CHINESE LEGAL TEXTS – QUANTITATIVE DESCRIPTION[*]

Ľuboš GAJDOŠ
Comenius University in Bratislava, Slovakia
lubos.gajdos@uniba.sk

Abstract

The aim of the paper is to provide a quantitative description of legal Chinese. This study adopts the approach of corpus-based analyses and is conducted on the Chinese monolingual corpus *Hanku*. It shows basic statistical parameters of legal texts in Chinese, namely the length of a sentence, the proportion of part of speech etc., and discusses the issues on statistical data processing from various corpora, such as the tokenisation and part of speech tagging, and their relevance to the study of register variations.

Keywords: Chinese language; written Chinese; legal texts; corpus linguistics

Povzetek

Namen članka je ponuditi kvantitativni jezikoslovni opis kitajskih pravnih besedil. V raziskavi avtor privzema metodologijo korpusnih analiz na kitajskem enojezičnem korpusu *Hanku*. Osredotoča na osnovne statistične parametre, kot so dolžina povedi, besedne vrste in njihov delež v besedilih idr. ter razpravlja o vprašanjih obdelave statističnih podatkov iz različnih korpusov, kot sta npr. navajanje in označevanje besednih vrst, ter njihov doprinos k raziskavam o raznolikosti registrov.

Ključne besede: kitajski jezik; pisna kitajščina; pravna besedila, korpusno jezikoslovje

[*] The study builds on my previous work Gajdoš, Ľ (2016). Quantitative description of written Chinese – a preliminary corpus-based study. *Studia orientalia: Victori Krupa dedicata*. Bratislava: Ústav orientalistiky, pp. 62-75.

Acta Linguistica Asiatica, 7(1), 2017.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.7.1.77-87

1 Chinese language

Chinese is one of the most widely used languages in the world (Sun, 2006). Genealogically, Chinese belongs to the Sino-Tibetan language family and is often classified as an ‘isolating’ or ‘analytic’ language (Li & Thompson, 2009, pp. 703-723). Chinese has a written tradition of more than 3000 years. The majority of Chinese words are mono- or disyllabic, the word order is relatively rigid with the SVO prototypic word order.

In a language with written tradition, the discrepancy between spoken and written registers[1] is a natural phenomenon. Chinese language is no exception, nevertheless, there are some issues that need to be considered when studying registers[2] variation, e.g. a relatively vague standard of the written language, the influence of *wenyan*[3] on the written language, missing qualitative and quantitative comparative study of language registers etc.

The study is based on the dichotomic model of Chinese language registers – colloquial and written. Hereinafter, the term ‘Chinese’ is reserved for the written standard of Chinese which is known as *putonghua* (普通话) ‘common language’.

2 Chinese legal language

The term *legal language/text* (hereinafter legal text) may be understood in many ways with respect to a research perspective, a research methodology, criteria of classification etc. Due to the content of the sub-corpus *zh-law*, the term legal text implies only monologic, prescriptive texts of legislation. Generally speaking, legislation is considered as the prototype of legal language (Biel, 2014, pp. 19-22).

Legal Chinese is a part of written language registers (sub-register) with unique lexis, syntax properties, terminology etc.

[1] In this article, I use the term *register* as “situationally-defined varieties described for their characteristic lexico-grammatical features” (BIBER, 2006, pp. 11-12).

[2] Both terms are defined very vaguely in Chinese linguistics. The spoken register is called *kouyu* (口语) ‘spoken language’ and written as *shumianyu* (书面语) ‘written or literary language’.

[3] *Wenyan* is known as ‘classical literary language‘ and “it looks to the style of writings prevalent in the period from the Spring and Autumn period to the Eastern Han dynasty for its grammatical and lexical norms” (Chen, 2004, p. 67).

3 Methodology

In this study, the Hanku corpus and corpus methodology is used as a systematic approach to the study of the register of legal Chinese. It is proposed that a variance across registers might be – to some extend – revealed by statistical data from a corpus.

By testing this hypothesis on the register of legal Chinese, the following criteria are taken as a part of the description.

