Glottometrics 45, 2019, 24-57

Distance between Chinese Registers Based on the Menzerath-Altmann Law and Regression Analysis

Renkui Hou[a,b]*, Chu-Ren Huang*[b]*, Mi Zhou*[b]*, Menghan Jiang*[b]
[a]College of Humanity, Guangzhou University, Guangzhou, China;
[b]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Kowloon, HongKong

Corresponding author: Renkui Hou, email address: hourk0917@163.com

Abstract. This paper proposes an innovative method/index to represent the formality of a register based on the Menzerath–Altmann law and regression analysis. This index also can be used to quantify the distance between two registers. Analysis demonstrates that average word length decreases with the increase of clause length in each register and that their relationship can be fitted by the formula $y = ax^b$. It can be shown that the link between average word length and clause length abides by the Menzerath–Altmann law. Texts were represented by the fitted parameters, *a* and *b*, and their positions were plotted in 2-dimensions. Linear regression can be used to fit the functional correlation between these two parameters in each register. We show that the *a*-intercept of this regression line can be used as an index to represent the formality degree of the register and to compute the distance between two registers.

Keywords: *Distance between Chinese registers, The Menzerath–Altmann law, Chinese word length, Chinese clause length, Regression analysis.*

1 Introduction

Variability is inherent in human language: people use different linguistic forms on different occasions and different speakers of a language convey the same messages in different ways. Register is often considered to be the most important perspective on text varieties (Biber and Conrad 2009). The register perspective combines an analysis of the linguistic characteristics that are common in particular text varieties with an analysis of the situations of use of those varieties.

The essential features of registers involve three factors: context, linguistic materials, and fixed ways of expressing objects, the combination of which forms a discourse. We will discuss the distances between different Chinese registers based on the Menzerath–Altmann law (henceforth, the MA law), which explores the relationship between language constructs and their immediate constituents, from the perspective of quantitative linguistics.

The MA law, which is one of the best known quantitative linguistic laws, originates from the fact that the length of a construct influences the lengths of its immediate constituents in different language domains. Paul Menzerath summarized the law as "the greater the whole, the smaller its parts" after he detected the dependency of syllable length on word length (Menzerath, 1954, p.101). Altmann generalized this hypothesis to all levels of linguistic analysis, formulating it as "The longer a language construct, the shorter its components" (Altmann, 1980). Hřebíček (1992, 1995, 1997) showed that the whole hierarchy of textual levels is based on this dependency, and called this the Menzerath–Altmann law.

The theoretical derivation and corresponding differential equation of the MA law were proposed by Altmann (1980) in his seminal 'Prolegomena to Menzerath's Law', as shown in Equation (1).

$$\frac{y'}{y} = -c + \frac{b}{x} \qquad \text{Equation (1)}$$

The solution to this differential equation is shown in the Formula (1):

$$y = ax^b e^{-cx} \qquad \text{Formula (1)}$$

where y is the mean size of the immediate constituents (average word length in this study), x is the size of the construct (clause length), and parameters a, b, and c depend mainly on the levels of the units under investigation, rather than on the language, the kind of text, or the author, as had previously been expected (quoted by Köhler, 2012). However, there is no convincing theoretical support for the substantiated interpretation of these parameters although it is a well-known distribution model in linguistics (Eroglu 2014). In this study, we will demonstrate that these parameter values are affected by the registers in Chinese.

It has previously been assumed that one of the two parameters, either b or c, can be neglected from the function. Then, two simplified forms are obtained:

$$y = ax^b \qquad \text{Formula (1a)}$$
$$y = ae^{-cx} \qquad \text{Formula (1b)}$$

A large number of observations have shown that parameter c is close to zero for higher levels of language whereas lower levels lead to very small values of parameter b; only for intermediate levels is the full formula needed (Köhler, 2012). Formula (1a) has become the most commonly used "standard form" for linguistic purposes (Grzybek, 2007).

This paper aims to establish an index to measure the formality of registers and to represent the distance between two Chinese registers based on the MA law and regression analysis.

1.1 Literature review

Generally speaking, a register is associated with a particular situation of use. It refers to the principles generated in communication and followed by speakers and listeners. Register and

linguistic performance are interdependent and are not tenable without each other as register is produced and shaped by linguistic performance and, in return, its rules regulate linguistic performance once it is formulated. Except for utterances with improper register, all utterances can be categorized into a register. Biber (2012) argued strongly that reference works that describe different linguistics levels, i.e., lexical, grammatical, and lexico-grammatical, should consider register difference. For example, Cacoullos (1999) provided evidence that reductive change in grammaticalizing forms may be manifested not only as a diachronic process but also as synchronic differences between formal and informal registers. The significance of comparing different registers in studies of Chinese grammar was introduced by Lv (1992). Zhang (2012) has shown that there is much variation of linguistic properties across written Chinese registers. Consequently, we should observe the differences of manifestation of quantitative linguistic laws in different registers. For example, Hou et al. (2017) showed that the relationship between sentences and their constituting clauses abides by the MA law in written formal register texts, but not in *TV Sitcom* and *TV Conversation*. Failing to take register into account can lead to inaccurate, even incorrect, conclusions.

Biber's (1994) observation of the lack of agreement on the definitions and taxonomy of registers also applies to the study of registers in Chinese. Yuan and Li (2005) took a discrete approach and proposed seven registers: conversational, officialese, scientific, news, literary and art, lectures, and advertisements. Similar to Biber and Conrad (2009), who regard register differences as a continuum of variation, Feng (2010) thought that register is generated in interpersonal communication and that the essence of register is to adjust the psychological distance between the communicators. He held formality to be the primary element of register and proposed that register is a polarized opposite continuum, with the written formal register being the most formal, the daily informal register being the most informal, and all other registers lying in between. However, the positions of other registers in this continuum and the distances between various registers were not discussed. We adopt Biber's (1994) position to reconcile the above differences: registers are varieties in a continuum, but they are still to be analytically identified as different categories.

Köhler (2012) pointed out that the mathematical methods are worth being integrated into linguistics. Register can also be studied using such mathematical methods. Biber (1986, 1988) is generally credited with introducing quantitative methods to the linguistic study of registers. Biber (1995) restated and underlined the role of computational, statistical, and interpretive techniques using multi-dimensional analysis. He pointed out that any text characteristic that is encoded in language and can be reliably identified and counted is a candidate for inclusion. Research on register characteristics has also been undertaken from the perspective of quantitative linguistics. For example, Hou, Huang, and Liu (2017) fitted the distribution of Chinese sentence lengths using nonlinear regression and used the fitted parameters as quantitative features of the corresponding Chinese registers. In this paper, we propose an index to represent the formality of registers and quantify the distance between two registers based on the MA law and regression analysis.

