Tradition and Modernity in 20th Century Chinese Poetry

Rob Voigt
Center for East Asian Studies
Stanford University
robvoigt@stanford.edu

Dan Jurafsky
Linguistics Department
Stanford University
jurafsky@stanford.edu

Abstract

Scholars of Chinese literature note that China's tumultuous literary history in the 20th century centered around the uncomfortable tensions between tradition and modernity. In this corpus study, we develop and automatically extract three features to show that the classical character of Chinese poetry decreased across the century. We also find that Taiwan poets constitute a surprising exception to the trend, demonstrating an unusually strong connection to classical diction in their work as late as the '50s and '60s.

1 Introduction

For virtually all of Chinese history through the fall of the Qing Dynasty, poetry was largely written in Classical Chinese and accessible to a small, educated fraction of the population. With the rise of the May Fourth Movement in 1919, prominent intellectuals such as Hu Shi and Lu Xun began to advocate for and produce a fresh vernacular literature.

This upheaval of tradition has been much discussed in literary studies; Michelle Yeh calls vernacular poetry "a self-proclaimed iconoclast struggling against a formidable predecessor: the heritage of three millennia of classical poetry" (Yeh, 1991).

While some propose that the May Fourth intellectuals "abolished the classical language and all of its literary genres" (Hockx and Smits, 2003), others make more measured claims: Mao Chen, for example, maintains that "a special relationship to tradition informs all phases of cultural activity during the May Fourth period" (Chen, 1997).

Julia Lin notes that the period following the May Fourth Movement through 1937 saw "the most exciting and diverse experimentation in the history of modern Chinese poetry" (Lin, 1973). Much of this experimentation was concerned with the question of modernity versus tradition, wherein some poets "adapt[ed] the reality of the modern spoken language to what they felt was the essence of the old classical Chinese forms" (Haft, 1989).

The founding of the People's Republic of China in 1949 was a second major turning point in the century, when "the Communists in one cataclysmic sweep [...] ruthlessly altered the course of the arts" and poetry "became totally subservient to the dictates of the party" (Lin, 1973). With the "physical removal of the old cultural leadership," many of whom fled to Taiwan, this period saw a substantial "vacuum in literature and the arts" (McDougall and Louie, 1997).

Post-Mao, publication restrictions gradually loosened and earlier cultural journals re-entered circulation. Poetry began to reclaim its audience, and a Chinese avant-garde associated with the "Misty Poets" developed (McDougall and Louie, 1997).

However, we lack broad-scale empirical evidence of the linguistic features that constituted the shift from tradition to modernity. Therefore, we propose a study that asks: To what extent were classical poetic forms and classical language immediately discarded with the advent of vernacular poetry? What is the status of classical language after 1949 and amidst the Maoist era, when we might expect its total absence? Does more contemporary poetry still draw connections to classical language?

2 Prior Work on Chinese Poetry in NLP

The majority of existing studies in NLP on Chinese poetry deal exclusively with the classical language.

Jiang and Zhou (2008) explore the problem of classical Chinese poetic couplets, and to develop a system to generate them automatically using techniques from machine translation.

Fang et al. (2009) use an ontology of imagery developed by Lo (2008) to identify imagery in classical Chinese poems, and develop a parser that is able to extract tree structures that identify complex imagistic language in the same.

More recent work develops useful resources for understanding classical poetry. Lee (2012) develops a corpus of classical Chinese poems that are word-segmented and annotated with nested part-of-speech tags that allow for different interpretations of "wordhood" - a non-trivial concept in considering Chinese texts classical and modern. Lee and Kong (2012) introduce a large-scale dependency treebank annotated on a corpus of 8th-century poems.

To our knowledge, there is no existing computational work that attempts to understand the development of modern Chinese poetry over time.

3 Data Collection

For this project, we use a corpus of modern poems collected on the site "Chinese Poetry Treasury" (中国诗歌库, www.shigeku.com) entitled the "Selected Database of Chinese Modern Poetry" (中国现代诗歌精品资料库). It is important to note that the poems in this collection were hand-selected by the group running the site for their canonicity, so our data are biased towards those poems that have, in a sense, "stood the test of time" in the eyes of a mainland Chinese readership.

This corpus is distributed through their site as a collection of html documents, one page per poet, which include brief biographical information for the poet and a collection of their works. We use unix command-line tools (sed, tr, iconv, grep) and basic python scripting to process these documents into a usable corpus with each poem as a separate, clean file, segmented character-by-character. [1]

The site categorizes poets by their "most active" decade, from the 1920s through the 1990s, and we extract this metadata to allow for comparisons over time. In our analysis, however, a methodological impediment arose: namely, the Cultural Revolution.

