A Corpus Analysis of Legal Chinese—Final Thought

Ľuboš Gajdoš

Korpusová analýza čínskeho právneho textu—zhrnutie

Resumé Článok voľne nadväzuje na predchádzajúce štúdie zaoberajúce sa kvantivatívnou deskripciou čínskeho právneho textu. Ponúka vybrané štatistické údaje (dĺžka vety, zatúpenie slovných druhov a pod.) platné pre čínsky právny text. Je zhrnutím výsledkov výskumu čínskeho právneho textu metódami korpusovej lingvistiky a príspevkom k precizovaniu metodológie korpusovej analýzy jazykových registrov.

Abstract The paper freely follows my previous studies on the quantitative description of the Chinese legal text. It offers selected statistical data (sentence length, a proportion of part of speech, etc.) for legal Chinese. It also summarizes the results of the previous research on this register and also intends to refine the methodology of corpus analysis when examining language registers.

Key words Chinese, legal Chinese, quantitative analysis · corpus linguistics, corpus-based approach · register variation, length of a sentence, word-length

Introduction

In this paper, we analyse our previous corpus-based research on legal Chinese and summarize the results with implication on a quantitative description and methodology. The article is divided into 2 parts in which we compare statistical data from 2 different corpora with the same language material but with different quantity, word segmentation and part-of-speech annotation (hereafter POS). In the second chapter, we examine differences in statistical data between two written registers—the sub-corpus of legal texts versus a general corpus.

1 *Different Corpora, Different Results?*

In this chapter, we compare statistical data from two different corpora of legal texts. For the comparison, we use the monolingual corpus Hanku and the parallel corpus Sihanku. The language material in both corpora is similar, however, the word segmentation and the POS taggers are different. The corpora have unequal sizes.

1.1 *Subcorpus Zh-Law*

The subcorpus Zh-law is part of the corpus *Hanku* which is a monolingual, synchronous Chinese corpus (in simplified Chinese characters) available via web interface. It can be accessed via the website of the Confucius Institute at Comenius University in Bratislava at: <konfuciovinstitut.sk>. The *Hanku* uses an open-source version of the Sketch Engine corpus manager (NoSketch Engine) as well as open-source tools for tokenization (ZPar)[1] and POS tagging (the Penn Chinese Treebank).[2] The subcorpus has a size of 7.2 million tokens (August 2017).[3] The language data in the sub-corpus contain texts of law and regulations from the People's Republic of China.

1.2 *Subcorpus Falv*

The subcorpus Falv is part of the corpus *Sihanku* which is a parallel synchronous Chinese corpus (in simplified Chinese characters) available via web interface. It can be accessed via the website of the Comenius University in Bratislava.[4] The *Sihanku* uses an open-source version of the Sketch Engine corpus manager (NoSketch Engine) as well as open-source tools for tokenization and POS tagging

1 See more at <http://people.sutd.edu.sg/~yue_zhang/doc/doc/joint_seg_tag.html> (last retrieval August 12, 2017).

2 See more at <http://www.cs.brandeis.edu/~clp/ctb/posguide.3rd.ch.pdf> (last retrieval August 12, 2017).

3 See Ľuboš Gajdoš, Radovan Garabík and Jana Benická, »The New Chinese Webcorpus Hanku— Origin, Parameters, Usage«, *Studia Orientalia Slovaca* 15,1 (2016), 21–33.

4 The corpus is accessible only to faculty's staff and students.

(the LCMC tagset).[5] The subcorpus has a size of 0.48 million tokens (August 2017).[6] The language data in the sub-corpus contain texts of law and regulations from the PRC.

1.3 *Comparing Data from two Corpora*

Here we compare only the data from 2 corpora that is easily accessible from the corpus user interface (UI)—most frequent words (content and function words), the length of a sentence etc. We provide all CQL[7] requests and research procedures. The results presented here are limited only to the 10 most frequent tokens.

Regarding the most frequent content words,[8] we searched in the *Hanku* by the CQL query: [tag="V.*|N.*|LC|CD|OD|JJ|AD|DT|PN|M"] and in the *Sihanku* by: [tag="v.*|n.*|a.*|b.*|d|f.*|m|q|r|t|s"]. The results are sorted by »node forms« and converted to IPM.[9]

5 The LCMC stand for Lancaster Corpus of Mandarin Chinese tagset. See more at <http://www.lancaster.ac.uk/fass/projects/corpus/LCMC/lcmc/lcmc_tagset.htm> (last retrieval August 12, 2017).