As one of the best-known laws of quantitative linguistics, the MA law establishes the interrelations between successive hierarchical levels of language, providing evidence that language is a self-organizing and self-regulating system. Previous research has validated the

MA law at different language levels. For example, Köhler (1982) conducted the first empirical test of the MA law at the sentence level, analyzing short stories in German and English and philosophical texts. In his investigation, Köhler counted clause lengths in terms of the number of constituent words. Statistical tests on the data confirmed the validity of the law with high significance. Tuldava (1995) examined the dependence of average word length on clause length, finding a statistically highly significant interdependence between average word length and clause length, indicating that there are other factors that influence average word length. Motalová et al. (2014) and Ščigulinská and Schusterová (2014) verified the validity of the MA law applied to contemporary written and spoken Chinese respectively. Benešová (2016) tested the potential validity of the MA law on samples in different languages and attempted to test the concept of this language universal. Wilson (2017) used the MA law to test the hypothesis that the intonation unit is a valid language construct whose immediate constituent is the foot.

Benešová & Čech (2015) proved the MA law from another perspective. They conducted that the data generated by random models does not fulfil the MA law. Consequently, they pointed out that the results can be viewed as another argument supporting the assumption considering that the MA law expresses one of important mechanisms controlling human language behavior.

In addition to applications of the MA law at different language levels, some researchers have studied the theory and formula of the law, which has been interpreted in various ways. For example, Köhler (1989) proposed that the mechanism of shortening is a consequence of memory limitations: the longer the construct, the more space must be reserved for structural information between the constituents, hence the size of the constituents must be reduced.

Hammerl and Sambor (1993) concluded that there is a negative correlation between the parameters of the MA law: the greater the value of *a*, the less the value of *b* (quoted in Kułacka, 2010). Cramer (2005) confirmed that the parameters, *a* and *b*, depend on the linguistic level of analysis and also showed that there is a functional correlation between *a* and *b*. This paper will also investigate the functional correlation between these two parameters in each register using linear regression.

1.2 Research question and methodology

This paper proposes an index to represent the formality of a register and the distance between two registers based on the MA law from the perspective of quantitative linguistics and regression analysis.

Effective register analyses are always comparative as it is virtually impossible to know what is distinctive about a particular register without comparing it to others. We have therefore selected texts from multiple registers to establish the corpus.

In contrast to Indo-European languages, it is difficult to define the terms “sentence” and “clause” in Chinese. Chinese sentences are often defined in terms of characteristics of speech (Huang and Shi, 2016; Lu, 1993). Chao (1968) and Zhu (1982) defined a sentence as an utterance with pauses and intonation changes at its boundaries. Huang and Liao (2002: P4) proposed that a sentence is a linguistic unit that has an intonation and can express a relatively complete meaning in Chinese. However, sentences are often defined using punctuation marks in corpus linguistics and quantitative linguistics. A common approach for identifying sentences

in syntactically annotated corpora (e.g., Chen et al., 1996; Chen et al., 2013; Huang and Chen, 2017 for Sinica TreeBank) is to mark all segments between punctuation marks that indicate utterance pauses as sentences. Such punctuation marks include commas, semicolons, colon, periods, exclamation marks, and question marks. Wang and Qin (2014) and Chen (1994) also adopted this operational definition and called such units *sentence segments*. Chen (1994) reported that about 75% of Chinese sentences are composed of more than two sentence segments separated by commas or semicolons by corpus analysis. Wang and Qin (2014) considered the lengths of sentence segments to be relevant to language use in Chinese. In fact, sentences (as defined by Chen et al., 2003; Huang and Chen, 2017) and sentence segments (as defined by Chen, 1994; Wang and Qin, 2014) are roughly equivalent to clauses. One sentence is composed of one or more clauses, which is called simple sentence or complex sentence (Huang and Liao, 2002: P5). The structures of the simple sentences and clauses are similar in Chinese, but the latter lack a complete intonation. In complex sentences, there are generally pauses represented by commas, semicolons and colons between clauses. Pauses at the boundaries of the sentences are represented by the periods, exclamation marks, and question marks (Huang and Liao, 2002: p 159). Thus, an operational definition of Chinese clauses can also be based on the written form, and the aforementioned punctuation marks determine the boundaries of the clauses.

It has become common in quantitative linguistics to measure the length of a linguistic entity as the number of its immediate constituents. We assume that the immediate constituents of Chinese clauses are words, hence clause length can be defined as the number of words. We consider words to be the segments delineated by blank spaces in the texts segmented by a Chinese lexical analysis system. There are various perspectives to define word length, for example, from the perspectives of pronunciation, duration, and syllable number. For Chinese, we define word length as the number of Chinese characters (*Hanzi*, 汉字) in the word (Hou, Yang and Jiang, 2014; Chen and Liu, 2016).

We selected Formula (1a) to fit the function between average word length and clause length in Chinese. Formula (1a) shows that this function is nonlinear. This nonlinear function can be transformed into a linear function in order to avoid the impact of the initial parameter estimates on the fitted result.

$$y = ax^b \qquad \text{Formula (1a)}$$

Taking the logarithm of both sides of Formula (1a) gives

$$\ln(y) = \ln(a) + b\ln(x)$$

Then, defining

$$Y = \ln(y); \quad X = \ln(x)$$

The linear function stated in Formula (1a-1) is obtained:

$$Y = bX + \ln(a) \qquad \text{Formula (1a-1)}$$

If the logarithm of average word length distribution can be fitted by this linear regression, as shown in Formula (1a-1), the average word length can be fitted by the non-linear regression, as shown in Formula (1a). We will show that the fitted result using linear regression is as well as that using nonlinear regression in later section. Thus the determination coefficient (R^2) was used to validate the fitted results of this linear regression as like residual sum-of-square for the validation of nonlinear regression result; it shows the goodness-of-fit of the model to the empirically collected data. It indicates the proportion of variance in the data that can be explained by the model (Conway & White, 2013). In quantitative linguistics, a fit is generally considered good if R^2 is greater than or equal to 0.9 (Popescu et al., 2009, p.16). A fit with $0.9 > R^2 > 0.7$ is tolerable. Our study will show that the residual sum-of-squares of nonlinear regression is small if the R^2 of linear regression is large. In addition, the different settings of initial parameter values affect the fitted result. Since the aim of the paper is to obtain the parameters, *a* and *b*, to represent the texts and then calculate the distance between the different registers, an approach that does not reliably yield constant parameters is not appropriate. We adopt the linear regression approach in this study because it can be used to fit the logarithm of average word length distribution and obtain the parameters.

