Glottometrics 48, 2020, 67-87

Function Words in Male and Female Authors: A Diachronic Investigation of Modern Chinese Prose

Xiaojin Zhang [1]，*Haitao Liu* [2]

Abstract. From the perspective of a blended micro- and macro-analysis, the present study mainly investigates the function words in modern Chinese prose since 1912 to 2019. First, we choose the cumulative frequencies of function words; this index is related to *h-point* as a quantitative indicator. We compare the cumulative frequencies of function words in 50 male and 50 female authors' books. The diachronic trend of the cumulative frequency of function words clearly shows that the proportion of male writers is higher than that of female writers. To be specific, the males use more numerals while females use more personal pronouns. Based on the visualized graphs, the macro-developmental trends of function words in modern Chinese prose from 1912 to 2019 are finally presented.

Keywords: *Modern Chinese prose, males and females, h-point, cumulative frequency of function words, network analysis*

1. Introduction

Being a window into the inner world of people, words are so fascinating and revealing that they link the language and the material world around us. For the authors, it might or might not be easy to conceal their gendered language. Women's vocabulary has been the focus to see, for example, which linguistic category is more "central" for the language of women than another (Jespersen, 1922). With the publication of *Language and Woman's Place* in 1975, Lakoff's influence on the study of gender and language has had its profound consequences. Researchers in the past 15 years have already been interested in gender and language. Correspondingly, many of them focused on the identification of text genre and authorship, including political speeches (Yu, 2014), novels (Rybicki, 2016; Weidman and O'Sullivan, 2018), scientific papers (Sarawgi et al., 2011), blogs and celebrity tweets (Schler et al., 2006), and, of course, written word (Koppel et al., 2003; Burrows, 2004; Newman et al., 2008; Mikros and Perifanos, 2013). Since gender is socially constructed (Crawford, 1995), the male and female authors do not inherently use words in the same way. Specifically, men are reported to use

[1] Department of Linguistics, Zhejiang University, Hangzhou,China; School of Foreign Studies, North Minzu University, Yinchuan, China.
[2] Department of Linguistics, Zhejiang University, Hangzhou, China; Institute of Quantitative Linguistics, Beijing Language and Culture University, Beijing, China; Centre for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou, China. Correspondence to: Haitao Liu. Email address: htliu@163.com, ORCID-No.: https://orcid.org/0000-0003-1724-4418.

more articles, nouns, long words, swear words, and numbers, while women are found to use more personal pronouns, verbs, and emotion words (Koppel et al., 2003; Newman et al., 2008). The common interpretation of these findings is, to some extent, that males and females tend to use function words differently.

There have been abundant studies about gender and language in English (Koppel et al., 2003; Stamatatos, 2009; Pennebaker, 2011; Jockers, 2013; Oakes, 2014; Rybicki, 2016). Comparison between English fiction and non-fiction (Koppel et al., 2003) extends the results of gender differences (Holmes, 1993; Biber et al., 1998). More importantly, it was found in English that women tend to use the first person singular more, cognitive and social words, and men tend to use more articles; there are no significant differences between men and women for the first person plural or positive emotion words in Pennebaker's research (2011). It is generally assumed that the methods used in English language should also work in other languages. If the stylistic features of function words above can also be applied to more languages, we may obtain a better sense of the stylistic feature distinguishing the usage of function words between males and females. Studies aiming at investigating the stylistic features of written or oral forms have been conducted in other languages, to name a few, in Greek (Mikros, 2013b), Italian (Bortolato, 2016), and Russian (Sboev et. al, 2016). Mikros (2013b) found the gender preference over the use of personal pronouns and coordinate conjunctions in Greek blogs, predicting author's gender with about 80% accuracy rate weighted by relative frequency. It is an attempt to make a distinction between the genders according to the frequencies.

As to the studies on frequency of words in texts, the frequencies of the determiners have been arrayed hierarchically to form a frequency profile for each text (Burrows, 1987). Kilgarriff (2001) noted that the differences among higher-frequency words play an important role in determining the (dis-)similarity for data in a text. However, analysis which is based on the most frequent word could not be necessarily identical to function words along. The most frequent words in a text take into account both content words and function words. Recently, word frequency related to *h-point* is adopted to compare the stylistic theme in a discourse (Wang and Liu, 2017). They discussed Trump and the other two candidates' political themes from their thematic words. "Writer's view", which is also related to *h-point*, is connected to the authors' control of function words and content words in the production processes of poetry (Pan et al., 2017). Obviously, all the aforementioned methods highlight the investigation into the *h-point*-related frequencies of function words in our study.

Pan et al. (2017) proposed that an author has his or her control of function words mainly above the so called *h-point* in the 'writer's view'. Although the concentration tendency of both function words and content words vary to the size of the texts, the cumulative frequency of the function words will give an overall trend of any function words in a text. If this kind of cumulative frequency distribution is observed in translated poems, will there be similar tendency of function words in prose? From a macro view, a significant leap from specific to general view on text features can be realized by network analysis (Jockers, 2013). As stated by Butts (2008), social network analysis studies are used to measure and analyze the relationship between social actors in order to understand and master the social structure in literary text. Jockers (2013) used macro-scale to investigate the influence of time and gender on theme and style. On the basis of distinct separation of gender markers of male and female writers (Pennebaker, 2011), Rybicki (2016) intensified the macro analysis of literary works from the

20th and 21st centuries, and compared them with the 18th- and 19th-century corpus. His finding verifies that distinctive gender markers may fade over time (Rybicki, 2016). These observations, in other words, show that the observed gender variable may become differentiable at different periods. Thus, if we observe the frequency of function words based on a corpus, it might be necessary to embrace the different stylistic features of males and females as well as stylistic evolution over time.

Therefore, for the issue of gender identification in multilingualism – particularly, Chinese in our study, a blended approach suggested by Jockers (2013), namely, and a sort of unification of the macro scales, could apply the quantitative methods and explain the use of function words of males and females in literary works better. Then, a blended macro- and micro-analysis is much better at recognizing the influence of gender and time on function words in Chinese prose, which leaves the following three questions:

Question 1: Is the function word a determinant differentiating male authors from female authors in Chinese prose?