6 See Ľuboš Gajdoš, »Slovensko-čínsky paralelný korpus« [The Slovak-Chinese Parallel Corpus], *Studia Orientalia Slovaca* 12,2 (2013), 313–317.

7 CQL—Corpus Query Language. See more at <www.sketchengine.co.uk/documentation/corpus-querying/> (last retrieval August 12, 2017).

8 Here, we have followed the traditional model of division of lexis into content words—nouns (space words, time words), verbs, adjectives, adverbs, pronouns, classifiers, numerals—and function words—conjunctions, prepositions, particles (auxilairies), interjections, onomatopoeias.

9 IPM—instance per million. All figures in the article are in IPM.

Rank	Token	Frequency in the Hanku	Token	Frequency in the Sihanku
1.	条	14 275	条	17 533
2.	应当	7 953	人	10 187 (1 175)[10]
3.	规定	7 004	应当	9 403
4.	管理	6 700	规定	8 080
5.	不	6 283 (3 844)	有	4 795 (3 204)
6.	部门	5 757	其他	4 634 (3 055)
7.	机构[11]	4 836 (2 350)	可以	4 580 (2 331)
8.	行政	4 712 (2 612)	管理	4 547
9.	人民	4 412 (1 525)	部门	4 223
10.	本	4 377 (1 932)	有关	3 851 (3 803)

Table 1
10 *Most Frequent Content Words.*

Regarding the most frequent function words, we searched in the *Hanku* by the CQL query: [tag="P|CC|CS|DE.|AS|SP|MSP|IJ|ON|LB|SB|BA"] and in the *Sihanku* by: [tag="u|y|o|p|c.*|e"]. The results are sorted by »node forms« and converted to IPM.

Rank	Token	Frequency in the Hanku	Token	Frequency in the Sihanku
1.	的	47 104	的	58 221
2.	和	11 994	或者	13 606

10 This is a good example of different approach to tokenization (word segmantation) that may result in the frequency shift. In the case of the Sihanku corpus, words like *guanli ren* 管理人 ('a manager') are tokenized as 2 separate tokens—*guanli* 管理 ('to manage') and *ren* 人 ('a person'). In the case of the Hanku corpus, such a word is tokenized as one token—*guanliren* 管理人.

11 For the different token, the ipm from the other corpus is provided in round brakets. A token is searched by the CQL query: [word="X"], here X stands for a concrete token.

3.	或者	6 575	和	11 741
4.	在	5 613	对	5 881
5.	对	5 286	在	5 761
6.	由	3 717	由	4 721
7.	并	3 395	并	3 863
8.	及	2 930 (398)	向	2 575 (2 426)
9.	或	2 612	按照	2 490 (2 119)
10.	与	2 449 (2 013)	被	2 306 (1 296)

Table 2
10 *Most Frequent Function Words.*

1.4 *Proportion of the Part-of-Speech*

As both corpora are POS tagged with separate software tools, it is interesting to observe can be accessed the different approaches to the POS tagging that might be seen in the proportions of the POS. We begin with a comparison of the proportion of content words vs. function words in both corpora.

The proportion of content words vs. function words in the monolingual corpus Hanku is 73% vs. 11% as the corpus contains other tokens as well (e.g. punctuations etc.). The proportion of content words vs. function words in the parallel corpus Sihanku is 70% vs. 14% as the corpus also contains other tokens (e.g. punctuations etc.).

To allow a more concise comparison, some POS are combined together (see the round brackets in the table) as shown below.

POS	Frequency in the Hanku	Frequency in the Sihanku
Nouns (NN+NR+LC+NT) (n+ng+nr+ns+nt+nx+nz+vn+t+s+f+l+b)	416 132	399 981
Verbs (VV+VC+VE) (v)	154 657	188 695
Particles DE (DEC+DEG+DEV+DER) (地+得+的)	48 209	59 377
Numbers (CD+OD) (m+mg)	41 884	38 468
Adjectives (VA+JJ) (a+ad+ag+an)	37662	22 337

Prepositions (P+BA+BEI+SB+LB) (p)	35 877	42 114
Adverbs (AD) (d+dg)	30 940	22 854
Conjunctions (CC+CS) (c+cg)	30 770	35 096
Measure words (M) (q+qg)	27 596	29 415
Pronouns (PN+DT) (r+rg)	18 524	19 892
Particles (AS+MSP+SP) (了+着+过+所+而+以)	3 124	4 165
Punctuation (PU) (ew+w)	151 566	139 113

Table 3
The POS Combined Together.