As discussed in the introduction, this tumultuous period severely disrupted the developmental path of modern Chinese literature. Indeed, we find in our corpus that almost none of the poets tagged as active in the '50s and '60s were mainland Chinese, but instead Taiwanese poets who fled to the island at the climax of the Chinese Civil War.

For this reason, combined with the potential noisiness induced by the fact that decade tags are per-poet instead of per-poem, we manually identify Taiwan poets and divide our corpus into three subsets for analysis: "early modern" poetry in the 1920s and '30s; "late modern" poetry in the '40s interrupted by the Maoist era but resuming in the late '70s, '80s, and '90s; and "Taiwan" poetry by Taiwan natives and transplanted mainlanders in Taiwan post-1949.

After pre-processing, our full corpus for analysis (denoted *Eval* in Table 1) contains 3,611 poems by 305 poets, with a total of 1,128,428 Chinese characters. This size is large enough for meaningful computational results, but small enough to allow for significant qualitative analysis.

We will later define metrics for evaluating the "classicality" of individual characters and radicals, so we process auxiliary corpora (denoted *Aux* in Table 1) of classical poetry and contemporary prose. For classical Chinese, we use a large corpus, from the same source (www.shigeku.com), of poems from the Tang Dynasty (618-907 AD), often considered the greatest classical era for Chinese poetry. For modern Chinese, we use a subset of a machine translation bi-text, comprised primarily of contemporary newswire, legal, and other prose texts.[2]

Since we aim to discover the overall "classicality" of association for individual characters, our auxiliary corpora are cross-genre to exaggerate the effects — a high "classicality" score will indicate both a period-specific classicality and a classical poetic genre association.

[1] Scripts and further information are available here: http://nlp.stanford.edu/robvoigt/chpoetry/

[2] From the BOLT Phase 1 Evaluation training data; see http://www.nist.gov/itl/iad/mig/bolt_p1.cfm

Table 1: Corpus inventory.

		Poems	Chars	Vocab
Eval	Early	351	89,226	3,299
	Taiwan	513	126,369	3,878
	Late	2,747	912,833	4,852
Aux	Classical		2,712,685	6,263
	Modern		9,405,549	5,517

4 Methodology

> *Speak in the language of the time in which you live.*
>
> — Hu Shi, 1917

As suggested in the introduction, modern poetry is distinguished linguistically from classical poetry in its explicit shift to the use of vernacular language. Classical poetry is formalized, concise, and imagistic. We propose three features to operationalize this classicality and computationally observe the shift to a poetic vernacular across the 20th century.

Final Rhyme Classical Chinese poetry in general has a highly regular structure, following strict metrical and rhyming conventions, and most prominently employs a highly consistent end-rhyme. We use the CJKLIB python library[3] to obtain the pronunciation for the last character in each line of each poem. The pronunciation of a given Chinese character may be divided into precisely one consonant (known as an "initial") and one vowel (known as a "final").

We therefore qualify a given line as "rhyming" if the last character of any line within a 3-line window shares its vowel final pronunciation, and for each poem calculate the proportion of rhyming lines.

Character-based Probability Ratio Inspired by the work of Underwood and Sellers (2012) in tracking shifts in literary diction in English poetry, we use our auxiliary corpora of Tang Dynasty poems and modern Chinese language text to create two simple metrics for understanding the "classicality" of poetic diction.

The extreme concision of classical poetry "focuses attention on the characters themselves" (Hinton, 2010), with common classical forms containing as few as ten or twenty characters. To analyze classical diction, for each character we aim to get a ratio describing how classical it sounds.

For this metric, we calculate the probability of each character occurring in its respective corpus using add-one smoothing. We then define the score for a given character as the difference of the character's log likelihood of occurring in the classical auxiliary corpus with its log likelihood of occurring in the modern auxiliary corpus. Scores range from -8 to +8, where a higher score indicates a more "classically"-tinged character.

We find these scores match up well with intuition. In the highly negative range, we find recently-invented, conversational, and grammatical characters unique to the modern vernacular. In the highly positive range, we find rareified literary, poetic characters. In the range surrounding 0.0, we find many common, persistent characters whose meanings have changed little over time. Selected examples of these scores can be seen in Table 2.

Table 2: Example classicality scores for selected characters on the Character-based Probability Ratio metric.

Character	Meaning	Score
	HIGHLY CLASSICAL	
遇 *yu*	To meet; to encounter	7.94
衾 *qin*	A thin quilt used to cover a corpse in a coffin	6.42
萧 *xiao*	A type of bamboo flute	5.99
柳 *liu*	Willow	4.68
	SIMILAR ACROSS PERIODS	
听 *ting*	Listen; hear	0.64
去 *qü*	To go; towards	0.61
直 *zhi*	Directly	-0.11
收 *shou*	To receive; to harvest	-0.53
	HIGHLY MODERN	
你 *ni*	Second-person pronoun	-4.49
够 *gou*	Sufficient; enough	-6.02
呢 *ne*	Sentence-final particle	-6.67
她 *ta*	Third-person female pronoun	-7.82

We calculate a score for a given poem on this metric by simply taking the average of the character-based probability ratio for each character in the poem. These results are denoted *Char* in Table 4.

