ARE WE DIFFERENT? MALE AND FEMALE "REGISTERS" IN CHINESE CORPUS DATA

https://doi.org/10.31577/aassav.2022.31.1.08

Ľuboš GAJDOŠ
Department of East Asian Studies, Faculty of Arts,
Department of German, Dutch and Scandinavian Studies, Comenius University,
Gondova 2, 811 02 Bratislava, Slovakia
lubos.gajdos@uniba.sk

Elena GAJDOŠOVÁ
Department of East Asian Studies, Faculty of Arts,
Department of German, Dutch and Scandinavian Studies, Comenius University,
Gondova 2, 811 02 Bratislava, Slovakia
gajdosova137@uniba.sk

The aim of this paper is to examine language data with regard to potential differences between male and female registers. Corpus linguistics is used as the basic methodological approach (mostly quantitative) to the topic and the *Hanku* corpus, more particularly the subcorpus *Litchi*, are used as the primary source of language data. The data are presented in the form of tables together with a brief analysis. The results indicate a considerable variation between male vs. female registers in some areas – lexicon, part-of-speech proportion etc. However in other areas (e.g. prosody) there exists no deviation at all. These indicators of variation will be subject of further, more detailed research.

Keywords: Chinese language, corpus, corpus linguistics, male register, female register, literary text

We all know, or at least firmly believe, that we are different individuals. Yet big data analysis also shows an opposite tendency: our lives, behaviour, language etc. conform to typical patterns or a predictable manner.

Let us move away from generalities and concentrate on language only. As for the first assumption, this is nothing new. Studies in stylometry, for example, have shown via a statistical approach that the "language" of every individual possesses certain distinctive features which may differentiate it from another individual. The sum of the linguistic properties of a particular person constitutes a so-called idiolect. On the

other hand, in some fields of language study the opposite tendencies are exhibited, that is to say, there may exist some (typical) properties of a group of individuals. To examine and challenge this assumption, this article, by using the methods of corpus linguistics, seeks evidence for the claim that a female group of authors possesses some unique features when compared to a male group and *vice versa*.

Before corpus linguistics was applied to this field of research, there were some other methods to address this kind of problem. Let us then discuss some issues related to the given subject from the perspective of the methodology used:

1. The data presented in this paper represent only the language of literary texts of the subcorpus *Litchi*.[1]
2. Male and female "registers"[2] may be manifested in different ways for different topics, that is to say, the language data presented here depend on topics as well.
3. The error rate of automatic tools (part-of-speech annotation, tokenization etc.) must be considered.
4. The results may not be generalized without further research.

1 The Criteria

As a corpus and corpus linguistics methods are used as the primary source and methodology in this study, there are certain criteria that may be used to gather empirical evidence on the subject.

To begin with, let us assume that there are some differences which might be reflected via quantitative data. The most basic indicators are probably (1) the length of a sentence and (2) the length of words. Using part-of-speech (hereafter POS) annotation brings us to the next criterion, namely the proportion of POS tags. This may be further broken down into the concrete part of speech, e.g. comparison of most frequent (3) verbs/adjectives, (4) nouns and (5) rest of the POS tags. At the end, the list of the most frequent words for the male and female subcorpora is presented.

Throughout the study, the *Hanku*[3] corpus and its subcorpora *zh-lit-1.1* are used; corpus queries are written in the CQL (Corpus Query Language). Frequencies, values, percentage corresponding to male authors are suffixed _M, female authors with _F.

[1] See PETROVČIČ, M., GARABÍK, R., GAJDOŠ, Ľ. The New Chinese Corpus of Literary Texts Litchi. In *Acta Linguistica Asiatica*, 2020, Vol. 10, No.2, pp. 65–81.
[2] See also BIBER, D. *Dimensions of register variation*, pp. 1–27.
[3] See also GAJDOŠ, Ľ., GARABÍK, R., BENICKÁ, J. The New Chinese Webcorpus Hanku – Origin, Parameters, Usage. In *Studia Orientalia Slovaca*, 2016, Vol. 15, No. 1, pp. 21–33.

Table 1. Basic parameters of the subcorpus zh-lit-1.1

	Whole subcorpus	Male	Female	N/A
Size in tokens	81398762	42427398	27295164	11676200
Only authors of Chinese origin	58419868	30909203	21461069	6049596

2 The Length of a Sentence

First, let us compare the length of a sentence (*Ls*). This is a quite straightforward operation, namely division of all tokens by the number of sentences.[4] Formula:

$$Ls = \frac{tokens}{number\ of\ sentence}$$

$$Ls\_M = \frac{30\ 909\ 203}{555\ 933} \qquad Ls\_F = \frac{21\ 461\ 069}{697\ 434}$$

Ls_M = 55.6
Ls_F = 30.8

As can be seen, the male authors prefer to use longer sentence. For comparison, the length of a sentence in the subcorpus *zh-law* is 29 tokens or 32 in the webcorpus *web-zh*.[5]

3 Length of Words

Now, let us compare the length of words in different corpora, namely the *web-zh* and the *zh-lit-1.1* (Litchi). As the tokens of punctation (PU), numbers (CD, OD) and foreign words (FW) might influence the length preferences, they are excluded from the queries ([tag!="PU|CD|OD|FW"]).

[4] For this purpose we use the following queries: number of tokens in male or female subcorpora: [word=".*"] within <doc gender="F"/> within <doc authors_origin="CN"/>, number of sentences: <s/> within <doc gender="M"/> within <doc authors_origin="CN"/>.

[5] See more GAJDOŠ, Ľ. Chinese legal texts – Quantitative Description. In *Acta Linguistica Asiatica*, 2017, Vol. 7, No. 1, pp. 77–87.

