Building a Corpus of Chinese Classical Poetry

Fangning Shao

Code and corpus is available at: https://github.com/fangningshao/chinese-poetry-corpus

1. Summary

Chinese classical poetry is one of the most important carrier of Ancient Chinese culture, and an extremely valuable collection of literatures that spans more than 2000 years. It is also a very intriguing source of language knowledge for linguistic studies. Unfortunately, there is no publicly available corpus of Chinese poetry for researchers to easily access all these content.

In this project, I built a corpus of Chinese classical poetry by crawling 搜韵 *Sou Yun* (http://sou-yun.com/), a comprehensive Chinese poetry website. This website carries one of the most complete datasets of Chinese poetry. It is also a well recognized tool for people to search poems by content, author and even rhyme.

The full scraped data is available in https://github.com/fangningshao/chinese-poetry-corpus/tree/master/data.

2. Structure of the corpus

The Sou Yun corpus we built contains 15 dynasties, from XianQin to current. The corpus contains a total of 28,596 authors and 747,521 poems.

On a high level, the corpus contains a TSV file listing dynasties (name and link), a list of authors (name and link) for each dynasty, and a TSV file of all poems for each author (poem title and content). See below hierarchy for the directory structure:

```
data/
      dynasty.tsv           # a list of all dynasties
      先秦/                  # for each dynasty
            authors.tsv     # all authors for that dynasty
            poems/          # all poems for that dynasty
                  诗经.tsv   # one file per author, one row per poem
                  屈原.tsv
                  宋玉.tsv
                  ...
```

3. How the corpus was built

The website of Sou Yun was strictly structured as following. There is an index page listing all dynasties, from which we can extract the link to each dynasty. In the page of each dynasty, there is a listing of links to all the authors. In an author's page, all the author's poems are

enumerated. This website structure makes it possible to start from the index page and crawl all the dynasties, authors, and poems.

When building the corpus, I initially tried to use the lab10 scrapy code, but I found it not obvious how to insert such specific logic to handle different hierarchies of pages. Therefore, I chose to build it with custom code using BeautifulSoup. I first wrote exploration code in a Jupyter Notebook and used Chrome to figure out the identifiers for the desired HTML elements, and then moved the code to a Python script main.py in my code base.

In sections below, we will discuss some potential studies we can perform with this corpus, and some of them will be approached. Note that there are more complex studies possible by refining the corpus schema, and below is just a set of initial studies.

4. Studies using this Corpus

Having a Chinese classical poetry corpus will enable studies in many topics: forms of poetry over time, interjection and particle usage, metaphor usage, sentiment analysis for different authors and ages, and many more.

4.1. Poetry Form Analysis

Chinese poetry has various forms, and the most popular ones are Lüshi (律诗) and Jueju (绝句).

Lüshi is a specific form of classical Chinese poetry, and it is also one of the most major poetry forms. Lüshi always has 8 lines, and in most Lüshi, each line can consist of 5 characters or 7 characters.

Jueju, another porm form, is most popular in Tang dynasty. It is similar with English Quatrains. It consists 4 lines with rhyme and each line can consist of 5 or 7 characters.

Using the corpus we built, we can split by commas and periods and try to estimate the form of the poetry. We can also use the annotation on the title, which we already crawled, to calculate the distribution of Lüshi and Jueju.

We take a sample of all the poems in different dynasties, used bash to count the occurrence of "绝句" or "律诗" annotation in each dynasty (see count-poetry-form.sh), and plot the result as distribution of Jueju and Lüshi in Figure 1 below:

