), “殘” (remnant), and “早” (early). It is apparent that modifier+headwinter is not only the most significant in frequential terms but also in terms of the variety of its collocates.

It is thus evident that the poetic expressions involving winter could have a more structured and therefore refined representation than what is currently available. We thus propose to distinguish the following types of imageries: primary, complex, extended, and textual.

Primary imageries refer to those head nouns that may have an imagery potential. Winter, for example, is a primary imagery.

Complex imageries refer to primary imageries that have either a premodifier or a determiner. By this definition, “嚴冬” (harsh winter, modifier + head) is a complex imagery and so is “三冬” (three winters, determiner + head).

Extended imageries are defined to include those complex imageries that either serve clausal functions with

an overt subject-verb-object structure or with other syntactic constructs that function as predicates.¹

Finally, textual imageries are represented in the poem as a system of extended imageries, carefully intended and designed by the poet as part of the artistic conception and articulation.

The four types of imageries thus correspond to four levels of linguistic analysis schematised in Figure 3.



Figure 3. A schematized correspondence between structured imagery analysis and different levels of linguistic analysis

The neat correspondence between structured imagery analysis and different levels of linguistic analysis shows that linguistic analysis can be deployed as a stepping stone between poems as raw texts and an ontology of structured imageries derived from the nominal groups. Computation can be performed on the expressions of imageries according to the clausal structure to derive extended imageries. If necessary, techniques for textual analysis can be applied to extended imageries to represent the raw text as a system of interrelated extended imageries.

The structured approach towards the analysis of imageries described in this section will have two immediate applications. The first has to do with the automatic extraction, analysis and classification of imageries, which practically means that ontology generation can be fully automated. The second application lies in the actual analysis of classical Chinese poems. The idea that there is an intrinsic structure within imageries makes it possible to stratify and hence better analyse imageries from lexical, grammatical, syntactic and textual perspectives (cf Figure 3). In the analysis of two poets for inter-poet stylistic differences and similarities, it might be possible that the two poets both make use of a similar set of primary imageries measured in terms of lexical use and semantic grouping. Their creative use of language, which marks them as two different poets or even two distinctive stylistic schools, comes from the creative manipulation of such primary imageries by way of complex imageries, extended imageries and textual imageries.

¹ In contemporary as well as classical Chinese, grammatical categories can be used interchangeably. The syntactic function of a predicate, for instance, can be performed variously by a verb phrase, an adjective phrase, or often a prepositional phrase. See Section 5 for additional information.

4. A Syntactic Parser for Classical Chinese Poems

The transfer of nominals to a structured representation of imageries thus requires a syntactic parser of classical Chinese poems. This section describes a parser that represents perhaps the very first effort to analyse poetic lines in a syntactic way.

The current version of the parser is driven by a phrase structure grammar (PSG) for the generation of syntactic trees. Written in Java, it takes two input files, one as a collection of PSG rules and the other containing a poetic line where each component character is tagged with a part-of-speech symbol.² It produces all of the possible syntactic analysis for the poetic line permissible by the grammar. Consider the following line from a song lyric (詞, *Ci*) written by Liu Yong (柳永) in the Song Dynasty.

殘日下，漁人鳴榔歸去。

(Under the setting sun, fisher men bang on the boat and leave for home.)

Each character in the above text is POS tagged to yield the following input text where a POS tag is assigned and associated with the character by an underscore:

殘_adj 日_n 下_marker , _punc 漁_n 人_n 鳴
_v 榔_n 歸去_v 。 _punc

A PSG grammar is written in the following manner:

```
S -> PP NP VP
AJP -> adj
NP -> NP n
NP -> AJP n
NP -> n
PP -> NP marker punc
VP -> VP VP
VP -> v NP
VP -> v punc
```

The parser then produces a syntactic tree shown in Figure 4.

² The parser is adapted from an earlier version written by Mr Norman Goalby as part of his MSc thesis (Goalby 2004 [6]) supervised by the first author of this article while lecturing at the Computer Science Department, University College London.

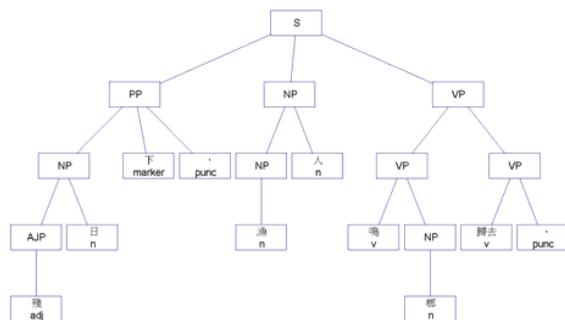


Figure 4: The syntactic tree automatically produced by the parser for the poetic line “殘日下，漁人鳴榔歸去。”.