1.5 *Length of a Sentence*

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*. As our previous research proved,[12] the length of a sentence might be an auxiliary indicator of the register variation.

As for the Hanku corpus, we use data from the previous research. The average length of a sentence in the Hanku is 29 tokens. It should be noted though that the number of tokens also include punctuation (the POS tag »PU«), e.g. ",、。()", therefore the length in words is shorter.

Regarding the legal Chinese in the Sihanku corpus, the average length of a sentence is simply calculated by dividing the number of tokens by the number of sentences in the sub-corpus. The average length of a sentence is then 25 tokens including e.g. symbols and non-sentential punctuation.

1.6 *Conclusion*

As it is obvious from the statistical data presented in this chapter; whereas both corpora use the different software solution for the tokenization and the POS tagging, proportions of the POS are very similar. It means that when comparing results from corpora with an unequal size, the size of a corpus alone does not play

12 Ľuboš Gajdoš, »Chinese Legal Texts—Quantitative Description«, *Acta Linguistica Asiatica* 7,1 (2017), 77–87.

a very important role in the proportion of the POS (content words vs. function words). This conclusion needs to be viewed as interim rather than final. Generally speaking, a small-sized corpus may provide basic statistical data for a quantitative description, and so provide a basis for further study. However, one should be aware of the limitations of this approach, for example, significant differences in the automatic POS tagging.[13]

By the frequency of occurrence of a given POS or a token, the size of a corpus may indeed lead to more relevant data. Regarding the length of a sentence, both corpora show different numbers. Further research in this area is required to reach more accurate data.

2 *Different Sub-corpora in the Corpus*

In this chapter, we compare two different language registers—legal Chinese versus an unstructured material of the general corpus Hanku. By comparing the registers, we provide the same statistical data as in the previous chapter, namely most frequent content words, function words, length of sentence plus word-length.

2.1 *The Sub-corpus Web-zh*

The subcorpus Web-zh is part of the corpus Hanku which is a monolingual, synchronous Chinese corpus (in simplified Chinese characters) available via web interface. It is available via the website of the Confucius Institute at Comenius University in Bratislava at: <konfuciovinstitut.sk>. The subcorpus has a size of 744 million tokens (August 2017).

Due to the lack of any other optimal solution, the Web-zh sub-corpus is in this study used as a general corpus in this study, [14] nevertheless, the only parameter for making this decision is the size of the sub-corpus. The language data of the sub-corpus are results of web crawling.

13 See e.g. the different approach to adjectives by both POS taggers—the most frequent token *youguan* 有关 is tagged as »JJ« in the Hanku corpus and as »vn« in the Sihanku.

14 For more on balance, representativeness and comparability of corpora, see Tony McEnery and Andrew Hardie, *Corpus Linguistics Method, Theory and Practice* (Cambridge: Cambridge University Press, 2012), 10–11.

2.2 *Most Frequent Words within one Corpus*

By comparing most frequent words within two sub-corpora of the same corpus, the Sketch Engine provides a very useful tool—Word list comparison (Word list options, Output type, Keywords). The results are shown in table 4 below.

Token	Freq in the Law-zh	Freq in the Web-zh
条	14276	389
应当	7954	101
规定	7005	191
和	12011	5239
或者	6576	248
管理	6701	536
部门	5757	261
机构	4836	215
行政	4712	101
人民	4413	258
有关	3922	190
单位	3891	199
本	4378	785
企业	4150	793
国家	3707	497
人员	3343	312
由	3725	695
申请	2917	92
其他	3055	397
机关	2673	94

Table 4
Most Frequent Words in the Law-zh Sub-corpus.

Based on empirical data, one would expect significant differences in the frequencies of tokens within different registers, in our case—legal texts vs. other registers. If these frequencies are compared for a maximum difference, most typical tokens for the given register—in this case legal texts—may be found. The maximum frequency difference is shown in the next table. This is a very simple and easy method to show the discrepancy in the lexis of different registers. The following table demonstrates the maximum frequency difference of a token in both registers.

Token	Freq in the Law-zh	Freq in the Zh-web	Difference in freq
条	14276	389	13887
应当	7954	101	7853
规定	7005	191	6814
和	12011	5239	6772
或者	6576	248	6328
管理	6701	536	6165
部门	5757	261	5496
机构	4836	215	4622
行政	4712	101	4611
人民	4413	258	4155
有关	3922	190	3732
单位	3891	199	3692
本	4378	785	3593
企业	4150	793	3357
国家	3707	497	3210
人员	3343	312	3031
由	3725	695	3030
申请	2917	92	2825
其他	3055	397	2659
对	5300	2691	2609

Table 5

Maximum Positive Difference in Frequency from the Perspective of the Law-zh.