Table 2. The length of words

	web-zh [word=".*"]	zh-lit-1.1 [word=".*"] within <doc authors_origin ="CN"/>	zh-lit-1.1 within <doc gender="M"/> within <doc authors_origin ="CN"/>	zh-lit-1.1 within <doc gender="F"/> within <doc authors_origin ="CN"/>
Tokens [word=".*"]	744721698	58419868	30909203	21461069
Tokens without PU, CD, OD [word=".*" & tag!= "PU|CD|OD|FW"]	623269903	52213106	28348252	18517925
Percentage	84%	89%	92%	86%
Monosyllabic without PU, CD, OD [word=".{1}" & tag!= "PU|CD|OD|FW"]	260624380	29539330	15782928	10740684
Percentage	42%	57%	56%	58%
Disyllabic without PU, CD, OD [word=".{2}" & tag!= "PU|CD|OD|FW"]	312998436	20165381	11085960	6982389
Percentage	50%	39%	39%	38%
Trisyllabic [word=".{3}" & tag!= "PU|CD|OD|FW"]	36993508	1900318	1141904	577924
Percentage	6%	4%	4%	3%
More than 4 syllabic [word=".{4,} " & tag!= "PU|CD|OD|FW"]	12653571	608077	337460	216928
Percentage	2%	1%	1%	1%

It is obvious from the table above that there are some differences between the subcorpora. However, the percentages between male and female authors hardly vary at all, that is to say, there is no evidence for word-length preference between male and female authors. Again, differences are only slight: the use of

punctuation, numbers, foreign words (6%) and the use of monosyllabic words (2%).

Now let us compare the language of translations (for the sake of simplicity, all literary works are considered as translations; except authors_origin!= “CN|TW”,) with those of Chinese origin (authors_origin=“CN|TW”).

Table 3. The comparison of the length of words for native authors and translations

	Frequency web-zh [word=".*"]	Frequency zh-lit-1.1 [word=".*"] within <doc authors_origin="CN|TW"/>	Frequency zh-lit-1.1 within <doc authors_origin!="CN|TW"/>
Tokens [word=".*"]	744721698	60683608	20715154
Tokens without PU, CD, OD [word=".*" & tag!="PU|CD|OD|FW"]	623269903	54306966	18856784
Percentage	84%	89%	91%
Monosyllabic without PU, CD, OD [word=".{1}" & tag!="PU|CD|OD|FW"]	260624380	30747198	10783743
Percentage	42%	57%	57%
Disyllabic without PU, CD, OD [word=".{2}" & tag!="PU|CD|OD|FW"]	312998436	20942867	7171848
Percentage	50%	39%	38%
Trisyllabic [word=".{3}" & tag!="PU|CD|OD|FW"]	36993508	1977121	647765
Percentage	6%	4%	3%
More than 4 syllabic [word=".{4,}" & tag!="PU|CD|OD|FW"]	12653571	536161	253428
Percentage	2%	1%	1%

As in the previous comparison, there are no pronounced differences here. This is just one parameter of many, and further research may reveal the existence of the register variation observed in other languages.

4 Part of Speech

Now let us observe part of speech (POS) variation. There are 52 tags in the *Hanku* corpus which correspond to part of speech.[6] As Petrovčič states, the key *doc.gender* represents the gender of the author but not the translator. That is why only male/female author from CN[7] is chosen.

The following table shows the proportion of the POS tags in the subcorpus *zh-lit-1.1* for Chinese authors only (CN). Also notice that the absolute frequencies are calculated for the first 10 million tokens in the subcorpus.

Table 4. The proportion of POS tags in the subcorpus zh-lit-1.1 CN.

Tag	Absolute frequency
VV verb	2008312
NN noun	1802656
AD adverb	1261847
PU punctuation	731558
PN pronoun	668587
AS aspect particle	332242
M measure word	317705
NR proper nouns	292285
P preposition	283170
CD cardinal number	282714
DEC particle DE 的	252117
VA predicative adjective	242028
DEG particle DE 的	235303
LC localizer	202666
DT determiner	199851
VC copula	196509
JJ noun-modifier	146354
VE verb [to have]	101758

[6] For more details see also GAJDOŠ, Ľ., GARABÍK, R., BENICKÁ, J. The New Chinese Webcorpus Hanku – Origin, Parameters, Usage. In *Studia Orientalia Slovaca*, 2016, Vol. 15, No. 1, pp. 21–33.

[7] PETROVČIČ, M., GARABÍK, R., GAJDOŠ, Ľ. The New Chinese Corpus of Literary Texts Litchi. In *Acta Linguistica Asiatica,* 2020, Vol. 10, No. 2, pp. 65–81.

Tag	Absolute frequency
SP sentence-final particle	77917
CC coordinating conjunction	68055
NT temporal noun	65283
DEV particle DE 地	54987
BA ba-construction	34892
MSP another particle	32152
CS subordinating conjunction	31583
DER particle 得	31389
OD ordinal number	12079
SB short bei-construction	11724
LB long bei-construction	8703
IJ interjection	5434
ETC etcetera	4921
FW foreign word	3219

The differences of the IPM (Instance Per Million)[8] and relative differences[9] are calculated in the following table. Let us calculate the relative percentage difference *df* of frequencies for each POS tag. The results are presented in the following table.

[8] The proportion of POS percentage is a division of IPM by 1000. The difference of percentage is calculated by subtraction.

[9] To calculate the relative percentage difference between male and female IPM frequencies, we have modified the formulae for relative difference $dr = \frac{|x-y|}{(\frac{|x+y|}{2})}$ to highlight the differences $df = \frac{x-y}{(\frac{|x+y|}{2})} \times 100$, i.e. the positive percentages are male deviations from the arithmetical mean and *vice versa*. As the texts in corpus *zh-lit-1.1* have only male or female value ("N/A" are still texts from male and female authors), we use arithmetic means in the denominator. It is necessary to bear in mind that the difference df (%) does not consider absolute frequency of a given POS tag.

Table 5. Frequency difference and relative difference between the male and female subcorpus

Tag	IPM_M	IPM_F	Frequency difference	dif
OD	1382,47	960,62	421,85	36,01
MSP	3607,73	2566,79	1040,94	33,72
ETC	562,91	405,80	157,10	32,44
CC	7475,54	5404,81	2070,73	32,15
NR	32079,22	24164,41	7914,81	28,14
CD	32262,79	25397,15	6865,64	23,81
JJ	16561,12	13059,41	3501,70	23,64
DT	21359,43	17118,25	4241,18	22,04
DEG	24929,92	20510,86	4419,06	19,45
M	34762,43	28938,77	5823,66	18,28
NN	196740,14	168558,98	28181,16	15,43
VC	20601,86	18120,63	2481,23	12,82
P	29071,12	25803,79	3267,32	11,91
IJ	390,53	346,72	43,81	11,88
VE	10771,06	9637,73	1133,33	11,11
BA	3536,20	3238,47	297,73	8,79
LC	21526,08	19741,00	1785,08	8,65
DEC	25656,79	23799,93	1856,86	7,51
NT	6624,27	6149,83	474,44	7,43
CS	3172,45	2958,47	213,98	6,98
VV	201896,79	200797,31	1099,48	0,55
AS	33453,60	33350,44	103,16	0,31
AD	125466,55	126915,53	-1448,98	-1,15
SB	1186,86	1214,15	-27,29	-2,27
DEV	5337,57	5601,86	-264,30	-4,83
VA	22785,45	25489,64	-2704,20	-11,20
PN	58272,71	67669,46	-9396,75	-14,92
DER	2711,10	3208,37	-497,27	-16,80
SP	5785,82	7039,63	-1253,81	-19,55
LB	820,76	1050,18	-229,42	-24,52
FW	281,73	430,64	-148,91	-41,81
PU	48927,01	110350,33	-61423,31	-77,13

The table above reveals some discrepancy (according to gender) in the proportions of the POS tags in the *Litchi* subcorpus.