The function between average word length and clause length was fitted by Formula (1a-1) in each text. Then the texts from various registers were represented by the fitted parameters, *a* and *b,* using a vector space model, allowing the positions of each register texts to be displayed on a coordinate graph. The positions of the texts in each register indicate that there is a systematic link between parameters *a* and *b* in the texts from each register, which can be fitted by linear regression. The point at which the regression line intersects the *a*-axis when *b* achieves its extreme maximum value, i.e., 0, is dependent on the particular register. The value of the *a*-intercept can be used as an index to represent the position of a register in the formality continuum and to quantify the distances between various registers.

We used the open source programming language and environment R (R Core Team, 2016) to realize the fitting procedure and for the computation of both clause length and average word length. The R function *lm*() was used to fit Formula (1a-1) in order to obtain the values of parameters *a* and *b*, and to carry out regression analysis on the link between parameters *a* and *b* in texts from the same register.

2. Corpus Establishment and Preprocessing

Texts from "*News Co-Broadcasting*", the situation comedy "*I Love My Family*", and "*Behind the Headlines with Wentao*" were selected to represent the *News Broadcasting*, *Sitcom Conversation*, and *TV Conversation* (i.e, *TV Talkshow*) registers respectively.

The Central China TV (CCTV) program, "*News Co-Broadcasting*", mainly consists of brief introductions of important state policies and events taking place both at home and abroad. It is characterized by formal use of language in non-interactive uni-directional speech. It is the representative of the *News Broadcasting* register.

"Behind the Headlines with Wentao" is a talk show of Phoenix Satellite TV in which the host discusses current hot issues and topics together with guests. Their dialogue is supposed to be

un-scripted with real time interaction. The speakers aim to entertain, inform, and even persuade the audience. The language use is representative of the *TV Conversation* register.

The situational comedy, "*I Love My Family*", tells the story of a family via well-constructed casual dialogues. Although the content is scripted, it is expected that the delivery should be informal and intimate. This is the representative of the *Sitcom Conversation* register.

Overall and intuitively, the *News Broadcasting* register is the most formal one, due both to its scripted nature, and the nature of being one-way communication aiming to inform. *TV Talkshow* is supposed to be less formal, due to its interactive and unscripted nature. Yet its discussion is still topical and the social inter-personal relation is only minimally expressed. Hence it is considered to be less formal. Lastly, even though *TV sitcom conversation* has to be scripted, it is scripted to reflect characteristics as well as the relation between the speaker and the addressee. And even though the conversation is meant to be heard by the audience, it doesn't need the audience to acquire information and gain information. Given that these contrasts, the register differences may be complex. We will use our result to explore whether the formality of register is dependent on one or more specific features.

The texts of *News Broadcasting* were obtained from the National Broadcast Language Resources Monitoring and Research Centre at the Communication University of China. Textual materials of "*Behind the Headlines with Wentao*" were collected from the website of Phoenix Satellite TV. The texts of "*I Love My Family*" were downloaded from the Internet. The names of speakers were deleted because they do not occur in either "*Behind the Headlines with Wentao*" or "*I Love My Family*".

The Chinese lexical analysis system created by the Institute of Computing Technology of the Chinese Academy of Sciences (ICTCLAS) was used for word segmentation. ICTCLAS has been acknowledged as having a high accuracy of 97.58%, a recall rate of over 90% for the recognition of unknown words based on role tagging, and a recall rate of approximately 98% for the recognition of Chinese names[1].

The segmented texts were screened manually. For example, words within bracket pairs in "*Behind the Headlines with Wentao*" were deleted if they were explanatory notes because explanatory notes are not considered to be parts of the texts. No special treatment was given to deal with isolated numbers and letters in the corpus.

The scales of the texts from these three registers are shown in Table 1.

Table 1
Scale of the texts from the different registers

	Number of Texts	Number of Types	Number of Tokens
News Co-Broadcasting	50	24,812	418,943
Behind the Headlines with Wentao	50	16,372	357,663
I Love My Family	60	14,107	317,661

[1] http://www.ict.ac.cn/jszy/jsxk_zlxk/mfxk/200706/t20070628_2121143.html

Having conducted preliminary research on the texts from these three registers, an index which can represent the formality degree and compute the distance between two registers was deduced. We then performed a test of validity of the index on the Lancaster Corpus of Mandarin Chinese (LCMC), which became available in 2003 (McEnery and Xiao, 2004). This corpus includes 500 texts of 2,000 word tokens each (i.e., totaling 1,000,000 words) from 15 written registers, taken from publications from mainland China between 1988 and 1992. We believe that this verification can make the conclusions that we draw here robust.

3 Experiments

3.1 Frequency distribution of clause length in terms of words

The frequency distributions of clause length in terms of words for each register were established, as shown in Figure 1. The occurrence frequency distributions and the relative occurrence frequencies of clauses with certain lengths are shown in Appendix 1 and 2 respectively. The figure demonstrates that the clause length distributions are similar in each register. In *Sitcom Conversation* texts, one-word clauses are more frequent than clauses with other lengths, reflecting the prevalence of such one-word clauses in daily conversation. The frequencies of clauses in texts from the other two registers, *News Broadcasting* and *TV Conversation*, first increase and then decrease with clause length.

The cumulative relative frequency distributions of clause lengths for each register are shown in Figure 2, from which we observe that most clauses are composed of few words. More than 98% of clauses in *TV Conversation* and *Sitcom Conversation* are composed of 1 to 15 words. About 99% of clauses in *News Broadcasting* are composed of fewer than 20 words. Figure 1 shows that the short clauses appear more frequently and longer clauses appear less frequently. Figure 2 shows that most clauses are short.