Question 2: On the micro level, with the indicator of h-point related to the cumulative frequency of function words, do male authors differ from female authors in Chinese prose?

Question 3: On the macro level, are there any developmental features concerning the usage of function words between the two genders from the year 1912 to 2019 in Chinese prose?

To be brief, as mentioned above, we attempt to compare and analyze the usage of function words in Chinese prose from the perspective of gender and time. First, we will display the materials and quantitative methods employed. Next, there are the results and a discussion related to both micro and a macro evolution analysis of usage of function word in Chinese males' and females' prose, after which is a brief conclusion.

2. Materials and Methods

2.1 Materials

We choose 100 authors' texts, spanning from the years 1912 to 2019, as the target material of our study (see Appendix 1). Zhang and Liu (2016) mentioned several drastic social changes that occurred during this period, namely, *the New Culture Movement*, *the Warlord Era*, *the Anti-Japanese War*, *the Chinese Civil War*, *the reunification in 1949*, along with *the reform and opening up policy* in the end of 1978. According to the publication years of the books, combined with the classified social changes, we finally set two main periods. The first period lasts from the *New Culture Movement* to 1962, and the second period from *the reform and opening up policy* in the end of 1978 to 2019. Very few books are available during the period 1962–1978. We skip *the Warlord Era*, *the Anti-Japanese War* and *the Chinese Civil War*, during which periods no published prose could be found available, due to the drastic wars and

social changes. All these events had impacts in human and social terms, and influenced literary works too. The authors' books are ranked according to their publication years (See Appendix 1).

2.2 Methods

First, we focus on the cumulative frequencies of the function words. The concept of the so-called *h-point* was conceived by Hirsch (2005) and later introduced into linguistics by Popescu (2007). If the word frequencies of a text are ranked in the descending order, the value of the *h-point* can thus be calculated when the rank of a word is equal to its frequency in its descending order (if it is not possible to find in this way, the calculation goes as shown below). H-point, as defined by Popescu et al. (2009a), is a fixed point in the rank-frequency distribution of words. For the rank-frequency distribution of words, *h-point* is important since it marks the fuzzy boundary between the content words and function words as in Figure 1. The function words get denser above the *h-point*, and content words increase in large quantities below the *h-point*.

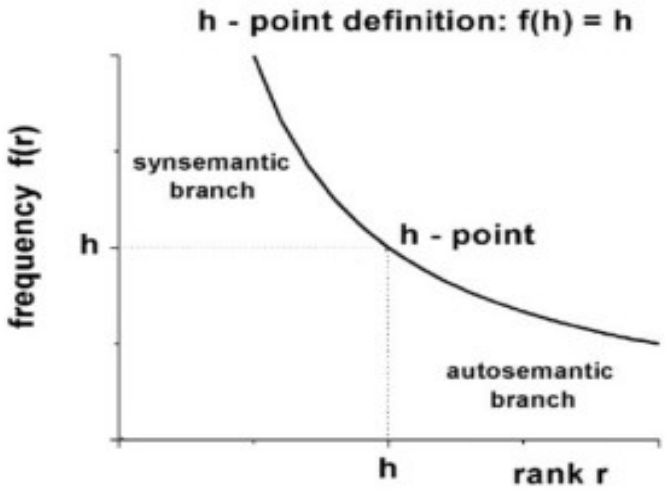

Figure 1. The position of the h-point on a rank-frequency distribution curve

Popescu et al. (2009a) demonstrated that synsemantic region and the autosemantic region, together building a text, are separated by the *h-point.* The synsemantic part includes function words like prepositions, pronouns, particles, and they occur before the *h-point*, indicating that they are the frequently used words by an author. The calculation of the *h*-point in the frequency distribution is shown below:

$$h = \begin{cases} r, & if \quad there \quad is \quad an \quad r = f(r) \\ \dfrac{f(i)r_j - f(j)r_i}{r_j - r_i + f(i) - f(j)}, & if \quad there \quad is \quad no \quad r = f(r) \end{cases}$$

In the above mathematical definition, r_i and r_j stand for any two neighboring words in a rank-frequency distribution; *f(i), f(j)* are the corresponding word frequencies of r_i and r_j re-

spectively. $R_1 = 1 - \underline{F(h)}$, in which R_1 is the cumulative frequency of content words in a text, and *F(h)* is the cumulative frequency of function words. In this way, we can derive the cumulative frequency of function words as $F(h)=1-R_1$.

Moretti (2005) pointed out that the macroanalytic analysis isn't a sum of individual cases but a collective system and should be grasped as a whole. This shows the advantages of a macroanalytic approach over the more traditional approach of studying literary periods and genres. In Jocker's opinion, a macroanalysis is conducted by means of a close study of "representative" texts (Jockers, 2013). From the perspective of the network analysis, a text can be processed and constructed out of two primary elements: nodes and edges.

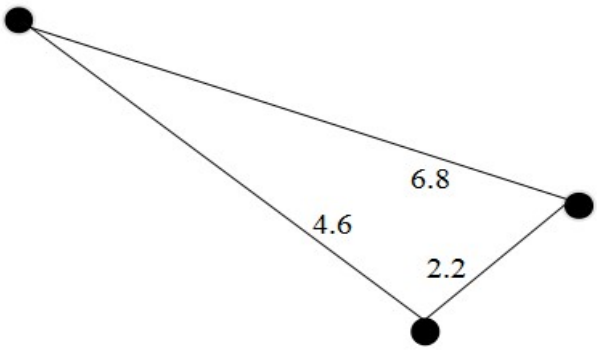

Figure 2. A simplified graph

Figure 2 offers a simplified example. The nods are all colored black with each nod representing an individual author. The distance between any two authors, which are viewed as two points in m-dimensional space are measured by the calculated Euclidean distance. Burrows reported that squared Euclidean distances and standardized variables yield the most accurate results for clustering texts by similarity (2004, p. 326). As in Figure 2, the edges between the nods could be 2.2, 4.6, or 6.8. Any two nods with smaller distances can be connected closer. When plotted, the nodes with larger distances are spread out farther in the graph. The software GEPHI provides a number of layout options and analysis routines for data, which makes the intricacies of the network more visible.