To conclude, from the above data one may assume that female authors (compared to male authors) use more punctuation (PU), sentence particles (SP), pronouns (PN), adjectives (VA) etc. and this might be described, in the context of lexis, as more emotional or personal. The differences are more noticeable in functional words (*xūcí* 虚词) and less in the group of notional words (*shící* 实词).

4.1 Most frequent verbs and adjectives

When comparing concrete words (tokens) two different metrics are used: IPM difference (frequency difference) and relative percentage difference *df*. The former might (to some extent) be compared to keywords in Sketch Engine[10]: the redder the word in the difference row, the more it is used by male authors and *vice versa*. The latter may in some cases be more relevant (see for example the token *ài* 爱 [to love]). In this and the following sections the IPM measure is used.

Let us start with verbs and adjectives[11] which are the most frequent POS of the whole corpus. CQL query:

[tag="VV|VA|VC|VE"] within <doc authors_origin="CN"/>

Table 6. The most frequent verbs in the male subcorpus compared to the female subcorpus

Token	Hanyu pinyin[12]	Trans-lation[13]	IPM_M	IPM_F	Frequency difference	dif
说道	shuōdao	say	449,19	135,50	313,69	107,30
能够	nénggòu	can	374,35	117,84	256,51	104,23
出现	chūxiàn	appear	489,72	226,97	262,76	73,32
使	shǐ	cause	387,17	187,41	199,76	69,53
当	dāng	become	389,75	245,09	144,66	45,57

[10] See more KILGARIFF, A. Simple Maths for Keywords. [online] [cit. 12 September 2021]. Available from https://www.sketchengine.eu/documentation/simple-maths/.
[11] So-called "adjectives" may function as predicates and have the POS tag "VA".
[12] The abbreviated name of *Hanyu pinyin fang'an* is used throughout the study.
[13] Many words in Chinese are polysemic in nature but, owing to limitations of space, only one translation is provided.

为	wèi	be	576,53	371,79	204,74	43,18
写	xiě	write	348,41	233,77	114,64	39,38
出	chū	go out	952,08	702,06	250,02	30,23
可能	kěnéng	may	615,51	454,26	161,25	30,15
死	sǐ	die	497,75	370,21	127,54	29,39
大	dà	big	576,79	431,62	145,17	28,79
发现	fāxiàn	find	491,67	382,13	109,53	25,07
到	dào	reach	2272,88	1770,60	502,28	24,84
了	liǎo	finish	302,14	239,36	62,78	23,19
成	chéng	become	406,06	322,63	83,43	22,90
说	shuō	say	7666,68	6101,42	1565,26	22,74
明白	míngbái	understand	344,04	275,71	68,33	22,05
应该	yīnggāi	should	503,41	411,40	92,01	20,12
没有	méiyǒu	not have	2987,01	2454,03	532,98	19,59
有	yǒu	have	7411,45	6236,50	1174,95	17,22
清楚	qīngchǔ	clear	271,70	235,22	36,48	14,39
能	néng	can	2979,15	2637,89	341,25	12,15
是	shì	be	20093,21	17828,38	2264,83	11,94
可	kě	can	1636,41	1482,27	154,14	9,89
要	yào	want	4272,16	3916,39	355,76	8,69
算	suàn	count	378,56	349,84	28,72	7,89
敢	gǎn	dare	604,35	563,49	40,87	7,00
需要	xūyào	need	272,09	257,44	14,64	5,53
无	wú	not have	759,29	718,79	40,50	5,48
喝	hē	drink	358,79	340,48	18,32	5,24
上	shàng	go up	434,08	412,47	21,61	5,11
多	duō	many	579,89	551,46	28,43	5,03
来	lái	come	1984,13	1887,93	96,20	4,97
开始	kāishǐ	start	533,34	508,83	24,51	4,70
感觉	gǎnjué	feel	266,98	257,72	9,25	3,53
离开	líkāi	leave	329,74	324,17	5,57	1,70
起来	qǐlái	get up	1253,02	1234,05	18,98	1,53
行	xíng	that's ok	312,95	310,10	2,85	0,92
让	ràng	let	1959,87	1963,93	-4,06	-0,21
起	qǐ	rise	429,55	433,62	-4,07	-0,94

可以	kěyǐ	may	995,98	1009,50	-13,52	-1,35
想到	xiǎngdào	think	497,23	507,34	-10,11	-2,01
该	gāi	should	372,45	380,74	-8,29	-2,20
请	qǐng	please	375,16	386,56	-11,40	-2,99
令	lìng	order	260,96	269,37	-8,41	-3,17
问	wèn	ask	908,34	940,31	-31,97	-3,46
下来	xiàlái	come down	624,80	646,99	-22,19	-3,49
用	yòng	use	785,49	816,59	-31,10	-3,88
告诉	gàosù	tell	378,66	393,78	-15,13	-3,92
般	bān	like	270,21	282,42	-12,21	-4,42
开	kāi	open	351,48	368,99	-17,51	-4,86
见	jiàn	see	1035,29	1104,98	-69,69	-6,51
打	dǎ	hit	675,53	722,38	-46,85	-6,70
给	gěi	give	594,45	642,70	-48,25	-7,80
快	kuài	fast	398,04	430,64	-32,60	-7,87
找	zhǎo	look for	672,55	728,99	-56,44	-8,05
知道	zhīdào	know	1891,67	2057,45	-165,78	-8,40
出来	chūlái	come out	991,06	1084,66	-93,60	-9,02
去	qù	go	2054,47	2263,87	-209,40	-9,70
叫	jiào	call	669,41	741,16	-71,74	-10,17
继续	jìxù	continue	269,08	300,82	-31,75	-11,14
一样	yīyàng	equally	516,67	577,98	-61,30	-11,20
下去	xiàqù	go down	303,50	340,52	-37,02	-11,50
知	zhī	know	581,57	654,49	-72,91	-11,80
会	huì	may	2924,08	3334,74	-410,66	-13,12
怕	pà	fear	439,29	505,89	-66,61	-14,09
带	dài	bring	522,69	604,02	-81,33	-14,44
得	dé	get	901,64	1046,36	-144,72	-14,86
过	guò	pass	490,31	572,62	-82,31	-15,49
想	xiǎng	think	2439,18	2850,14	-410,96	-15,54
放	fàng	release	318,00	382,46	-64,46	-18,41
住	zhù	live	672,52	812,12	-139,60	-18,81
坐	zuò	sit	605,16	732,68	-127,52	-19,06
听	tīng	hear	819,21	1003,72	-184,52	-20,24
回来	huílái	come back	409,00	501,28	-92,28	-20,27