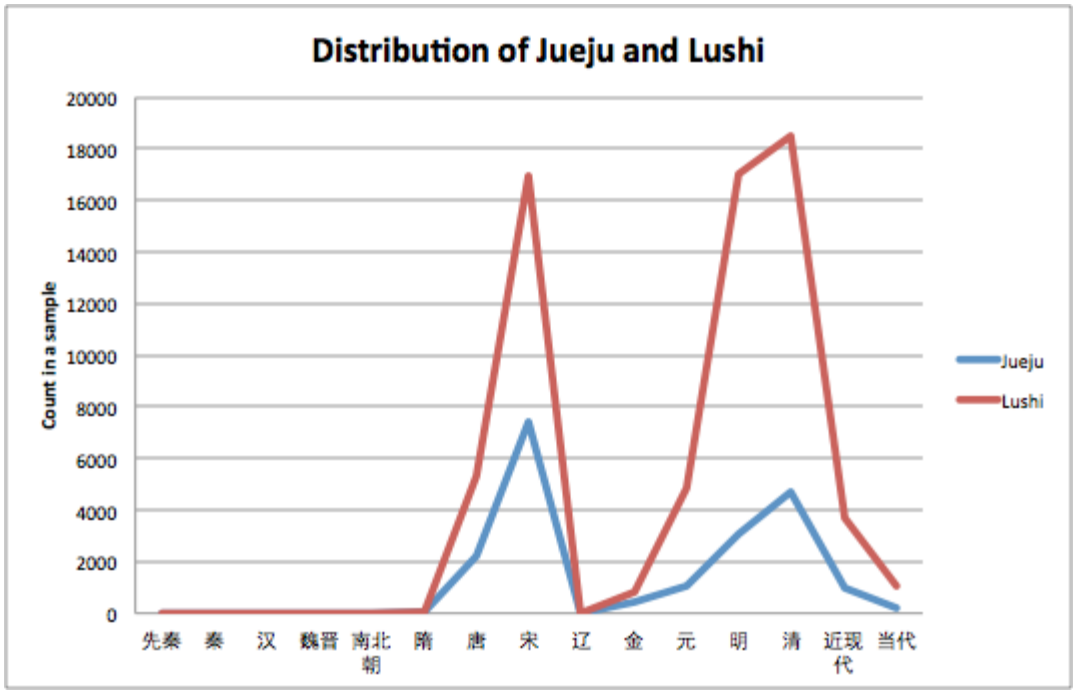


Figure 1: Distribution of Jueju and Lüshi

From the figure above we can see, that Lüshi is more common than Jueju almost in every dynasty, and they both about to start become very popular in Tang (唐). We can also find that there are not a lot of poetry in Liao, Jin and Yuan dynasty. This result actually accords with the history. In Liao, Jin and Yuan dynasty, the emperors came from ethnic minorities which don't speak Chinese. The official languages of these dynasties are not Chinese neither. Therefore the chinese literature in those years did not have substantial development.

4.2. Interjection analysis

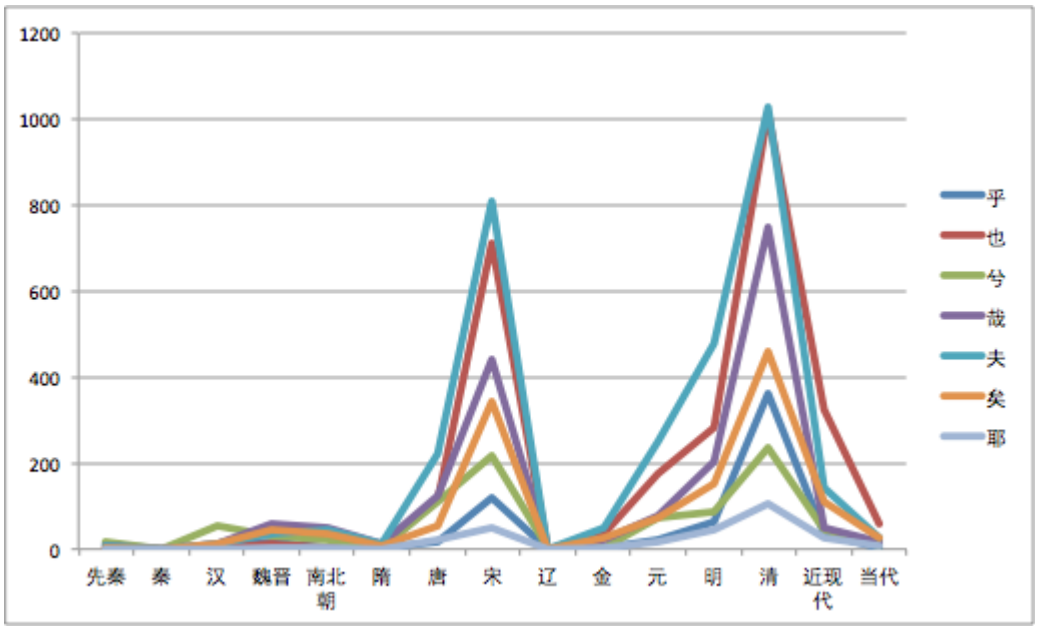
In ancient Chinese language, there are different interjection words and the usage of them changes over time. Common ones include: Hu(乎), Ye(耶), Xi(兮), Zai(哉), Fu(夫), Yi(矣) and Ye(也).

The distribution of interjections across dynasties in our sample is shown in the table below.

Dynasty	乎	也	兮	哉	夫	矣	耶	Grand Total
先秦	11	5	19	8	9	5	0	57
秦	0	0	0	0	0	0	0	0
汉	8	2	53	11	12	11	0	97
魏晋	19	12	30	59	36	44	2	202
南北朝	5	5	21	50	47	34	1	163
隋	2	3	5	14	11	9	0	44

唐	16	117	110	122	224	56	20	665
宋	119	709	217	443	810	342	49	2689
辽	0	0	2	0	0	0	0	2
金	5	36	3	20	50	26	2	142
元	22	174	73	79	249	73	17	687
明	64	284	87	205	476	154	44	1314
清	360	1013	236	749	1028	458	107	3951
近现代	36	326	43	48	142	108	26	729
当代	5	58	9	16	25	26	8	147
Grand Total	672	2744	908	1824	3119	1346	276	10889

And the trend of change is shown in the chart below.



From the table and figure above, we can identify a few patterns: (1) Fu (夫) is the most common interjection in most dynasties except contemporary poems (近现代, 当代). Ye (也)

is second to Fu. Xi (兮) is the most popular interjection in early times until Weijin (魏晋). Hu (乎) gains popularity in later ages like Qing (清).

4.3 Other usages of this corpus

I wrote two examples of analysis in linguistics field above. These are only a small part of the potential uses. This corpus can be used in many other analysis as well. For example, researchers in the literature field can use it to analyse the usage of metaphors, the writing style of some authors and the changement of popular topics. Researchers in computer science field can use it to do the sentimental analysis, classic poetry automatic translation and so on.