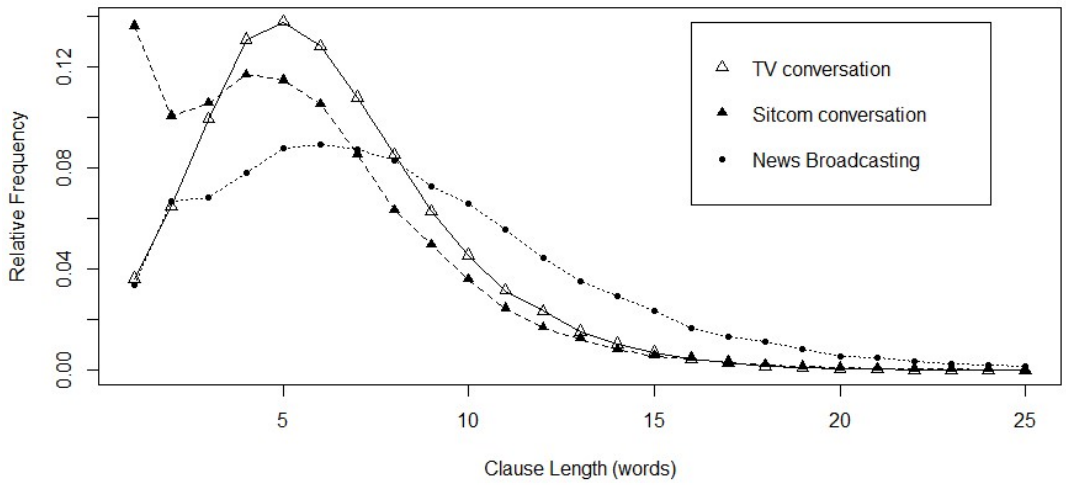

Figure 1: Frequency distributions of clause length in terms of words

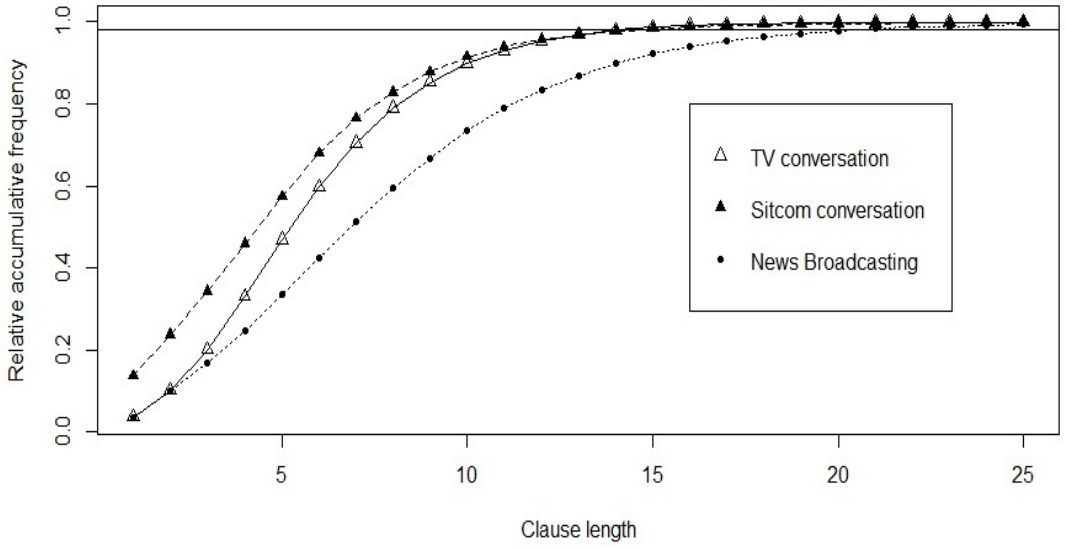

Figure 2: Cumulative relative frequency distributions of clause length in terms of words

3.2 Average word length distribution in clauses

The average word length in clauses with a certain length was calculated as the number of Chinese characters in the given clauses divided by the number of words in those clauses, which is shown in Appendix 3. As well as for texts from these three registers, we also calculated the average word length in the clauses having a certain length across texts from all registers.

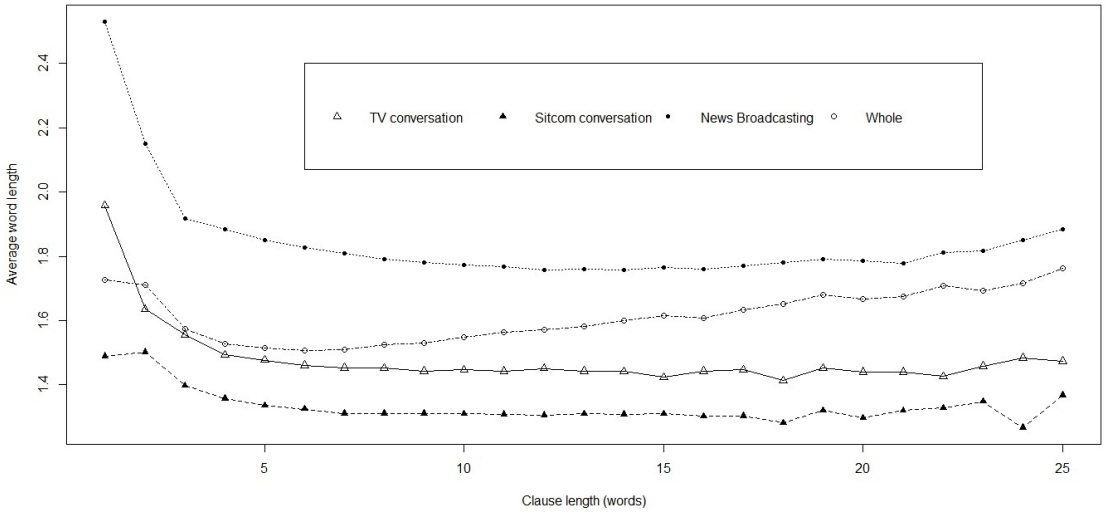

Figure 3: Average word length distributions in clauses

Figure 3 shows the negative relationship between average word length and clause length in each register. The average word length decreases with the increases of clause length in most clauses. The reason for the irregular change of average clause length in few long clauses needs to be explored in Chinese. From the figure, we observe that average word length in *News*

Broadcasting and *TV Conversation* texts decreases with clause length for most clauses. In *Sitcom Conversation*, the average word length in one-word clauses is smaller than in two-word clauses due to the large frequency of one-character words in one-word clauses, which are mostly interjections. In clauses with more than 1 word, the average word length decreases with increase of clause length. However, for all texts across registers, the average word length decreases with clause length only for short clauses of 1 to 6 words, accounting for 57.3% of all clauses. For longer clauses, the average word length increases with clause length. It is necessary to examine the distribution of average word length separately in each register in Chinese; otherwise, an incorrect conclusion would be obtained.

3.3 Regression analysis

Formula (1a-1) was selected to fit the relationship between average word length and clause length. In the fitting process, the clauses whose lengths are 15, 15 and 21 words in *TV Conversation*, *Sitcom Conversation* and *News Broadcasting* were fitted respectively. The fitted results are shown in Table 2 and Figure 4.

In Table 2, the values of determination coefficient, R^2, show that the link between the logarithm of average word length and the logarithm of clause length can be fitted by Formula (1a-1) for each of the three registers: *News Broadcasting*, *TV Conversation*, and *Sitcom Conversation*. The *p*-values, which are all smaller than 0.05, indicate the presence of a significant linear relationship between Y (the logarithm of average word length) and X (the logarithm of clause length).

The residual sum-of-squares is considered the measure to validate the result of nonlinear regression. We also calculated the residual sum-of-squares of the result of linear regression, which is the sum of squares of the difference between the predicted values and the observed values, in order to compare the results between linear regression and nonlinear regression.