Such a tree structure would allow the identification of a set of useful units summarized in Table 3.

Table 3. Useful imagery units extracted from the parse tree in Figure 4.

Imagery Units	Lexical Units
NPs	殘日 (remnant sun), 漁人 (fisher men), 榔 (boat)
Primary imageries	日 (sun), 漁 (fisher), 人 (men), 榔 (boat)
Complex imageries	殘日 (setting sun), 漁人 (fisher men)
Extended imageries	漁人鳴榔 (fisher men bang on the boat)

Of course, as is also evident from the sample parse tree, syntactic functions are inferred implicitly. The pre-VP NP is interpreted as the subject of the sentence and the post-VP NP as the direct object. The pre-head NP “漁” (fisher) is interpreted as an NP that functions as a premodifier of the head “人” (men) because of its position within the NP proper. If two nouns are juxtaposed or coordinated within the same NP with or without an overt coordinator, the first conjoin will not be promoted to a NP in its own right. It is thus possible to distinguish between NP as a premodifier and a noun as part of a coordinated construction.

While syntactic relations can be implicitly inferred without much problem for this particular example, we envisage the further development of the current parser into one that not only produces a tree structure labelled by phrasal categories such as NP and VP but will also explicitly indicate the syntactic functions of such phrasal categories such as subject and direct object. A good example can be found in the Survey Parser, which was used to complete the one-million-word International Corpus of English (Fang 1996 [2] and 2000 [53]; Greenbaum 1996 [7]) with a grammar of fine granularity (Fang 2005 [4]).

5. A Syntactic Study of Two Lyrics Written in the Song Dynasty

As was mentioned at the end of Section 4, the creative use of language by two poets, which marks them as two different poets or even as belonging to two distinctive stylistic schools, comes from the creative manipulation of primary imageries by way of complex imageries, extended imageries and textual imageries. Naturally, a range of linguistic devices, within the paradigm of syntax, will have to be exploited in order to achieve the intended poetic articulation. For this matter, it is also of great benefit to perform syntactic analysis for extractions of the syntactic relations preferred or typically used by individual poets. This section describes a case study.

Two lyrics (詞, *Ci*) were selected. They are *Yong Yu Yue* (永遇樂. See Appendix A for original text in Chinese) and *Ye Ban Yue* (夜半樂. See Appendix B for original text in Chinese). The two lyrics were written in the Song Dynasty respectively by Su Shi (蘇軾) and Liu Yong (柳永) as representatives of the two stylistically distinctive schools, namely, the *Hao-fang* School (豪放派) and the *Wan-yue* School (婉約派).³ Su Shi was selected also for a second reason: he also wrote substantially in the wan-yue style. It is thus possible to formulate a comparative framework whereby Shu Shi and Liu Yong can be compared in the first instance for their differences and similarities in the manipulation of imageries as representatives of two contrastive stylistic schools. The same procedure can be applied to Shu Shi as a single poet for differences and similarities in his use of imageries across the two stylistically different groups of poems.

Table 4 summarises the two lyrics in terms of tokens, types, and type-token ratios (TTR).

Table 4. Basic statistics about the two lyrics.

Poet	Title	Token	Type	TTR
Su Shi	<i>Yong Yu Le</i>	127	87	68.5%
Liu Yong	<i>Ye Ban Le</i>	167	130	77.8%

The TTR for Su Shi is 68.5% while the TTR for Liu Yong is 77.8%, a difference of nearly 10%. The difference seems to indicate different degrees of lexical density but since the two poems have a different number of characters, interpretations here remain inconclusive. Our focus of study will be described in Sections 5.1 and 5.2.

³ To quote Owen (1996: 582 [9]), “[t]he ‘masculine’ style was called *hao-fang*, loosely translated as ‘bold and extravagant’; the ‘feminine’ style was called *wan-yue*, something like ‘having a delicate sensibility’.” The two terms are also translated as *the heroic style* and *the devious-evocative* in Yeh (1998 [10]).

5.1 A grammatical analysis of the two lyrics

Both lyrics were segmented and duly analysed by hand at grammatical, phrasal and clausal levels according to a fine-grained formal grammar noting both category types and syntactic functions. Table 5 is a summary of the two lyrics in terms of the four open classes, namely, adjectives, adverbs, nouns and verbs. Column F lists the raw frequency of occurrence for the classes and Column % the proportion of the four classes in all the word tokens.

Table 5. Summary statistics about the two lyrics' use of adjectives, adverbs, nouns and verbs.