The selection of the 17 types of function words in the study is based on the criterion of Chinese word segmentation NLPIR.[3] Then, we obtained the rank-frequency distribution data of each text with software QUITA.[4] Based on the processed texts, we calculated the value of *h-point*, R_1, and the value of $\underline{F(h)}$ (cumulative frequency of function words) for each text (see Appendix 2). The final network is generated from GEPHI[5], based on the results of Euclidean distances between the authors (see Appendix 3).

[3] https://code.google.com/p/oltk/.

[4] http://oltk.upol.cz/software.

[5] http:// gephi.org.

3. Results and Discussion

3.1 Comparison of function words of two genders

When we try to figure out the different uses of function words between male and female authors, we need to extract their specific types according to the selected parameters. Two traits taken into account from Oakes (2014) are: they have to be frequent enough to support a statistical approach and to become statistically evident, and they need to be objectively countable. When it comes to the Chinese prose, as a literary genre, it was once used to compare with Chinese novels (Zhang and Liu, 2015), due to the historic and literary features.

Dramatic changes in China give Chinese women an opportunity to step into the stage. Lots of excellent female authors sprang up like Ailing Zhang, Bingxin, Weiyin Lin, Shuting, Murong Xi, etc. Within such linguistic context, how do male author and female author use function words? Is it possible that women writers are in fact writing with a different literary style compared with men? We obtain the following valid 17 categories in Table 1. For a better understanding of some Chinese function words which have no equivalence in English, we give the English meaning of Chinese function words: "yu"(same as to "和" with its equivalent in English "and") , "ule", "udeng", "uzhi", "uzhe", "uguo", "ude" in the last column.

Table 1
17 function words in 100 items of Chinese prose

Abbreviation	Function word	Abbreviation	Function word	English meaning
m	numeral	yu	"与"	and
q	quantity	ule	"了"	finished
d	adverb	udeng	"等"	and so on
p	preposition	uzhi	"之"	this
r	pronoun	uzhe	"这"	this
rr	personal pronoun	uguo	"过"	finished
c	conjunction	ude	"的"	of
e	exclamation			
u	auxiliary			
o	onomatopoeia			

We must make sure what types of the selected function words could be used in the final results. Otherwise, if there is any individual variable not working, the function word should be removed. Generally, this process is judged by one-way analysis of variance (ANOVA). In our study, k-means ANOVA, a useful statistical test verifying the statistic hypotheses, is adopted to observe whether the selected function word is capable of detecting Chinese male author and female author, or not.

Table 2
k-means of ANOVA's Results of the Selected Function Words

	k-means		Std. Error		F	*p* <
	MS	df	MS	df		
numeral	68112.092	1	1042.008	38	65.366	0.001
quantity	22594.852	1	468.340	38	48.245	0.001
adverb	566966.410	1	6549.989	38	86.560	0.001
preposition	58499.878	1	1239.892	38	47.181	0.001
personal pronoun	47.426	1	6.341	38	7.479	0.001
pronoun	195171.539	1	2116.827	38	92.200	0.001
conjunction	23758.810	1	261.094	38	90.997	0.001
exclamation	0.108	1	0.754	38	0.144	.
yu "与"	24545.733	1	166.123	38	147.757	0.001
onomatopoeia	0.656	1	3.520	38	0.187	.
auxiliary	0.023	1	.344	38	0.067	.
ule"了"	2026.464	1	167.196	38	12.120	0.001
udeng "等"	0.000	1	0.000	38	.	.
uzhi "之"	0.000	1	0.000	38	.	.
uzhe "这"	3579.339	1	148.669	38	24.076	0.001
uguo "过"	7.477	1	6.814	38	1.097	.
ude "的"	198293.426	1	7331.605	38	27.046	0.001

As seen in Table 2, function categories with the values of k-means less than 0.001 are: numeral = 0.000 < 0.001, quantity = 0.000 < 0.001, adverb = 0.000 < 0.001, preposition = 0.000 < 0.001, personal pronoun = 0.000 < 0.001, conjunction = 0.000 < 0.0001, yu = 0.000 < 0.001, ule = 0.000 < 0.001, uzhe = 0.000 < 0.001 and ude = 0.001 ≤ 0.001, which means the 10 variables have significant effect on the final result. Among the 17 categories, function categories with the kmeans larger than 0.05 are: pronoun = 0.009 > 0.001, exclamation = 0.707 > 0.05, onomatopoeia = 0.668 > 0.05, u = 0.797 > 0.05, uguo = 0. 301 > 0.05, indicating that the left 7 categories of function words have no significant effect, and they are filtered out.

The one-way between-groups (ANOVA) test was performed on two genders. There are significant differences ($F_{(2, 37)} = 40.651$, $p = 0.000$). Results of post hoc comparisons test indicates that the cumulative frequency level of females (M = 0.3000, SD = 0.03266) demonstrates a significant difference from those of the males (M = 0.4085, SD = 0.04368, $p = 0.001$, respectively). Therefore, the statistical results confirm that there is significant difference in the usage of function words between the male and female authors.

In Table 3, we rank the frequencies of function words above *h-point* to see the most frequently function words used by males and females.

Table 3
Function words above *h-point* in the studies corpus

Rank	Males	Frequency	Females	Frequency
1	adverb	5654	adverb	5791
2	numeral	2573	numeral	2489
3	preposition	2418	preposition	2516
4	quantity	1959	quantity	1746
5	conjunction	1440	personal pronoun	1582
6	personal pronoun	151	conjunction	1127

As shown in Table 3, male authors use more numerals and female authors use more adverb and personal pronouns. Based on the analysis of the general trends of function words of the two genders, as well as their preference for frequently used function words, our finding confirms the one of Pennebaker (2011) that men use articles more than women do. Besides, a higher percentage of pronouns were seen as a strong female language indicator (e.g., Biber et al. 1998; Koppel et al., 2003; Pennebaker, 2011) and the social variable of 'class' (Bernstein, 1971). Pronoun, according to our statistical results, does not exhibit a strong female language characteristic. However, the results confirm that men have been reported to use more numbers, while women have been found to use more personal pronouns (Koppel et al., 2003; Newman et al., 2008; Pennebaker, 2011). This kind of restriction still leaves room for the traits of an author's style to emerge in each piece of prose.