望	wàng	look towards	292,02	358,56	-66,54	-20,46
看看	kànkan	have a look	289,53	355,53	-66,00	-20,46
站	zhàn	stand	476,36	585,67	-109,30	-20,58
吃	chī	eat	638,09	789,62	-151,52	-21,23
走	zǒu	go	997,83	1240,48	-242,65	-21,68
做	zuò	do	1093,33	1365,87	-272,54	-22,16
看	kàn	see	3239,16	4111,68	-872,51	-23,74
出去	chūqù	get out	336,21	440,19	-103,98	-26,79
看见	kànjiàn	see	282,63	377,33	-94,70	-28,70
听到	tīngdao	hear	245,07	334,93	-89,86	-30,99
买	mǎi	buy	217,70	298,59	-80,89	-31,33
说话	shuōhuà	speak	327,77	456,13	-128,36	-32,75
没	méi	not have	1373,35	1912,91	-539,56	-32,84
看到	kàndao	notice	564,69	790,64	-225,95	-33,34
以为	yǐwéi	think	259,02	363,87	-104,85	-33,67
好	hǎo	good	1224,42	1742,97	-518,55	-34,95
拿	ná	take	333,33	485,16	-151,83	-37,10
过来	guòlái	come over	435,79	644,33	-208,54	-38,61
觉得	juéde	feel	632,47	940,68	-308,21	-39,18
笑	xiào	laugh	1037,23	1579,32	-542,09	-41,44
等	děng	wait	221,23	344,11	-122,88	-43,47
帮	bāng	help	213,04	340,99	-127,95	-46,19
爱	ài	love	196,51	373,65	-177,14	-62,14
喜欢	xǐhuan	like	274,74	578,26	-303,52	-71,16
道	dào	say	388,72	1162,76	-774,04	-99,78

The results show that some verbs are preferred by male or by female authors. The redder the cell is the more male authors use these words and *vice versa*. For example, female authors more often use verbs such as *xǐhuan* 喜欢 [to like], *ài* 爱 [to love], *xiào* 笑 [to smile], *juéde* 觉得 [to feel] etc.

4.2 Most frequent nouns

As in the previous section, let us compare concrete nouns in the male and female subcorpus. To avoid specific tokens of particular literary works, the following CQL query is used:

[tag="NN|NT"] within <doc authors_origin="CN"/>

The original intention was to compare the 100 most frequent nouns in both subcorpora. However, because of some noisy data (inadequate tokenization and POS annotation) only 99 are presented.

Table 7. Most frequent nouns in the male subcorpus compared to the female subcorpus

Token	Hanyu pinyin	Translation	IPM_M	IPM_F	Frequency difference	dif
书记	shūjì	secretary	282,93	47,29	235,63	142,71
人物	rénwù	person	243,52	50,74	192,78	131,02
力量	lìliàng	strength	297,16	87,23	209,93	109,23
情况	qíngkuàng	situation	402,40	133,87	268,53	100,15
剑	jiàn	sword	272,80	97,62	175,18	94,59
道	dào	way	229,32	126,09	103,23	58,09
问题	wèntí	problem	587,14	334,84	252,30	54,73
其中	qízhōng	among	217,18	124,83	92,35	54,01
工作	gōngzuò	work	429,74	261,64	168,11	48,63
山	shān	mountain	217,57	133,17	84,40	48,13
事情	shìqíng	thing	631,98	396,81	235,17	45,72
身体	shēntǐ	body	470,86	297,47	173,39	45,14
世界	shìjiè	world	308,68	206,98	101,70	39,44
此时	cǐshí	now	283,51	192,95	90,55	38,01
父亲	fùqīn	father	354,26	241,93	112,34	37,68
先生	xiānsheng	Mr.	266,39	182,94	83,46	37,15
当年	dāngnián	that year	194,60	136,15	58,45	35,34
当时	dāngshí	at that time	260,54	182,33	78,21	35,32
消息	xiāoxi	news	184,35	134,76	49,59	31,08

关系	guānxì	relationship	304,34	222,96	81,38	30,87
地方	dìfang	local	473,06	364,99	108,08	25,79
机会	jīhuì	opportunity	240,77	186,80	53,97	25,24
天	tiān	day	233,30	184,19	49,10	23,52
路	lù	road	391,99	312,01	79,98	22,72
书	shū	book	206,38	166,39	39,98	21,45
酒	jiǔ	alcohol	250,48	203,21	47,27	20,84
地	dì	land	438,96	368,81	70,16	17,37
人	rén	person	6643,43	5638,49	1004,94	16,36
目光	mùguāng	view	414,28	360,98	53,30	13,75
生活	shēnghuó	life	225,66	196,82	28,84	13,65
办法	bànfǎ	way	234,88	208,33	26,55	11,98
此刻	cǐkè	this moment	162,54	144,26	18,28	11,92
感觉	gǎnjué	feel	212,17	192,35	19,82	9,80
时间	shíjiān	time	655,92	596,34	59,59	9,52
现在	xiànzài	now	1126,69	1031,92	94,77	8,78
如今	rújīn	nowadays	230,38	211,27	19,12	8,66
女人	nǚrén	woman	459,31	432,27	27,04	6,07
话	huà	word	1108,63	1055,59	53,05	4,90
丝	sī	silk	217,64	211,17	6,46	3,02
月	yuè	month	267,56	259,77	7,79	2,95
意思	yìsi	meaning	245,72	240,99	4,73	1,94
晚上	wǎnshang	night	227,57	225,90	1,67	0,74
东西	dōngxi	thing	499,46	499,79	-0,33	-0,07
瞬间	shùnjiān	moment	147,72	150,04	-2,32	-1,56
事	shì	matter	1422,23	1462,09	-39,86	-2,76
脸色	liǎnsè	look	175,16	180,75	-5,59	-3,14
钱	qián	money	596,55	619,35	-22,80	-3,75
脚	jiǎo	foot	210,36	220,49	-10,13	-4,70
人家	rénjiā	family	195,99	205,77	-9,77	-4,87
今天	jīntiān	today	457,40	481,34	-23,93	-5,10
声	shēng	sound	223,59	239,60	-16,01	-6,91
时候	shíhòu	time	1185,60	1276,73	-91,13	-7,40
儿子	érzi	son	257,56	279,90	-22,34	-8,31
别人	biérén	others	248,63	270,54	-21,90	-8,44