POS	Su Shi		Liu Yong	
	F	%	F	%
Adjectives	21	16.5	20	11.9
Adverbs	9	7.1	17	6.6
Noun	41	32.3	55	32.9
Verbs	13	10.2	21	16.2

Grammatically speaking, according to Table 5, the two lyric writers seem to be similar in their use of nouns, which account for 32.3% and 32.9% respectively for the two. Adverbs also seem to show a good degree of similarity between the two poets.

Significant differences between the two writers arise from their use of verbs and adjectives. Liu Yong employs a much higher proportion of verbs, which accounts for 16.2% of his total use of the word tokens. Su Shi, in contrast, has a much smaller proportion of verbs of only 10.2%, 6% lower than the other writer. Regarding adjectives, Su Shi has a more frequent use of adjectives (16.5%) than Liu Yong, whose adjective use has a lower proportion of 11.9%. That Liu Yong has a relative higher use of verbs but lower use of adjectives seems to suggest that this poet is perhaps stylistically more colloquial and more intended for freer vocal rendition by those who liked him. The relatively lower verb proportion and higher adjective use by Su Shi, on the other hand, seems to relate his writing style to one that is more intended for the reading eye and thus more formal, densely expressed and hence scholarly.⁴

5.2 A syntactic analysis of the two lyrics

Syntactic properties for the two poets are summarised in Table 6, which lists six syntactic functions: subject, predicate, object, premodifer, adverbial and complement. Each property is described by its raw frequency F and its associated proportion % for the two poets.

Table 6. Summary statistics about the two lyrics' use of syntactic constructions.

Function	Su Shi		Liu Yong	
	F	%	F	%
Subject	15	11.8	15	8.90
Predicate	19	14.9	31	18.6
Object	4	3.1	1	0.59
Premodifer	18	14.2	30	17.9
Adverbial	16	12.6	17	10.2
Complement	1	0.79	0	0.0

According to Table 6, Su Shi outperforms Liu Yong in his use of subjects, objects, adverbials and complement. Conversely, Liu Yong outperforms Su Shi in his use of predicates and premodifiers. The pattern of differences here as evidenced by Table 6 seems to suggest that Su Shi deploys a wider range of syntactic devices than Liu Yong, which is also supported by our observations based on grammatical properties summarised in Table 5. Again, a higher proportion of predicates in Liu Yong's writing seems to suggest a more casual, colloquial style while the lower proportion of predicates by Su Shi is accompanied by a relatively higher proportion of objects, thus indicating a preference for SVO constructions, a feature typically found in formal prose. Su Shi's more frequent use of adverbials also seems to confirm this as adverbials have been found to correlate with degrees of formality albeit reported for contemporary British writing. See Fang (2006 [5]) for a detailed report of the empirical study.

That Liu Yong's works tend to be more casual and colloquial is also pointed out by literary critics. As Yeh (2000:1-12 [11]) points out, Liu Yong was a popular song writer at the time and 'every common person by the community water well could sing his songs'. Liu Yong as a lyric writer is innovative in that he abandoned the poetic convention and adopted a more life-like tone and voice in the description and presentation of women in a more realistic style (ibid [11]).

Although only a case study based on very limited evidence, the analysis results presented in this section already speak strongly and favourably for contrastive stylistic analysis of different poets within a framework where linguistic properties at grammatical and syntactic levels can be usefully retrieved and computed. The authors believe that such a computational framework represents a powerful instrument in the automatic analysis of literary texts in general and classical Chinese poems in particular.

6. Conclusion

This paper described a computational framework for the analysis of classical Chinese poems. In particular, it presented an ontology of imageries that has been empirically generated from the complete collection of

⁴ This is partially evidenced by a recent study reported in Cao and Fang (2009 [1]), which reveals that adjectives tend to occur more frequently in formal academic writing than in informal casual speech though the primary data for the analysis comes from contemporary English.

poems written in the Tang Dynasty. The ontology is now accessible on the Internet and has been a major instrument for the analysis of imageries.

The article then argued for a structured analysis of imageries and proposed a system of imageries sub-categorised as primary, complex, extended and textual. Such a system usefully relates linguistic analysis at lexical, grammatical, syntactic and textual levels to the analysis and evaluation of imageries.

We then described an automatic parser of classical Chinese poems and demonstrated that linguistic analysis could be effectively automated to enable the automatic generation of structured imageries for poetic studies.

Finally, a case study was described that investigated the grammatical and syntactic properties in two lyric poems written by two different poets in the Song Dynasty. The study demonstrated that the two poets as seen in the two lyrics made use of different linguistic devices and that a frequential account of such devices seemed to support literary as well as linguistic interpretations of the two distinctive composing styles.