Here, we would like to point out that this research has been conducted only on the 1 000 most frequent words. It is interesting to see that the most frequent words in legal Chinese are, at the same time, the most typical words for the register. What other information may be observed from the statistical data? Except for some formal means, such as the classifier *tiao* 条 which marks a section of legal code (§); there are some typical words of legal texts, e.g. the modal verb *yingdang* 应当 ('should') is almost exclusively used in legal texts for deontic modality, some function words are several times more frequent in legal texts (the conjunction *he* 和, 'and', or the preposition *you* 由, 'by, via', just to mention a few) etc.

If we conduct the observation of frequency from the other side of the spectrum (from the side of the general corpus), it seems that the occurrence of words that are typical for the general register, is very rare in legal texts (see table 6).

Token	Freq in the Web-zh	Freq in the Law-zh	Freq difference
是	13426	1404	12022
了	10226	246	9980
一	10639	3989	6650
我	5529	185	5344
这	5352	45	5307
个	6208	1273	4935
就	4892	196	4696
你	3833	9	3824
也	4052	306	3747
不	9911	6283	3627
在	9211	5665	3546
他	3178	64	3114
都	3171	111	3060
上	4076	1190	2886
人	4038	1176	2862
着	2498	33	2466

有	5541	3204	2337
大	2552	249	2303
要	2824	605	2219
说	2228	29	2199

Table 6
Maximum Difference in Frequency from the Perspective of the Web-zh.

2.3 *Proportion of a Part-of-Speech*

As the situation in one corpus differs from the situation described in the chapter 1.4, we compare two sub-corpora of the same corpus for the POS proportion.

Our previous research has revealed that the ratios between some POS of different registers may vary and might be a good indicator of the tendency of texts to be more formal or colloquial.[15] In this respect, the ratio of nouns vs. verbs and nouns vs. pronouns seems to be relevant. The next table depicts the proportions of various POS in the given registers.

POS	Freq in the Zh-law	Freq in the Web-zh	Zh-law %	Web-zh %
Nouns (NN+NR+LC+NT)	416132	338955	42	34
Verbs (VV+VC+VE)	154657	185434	16	19
Particles DE (DEC+DEG+DEV+DER)	48209	47916	5	5
Numbers (CD+OD)	41884	41311	4	4
Adjectives (VA)	5387	17282	1	2
Adjectives (JJ)	32274	26018	3	3
Prepositions (P+BA+BEI+SB+LB)	35877	31438	4	3
Adverbs (AD)	30940	87763	3	9
Conjunctions (CC+CS)	30770	12977	3	1
Measure words (M)	27596	28203	3	3

15 See Gajdoš, »Quantitative Description of Written Chinese«, 68–69.

Pronouns (PN+DT)	18524	41106	2	4
Particles (AS+MSP+SP)	3124	18251	0,3	2
Punctuation (PU)	151566	121657	15	12

Table 7
The Proportion of the POS.

It is apparent that nouns are more preferred in legal Chinese than in the Web-zh corpus, some other POS are less frequently used in legal Chinese, e.g. adjectives (VA), adverbs, pronouns and some are almost absent—particles.

Let us compare the proportions of nouns and verbs, i.e. the ratio R1:[16]
legal: R1 = nouns/verbs = 42/16 = 2,62
general: R1 = nouns/verbs = 34/19 = 1,79

R2 is the ratio between nouns and pronouns:
legal: R2 = nouns/pronouns = 42/2 = 21
general: R2 = nouns/pronouns = 34/4 = 8,5

From the table above, a third ratio may be added, namely verbs vs. adverbs:
legal: R3 = verbs/adverbs = 16/3 = 5,4
general: R3 = verbs/adverbs = 19/9 = 2,12

If we consider legal texts as a closed register (variety) and the unstructured language material of the general corpus as open, the R3 indicators may, to some extent, reflect the differences. On the other hand, for the complete description of register variation, other factors and language registers (e.g. of newspapers, literature) must be added—for example, the absence of modal particles,

16 The linguistics literature indicates that professional texts are more information-saturated, that is, nouns prevail over other POS words in these texts. On the other hand, it also argues that informal texts (speech) contain more verbs compared to written language. It is also stated that e.g. in Slovak the frequency ratio between nouns vs. pronouns may indicate the registers of professional texts (the ratio in favour of the former) and informal texts (vice versa). For more information, see Ján Findra, *Štylistika slovenčiny* [Stylistics of Slovak] (Martin: Osveta, 2004), 61–62.

onomatopoeias and interjections might serve as auxiliary indicators of formal texts. Only then, one may consider the quantitative description as detailed and accurate.