身	shēn	body	1118,50	1219,60	-101,10	-8,65
水	shuǐ	water	351,93	384,28	-32,34	-8,79
里面	lǐmian	inside	201,62	220,91	-19,29	-9,13
眼	yǎn	eye	497,33	554,63	-57,30	-10,89
面前	miànqián	before	171,92	191,93	-20,01	-11,00
头	tóu	head	638,74	716,51	-77,76	-11,48
过去	guòqù	past times	403,41	454,45	-51,04	-11,90
家	jiā	home	618,07	698,29	-80,22	-12,19
名字	míngzì	name	158,56	181,96	-23,40	-13,74
字	zì	Chinese character	220,03	253,39	-33,36	-14,09
女儿	nǚ'ér	daughter	177,62	205,16	-27,55	-14,39
车	chē	vehicle	277,20	320,77	-43,57	-14,57
外面	wàimiàn	outside	164,64	190,86	-26,21	-14,75
老师	lǎoshī	teacher	169,04	199,62	-30,57	-16,59
小姐	xiǎojiě	miss	172,25	206,47	-34,22	-18,07
手	shǒu	hand	1410,49	1701,08	-290,59	-18,68
心	xīn	heart	1161,08	1412,60	-251,53	-19,55
姑娘	gūniáng	girl	145,33	180,51	-35,18	-21,60
老板	lǎobǎn	boss	154,65	192,26	-37,61	-21,68
门	mén	door	316,96	394,25	-77,29	-21,73
面	miàn	noodles	194,93	244,68	-49,75	-22,63
母亲	mǔqīn	mother	201,72	256,88	-55,16	-24,06
公司	gōngsī	company	206,57	266,62	-60,05	-25,38
气	qì	breath	150,08	195,61	-45,53	-26,34
女孩	nǚhái	girl	161,38	211,50	-50,12	-26,88
嘴	zuǐ	mouth	232,84	305,76	-72,92	-27,08
声音	shēngyīn	voice	468,53	631,14	-162,61	-29,57
房间	fángjiān	room	140,22	189,55	-49,34	-29,92
医院	yīyuàn	hospital	129,99	180,65	-50,66	-32,62
女子	nǚzǐ	woman	153,45	217,04	-63,59	-34,33
样子	yàngzi	a look	221,94	315,45	-93,51	-34,80
皇帝	huángdì	emperor	162,73	231,63	-68,89	-34,94
朋友	péngyou	friend	269,85	384,23	-114,38	-34,97
电话	diànhuà	telephone	435,11	626,44	-191,32	-36,05

眼睛	yǎnjīng	eye	420,42	617,07	-196,65	-37,91
脸	liǎn	face	763,91	1126,64	-362,73	-38,37
床	chuáng	bed	159,79	237,27	-77,48	-39,03
衣服	yīfu	clothes	146,30	220,45	-74,15	-40,43
一会儿	yīhuìr	a little while	136,72	212,66	-75,94	-43,47
表情	biǎoqíng	expression	141,03	220,31	-79,28	-43,88
男人	nánrén	man	325,50	512,79	-187,29	-44,68
孩子	háizi	children	384,48	626,72	-242,24	-47,91
爸爸	bàba	dad	131,22	217,18	-85,96	-49,35
手机	shǒujī	mobile phone	130,06	279,53	-149,47	-72,99
妈妈	māmā	mum	145,04	489,72	-344,69	-108,60

Again, as may be seen from the table above, female authors use nouns (lexicon) related to family or more personal things e.g. *nánrén* 男人 [man], *háizi* 孩子 [child], *bàba* 爸爸 [father], *shǒujī* 手机 [mobile phone], *māma* 妈妈 [mum]. On the other hand, male authors tend to use nouns related to non-personal matters, e.g. *shūjì* 书记 [secretary], *rénwù* 人物 [person], *lìliang* 力量 [strength], *qíngkuàng* 情况 [situation], *jiàn* 剑 [sword]. As already stated, the lexicon is topic-related and this point will be developed in further, more detailed research.

4.3 Rest of the POS tags

For the rest of the tokens (except punctuation PU, numbers CD, OD, proper nouns NR), the following CQL query is used:

[tag="AD|PN|M|AS|P|DEC|DEG|VA|DT|LC|JJ|CC|SP|DEV|BA|MSP|CS|DER|SB|LB|ETC|IJ|FW"] within <doc authors_origin="CN"/>

We also set the frequency limit to 5000 of the absolute frequency in *zh-lit-1.1*.