(1) the length of a sentence in a register
(2) frequency of every part of speech (hereinafter the POS) in a register
(3) the relative representation of some POS and their comparison
(4) markers of passive voice

The statistical data presented in this study are given in two values – absolute frequency and frequency in IPM.[4]

4 The Chinese corpus *Hanku*

The *Hanku* corpus is available at: http://konfuciovinstitut.sk/corpus-hanku/, *NoSketch Engine*[5] is used as the corpus manager. The basic block of the corpus is a token which basically corresponds to one word. A token is annotated for the part of speech (POS labelling), its composition into characters and the *Hanyu pinyin* transcription.[6] The corpus is divided into two subcorpora (May 2017):

- *web-zh* – texts from the PRC
- *zh-law* – legal texts from the PRC; texts of laws and regulations

The statistical data used in this study was obtained by writing CQL queries in the *NoSketch Engine* user interface.

[4] IPM: Instances Per Million, the number of occurrences normalized by the size of the corpus.

[5] Nosketch Engine is an open-source version of the Sketch Engine. See more at: https://www.sketchengine.co.uk. For other Chinese corpora available in Sketch Engine, see Petrovčič (2016).

[6] The POS annotation, tokenization and *Hanyu pinyin* transcription are results of automatic processing.

Table 1: Subcorpus Zh-law – parameters

Parameters	Status	Notes
Type	synchronous	legal texts from the PRC
Language of interface	English, Chinese	
Size (June 2017)	7.2 million	size referred in tokens
Tokenisation	✓	
POS annotation	✓	Penn Chinese Treebank[7]
Bibliographic annotation	✓	
Phonetic annotation	✓	
Statistic tools	✓	frequency in IPM, average reduced frequency
Save results directly from the interface	✓	in text or XML format
KWIC	✓	
Collocations search	✓	many collocation measures
Advanced search options	✓	Boolean operators – conjunction, disjunction, negation; possibility to use regular expressions at the character, word, pinyin, and metadata level; full CQL[8] etc.
Sorting by	✓	Left, right, node, references etc.

5 The length of a sentence in the subcorpus *zh-law*

It is generally believed that there is a positive correlation between the length of a sentence and the register affiliation – the more formal a text is, the longer the sentences are, and *vice versa*.

Our previous research on written Chinese has confirmed this tendency, yet with more accurate statistical data showing that the length tends to have more than 29 tokens.[9] It should be noted though that the number of tokens also include punctuation (the POS tag “PU”), e.g. “，、。（）”, so that the length in words is shorter.[10]

[7] See more at: http://www.cs.brandeis.edu/~clp/ctb/posguide.3rd.ch.pdf.

[8] CQL – Computer Query Language.

[9] Our previous research indicated slightly different figures, i.e. 20 words (tokens without punctuations).

[10] The length of a sentence was simply calculated by division of a number of tokens by a number of sentences in the sub-corpus.

The results indicate that a sentence in legal Chinese tends to have the length of approximately 29 tokens or 25 words (tokens without punctuations). It is also evident that the more information-saturated a text (a text in nominal style) is, the longer it is. When analysing the corpus data here, one must also pay attention to the fact that the name of a law or a regulation is tokenized as a sentence as well. That is to say, the length of a sentence may even be longer.

6 Parts of speech in the sub-corpus zh-law

The proportion of individual POS was directly retrieved from the corpus with a query written in CQL,[11] then, the result was sorted by "node tags" and converted to IPM.

Table 2: A proportion of POS in Zh-law

Part of Speech	Examples	Tag	Frequency in the corpus	IPM
Nouns		NN	2 804 164	389 236
Punctuations		PU	1 091 923	151 566
Verbs		VV	1 070 286	148 562
Prepositions		P	242 165	33 614
Non-predicative adjectives	共同，女	JJ	232 514	32 274
Adverbs		AD	222 902	30 940
Coordinating Conjunctions	与，和，或者	CC	219 759	30 504
Particle DE	的	DEC	215 381	29 896
Measure words		M	198 809	27 596
Cardinal numbers		CD	162 183	22 512
Ordinal numbers		OD	139 560	19 371
Particle DE as genitive marker	的	DEG	129 286	17 945
Localizer		LC	109 661	15 221
Determiners	这，那	DT	99 523	13 814
Proper nouns		NR	48 931	6 792
Adjectives		VA	38 810	5 387
Temporal nouns		NT	35 166	4 881

[11] CQL – Corpus Query Language; [tag=".*"].