Writing style, as Pennebaker (2011) said, was revealed through function words. As many researchers (Pennebaker, 2011; Jockers, 2013; Rybicki, 2016) have suggested, women and men differ in literary styles from a period-based perspective, and the evolution of style over periods isolates the gender groupings. It is easy to understand that great changes have taken place in China especially since the *reform and opening up policy* in the end of 1978. Correspondingly, the traces of the impact of social and economic changes on literature are revealed in the authors' writing. So far, for question one—function word is effective in isolating the male authors from the female authors. As to the second question, the two genders differ from each other on the grounds of the numerals and personal pronouns in Chinese prose. After the discussion on the gender and time from a micro view, we will study their stylistic features from the macro view.

3.2 Macro development of function words with genders and periods

Figure 3 provides a visualization of the different trends of function words used by males and females. In this plot, the x-axis symbolizes the 100 authors (50 males and 50 females); and the y-axis represents the values of $\underline{F(h)}$ of the 10 categories of function words. The gray line stands for the trends of male authors, and the black one for the female authors. Obviously, the trend of females, denoted by the data, is lower than the one of the males. This shows that

Chinese male authors tend to use function words more than female authors.

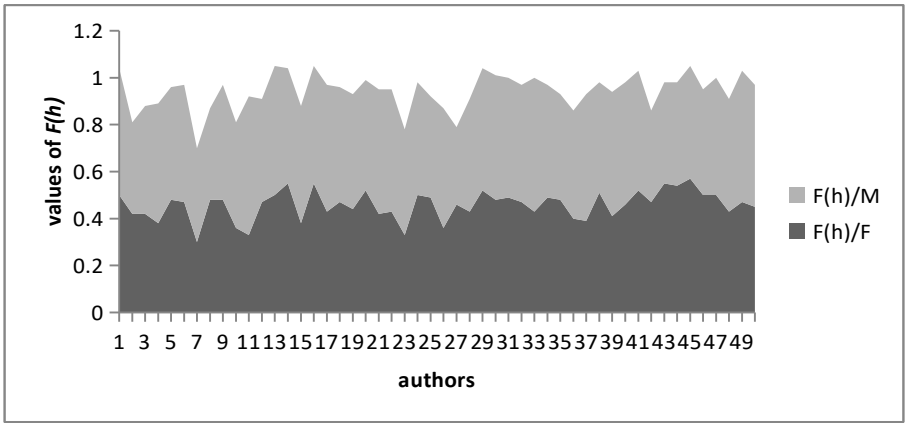

Figure 3. Results of diachronic trends of *F(h)* between two genders

Slight difference of function words between males and females is presented in Figure 3. Jockers (2013) drew the macro network of nineteenth-century novels reflecting that the style and theme evolve chronologically; therefore, comparisons of stylistic feature between gender and time in prose may present the distribution of function words between male and female authors better from this perspective. A further close reading of the network analysis across two genders and two periods are shown in Figure 4 and Figure 5. The network graph of authorial gender is shown in Figure 4, in which the lighter gray colour represents nods and edges of the males, and darker colour for nodes and edges of the females. In Figure 5, the lighter gray colors represent nods and edges of the earlier published prose from 1912 to 1962, and darker colors represent nodes and edges for more recently published prose from 1978 to 2019.

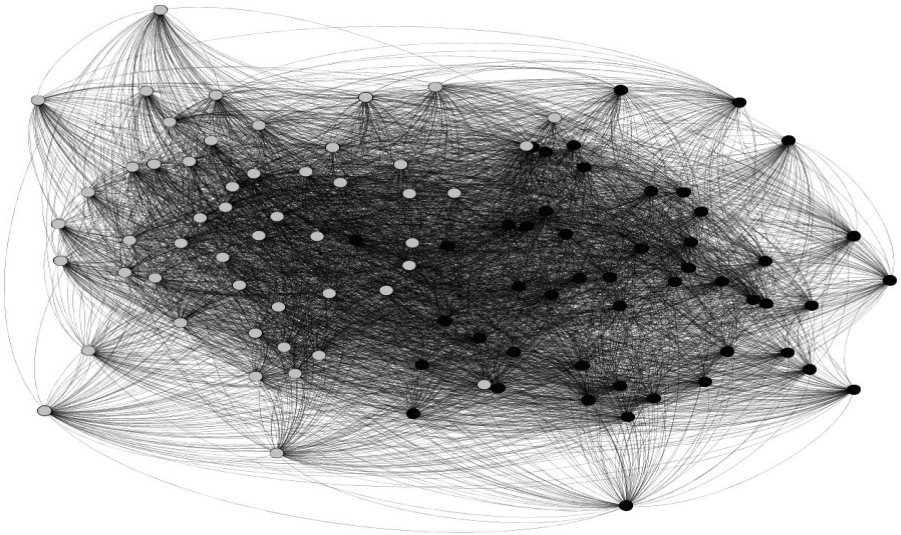

Figure 4. Development of function word according to the gender

In Figure 4, it can be seen that the two colours in network is clearly divided into two main portions—the lighter on the left and darker on the right. Prose written by female authors

is more stylistically similar to each other, and obviously female authors cluster together in the east portions of the main network. Males cluster together mainly in the west portion in the network. In Figure 5, each book's node and edge has been colored according to its publication year and they are coloured in the same shade. Specifically, the earlier the period of time is, the lighter gray the nodes and the edges are; the later period of time is, the darker the nodes and edges are. In other words, the lightness of the colour changes with the time. In this way, a clear time signature to the stylistic data can be shaded by the nodes by year. Again, the network is divided into two portions. Prose published between the years 1912 to 1962 are more stylistically similar to each other, and cluster together in the east portions of the main network. While, prose published between the years 1978 to 2019 clusters together in the west portions of the network.