Table 8. Rest of the most frequent words in the male and female subcorpus

Token	Hanyu pinyin	Translation	IPM_M	IPM_F	Frequency difference	dif
位	wèi	*measure word*	893,29	425,37	467,92	70,97
之	zhī	*of*	2242,83	1148,92	1093,91	64,50
以	yǐ	with	846,90	493,22	353,68	52,78
老	lǎo	old	823,19	490,84	332,34	50,58
此	cǐ	this	830,17	500,67	329,50	49,52
它	tā	it	729,07	439,96	289,11	49,46
向	xiàng	towards	1076,54	666,56	409,98	47,04
这些	zhè xiē	these	751,88	469,55	282,33	46,23
所	suǒ	*marker*	720,08	463,49	256,59	43,36
我们	wǒmen	we	2464,57	1642,00	822,58	40,06
大	dà	large	2748,47	1853,40	895,07	38,90
他们	tāmen	they	2293,59	1588,36	705,22	36,33
种	zhǒng	species	1514,73	1067,79	446,93	34,61
中	zhōng	in	3331,86	2407,10	924,75	32,23
你们	nǐmen	you	1008,70	735,47	273,23	31,33
来	lái	come	1156,06	857,55	298,51	29,65
并	bìng	and	1005,53	754,72	250,81	28,50
这	zhè	this	9827,62	7381,37	2446,26	28,43
和	hé	and	3992,66	3013,92	978,74	27,94
这里	zhèlǐ	here	641,36	509,01	132,35	23,01
但	dàn	but	1915,38	1526,30	389,09	22,61
已	yǐ	already	1015,49	816,27	199,22	21,75
如果	rúguǒ	if	769,29	626,11	143,17	20,52
声	shēng	sound	767,60	641,02	126,58	17,97
次	cì	times	1371,92	1158,19	213,73	16,90
就	jiù	just	8316,94	7099,88	1217,06	15,79
将	jiāng	*marker*	1515,12	1295,18	219,93	15,65
对	duì	right	2596,28	2221,18	375,10	15,57
而	ér	and	2233,51	1914,58	318,93	15,38
年	nián	year	1160,95	998,23	162,72	15,07
从	cóng	from	1968,18	1704,67	263,52	14,35
有些	yǒuxiē	some	785,56	683,89	101,67	13,84

在	zài	be	10905,00	9721,42	1183,59	11,48
的	de	of	49311,37	44181,12	5130,25	10,97
正	zhèng	just	805,07	726,80	78,26	10,22
那	nà	that	5188,07	4707,17	480,89	9,72
已经	yǐjīng	already	1541,19	1398,53	142,66	9,71
为	wèi	for	718,20	652,20	66,00	9,63
个	gè	*measure word*	9855,32	8969,36	885,96	9,41
真	zhēn	really	929,21	851,64	77,57	8,71
最	zuì	most	1014,20	930,62	83,58	8,60
把	bǎ	BA	2755,13	2536,92	218,21	8,25
下	xià	down	1297,51	1195,37	102,14	8,19
前	qián	front	1115,78	1037,88	77,90	7,23
也	yě	also	6866,98	6407,65	459,33	6,92
上	shàng	upper	5181,11	4836,76	344,35	6,87
一下	yīxià	once	591,70	557,38	34,32	5,97
了	le	*marker*	24416,16	23213,48	1202,68	5,05
呢	ne	*particle*	1265,71	1206,18	59,52	4,82
更	gèng	more	956,67	912,77	43,90	4,70
天	tiān	day	1234,81	1186,61	48,20	3,98
又	yòu	also	3040,62	2930,19	110,43	3,70
这样	zhèyàng	such	1155,25	1118,12	37,14	3,27
多	duō	many	1217,63	1200,08	17,55	1,45
谁	shéi	who	924,00	915,52	8,48	0,92
与	yǔ	with	1284,54	1275,01	9,53	0,74
还	hái	still	3803,50	3779,96	23,53	0,62
因为	yīnwèi	because	926,62	928,29	-1,67	-0,18
时	shí	time	1058,42	1060,90	-2,48	-0,23
一	yī	one	1080,33	1084,99	-4,66	-0,43
像	xiàng	like	890,93	897,07	-6,13	-0,69
不过	bùguò	however	731,14	739,34	-8,20	-1,11
每	měi	each	600,44	611,11	-10,67	-1,76
过	guò	*marker*	1789,47	1827,36	-37,89	-2,10
他	tā	he	12621,10	13003,03	-381,94	-2,98
太	tài	too	762,98	787,75	-24,78	-3,20
给	gěi	give	938,49	972,41	-33,92	-3,55

里	lǐ	in	3814,66	4001,29	-186,64	-4,78
都	dōu	all	4752,08	5012,01	-259,93	-5,32
吧	ba	*particle*	1348,01	1428,03	-80,01	-5,76
地	de	*marker*	4527,29	4821,20	-293,90	-6,29
什么	shénme	what	2836,47	3031,68	-195,21	-6,65
再	zài	again	1766,40	1888,02	-121,62	-6,66
不	bù	*negative*	17138,07	18338,70	-1200,63	-6,77
便	biàn	then	971,52	1049,48	-77,96	-7,71
只	zhǐ	only	2644,26	2892,31	-248,05	-8,96
一样	yīyàng	equally	610,01	668,47	-58,45	-9,14
自己	zìjǐ	self	2612,20	2865,84	-253,64	-9,26
后	hòu	after	1522,78	1720,93	-198,15	-12,22
很	hěn	very	2233,32	2533,89	-300,58	-12,61
被	bèi	BEI	1966,70	2235,49	-268,79	-12,79
那么	nàme	that	705,68	802,85	-97,17	-12,88
还是	háishì	still	941,05	1081,82	-140,77	-13,92
小	xiǎo	small	1992,45	2315,45	-323,00	-15,00
句	jù	sentence	588,43	694,28	-105,85	-16,50
着	zhe	*marker*	7698,23	9167,16	-1468,93	-17,42
你	nǐ	you	9106,16	10926,81	-1820,65	-18,18
却	què	but	1710,46	2084,84	-374,38	-19,73
怎么	zěnme	how	1329,77	1644,75	-314,98	-21,18
吗	ma	*particle*	1180,52	1462,88	-282,36	-21,36
这么	zhème	such	897,79	1115,79	-218,00	-21,65
得	de	*marker*	2216,20	2766,08	-549,88	-22,07
才	cái	just	1496,19	1876,00	-379,81	-22,53
啊	a	ah	1093,59	1386,56	-292,97	-23,62
边	biān	*suffix*	819,27	1062,06	-242,79	-25,81
我	wǒ	I, me	12620,16	16553,98	-3933,82	-26,97
好	hǎo	good	1900,31	2498,99	-598,68	-27,22
没	méi	not have	771,23	1076,55	-305,33	-33,05
跟	gēn	with	833,57	1314,85	-481,28	-44,80
她	tā	she	4793,49	11044,23	-6250,74	-78,93

As expected from the previous subsections, female authors tend to use a more personal (and to some extent a more natural) approach (e.g. personal pronouns in singular vs. personal pronouns in plural by male authors, interjections *a* 啊 [interjection], e.g. *wǒ* 我 [I, me], *hǎo* 好 [good], *méi* 没 [*negative*], *gēn* 跟 [with], *tā* 她 [she, her]). Male authors on the other hand seem to use more lexis from the so-called written register (<*written*>) *shūmiànyǔ* 书面语[14] (e.g. *zhī* 之 [marker <*written*>], *cǐ* 此 [this <*written*>], *yǐ* 以 [using <*written*>], *suǒ* 所 [marker <*written*>]).