Part of Speech	Examples	Tag	Frequency in the corpus	IPM
Pronouns		PN	33 930	4 709
Verbs	有，没有，无	VE	28 195	3 913
Etcetera	等等	ETC	21 082	2 926
Particles	所，以，而	MSP	18 021	2 501
Copulas		VC	15 708	2 180
Preposition BA		BA	6 586	914
Preposition BEI		SB	6 378	885
Preposition BEI		LB	3 338	463
Aspect particles	了，着，过	AS	2 898	402
Particle DE	地	DEV	2 489	345
Subordinating conjunctions	如果，要是	CS	1 917	266
Modal particles	了，吧，吗	SP	1 591	220
Foreign words		FW	948	131
Particle DE	得	DER	157	22

Data in Table 2 reveal that interjections and onomatopoeias are not present in legal texts at all. By comparing statistical data with other registers, there are some factors that should be taken into consideration – due to the fact that the form of legal texts differs from other registers, the frequency of punctuation, cardinal and ordinal numbers, or measure words is much higher. See Chapter 7.

To allow a more concise comparison, some POS are combined together as shown below in Table 3.

Table 3: POS combined together

POS	Frequency	IPM
Nouns (NN+NR+LC+NT)	2 997 922	416 132
Verbs (VV+VC+VE)	1 114 189	154 657
Particles DE (DEC+DEG+DEV+DER)	347 313	48 209
Numbers (CD+OD)	301 743	41 884
Prepositions (P+BA+BEI)	258 467	35 877
Non-predicative adjectives (JJ)	232 514	32 275
Adverbs (AD)	222 902	30 940

POS	Frequency	IPM
Conjunctions (CC+CS)	221 676	30 770
Measure words (M)	198 809	27 596
Pronouns (PN+DT)	133 453	18 524
Particles (ETC+AS+MSP+SP)	43 592	6 051
Adjectives (VA)	38 810	5 387
Passive markers (SB+LB)	9 716	1 349
Punctuation (PU)	1 091 923	151 566

The chart shows that the nominal style is preferred in legal Chinese with a dominance of nouns.

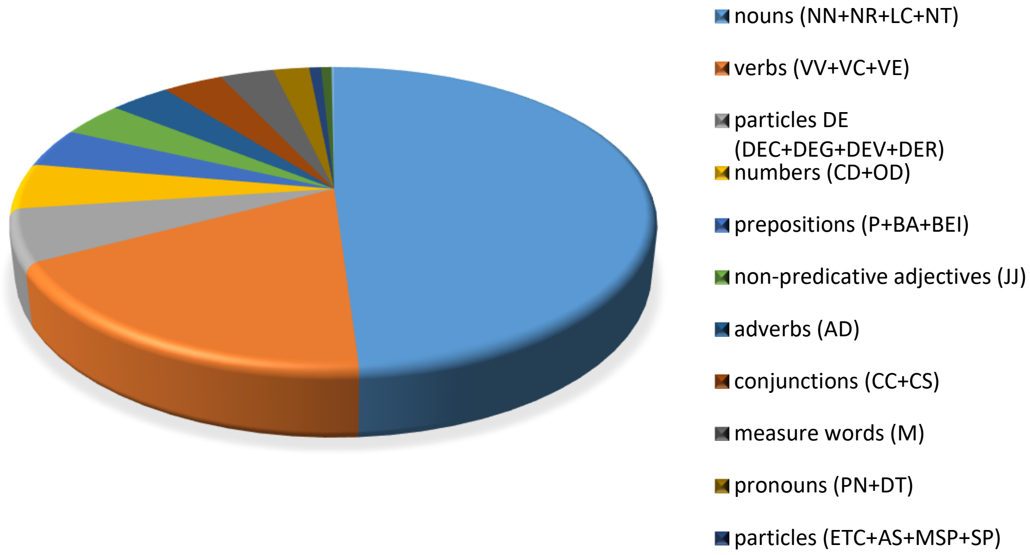

Figure 1: Proportion of part of speech in the zh-law

Mainly because of the different approach to the tokenization and the POS labelling, statistical data have exhibited a considerable divergence between the proportion of verbs in the sub-corpus *zh-law* (18%) versus the Sihanku corpus (26%).