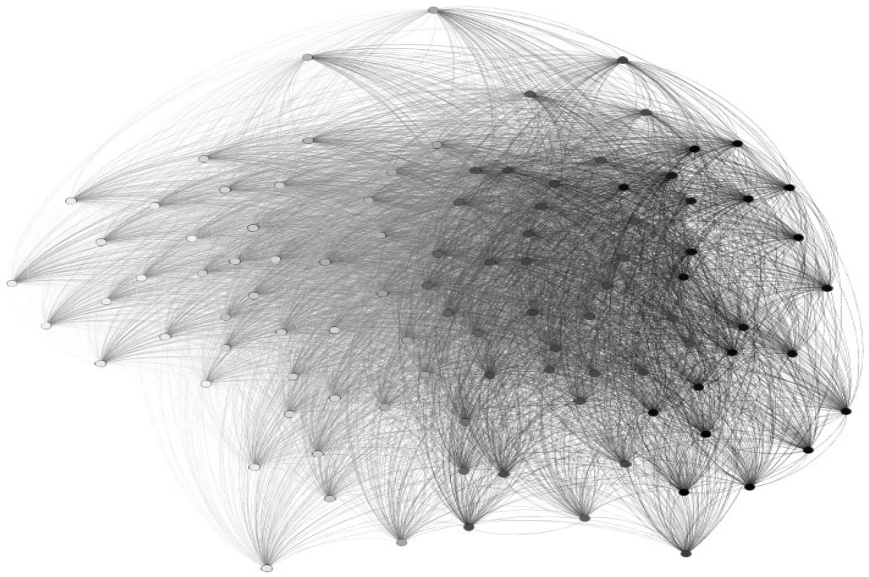

Figure 5. Twentieth century development of function word

Thus, the third question is answered. In both the networks, we see evidence of time and gender influences on the usage of function words at the macro scale. While processing, the graphing software does not know who the authors are, and the macroanalysis offers us a way of finding them in the visualized graphs in this period. In Figure 4, there are male authors and female authors that are placed firmly in their gender-dominated regions of the graph, and in Figure 5, there are books from the earlier period in the century and the later period that cluster firmly in their dominated portions, respectively.

4. Conclusion

The research has aimed at testing function words as differential elements for gender and time. In particular, we have started from a comparison of general trends of function words between the two genders.On the grounds of the quantitative indicator *F(h)*, altogether with the

visualized network analysis, we have identified trends and correlated changes in Chinese prose. China underwent an irregular course of development in the 20th century. The authors' writing styles reflect the process of this transformation and reveal traces of the impact of social metamorphosis through function words. The outcome of the micro analysis and macro network graphs successfully differentiate the authors according to gender and time periods. We were hence able to answer the main questions we proposed in the following way: Male authors tend to use many more function words than female authors. From a micro point of view, male authors use more numerals, and female authors use more personal pronouns. From the macro point of view, the writing styles of authors of the same gender and time period appear much more similar. The results confirm the predictability of the use of function words to identify authors' genders.

The above research and results are still somewhat insufficient and 'in need of development' (Savoy, 2012). The corpus on which our study was based provides diachronic, social, and literary characteristics of 100 Chinese authors. From a statistical perspective, the analysis is less stable if it concentrates only on a limited number of forms. Further methods need to be applied in order to evaluate the stability of this approach. For instance, using other stylistic features and combining a variety of procedures so far employed in other pieces of research. Thus, it will be possible to enhance the understanding on gender identification by means of more statistical methods, with attention paid to the interrelations between properties which may differ in texts written by female and male authors. Moreover, it should be possible to ascertain more quantitative indicators that can be better used for the classification of gender identification.

References

Bernstein, B. (1971). *Class, Codes and Control* (volume 1). London: Routledge and Kegan Paul.

Biber, D., Conrad, S., Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge, UK: Cambridge University Press.

Burrows, J. F. (1987). Word-patterns and story-shapes: The statistical analysis of narrative style. *Literary and Linguistic Computing*2 (2): 61–70.

Burrows, J. F. (2004). Textual analysis. In: Scheibman, S., Siemans, R., Unsworth, J. (eds), *A Companion to Digital Humanities*. Oxford: Blackwell.

Butts, C. T. (2008). Social Network Analysis: a Methodological Introduction. *Asian Journal of Social Psychology* 11:13–41.

Bortolato, C. (2016). Intertextual Distance of Function Words as a Tool to Detect an Author's Gender. A Corpus-Based Study on Contemporary Italian Literature. *Glottometrics* 34, 28−43.

Crawford, M. (1995). *Talking difference: on gender and language.* London: Sage.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America 102*(46), 16569−16572.

Holmes, J. (1993). Women's talk: the question of sociolinguistic universals. *Australian Journal of Communications*20 (3), 125–49.

Jespersen, O. (1922). *Language. Its Nature, development and origin*. London: Allen & Unwin.

Jockers, M. (2013). *Macroanalysis. Digital Methods and Literary History*. Champaign: University of Illinois Press.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6, 97–133.

Koppel, M., Argamon, S., and Shimoni, A. R. (2003). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing* 17, 101–8.

Lakoff, R. (1975). *Language and woman's place*. New York: Harper and Row.

Mikros, G. K. and Perifanos, K. (2013). Authorship Attribution in Greek Tweets Using Multilevel Author's N-gram Profiles. In Hovy, E., Markman, V., Martell, C. H., Uthus, D. (eds), *Papers from the 2013 AAAI Spring Symposium "Analyzing Microtext"*, Stanford, CA, 25-27 March 2013. Palo Alto, CA: AAAI Press, pp. 17–23.