2.4 *Length of a Sentence*

As we have already mentioned in the chapter 1.5, the average length of a sentence for legal texts in Chinese is in the interval between 25 to 29 tokens, considering the second figure as more precise.

The average length for the sub-corpus Web-zh is surprisingly even longer with the figure of 32. As there is no available analysis of the language data in the Web-zh sub-corpus, we are not able to evaluate these data; but our previous research in this field has shown that the more colloquial the text is, the shorter is the average length of a sentence with the figures of 16 and 19 words for literary texts and dialogues, respectively.[17]

2.5 *Word-length in Different Registers*

Before quantitative analysis began in the linguistics research, it was generally believed that the length of a word (token) in Chinese varied across language registers so that the word-length might, to some extent, correlate with the language variety.[18]

In this chapter, we compare word-length preferences for different language registers.[19] I do not compare the so-called synonymous pairs of words (mono- vs. disyllabic), but only the preference of the POS in 2 sub-corpora. As there are some POS which are not very frequent and mostly have only one form (modal particles); we only compare nouns, verbs, adjectives, conjunctions, adverbs and prepositions. The next table shows the results.

We might also compare the data for the different corpora, but as the word-length is very sensitive to tokenization, we will conduct this research in the future.

17 See Gajdoš, »Quantitative Description of Written Chinese«, 67–68. In this article, the length of a sentence was counted in words, not in tokens.

18 See e.g. San Duanmu, »Word-length preferences in Chinese: a corpus study«, *East Asian Linguist*, 21 (2012), 89–114.

19 The CQL query, for instance, for all disyllabic nouns is: [tag="NN" & word="(?i)..."].

	zh-law (IPM)	Web-zh (IPM)	Zh-law %	Web-zh %
Nouns (only NN)	389236	272974		
monosyllabic nouns	13665	28892	4	11
2-syllabic	331503	209410	85	77
3-syllabic	43020	29519	11	11
4-syllabic	942	2404	0.2	1
Verbs (only VV)	148562	163752		
monosyllabic	19757	53049	13	32
2-syllabic	127324	105146	86	64
3-syllabic V	732	1967	0.5	1
Adjectives (VA)	5387	17282		
monosyllabic	826	5995	15	35
2-syllabic	4529	11029	84	64
3-syllabic	20	156	0.4	1
Adjectives (JJ)	32274	26018		
monosyllabic	2658	9459	8	36
2-syllabic	28005	14923	87	57
3-syllabic	1478	1132	4.6	4
Prepositions (P)	33614	28084		
monosyllabic	26857	23694	80	84
2-syllabic P	6752	4336	20	15
Adverbs (AD)	30940	87764		
monosyllabic	15989	54836	52	62
2-syllabic	14654	31477	47	36
3-syllabic	279	1359	1	2
Conjunctions (CC)	30504	10899		
monosyllabic	22727	9573	75	88

2-syllabic	7756	1210	25	11
Conjunctions (CS)	266	2077		
monosyllabic	115	443	43	21
2-syllabic	150	1630	56	78
Pronouns (PN)	4709	26065		
monosyllabic	3545	18190	75	70
2-syllabic	1162	7758	25	30

Table 8
Word-length Preference in Subcorpora.

In terms of the word-length preference in legal texts, there are no considerable (clear-cut) differences between both sub-corpora. However, there are some differences that may become more evident from the next chart.

From the chart presented below, it can be seen that there is a stronger preference for disyllabic words in legal texts and a tendency for monosyllabic words in the general corpus, that is to say, there is a positive correlation between the proportions of disyllabic words in legal texts. The chart also shows that this tendency is in positive numbers for the Zh-law and in negative numbers for the Web-zh. By the 3- and 4-syllabic words, the differences are minimal (from 0 to 1 percent). From a word-length comparison, the tags CS (subordinating conjunctions) and PN (pronouns) as well as the 3- and 4-syllabic words are excluded.[20]

20 Subordinating conjuctions CS have shown very low frequency (266 ipm compared to 30 504 ipm of coordinating conjuctions CC). As for the pronouns, the pronoun *qi* 其 represents 67 % of all pronouns, and thus affects the statistical data used for comparison.