Non-linear regression was used to fit the average word length distribution in *TV Conversation* text. We used the values of parameters, which obtained from the linear regression of the logarithm of average word length distribution, as the initial values of them. The residual sum-of-squares is 0.053 in the nonlinear regression result of the average word length distribution in *TV Conversation* text. In the meantime, the residual sum-of-squares is 0.054 using the fitted result of linear regression in *TV Conversation* text. The difference is 0.001 between them, which means the result of linear regression is as well as that of the nonlinear regression.

Similarly, the residual sum-of-squares is 0.009 in the nonlinear regression of the average word length distribution in *Sitcom Conversation* text. In the meantime, the residual sum-of-squares is also 0.009 when the linear regression was used to fit the logarithm of the average word length distribution in *Sitcom Conversation* text. The same values of residual sum-of-squares means the results of linear and nonlinear regressions are both well. In addition, the residual sum-of-squares in the regression result of average word length distribution in *Sitcom conversation* is less than that in *TV Conversation*. It means the regression result of the average word length distribution in *Sitcom Conversation* is better than that in *TV Conversation*. In the meantime, the R^2 of the linear regression result of average word length distribution in

Sitcom Conversation is more than that in *TV Conversation*. The linear regression result in *Sitcom Conversation* is better than that in *TV Conversation*. The conclusion is as same as that from the residual sum-of-squares.

The values of residual sum-of-squares are 0.153 in nonlinear regression of average word length distribution and 0.158 in linear regression of the logarithm of average word length distribution in *News Broadcasting*. The little difference between these two values showed that the results of linear regression is as similar as that of nonlinear regression. This residual sum-of-squares is more than that in *TV Conversation* and *Sitcom Conversation*. In the meantime, the R^2 is less than that in *TV Conversation* and *Sitcom Conversation*. They all showed that the fitted result of average word length distribution in *News Broadcasting* is not as well as that in *TV Conversation* and *Sitcom Conversation*.

We can see that the linear regression result of the logarithm of average word length distribution is similar with the nonlinear regression result of average word length from the comparison of the residual sum-of-squares. The more R^2 means the smaller residual sum-of-squares, which means that the good fitted result. The R^2 in line regression can also validate the fitted result of nonlinear regression result indirectly.

Hence we used linear regression to fit the average word length distribution because its result is similar with the nonlinear regression and the values of parameters is not set beforehand.

Table 2

Fitted results of link between average word lengths and clause length

	a	*b*	R^2	*p*-value
TV Conversation	1.784	−0.093	79.38%	8.352×10^{-6}
Sitcom Conversation	1.490	−0.055	84.74%	1.148×10^{-6}
News Broadcasting	2.240	−0.091	75.28%	3.513×10^{-7}
Whole	1.626	−0.013	6.94%	0.291

For each register, the value of parameter *b* is negative, which indicates that average word length decreases with clause length. Thus, as can be seen from Table 2 and Figure 4, the relationship between clauses and their constituent words abides by the MA law in each register. For texts across all three registers combined, $R^2 = 6.94\%$, indicating that the link between average word length and clause length cannot be fitted by Formula (1a-1), and the *p*-value, 0.291 (which is greater than 0.05), shows that there is not a linear relationship between *Y* and *X*, indicating that the relationship between clauses and their constituent words does not abide by the MA law.

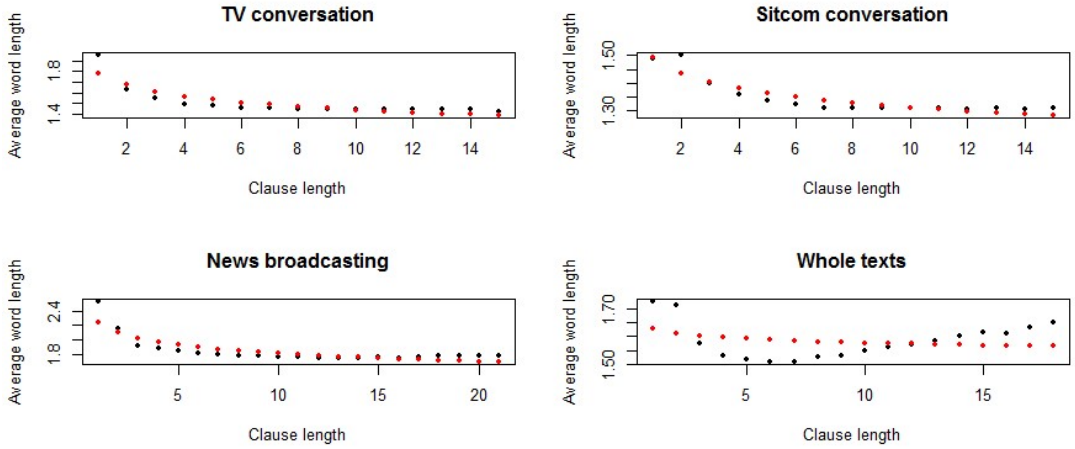

Figure 4. Fitted results of link between average word length and clause length (black dots represent the observed values of average word length; red dots represent the fitted values of average word length)

The long clauses have to be included in this experiment in order to consider as many clauses as possible, especially in the texts from *News Broadcasting*, as indicated by Figures 1 and 2. Figure 4 and Table 2 show that the link between average word length and clause length across the three registers combined cannot be fitted by Formula (1a-1) and, therefore, does not abide by the MA law. Thus, it is necessary to focus on particular registers in exploring this link based on the MA law.

3.4 Method to compute the distance between two registers

The average word length in clauses was calculated for each text in the corpus. The links between average word length and clause length were fitted by Formula (1a-1), allowing each text to be represented by its fitted parameters, *a* and *b* of the MA law (the values of these two parameters in all texts are shown in Appendix 4). The distributions of these two parameters among texts from each register are shown in Figure 5 using box plots. Box plots provide a graphical way to display median, quartiles, and extremes of a data set on a number line to summarize the distribution of the data. As can be seen from Figure 5, there are significant differences among the values of parameters *a* and *b* across the registers.

Correlation analysis examines possible correlations, such as direction and degree, between different phenomena. Pearson's correlation coefficient, the most widely used measure of dependence, was selected to compute the correlation direction and degree between parameters *a* and *b* of the MA law both within each register and across registers. Different values of the correlation coefficient indicate different directions and degrees of relevancy between the two variables. In the extreme case, a correlation coefficient value of 1 (or −1) indicates a perfectly linear positive (or negative) correlation between them. The closer the coefficient is to either −1 or 1, the stronger the correlation is between the two variables.