Mikros, G. K. (2013). Systematic stylometric differences in men and women authors: a corpus-based study. In: Köhler, R., Altmann, G. (eds.), *Issues in Quantitative Linguistics* 3,206–223. Lüdenscheid: RAM – Verlag.

Moretti, F. (2005). Graphs, Maps, Trees: Abstract Models for a Literary History. London and New York: Verso.

Newman, M. L., Groom, C. J., Handelman, L. D., and Pennebaker, J. W. (2008). Gender differences in language use: an analysis of 14,000 text samples. *Discourse Processes* 45, 211–36.

Oakes, M. P. (2014). *Literary Detective Work on the Computer*. Amsterdam/Philadelphia: John Benjamins Publishing.

Pan, X., Cheng, X.-Y., and Liu, H.-T. (2017). Harmony in diversity: The language codes in English–Chinese poetry translation. *Digital Scholarship in the Humanities* 1[33], 128–132.

Pennebaker, J. W. (2011). *The Secret Life of Pronouns: What Our Words Say about Us*. London, UK: Bloomsbury Press.

Popescu, I. I. (2007). Text ranking by the weight of highly frequent words. In: Grzybek P. (ed)**,** *Exact Methods in the Study of Language and Text***.** Berlin: Mouton de Gruyter, 555–566.

Popescu, I., Mačutek, J., Altmann, G. (2009). *Aspects of Word Frequencies*. Lüdenscheid: RAM-Verlag.

Rybicki, J.(2016). Vive la diffe´rence: Tracing the (authorial) gender signal by multivariate analysis of word frequencies. *Digital Scholarship in the Humanities* 31(4), 746–761.

Sarawgi, R., Gajulapalli, K., Choi, Y. (2011). Gender attribution: Tracing stylometric evidence beyond topic and genre. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning.* Stroudsburg, PA: Association for Computational Linguistics (ACL), 78–86.

Savoy, J. (2012). Authorship attribution: a comparative study of three text corpora and three languages. *Journal of Quantitative Linguistics* 19(2), 132–161.

Sboev. A., Litvinova, T., Gudovskikh, D., Rybka, R., Moloshnikov, I. (2016). Machine

Learning Models of Text Categorization by Author Gender Using Topic-Independent Features. *Procedia Computer Science* 101, 135–142.

Schler, J., Koppel, M., Argamon, S., Pennebaker, J. (2006). Effects of Age and Gender on Blogging. *Computational Approaches to Analyzing Weblogs: Papers from the AAAI Spring Symposium*. Menlo Park: The AAAI Press, 199–205.

Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology* 60(3), 538–556.

Van Duijn, M. A. J., Zeggelink, E. P., Huisman, M., Stokman, F. N., Wasseur, F. W. (2003). Evolution of sociology freshmen into a friendship network, *Journal of Mathematical Sociology* 27(2–3), 153–191.

Wang, Y. -Q, Liu, H.- T. (2017). Is Trump always rambling like a fourth-grade student? An analysis of stylistic features of Donald Trump's political discourse during the 2016 election. *Discourse & Society* 29 (3), 299–323.

Weidman, S. G., O'Sullivan, J. (2018). The limits of distinctive words: Re-evaluating literature's gender marker debate. *Digital Scholarship in the Humanities* 33(2), 374–390.

Yu, B. (2014). Language and gender in congressional speech. *Literary and Linguistic Computing* 29(1), 118–32.

Zhang, C., and Liu, H. -T. (2017). A Quantitative Investigation of the Genre Development of Modern Chinese Novels. *Glottometrics* 32, 9–20.

Appendix 1

100 authors and their prose spanning the years 1912 to 2019

Author/Gender	Text Name (in Chinese pinyin / in English)	Publication Year
Dazhao Li/M	Jin *Today*	1918
Bannong Liu/M	Yu *Rain*	1920
Disheng Xu/M	Luohuashan *The Peanut*	1922
Pingbo Yu/M	Qinghefang *Qinghe Street*	1925
Luxun/M	Congbaicaoyuandaosanweishuwu *From Baicao Garden to Sanwei Private School*	1926
Zhimo Xu/M	Luoye *The Fallen Leaves*	1926
Zuoren Zhou/M	Wupengchuan *The Black Ship*	1926
Zhenduo Zheng/M	Haiyan *Sea Swallow*	1927
Ziqing Zhu/M	Hetangyuese	1927

	Lotus and the Moonlight	
Shengtao Ye/M	Qianniuhua *Morning Glory*	1931
Laoshe/M	Jinandeqiutian *Autumn in Jinan*	1933
Xuefeng Fen/M	Yangdejiayuan *Family of the Sheep*	1936
Wangshu Dai/M	Balideshutan *A Bookstall in Paris*	1938
Yan Xia/M	Yecao *Grass*	1940
Bajin/M	Feiyuanwai *Outside the Waste Garden*	1941
Maodun/M	Baiyanglizan *Ode to the White Poplar*	1941
Zhongshu Qi-an/M	Lunkuaile *On Happiness*	1941
Weiwei/M	Shuishizuikeaideren? *Soldier in Chinese Federation of Korean War*	1951
Baiyu Liu/M	Richu *Sunrise*	1959
Jianwu Li/M	Yuzhongdengtaishan *Climbing in the Rain*	1961
Boxiao Wu/M	Jiyiliangfangche *A Wheel*	1962
Guangtian Li/M	Huachao *Take the Tide*	1962
Aiqing/M	Nachangyu *Rain*	1980
Dafu Yu/M	Jiangnandedongjing *Winter in Jinan*	1982
Yong Liu/M	Renshengdeqiju *The Game of Life*	1986
Meng Wang/M	Suzhoufu *To Suzhou*	1988
Congwen Shen/M	Lunyouqing *On Friendship*	1990
Zikai Feng/M	Shanzhongbiyu *Rain in the Mountain*	1991
Qiuyu Yu/M	Mogaoku *Mogao Grottoes*	1992
Zengqi Wang/M	Zaijian, Hutong *Say Farewell to Hutong*	1993