We conclude by reiterating that such a computational framework represents a useful and insightful instrument in the automatic analysis of literary texts in general and classical Chinese poems in particular.

7. Acknowledgments

Research reported in this article was supported in part by research grants (Nos 7002190, 7200120 and 7002190) from the City University of Hong Kong. The authors acknowledge the support received from Prof Shi-wen Yu (俞士汶) at the Institute of Computational Linguistics, Peking University, PR China. The authors also thank Dr Chunshen Zhu at City University of Hong Kong for comments on an earlier draft of the article.

8. References

- [1] J. Cao and Alex C. Fang. 2009. Investigating Variations in Adjective Use across Different Text Categories. In *Advances in Computational Linguistics, Journal of Research In Computing Science Vol 41*. pp 207-216.
- [2] A.C. Fang. 1996. The Survey Parser: Design and Development. In *Comparing English World Wide: The International Corpus of English*, ed. by Sidney Greenbaum. Oxford: Oxford University Press. pp 142-160.
- [3] A.C. Fang. 2000. From Cases to Rules and Vice Versa: Robust Practical Parsing with Analogy. In *Proceedings of the Sixth International Workshop on Parsing Technologies*, 23-25 February 2000, Trento, Italy. pp 77-88.
- [4] A.C. Fang. 2005. Evaluating the Performance of the Survey Parser with the NIST Scheme. In *Lecture Notes in Computer Science 3878: Computational Linguistics and Intelligent Text Processing*, ed. by A. Gelbukh. Berlin Heidelberg: Springer-Verlag. pp. 168-179.
- [5] A.C. Fang. 2006. A Corpus-Based Empirical Account of Adverbial Clauses across Speech and Writing in Contemporary British English. In *LNCS 4139/2006: Advances in Natural Language Processing*. Berlin and Heidelberg: Springer. pp 32-43.
- [6] N. Goalby. 2004. *Using Instance-Based Learning to Supplement Early Parsing*. MSc thesis, Computer Science Department, University College London.
- [7] S. Greenbaum. 1996. *Comparing English World Wide: The International Corpus of English*. Oxford: Oxford University Press.
- [8] F. Lo. 2008. The Research of Building a Semantic Category System Based on the Language Characteristic of Chinese Poetry. In *Proceedings of the 9th Cross-Strait Symposium on Library Information Science*, 3-6 July 2008, Wuhan University, China.
- [9] S. Owen. 1996. *An Anthology of Chinese Literature: Beginnings to 1911*. New York and London: W.W. Norton & Company, Inc.
- [10] C.-Y. Yeh. 1998. Ambiguity in the Song Lyric: The Influence of the Female Voice in Huajian Songs. In *Studies of Chinese Poetry*, ed. by James Hightower and Chia-ying Yeh. Harvard: Harvard University Asian Center.
- [11] C.-Y. Yeh. 2000. *Liu Yong and Zhou Bang-yan*. Taiwan: Daan Press.
- [12] X.-P. Yuan. 1989. *Research in the Art of Poetry*. Taiwan: Wunan Book Publishing Co.

Appendix A. Yong Yu Yue by Su Shi

蘇軾《永遇樂》

徐州夜夢覺，北登燕子樓作。

明月如霜，
好風如水，
清景無限。
曲港跳魚，
圓荷瀉露，
寂寞無人見。
絃如三鼓，
鏗然一葉，
黯黯夢雲驚斷。
夜茫茫，
重尋無處，
覺來小園行遍。

天涯倦客，
山中歸路，
望斷故園心眼。
燕子樓空，
佳人何在，

空鎖樓中燕。
古今如夢，
何曾夢覺，
但有舊歡新怨。
異時對、
黃樓夜景，
為余浩歎。

Appendix B. *Ye Ban Yue* by Liu Yong

柳永《夜半樂》

凍雲黯淡天氣，
扁舟一葉，
乘興離江渚。
渡萬壑千巖，
越溪深處。
怒濤漸息，
樵風乍起，
更聞商旅相呼。
片帆高舉。
泛畫鷁、
翩翩過南浦。
望中酒旆閃閃，
一簇煙村，
數行霜樹。
殘日下，
漁人鳴榔歸去。
敗荷零落，
衰楊掩映，
岸邊兩兩三三，
浣沙遊女。
避行客、
含羞笑相語。

到此因念，
繡閣輕拋，
浪萍難駐。
歎後約丁寧竟何據。
慘離懷，
空恨歲晚歸期阻。
凝淚眼、
杳杳神京路。
斷鴻聲遠長天暮。