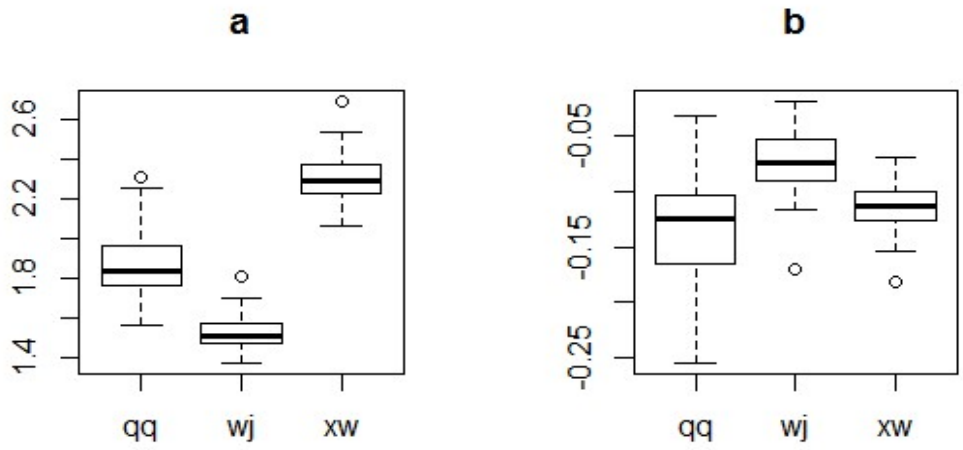

Figure 5: The distribution of fitted parameters, *a* and *b*, in the texts from different registers ("qq" refers to *TV Conversation*, "wj" refers to *Sitcom Conversation*, "xw" refers to *News Broadcasting*)

For texts across registers, the correlation coefficient between *a* and *b* is −0.634, which shows a negative correlation between them. The smooth trend line in Figure 6 shows that there is no regular functional relationship between parameters, *b* and *a*, across registers, although they are negatively correlated.

The correlation coefficients between the parameters are −0.870, −0.983, and −0.917 for texts in the *News Broadcasting*, *TV Conversation*, and *Sitcom Conversation* registers, respectively. The strong negative correlation between the parameters can be fitted by linear regression in each register. Kelih (2010) also proposed that there is a functional correlation between *a* and *b* of the MA law. On the basis of that interpretation, Köhler predicted that the borderline case forms a straight line (according to Kelih 2010).

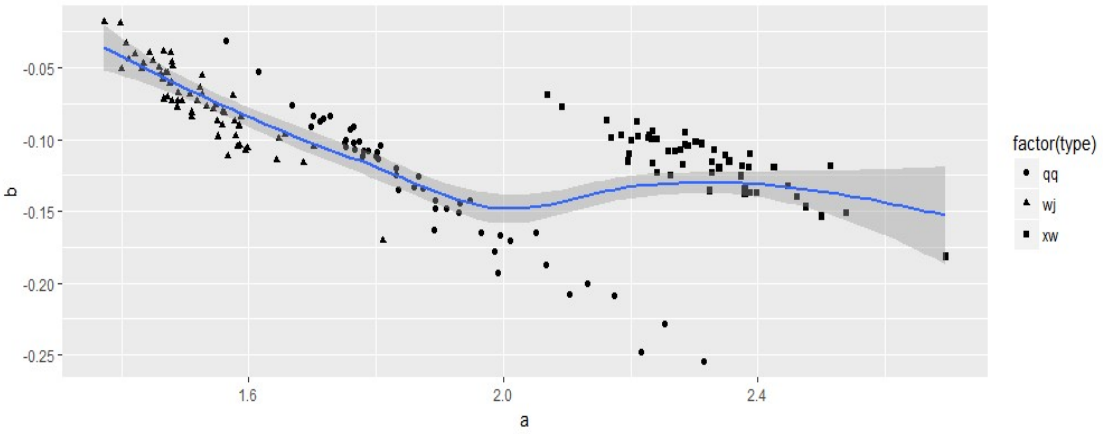

Figure 6: The negative correlation between parameters *b* and *a* across various registers ("qq", "wj", and "xw" refer to *TV Conversation*, *Sitcom Conversation*, and *News Broadcasting*, respectively)

Figure 6 shows that there are obvious boundaries among the texts from each register. In particular, the distance between the *News Broadcasting* texts and other register texts is large. The *Sitcom Conversation* and *TV Conversation* texts are close together, but far from the *News Broadcasting* texts, reflecting their different degrees of formality. From Figure 6, we also observe that parameter *b* is strongly negatively correlated with parameter *a* in each register.

Linear regression, realized by function *lm*() in R, was used to fit the functional link between these two parameters in each register. The fitted results are shown in Table 3 and Figure 7. The values of R^2 show that the fitted results are good and that there is negative linear relationship between parameters, *b* and *a*, of the MA law in each register.

Table 3

Fitted results of the relationship between parameters *b* and *a* of the MA law in each register

	Slope	*intercept*	R^2	*a*-intercept
TV Conversation	−0.288	0.405	96.53%	1.408
Sitcom Conversation	−0.304	0.389	84.01%	1.281
News Broadcasting	−0.153	0.238	75.69%	1.561

As mentioned in section 2, *News Broadcasting* is the most formal register whereas *Sitcom Conversation* is the most informal. In Table 3, for each register, the intercept is the value of the intersection of the fitted line with the *b*-axis. The *a*-axis intercept of the fitted line is obtained when *b* is equal to 0. The *a*-axis intercepts are 1.561, 1.408 and 1.281 in *News Broadcasting*, *TV Conversation*, and *Sitcom Conversation* respectively. It can be seen that the order of these values from large to small is consistent with the formality rank of the corresponding registers from formal to informal.

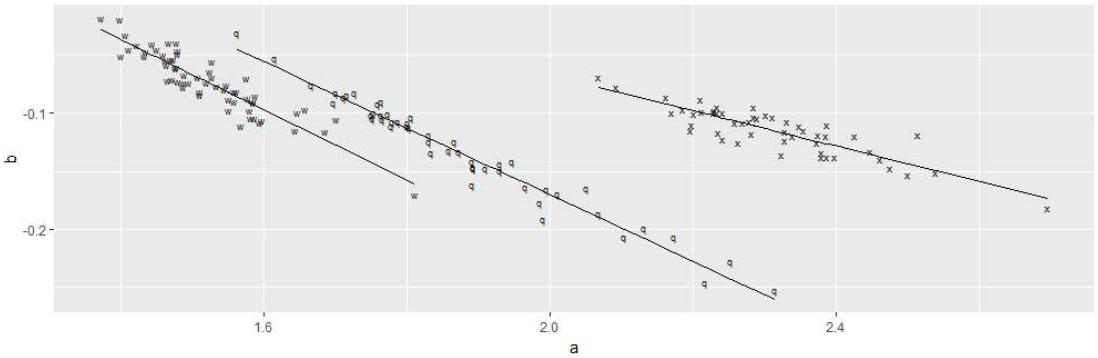

Figure 7: Regression line between fitted parameters, *b* and *a*, in each register ("*q*", "*w*", and "*x*" represent *TV Conversation*, *Sitcom Conversation*, and *News Broadcasting*, respectively)

We propose that the *a*-axis intercept can be used as an index to evaluate the formality degree of the register. For example, the formality degree of the *News Broadcasting* register is 1.561, and it is the most formal of the three registers. The distance between two registers can be quantified using the difference between their formality degrees, i.e., the *a*-axis intercepts of their fitted

lines. For example, the distance between *News Broadcasting* and *TV Conversation* is 0.153, with the former register more formal than the latter.