Xianlin Ji/M	Qingtangheyun *The Lotus*	1995
Pingwa Jia/M	Tianma *Painting of a Horse*	1997
Hua Yu/M	Yiyuanlidetongnian *Childhood in the Hospital*	1998
Qifang He/M	Qiuhaitang *Begonia*	1999
Zhongxing Zhang/M	Jiuyan *Swallow in the Past*	2000
Guangzhong Yu/M	Tingtingnalengyu *Voice of the Rain*	2003
Keling/M	Yexing *In the midnight*	2004
Shuo Yang/M	Lizhimi *The Lychee*	2006
Moyan/M	Muqin *My Mother*	2008
Shixi Sheng/M	Banlingfeidu *The Deer*	2008
Tiesheng Shi/M	Hehuanshu *Silktree*	2008
Mu Qin/M	Zaixianrenzhangcunshengdedifang *Cactus*	2009
Qingxuan Lin/M	Meiguiyuci *Roses and Thorns*	2009
Xinwu Liu/M	Congyigeweixiaokaish *From a Smile*	2009
Zhongshi Chen/M	Zaihezhizhou *On the Riverbank*	2010
Guozhen Wang/M	Danbo *Simplicity in Life*	2011
Luyao/M	Zaochengcongzhongwukaishi *Start form the Noon*	2012
Jicai Feng/M	Shiguang *Time*	2014
Xiaosheng Liang/M	Penhu *Sprinkling Can*	2016
Alai/M	Yidishuijingguolijiang *A Drop of Water in Lijiang*	2018
Changying Yu-an/F	Baludeyiye *One Night in Paris*	1912
Pingmei Shi/F	Zuihouyimu *The Last Scene*	1915

Weiying Lin/F	Zhusiyumei *Spider Silk and the Plum*	1925
Bingxin/F	Yizhimuji *A Clog*	1926
Xuelin Su/F	Shouhuo *Harvest*	1928
Wei Bai/F	Qingshu *Love Letter*	1933
Luying/F	Chuangwaidechunguang *Spring outside the Window*	1934
Xiaohong/F	Chunyiguashangleshushao *Green Leaves of Spring*	1940
Ling Ding/F	FengyuzhongyiXiaohong *Recalling Xiaohong*	1942
Ailing Zhang/F	Gengyiji *A Chronicle of Changing Clothes*	1943
Fengyuanjun/F	Qingyin *Silence*	1949
Haiyin Lin/F	Chuang *The Window*	1972
Jie Zhang/F	Meng *Dream*	1981
Yangmo/F	Xiaoxi *The River*	1981
Zongpu/F	Zitengluopubu *Wisteria Falls*	1981
Lu Guan/F	Haidemeng *The Dream of the Sea*	1986
Murong Xi/F	Xiegeiyuanfang *Write to the Place I Long for*	1989
Ning Tie/F	Hezhinv *Daughter of the River*	1994
Zijian Chi/F	Shuanghuaizhimei *Beauty of Sadness*	1995
Qijun/F	Ran *Hair*	1996
Canxue/F	Zuizuichunjingdeyuyan *The Most Beautiful Words*	1999
Xiaoyun Yang/F	Shengmingdejiazhi *Meaning of Life*	1999
Jiang Yang/F	Feng *The Wind*	2004
Sanmao/F	Qingniaobudaodedifang *Somewhere beyond the Bluebird*	2004

Fangfang/F	Xihuansudongpo *To Su Dongpo*	2005
Linbai/F	Fenghuang *Phoenix*	2005
Yueran Zhang/F	Jiushiguangshigemeiren *AsBeautiful as the Old Time*	2005
Shuting/F	Xinyan *On my Mind*	2006
Jiangyun/F	Xiaoshiqing *Daily Event*	2008
Ling Shuhua/F	Dengfushishan *Climbing Fuji Mountain*	2009
Xukun/F	Zhuangzaihongqiqu *Great Hongqiqu*	2009
Yingtai Long/F	Musong *Seeing off My Son*	2009
Meijie/F	Leishuizhihua *Tears*	2010
Li Chi/F	Renshengsanjingjie *Three Realms in Life*	2012
Xuexiaochan/F	Henwan *Too Late*	2012
Jianzheng/F	Meilidejian *The Beautiful Cocoon*	2014
Kangkang Zhang/F	Shouwangxihudeqingteng *Guard the Ivy of the West Lake*	2014
Shumin Bi/F	Qingchongzhiai *Love of A Green Grub of a Butterfly*	2014
Chenran/F	Gududenengli *Enjoy Being Longly*	2015
Qiongyao/F	Yuanyuandixinshang *Another way toLove*	2015
Yemi/F	Shanggaoshuiyuan *Travel*	2015
Yishu/F	Burujiuzaijintian *Why not Today?*	2015
Wanganyi/F	Yaoyan *Rumour*	2016
Yangeling/F	Muqingyuxiaoyu *My Mother and the Fish*	2016
Yeqingcheng/F	Tiandingdeyueniang *The Moon*	2016
Xiaoqing Fang/F	Yigerendechezhan *Along at the Station*	2017

Jiangfangzhou/F	Zhongyuneixin Qingxingchengzhang *Be True to One's Heart*	2018
Zhangxiaoxian/F	Meilideyueding *A Date*	2018
Zhangxiaofeng/F	Yujian *Encounter*	2019