7 Most frequent content and function words in the sub-corpus zh-law

Following the traditional model[12] of division of lexis into content and function words (*shici* 实词, *xuci* 虚词, respectively), I have searched for the most frequent content words by the CQL query: [tag=”V.*|N.*|LC|CD|OD|JJ|AD|DT|PN|M”]. The result was then sorted by “node forms” and converted to IPM. As for the function words, the CQL expression is: [tag="P|CC|DE.|ETC|MSP|BA|SB|LB|AS|CS|SP"].

Table 4: 30 most frequent content and function words in legal Chinese

Content word	Frequency	IPM	Function word	Frequency	IPM
条	102845	14276	的	339351	47104
应当	57301	7954	和	86413	11995
规定	50465	7005	或者	47375	6576
管理	48272	6700	在	40438	5613
不	45266	6283	对	38084	5286
部门	41477	5757	由	26783	3718
机构	34843	4836	并	24460	3395
行政	33949	4712	及	21111	2930
人民	31789	4413	等	21060	2923
本	31540	4378	或	18820	2612
企业	29894	4149	与	17647	2450
一	28734	3988	向	17483	2427
单位	28029	3891	按照	15267	2119
有关	27402	3804	所	14866	2064
国家	26708	3707	根据	10988	1525
人员	24086	3343	经	10928	1517
工作	23779	3301	以	10130	1406
有	23085	3204	被	9338	1296
其	22666	3146	以及	7595	1054
其他	22012	3055	为	7355	1021
进行	21269	2952	之	7216	1002
申请	21018	2917	按	7176	996

[12] E.g. Liu, Y. et al. (2004). *Practical Chinese Grammar*.

Content word	Frequency	IPM	Function word	Frequency	IPM
应	20085	2788	依照	6732	934
机关	19254	2673	自	6691	929
内	18847	2616	将	6160	855
公司	18130	2517	因	4353	604
监督	17920	2487	于	4042	561
主管	17711	2458	关于	3418	474
二	17333	2406	通过	2958	411
三	17096	2373	除	2627	365

Based on the content of legal texts, it is no surprise that the most frequent content word is the measure word *tiao* 条 as it stands for an article of a law (§). A second most frequent word is the modal verb *yingdang* 应当 which is a mean of expressing deontic modality.

Our previous research has also proven the tendency that on the part of function words, there is a high frequency of conjunctions compared to unstructured texts (e.g. language data from the sub-corpus web-zh) and this might be an indication of formal, written texts. There is also a high figure of prepositions which are considered as a formal expression, e.g. *yu* 与 or *yi* 以.

8 Passive voice in legal Chinese

In this chapter, the result of Straňák's study (2015) who has conducted his research on the relatively small sub-corpus[13] of legal texts with a different tagset and tokenizer are compared.

Passive voice in Chinese may be marked with prepositions e.g. *bei* 被 or unmarked. Since Straňák has only searched for the passive voice marked by preposition, I adopt this approach here as well. In the Hanku corpus, to search for the passive voice is quite straightforward with the CQL query as follows: [tag="LB|SB"].

The results in the Table 5 show that the frequency of passive markers in all corpora varies. Comparison of the passive marker *bei* 被 from the sub-corpora *zh-law* and *web-zh* also indicate slightly opposite tendency as it was the case of Straňák's research in which the frequency of the passive marker in legal Chinese was significantly higher that

[13] The sub-corpus had a size of 480.000 tokens.

the frequency in other sub-corpora of written language. It is worth noting that the statistical data is not sufficiently large to enable comparison to be performed.