3.5 Test of Hypothesis

We aim to test the following three hypotheses: (1) that the link between average word length and clause length abides by the MA law; (2) that there is a linear relationship between the fitted parameters, *a* and *b*, in each register; and (3) that the *a*-axis intercepts of the fitted lines can be used to represent the formality degree of Chinese registers and to quantify the distances between two registers. The Lancaster Corpus of Mandarin Chinese (LCMC) was used to verify the above conclusions. A summary of the LCMC corpus is presented in Table 5 (McEnery and Xiao)..

Table 5
Text type and number in the LCMC

Text type	Text Number	Text type	Text Number
Press reportage (A)	44	Academic prose (J)	80
Press editorial (B)	27	General fiction (K)	29
Press reviews (C)	17	Mystery/detective fiction (L)	24
Religious writing (D)	17	Science fiction (M)	6
Instructional writing (E)	38	Adventure fiction (N)	29
Popular lore (F)	44	Romantic fiction (P)	29
Biographies/essays (G)	77	Humor (R)	9
Official documents (H)	30		

We selected texts from the press reportage (A), press editorial (B), press reviews (C), official documents (H), academic prose (J), general fiction (K), science fiction (M), and adventure fiction (N) text types in LCMC. Texts from the press editorial and press reviews represent the *Press Editorials* register. Texts from general, adventure, and science fiction represent the *Fiction* register. Texts from academic prose represent the *Science* register. These registers are chosen for their variety in formality and also in terms of differences in media and modes of communication.

The cumulative relative frequencies of clause lengths, shown in Figure 8, indicate that 96% of clauses in the *Fiction* register, in the *Press Reportage* and *Press Editorials* registers, and in the *Officialese* and *Science* registers contain up to 12, 15, and 18 words, respectively.

As can be seen from Figure 9, the average word length decreases with clause length, except when the clause is very long. The average word length distributions are shown in Appendix 5. Figure 8 shows that these long clauses account for a very small proportion of clauses. We therefore infer that there is an inverse relationship between average word length and clause length.

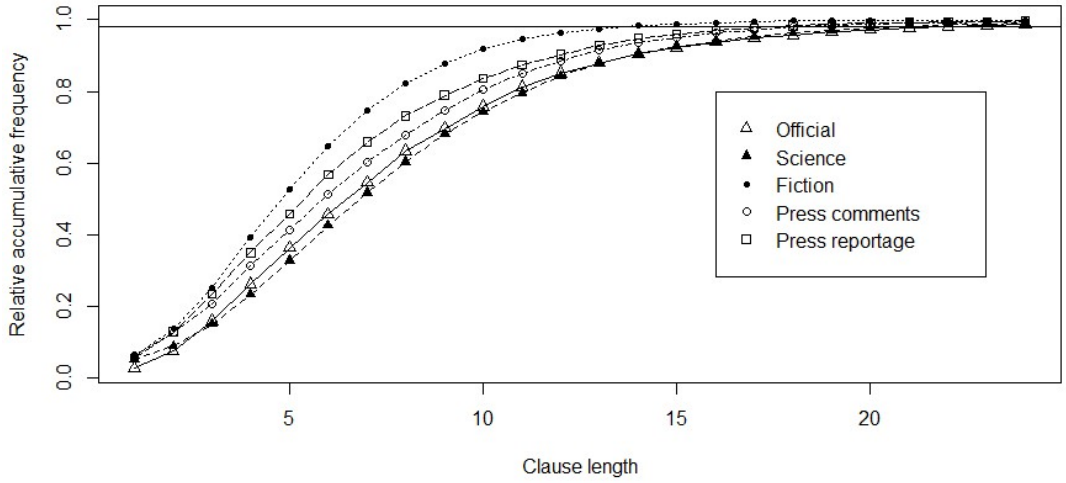

Figure 8. Cumulative relative frequencies of clause length in terms of words

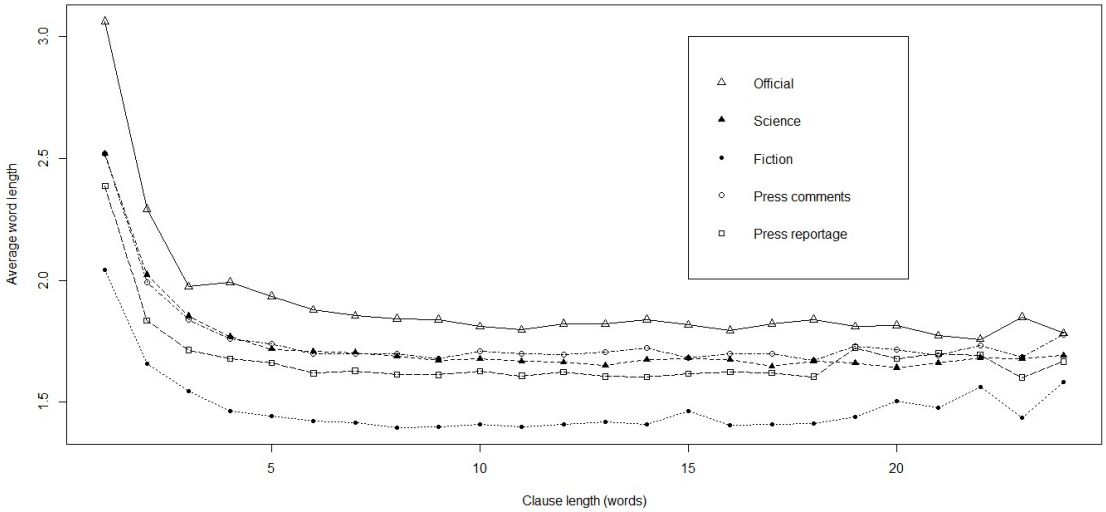

Figure 9. Distribution of average word length in clauses

Table 6

Fitted parameters of average word length distributions

	a	*b*	R^2	p-value
Officialese	2.697	−0.184	83.92%	2.847×10^{-5}
Science	2.295	−0.149	85.96%	1.430×10^{-5}
Fiction	1.869	−0.136	84.40%	2.437×10^{-5}
Press Editorials	2.266	−0.139	80.38%	7.825×10^{-5}
Press Reportage	2.117	−0.129	75.92%	2.228×10^{-4}

Formula (1a-1) was used to fit the average word length distribution for the texts from each of these five registers. The range of clause length was set to be 1:12. The fitted results are shown in

Table 6. The R^2 values demonstrate that the fitted results are good and the *p*-values indicate that the inverse relationships are significant. Thus, the link between average word length and clause length for the texts from each of these five registers abides by the MA law.