Radical-based Probability Ratio This metric is fundamentally similar to the above character-based method, but offers the potential to provide a different kind of insight. The majority of Chinese characters are compositional, with a semantic component and a phonetic component.

[3] http://code.google.com/p/cjklib/

We start from the intuition that contemporary texts will be more likely to use characters that contain the 口 (*kou*, “mouth”) radical as their semantic component, because this radical is commonly found in modern conversational particles that were not used in ancient texts. We generalize this hypothesis and consider that the use of characters with certain semantic radicals is correlated with the classicality of a text.

We again use the CJKLIB python library to process our auxiliary corpora, extracting the semantic component radical from each character and calculating the ratio of its probability of occurrence, with add-one smoothing, in the auxiliary classical and modern corpora. As above, we obtain the ratio scores for each radical, and score each poem in our corpus by averaging these scores for each character in the poem.

While these scores are less immediately accessible to intuition than those of the character-based metric, the radical-based scores, with examples seen in Table 3, demonstrate a consistency that parallels the character-based scores.

The semantic radicals most prevalent in classical poetry include those signifying bird, horse, valley, mountain, ghost, dragon, and so on; classical poetry has a pastoral and mythological aesthetic that is directly reflected in the distribution of its radicals. Conversely, modern prose is more likely to use semantic radicals related to work, family, money, speech, and movement; they convey the practical realism of contemporary conversational speech.

Table 3: Example classicality scores for selected semantic radicals on the Radical-based Probability Ratio metric.

Radical	Meaning	Score
	HIGHLY CLASSICAL	
鬼 *gui*	Ghost	2.18
山 *shan*	Mountain	2.09
虫 *chong*	Insect	1.43
	SIMILAR ACROSS PERIODS	
女 *nü*	Female	0.01
文 *wen*	Culture; language	-0.02
生 *sheng*	Life; birth	-0.01
	HIGHLY MODERN	
手 *shou*	Hand	-0.48
言 *yan*	Words; speech	-0.61
力 *li*	Force; work	-0.94

4.1 Diachronic Statistical Analysis

We began from the hypothesis that each of the metrics described above will demonstrate, broadly, that the classical nature of Chinese poetry decreased over the course of the 20th century. The raw statistical counts for our features can been seen in Table 4.

Table 4: Raw feature statistics across sub-corpora. Higher values in the AVG rows indicate a greater “classicality.” For all three features, classicality decreased over the century, with the exception of Taiwan.

		Early	Taiwan	Late
Rhyme	AVG	0.281	0.244	0.226
	STDDEV	0.193	0.169	0.152
Char	AVG	-0.695	-0.620	-0.882
	STDDEV	0.494	0.446	0.404
Radical	AVG	-0.072	-0.081	-0.116
	STDDEV	0.121	0.105	0.097

We calculate the presence of the “classical” features defined above for each subset, and compute a binary logistic regression with the scikit-learn python library (Pedregosa et al., 2011)[4] to find correlation coefficients for those features between the “early modern” and “late modern” subsets.

5 Results and Discussion

Several claims from the literary community are well-supported by our results.

Logistic regression reveals a significant downward trend for our features as we shift from “early modern” to “late modern” poetry ($R^2 = 0.89$), indicating decreased use of end-rhyme, increased use of modern characters, and increased prevalence of modern semantic radicals over the course of the century.

Though the early works use more classical characters on the whole, we also observe a higher statistical variance for all metrics in the ’20s and ’30s, supporting the literary hypothesis that the May Fourth period was one of increased experimentation that later settled into a somewhat more consistent modernity.

We find, however, less support for the idea that Chinese modern poets “abolished the classical language” in their work (Hockx and Smits, 2003).

[4] http://scikit-learn.org

Throughout the century we find repeated instances of highly classical language, with individual poems reaching a maximum character-based probability ratio of 0.70 in the "early" works, 0.76 in the "late" works, and 0.87 in the "Taiwan" works; compare these with an average score of 1.20 for the auxiliary classical dataset overall. Considering that a score of 0.0 would indicate an equal distribution of weight between "classical" and "modern" characters, it's clear that these 20th-century poems still contain a substantial proportion of characters drawn from the classical language.

Poems from Taiwan in the '50s and '60s offer perhaps the most interesting results in this study. It's notable in the first place that poets in our corpus selected as worth remembering by contemporary mainland Chinese from the most authoritarian period of Communist control are almost exclusively from Taiwanese authors. Furthermore, the dip towards modernity we see in '40s mainland poetry was rejected in the next decade by those mainland poets who found themselves in Taiwan after 1949; the Taiwan poems bear far greater resemblance to the early subset of our data than to the late.