Appendix 2

The Value of $\underline{F(h)}$ for Male Authors and Female Authors

Text Name/Male	R_1	$\underline{F(h)}$	Text Name/Female	R_1	$\underline{F(h)}$
Feiyuanwai *Outside the Waste Garden*	0.51	0.49	Qingshu *Love Letter*	0.52	0.48
Congbaicaoyuandaosan-weishuwu *From Baicao Garden to San-wei Private School*	0.50	0.50	Qingchongzhiai *Love of A Green Grub of a Butterfly*	0.61	0.39
Balideshutan *A Bookstall in Paris*	0.45	0.55	Yizhimuji *A Clog*	0.63	0.37
Qingtangheyun *The Lotus*	0.49	0.51	Renshengsanjingjie *Three Realms in Life*	0.60	0.40
Jinandeqiutian *Autumn in Jinan*	0.41	0.59	Shuanghuaizhimei *Beauty of Sadness*	0.56	0.44
Tianma *Painting of a Horse*	0.50	0.50	Feng-yuzhongyixiaohong *Recalling Xiaohong*	0.70	0.30
Meiguiyuci *Roses and Thorns*	0.61	0.39	Zhusiyumei *Spider Silk and the Plum*	0.58	0.42
Baiyanglizan *Ode to the White Poplar*	0.50	0.50	Musong *Seeing off My Son*	0.57	0.43
Muqin *My Mother*	0.49	0.51	Chuangwaidechun-guang *Spring outside the Window*	0.53	0.47
Lunkuaile *On Happiness*	0.50	0.50	Qingniaobu-daodedifang *Somewhere beyond the Bluebird*	0.50	0.50
Zaijian, Hutong *Say Farewell to Hutong*	0.47	0.53	Zuihouyimu *The Last Scene*	0.58	0.42

Danbo *Simplicity in Life*	0.55	0.45	Xinyan *On my Mind*	0.57	0.43
Luoye *The Fallen Leaves*	0.60	0.40	Shouhuo *Harvest*	0.62	0.38
Lizhimi *The Lychee*	0.51	0.49	Hezhinv *Daughter of the River*	0.53	0.47
Tingtingnalengyu *Voice of the Rain*	0.46	0.54	Xiegeiyuanfang *Write to the Place I Long for*	0.57	0.43
Maogaoku *Mogao Grottoes*	0.48	0.52	Chun-yiguashangleshushao *Green Leaves of Spring*	0.64	0.36
Jiangnandedongjing *Winter in Jinan*	0.52	0.48	Feng *The Wind*	0.67	0.33
Jiuyan *Swallow in the Past*	0.54	0.46	Gengyiji *A Chronicle of Changing Clothes*	0.52	0.48
Hetangyuese *Lotus and the Moonlight*	0.52	0.48	Shouwang-xihudeqingteng *Guard the Ivy of the West Lake*	0.59	0.41
Lunyouqing *On Friendship*	0.60	0.40	Zitengluopubu *Wisteria Falls*	0.62	0.38
Shanzhongbiyu *Rain in the Mountain*	0.52	0.48	Haidemeng *The Dream of the Sea*	0.45	0.55
Suzhoufu *To Suzhou*	0.2	0.8	Meng *Dream*	0.45	0.55
Shiguang *Time*	0.52	0.48	Xihuansudongpo *To Su Dongpo*	0.54	0.46
Banlingfeidu *The Deer*	0.48	0.52	Ditingshuisheng *The Sound of Water*	0.60	0.40
Penhu *Sprinkling Can*	0.44	0.56	Yigerendechezhan *Along at the Station*	0.58	0.42
Hehuanshu *Silktree*	0.47	0.53	Fenghuang *Phoenix*	0.51	0.49
Congyigeweixiaokaishi *From a Smile*	0.56	0.44	Zhuangzaihongqiqu *Great Hongqiqu*	0.51	0.49
Yiyuanlidetongnian *Childhood in the Hospital*	0.44	0.56	Shanggaoshuiyuan *Travel*	0.45	0.55
Zaixianrenzhangcunsheng-dedifang *Cactus*	0.57	0.43	Gududenengli *Enjoy Being Longly*	0.54	0.46
Zaochengcongzhongwukaishi	0.50	0.50	Tiandingdeyueliang	0.50	0.50

Start form the Noon			*The Moon*		
Zaihezhizhou *On the Riverbank*	0.52	0.48	Henwan *Too Late*	0.52	0.48
Qianniuhua *Morning Glory*	0.55	0.45	Meilideyueding *A Date*	0.57	0.43
Renshengdeqiju *The Game of Life*	0.24	0.76	Yuanyuandixinshang *Another Way to Love*	0.48	0.52
Yecao *Grass*	0.53	0.47	Muqingyuxiaoyu *My Mother and the Fish*	0.46	0.54
Nachangyu *Rain*	0.47	0.53	Zhongyuneixin Qingxingchengzhang *Be True to One's Heart*	0.54	0.47
Wupengchuan *The Black Ship*	0.52	0.48	Yujian *Encounter*	0.55	0.45
Yidishuijingguolijiang *A Drop of Water in Lijiang*	0.48	0.52	Yaoyan *Rumour*	0.43	0.57
Jin *Today*	0.46	0.54	Xiaoshiqing *Daily Event*	0.52	0.48
Shuishizuikeaideren? *Solider in Chinese Federation of Korean War*	0.45	0.54	Qingyin *Silence*	0.52	0.48
Jiyiliangfangche *A Wheel*	0.48	0.52	Xiaoxi *The River*	0.50	0.50
Luohuasheng *The Peanunt*	0.54	0.46	Meilidejian *The Beautiful Cocoon*	0.49	0.51
Yangdejiayuan *Family of the Sheep*	0.56	0.44	Dengfushishan *Climbing Fuji Mountain*	0.53	0.47
Yuzhongdengtaishan *Climbing in the Rain*	0.53	0.47	Shengmingdejiazhi *Meaning of Life*	0.57	0.43
Qinghefang *Qinghe Street*	0.49	0.51	Chuang *The Window*	0.53	0.47
Richu *Sunrise*	0.51	0.49	Zuizuichunjing-deyuyan *The Most Beautiful words*	0.58	0.42
Yexing *In the midnight*	0.53	0.47	Ditingshuisheng *The Sound of the Water*	0.51	0.49
Qiuhaitang *Begonia*	0.55	0.45	Ran *Hair*	0.48	0.52
Huachao	0.47	0.53	Burujiuzaijintian	0.53	0.47

Take the Tide			Why not Today?		
Haiyan *Sea Swallow*	0.51	0.49	Balideyiye *One Night in Paris*	0.50	0.50
Yu *Rain*	0.61	0.39	Jiushiguang-shigemeiren *As Beautiful as the Old Time*	0.64	0.36