4.4 Absolute IPM difference

When searching for the most frequent tokens in the *zh-lit-1.1* (*Word list* User Interface), the following CQL query is used:

[tag!="PU|CD|OD|NR"] within <doc authors_origin="CN"/>

It should be noted that IPM is calculated for the male/female subcorpus separately and not for the whole corpus *zh-lit-1.1*. Also, the IPM for e.g. a particular verb may differ from the IPM presented in the following table. The tokens in the table below are not restricted to any tag, i.e. a certain verb may belong to two or more part of speech tags.

Table 9. Most frequent tokens in the male and female subcorpus

Token	Hanyu pinyin	Translation	IPM_M	IPM_F	Frequency difference	dif
的	DE	of	49373,94	44181,12	5192,82	11,10
一	yī	one	18319,59	15461,72	2857,87	16,92
这	zhè	this	9840,56	7381,51	2459,06	28,56
是	shì	be	20098,48	17832,71	2265,77	11,95
说	shuō	say	7684,73	6113,12	1571,62	22,78
了	le	*marker*	24748,81	23453,49	1295,32	5,37
就	jiù	just	8336,29	7111,30	1224,99	15,86

[14] See more at GAJDOŠ, Ľ. The Discrepancy between Spoken and Written Chinese Methodological Notes on Linguistics. In *Studia Orientalia Slovaca*, 2011, Vol. 10, No. 1, pp. 155–159.

在	zài	in	11152,21	9976,25	1175,96	11,13
有	yǒu	have	7417,86	6242,23	1175,62	17,21
之	zhī	of	2245,58	1148,92	1096,66	64,61
人	rén	person	6643,43	5638,49	1004,94	16,36
和	hé	and	4000,43	3016,21	984,22	28,05
中	zhōng	in	3438,72	2459,80	978,91	33,19
大	dà	big	2768,24	1871,71	896,52	38,64
个	gè	*measure word*	9865,44	8969,36	896,09	9,52
我们	wǒmen	we	2468,16	1642,00	826,17	40,20
他们	tāmen	they	2295,66	1588,36	707,29	36,42
到	dào	arrive	2762,25	2220,49	541,77	21,75
没有	méiyǒu	not have	3003,99	2472,20	531,79	19,42
那	nà	that	5194,86	4707,73	487,13	9,84
也	yě	also	6875,56	6407,65	467,91	7,05
种	zhǒng	kind	1551,51	1094,07	457,44	34,58
来	lái	come	3145,92	2750,05	395,87	13,43
但	dàn	but	1918,00	1526,30	391,71	22,75
上	shàng	upper	5660,29	5278,67	381,61	6,98
要	yào	want	4286,85	3931,26	355,59	8,65
能	néng	can	2987,78	2641,76	346,02	12,29
对	duì	right	2673,41	2330,83	342,59	13,69
而	ér	and	2235,94	1914,58	321,35	15,48
为	wèi	for	1295,86	1023,99	271,87	23,44
从	cóng	from	1970,71	1705,09	265,62	14,45
将	jiāng	*marker*	1519,61	1296,35	223,26	15,86
把	bǎ	BA	2759,40	2537,76	221,65	8,37
次	cì	times	1374,87	1159,40	215,46	17,00
年	nián	year	1199,38	1042,45	156,94	14,00
已经	yǐjīng	already	1542,71	1398,53	144,18	9,80
可	kě	can	1890,31	1759,14	131,17	7,19
下	xià	down	1549,21	1428,40	120,81	8,11
又	yòu	again	3044,11	2930,19	113,92	3,81
天	tiān	sky	1470,86	1371,23	99,63	7,01
现在	xiànzài	now	1126,72	1031,92	94,80	8,78

前	qián	forward	1122,48	1041,70	80,78	7,47
话	huà	talk	1132,64	1071,48	61,16	5,55
呢	ne	*particle*	1267,36	1206,42	60,94	4,93
多	duō	many	1545,24	1505,47	39,77	2,61
时	shí	time	1139,18	1106,79	32,38	2,88
起来	qǐlái	stand up	1253,02	1234,05	18,98	1,53
还	hái	also	3834,81	3817,33	17,48	0,46
这样	zhèyàng	so	1191,20	1175,90	15,30	1,29
与	yǔ	with	1286,77	1275,01	11,76	0,92
让	ràng	let	1960,29	1964,02	-3,73	-0,19
用	yòng	use	1151,79	1158,33	-6,54	-0,57
可以	kěyǐ	may	995,98	1009,50	-13,52	-1,35
事	shì	thing	1422,59	1462,23	-39,64	-2,75
见	jiàn	see	1036,94	1106,00	-69,06	-6,45
便	biàn	just	972,98	1050,23	-77,25	-7,64
吧	ba	*particle*	1349,82	1428,54	-78,72	-5,67
给	gěi	give	1533,82	1615,11	-81,30	-5,16
时候	shíhòu	when	1185,63	1276,78	-91,14	-7,40
出来	chūlái	come out	991,06	1084,66	-93,60	-9,02
身	shēn	body	1183,85	1284,84	-100,98	-8,18
过	guo	*marker*	2282,62	2400,30	-117,68	-5,03
再	zài	again	1768,73	1888,35	-119,62	-6,54
还是	háishì	still	941,92	1081,82	-139,90	-13,83
知道	zhīdào	know	1892,25	2057,87	-165,61	-8,39
里	lǐ	inside	3871,76	4038,01	-166,25	-4,20
去	qù	go	2490,94	2672,14	-181,20	-7,02
后	hòu	rear	1538,73	1729,69	-190,96	-11,68
什么	shénme	what	2840,00	3031,72	-191,73	-6,53
地	de	*marker*	4971,53	5190,00	-218,47	-4,30
走	zǒu	go	997,92	1240,53	-242,60	-21,68
只	zhǐ	only	2647,56	2892,31	-244,75	-8,84
自己	zìjǐ	self	2618,93	2867,15	-248,22	-9,05
心	xīn	heart	1164,48	1414,70	-250,23	-19,40
都	dōu	all	4761,14	5014,48	-253,34	-5,18
被	bèi	BEI	1970,38	2236,05	-265,66	-12,63