Table 5: Passive markers in the sub-corpora zh-law, web-zh and the corpus Sihanku

Marker	IPM zh-law	IPM web-zh	IPM Sihanku
被	1296	1458	2206
受	47	20	1312
为	5	20	?

The statistical data also has not proven that the passive markers might be one of the indicator of legal texts. The figures of passive markers do not indicate any significant differences between two sub-corpora of the Hanku.

9 Implications for language pedagogy

The quantitative study and its result may be used in language pedagogy too, i.e. by learning Chinese legal texts, one might choose to study not only verbs (according their occurrence in a corpus) but the collocational preferences of verbs with a subject/object or prepositional phrases, which is known as “wordsketch” in corpus linguistics or as “valency” in general linguistics.

It is only a matter of practise for students of Chinese language to write a CQL query or a regular expression and search for the concrete POS or words, even for patterns of sentences. I assume that this may help to improve language teaching methods and materials.

10 Conclusion

In this article, I have presented the results of the corpus-based approach to the study of register variation in Chinese. The research was conducted on a relatively small corpus yet the language data in it may be described as complete and closed. The statistical data of legal Chinese reveals that there are evident differences, e.g. in lexis or syntax. Among the above indicators, the absence of modal particles, onomatopoeias, interjections may be clear evidence of a formal, written register. I have also presented an amount of statistical data in support of the hypothesis in which the proportion of nouns in written formal register (here *zh-law)* prevails over other POS.

Some caution should be applied, when comparing the statistical data in this study with other corpus data. Let us here just highlight some issues:

- The size of a corpus matters, e.g. as it is the case of passive markers
- Different approach to tokenization may result in different statistical data
- It is not always a simple task to compare results from two corpora with different tagsets.

To conclude, a quantitative description is as accurate as precise the automatic process of tokenization and POS labelling is. Over the past few years, we have witnessed steady improvement in automatic tokenisation and the POS tagging processes, nevertheless problems still remain with regards of quantitative comparison of results from different corpora.

References

Biber, D. (2006). *Univesity Language*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Biel, L. (2014). *Lost in the Eurofog*. Frankfurt am Main: Peter Lang AG, pp. 19-22.

Chen, P. (2004). *Modern Chinese: History and Sociolinguistics*. Cambridge: Cambridge University Press.

Gajdoš, Ľ. (2011). Discrepancy Between Spoken and Written Chinese – Methodical Notes on Linguistics. *Studia Orientalia Slovaca, 10*(1)*,* 155-159.

Gajdoš, Ľ. (2013). Slovensko-čínsky paralelný korpus [The Slovak-Chinese parallel corpus]. *Studia orientalia Slovaca, 12*(2), 313-317.

Gajdoš, Ľ (2016). Quantitative description of written Chinese – a preliminary corpus-based study. *Studia orientalia: Victori Krupa dedicata*. Bratislava: Ústav orientalistiky, 62-75.

Gajdoš, Ľ., Garabík, R., Benická, J. (2016). The New Chinese Webcorpus Hanku – Origin, Parameters, Usage. *Studia Orientalia Slovaca, 15*(1), 53-65.

Li, N. Ch., Thompson, S. (2009). Chinese. *The World's Major Languages* (Second Edition), ed. by Comrie, B. Oxon: Routledge.

Liu, Y. [刘月华] et al. (2004). *Practical Chinese Grammar* [实用汉语语法]. Beijing: Shangwu yishuguan.

NoSketch Engine [online] [10 May 2017]. Retrieved from Sketch Engine: https://www.sketchengine.co.uk.

Petrovčič, M. (2016). Word Sketches of Separable Words Liheci in Chinese. *Acta Linguistica Asiatica, 6*(1), 47-57.

Straňák, I. (2015). Quantitative Analysis of Function words in Chinese Legal Texts. *Studia Orientalia Slovaca, 14*(2), 165-182.

Sun, Ch. (2006). *Chinese: A Linguistic Introduction.* Cambridge: Cambridge University Press.