This finding parallels work on this period from literary scholars. Yvonne Chang writes that in '50s and '60s Taiwan, valorization of traditional Chinese culture and romanticization of the early 20th-century Nationalist period in mainland China was heavily encouraged. In particular, the concept of "纯文学" (*chun wenxue*, "pure literature") gained popularity in Taiwan's literary circles, and with it came a resurgence of more traditional diction and forms (Chang, 1993).

Fangming Chen further describes poetry in postwar Taiwan as a political outlet for the Kuomintang, the sole ruling party of Taiwan at the time, as they "forcefully brought Chinese nationalism" to the island. Poets who demonstrated a deep "nostalgia" for the "motherland" of mainland China were far more likely to be rewarded with cultural resources such as grants and publication money, being that the government had a vested interest in keeping the public on board with plans to "reclaim the homeland" (Chen, 2007). It is fascinating, then, that we observe this tendency computationally with a return to the levels of classicality seen in '20s and '30s mainland China.

In spite of these encouraging results, this work has several limitations. Our reliance on decade-based labels applied to poets, rather than poems, introduces significant noise. The outlier behavior observed in Taiwan poets is indicative of the need for a better understanding of regional differences, and a comparison with a similarly isolated Sinophone region such as Hong Kong would be productive in this regard. In both cases, information extraction techniques might allow us to tag poems with their date of publication and poets with their hometown, facilitating fine-grained analysis, as would a broader dataset that goes beyond the modern canon.

6 Conclusion

In this paper, we computationally operationalized three features that successfully track the declining influence of classical poetic style and language in 20th-century Chinese poetry. We identified Taiwan poets as an outlier in the dataset, and found empirical evidence for the political externalities of the '50s and '60s that called for a return to a nostalgic classicism. In this way, this work presents a promising first step to a thorough empirical understanding of the development of modern Chinese poetry.

Acknowledgments

Thanks to three anonymous reviewers for detailed and insightful comments. This research was supported in part by the Foreign Language and Area Studies Fellowships, United States Department of Education.

References

Sung-sheng Yvonne Chang. 1993. *Modernism and the Nativist Resistance*. Duke University Press: Durham and London.

Fangming Chen. 2007. Postmodern or Postcolonial? An Inquiry into Postwar Taiwanese Literary History. In *Writing Taiwan*, David Der-wei Wang and Carlos Rojas, eds. Duke University Press, Durham and London.

Mao Chen. 1997. *Between Tradition and Change*. University Press of America, Lanham, MA.

Alex Chengyu Fang, Fengju Lo, and Cheuk Kit Chinn. 2009. Adapting NLP and Corpus Analysis Techniques to Structured Imagery Analysis in Classical Chinese Poetry. In *Workshop Adaptation of Language Resources and Technology to New Domains*, Borovets, Bulgaria.

Lloyd Haft. 1989. *A Selective Guide to Chinese Literature: 1900-1949*. E.J. Brill, New York.

David Hinton, ed. 2010. *Classical Chinese Poetry: An Anthology*. Farrar, Straus, and Giroux.

Michel Hockx and Ivo Smits, eds. 2003. *Reading East Asian Writing: The Limits of Literary Theory*. RoutledgeCurzon, London and New York.

Long Jiang and Ming Zhou. 2008. Generating Chinese Couplets using a Statistical MT Approach. In *COLING*.

John Lee. 2012. A Classical Chinese Corpus with Nested Part-of-Speech Tags. In *Proceedings of the 6th EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Avignon, France.

John Lee and Yin Hei Kong. 2012. A Dependency Treebank of Classical Chinese Poems. In *NAACL-HLT*, Montreal, Canada.

Julia Lin. 1973. *Modern Chinese Poetry: An Introduction*. University of Washington Press, Seattle, WA.

Fengju Lo. 2008. The Research of Building a Semantic Cetegory System Based on the Language Characteristic of Chinese Poetry. In *Proceedings of the 9th Cross-Strait Symposium on Library Information Science*.

Lu Zhiwei. 1984. *Five Lectures on Chinese Poetry*. Joint Publishing Co., Hong Kong.

Bonnie McDougall and Kam Louie, eds. 1997. *The Literature of China in the Twentieth Century*. Hurst and Company, London.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 12:2825-2830

Ted Underwood and Jordan Sellers. 2012. The Emergence of Literary Diction. *The Journal of Digital Humanities*, 1(2). http://journalofdigitalhumanities.org/1-2/the-emergence-of-literary-diction-by-ted-underwood-and-jordan-sellers/

Michelle Yeh. 1991. *Modern Chinese Poetry: Theory and Practice since 1917*. Yale University Press, New Haven, CT.