做	zuò	do	1093,40	1365,92	-272,52	-22,16
吗	ma	*particle*	1182,20	1463,21	-281,00	-21,24
手	shǒu	hand	1423,17	1712,22	-289,05	-18,44
啊	a	ah	1095,40	1386,84	-291,43	-23,48
怎么	zěnme	how	1332,77	1645,63	-312,86	-21,01
很	hěn	very	2237,78	2553,65	-315,87	-13,18
小	xiǎo	small	2029,04	2377,75	-348,71	-15,83
他	tā	he	12635,75	13003,03	-367,28	-2,87
却	què	but	1712,27	2084,84	-372,57	-19,62
才	cái	just	1507,09	1886,02	-378,93	-22,34
会	huì	may	2959,93	3352,86	-392,93	-12,45
想	xiǎng	think	2441,31	2851,68	-410,36	-15,51
道	dào	say	1023,38	1457,43	-434,05	-34,99
跟	gēn	with	1014,33	1536,97	-522,64	-40,97
笑	xiào	laugh	1094,08	1701,55	-607,47	-43,46
好	hǎo	good	1949,81	2558,87	-609,06	-27,02
得	de	*marker*	3122,73	3815,98	-693,25	-19,98
没	méi	not have	2147,90	2990,58	-842,67	-32,80
看	kàn	look	3239,78	4111,91	-872,13	-23,73
不	bù	*negative*	17168,90	18347,41	-1178,51	-6,64
着	zhe	*marker*	7839,96	9290,12	-1450,16	-16,93
你	nǐ	you	9169,05	10987,52	-1818,47	-18,04
我	wǒ	I, me	12636,37	16553,98	-3917,61	-26,84
她	tā	she, her	4797,28	11044,51	-6247,23	-78,87

Again, there is no surprise: the main differences occur for the most frequent words, which are, in most cases, pronouns or functional words, e.g. in the male register *de* 的 [attributive marker], *yī* 一 [one], *zhè* 这 [this], *shì* 是 [to be], *shuō* 说 [to speak]; in the female register *tā* 她 [she, her], *wǒ* 我 [I, me], *nǐ* 你[you], *zhe* 着 [grammatical marker].

5 Conclusion

To conclude, from the language data presented in this paper, it is apparent that there are several differences in the male and female registers of literary texts. Let us highlight these.

Criteria	*Male*	*Female*
the length of a sentence	longer sentence (around 55 tokens)	shorter sentence (around 31 tokens)
the length of words	more disyllabic (proportion in the male subcorpus)	more monosyllabic (proportion in the female subcorpus)
punctuation	less (proportion in the male subcorpus)	more (double the IPM compared to the male subcorpus)
nouns	more (in proportion in the male subcorpus)	less (in proportion in the female subcorpus)
pronouns	less (in proportion in the male subcorpus)	more (in proportion in the female subcorpus)
keywords	de 的 [attributive marker], yī 一 [one], zhè 这 [this], shì 是 [to be], shuō 说 [to say], le 了 [grammatical marker], jiù 就 [just], zài 在 [to exist], yǒu 有 [to have], zhī 之 [grammatical marker], rén 人 [person]	tā 她 [she, her], wǒ 我 [I, me], nǐ 你[you], zhe 着 [grammatical marker], bù 不 [not], kàn 看 [to look], méi 没 [not have], de 得 [grammatical marker], hǎo 好 [good], xiào 笑 [to laugh], gēn 跟 [with]
noun keywords	rén 人 [person] qíngkuàng 情况 [condition] wèntí 问题 [problem] shūjì 书记 [secretary]	liǎn 脸 [face] māma 妈妈 [mom] shǒu 手 [hand] xīn 心 [heart]

	shìqíng 事情 [matter]	háizi 孩子 [child]
	lìliang 力量 [strength]	yǎnjīng 眼睛 [eye]
	rénwù 人物 [figure, person]	diànhuà 电话 [telephone]
	jiàn 剑 [sword]	nánrén 男人 [man]
	shēntǐ 身体 [body]	shēngyīn 声音 [sound]
	gōngzuò 工作 [work]	shǒujī 手机 [cell phone]
verb and adjective keywords	shì 是 [to be]	kàn 看 [to look]
	shuō 说 [to say]	dào 道 [*to say*]
	yǒu 有 [to have]	xiào 笑 [to laugh]
	méiyǒu 没有 [to have not]	méi 没 [to have not]
	dào 到 [to arrive]	hǎo 好 [good]
	yào 要 [to want]	xiǎng 想 [to think]
	néng 能 [can]	huì 会 [can]
	shuōdao 说道 [to say]	juéde 觉得 [to feel]
	chūxiàn 出现 [to appear]	xǐhuān 喜欢 [to like]
	nénggòu 能够 [to be able]	zuò 做 [to do]

It is worth noting that the results and conclusions support our basic assumption but further research on a larger data set must be conducted to prove it.

Finally, in answer to the question; are we different? Yes, to some extent…

REFERENCES

BIBER, Douglas. *Dimensions of Register Variation*. Cambridge: Cambridge University Press, 1995.

GAJDOŠ, Ľuboš. The Discrepancy between Spoken and Written Chinese Methodological Notes on Linguistics. In *Studia Orientalia Slovaca*, 2011, Vol. 10, No. 1, pp. 155–159.

GAJDOŠ, Ľuboš, GARABÍK, Radovan, BENICKÁ, Jana. The New Chinese Webcorpus Hanku – Origin, Parameters, Usage. In *Studia Orientalia Slovaca*, 2016, Vol. 15, No. 1, pp. 21–33.

GAJDOŠ, Ľuboš. Chinese Legal Texts – Quantitative Description. In *Acta Linguistica Asiatica*, 2017, Vol. 7, No. 1, pp. 77–87.

KILGARIFF, Adam. Simple Maths for Keywords. [online] [cit. 12 September 2021]. Available from https://www.sketchengine.eu/documentation/simple-maths/

PETROVČIČ, Mateja. 2018. Distribution of "Young Words" in the Chinese Web 2011. Corpus and the Hanku Corpus. In *Studia Orientalia Slovaca*, vol. 17, No. 2, pp. 171–180.

PETROVČIČ, Mateja, GARABÍK, Radovan, GAJDOŠ, Ľuboš. The New Chinese Corpus of Literary Texts Litchi. In *Acta Linguistica Asiatica*, 2020, Vol. 10, No. 2, pp. 65–81.