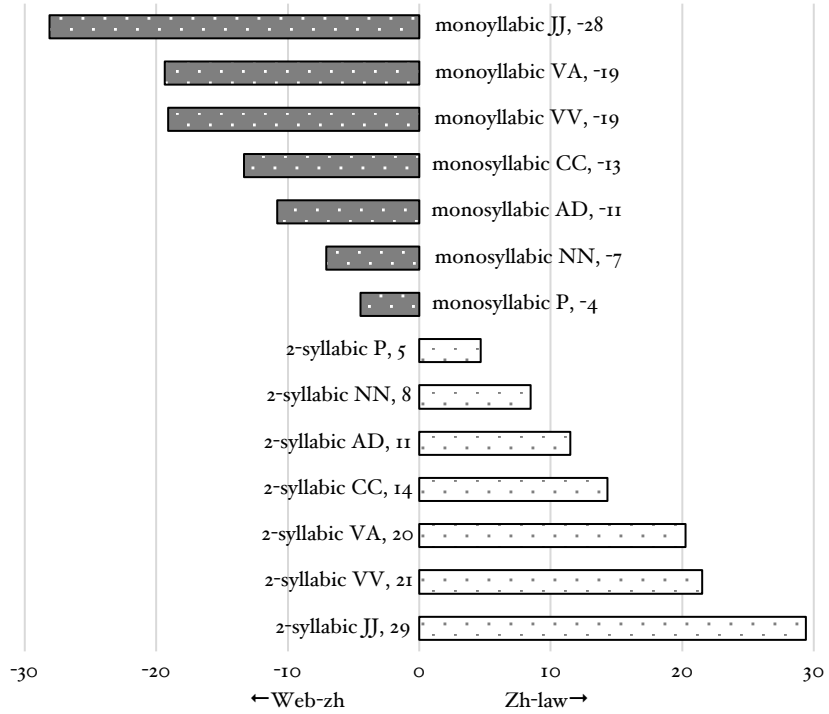

Chart 1

Differences between Word-length in both Sub-corpora.

2.6 *Conclusion*

The problem by comparing between 2 or more registers is the so-called general corpus. Due to the lack of such a corpus, we have used only the available Web-zh corpus. Quantitative description and comparison are then limited to the 'quality' of the corpora used. From the statistical data presented in the 2nd chapter, we may only outline statistical data for the legal texts. To obtain a precise measurement of statistical data, a balanced corpus is needed. Nevertheless, despite these limitations and potential problems, we personally believe that the survey data from our study is helpful in describing register variation in Chinese using a quantitative approach.

3 *Final Thought*

This study offers quantitative description of legal Chinese mostly by corpus-based approach.

Based on the statistical data from 2 corpora with different word segmentation and POS labelling but with similar language data, it may be concluded that despite the very different sizes (the Hanku sub-corpus Zh-law is 15 times larger), we have obtained very similar statistical values for the POS, and to some extent, for the frequency therefore for both the corpora. It can also be seen that the slightly different frequency values may be explained by different approach to the word segmentation and the POS tagging. From the point of building a small corpus, this is a very positive result. On the other hand, the larger a corpus is, the more accurate is the statistical description. In the future, it will be very interesting to compare the statistics of the sub-corpus Zh-law with that of a corpus of legal texts, which may be 10 times larger.

To conclude, it has been the intention of this paper to work with statistical data obtained from the corpus Hanku, Sihanku, and to provide empirical evidence for the quantitative description of legal Chinese. The following parameters may help to identify the register of legal Chinese:

- the average sentence-length preference is about 29 tokens (including punctuations)
- strong preference for information-saturation indicated by the presence of nouns and verbs in 2:1 ratio / where there are at least twice as many nouns as verbs
- absence of onomatopoeias, very low usage of particles
- a significant divergence between the frequency of adverbs and pronouns versus verbs and nouns, respectively
- the use of adjectives (VA) is only half of that seen in the unstructured material of the general corpus
- a stronger preference for disyllabic words across most examined part-of-speech.

For further research, it is very important and relevant to have a general corpus and sub-corpora of other language registers as well. Only in this way will it be possible to objectify the differences between registers and produce sufficient statistical evidence for the discrepancy. Another important aspect of describing

the language registers is the identification of specific parameters such as typical context of a given word, word collocations and sentence patterns.

Comenius University in Bratislava, Department of East Asian Studies