Next, pairs of texts in each register were merged to form a single text in the corpus — this was done because the numbers of clauses in the original texts were not enough to assess the clause frequencies of certain lengths. The average word length in clauses was calculated in this corpus. The relationships between average word length and clause length were fitted by Formula (1a-1). The texts were represented by the fitted parameters *a* and *b*, whose values are shown in Appendix 6.

Similar to section 3.3, linear regression was used to determine the systematic correlation between these two parameters, *b* and *a*, in each register. The fitted results are shown in Table 7 and the regression lines are shown in Figure 10.

Table 7
Fitted parameters of the function between parameter *b* and *a* in each register

	Slope	*b-intercept*	R^2	*a*-intercept
Officialese	−0.149	0.234	96.79%	1.570
Science	−0.189	0.281	86.62%	1.487
Fiction	−0.250	0.332	79.84%	1.328
Press Editorials	−0.238	0.401	80.36%	1.685
Press Reportage	−0.189	0.275	81.81%	1.455

The *a*-intercepts of fitted lines were calculated, which are 1.328, 1.455, 1.487, 1.570, and 1.685 in the *Fiction*, *Press Reportage*, *Science*, *Officialese*, and *Press Editorials* registers respectively, as shown in Table 7. These numbers show that the formality degree increases from *Fiction* to *Press editorials*. Hence, the *a*-intercept can be used as an index to represent the formality degree of a register and to quantify the distance between two registers. For example, the distances between *Press reportage* and *Fiction*, and between *Press reportage* and *Science* are 0.127 and -0.032 respectively. Hence, we can say that *Press reportage* is closer to *Science* than to *Fiction* in terms of formality degree and *Press reportage* is more formal than *Fiction*, while *Press reportage* is less formal than *Science*. This is consistent with our intuitive experience.

Table 8
Formality Grouping of Registers according to *a*-intercept

Formality	Register	*a*-intercept
Informal	*Sitcom Conversation*	1.281
	Fiction	1.328
	TV Conversation	1.408
Semi-formal	*Press Reportage*	1.455
	Science	1.487
High-formal	*News Broadcasting*	1.561
	Officialese	1.570
	Press Editorials	1.685

As stated in section 3.3, the *a*-axis intercepts of the regression lines are 1.281, 1.408, and 1.561 in the *Sitcom Conversation*, *TV Conversation*, and *News Broadcasting* registers respectively. Combining two studies covering eight registers from different sources, we have the following result based on *a*-intercept, as in Table 8.

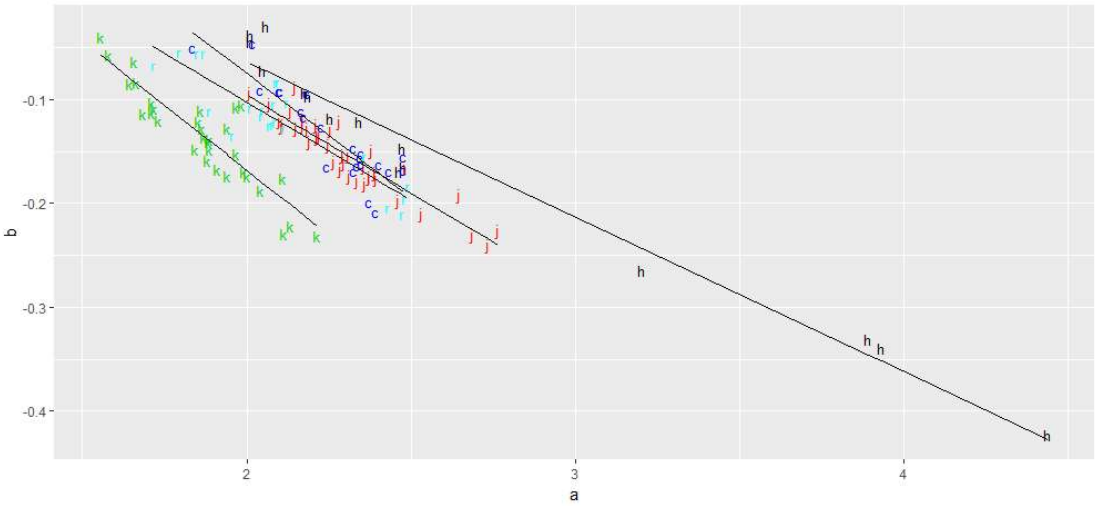

Figure 10- The regression lines for the link between *b* and *a* in each register ("h" represents *Officialese*, "j" represents *Science*, "k" represents *Fiction*, "c"represent *News Comments*, "r" represent *News Reports*)

It is interesting to observe the three clusters formed according to *a*-intercept values can be characterized by differences in degree of formality in terms of *informal*, *semi-formal* and *high-formal*. In addition, the nature of these three clusters can also be attributed to different modes of communication. The three informal registers all involve dialogue or descriptive style and could involve more than one speaker. This analysis supports the theoretical view that fictions are dialogues between the author and the reader (Bakhtin 1981). As the distributional analysis we undertake here does not consider turns and different speakers, what we capture is the planning of each text in response to and expecting responses from the other dialogue partner. This is where fiction writing is similar to the conversation and dialogue. The two semi-formal registers are conveying information with specific target audience: either to persuade (*Science*) or to inform (*Press Reportage*). In other words, although there is no direct dialogue, the speakers are aware of needs to persuade/inform when they plan their speech. The three high-formal registers involve pronouncement. I.e. the speaker is making a statement that is expected to be taken for granted. This is clear for *Officialese*, and *Press Editorials* (as news-paper editorials are considered as formal policy statement by the government in China). The somewhat surprising member of this group is *News Broadcasting*. We consider that there are two important characteristics to differentiate it from *Press Reportage*. On one hand, the person delivering *News Broadcasting* is typically different from the one who wrote it. Hence the nature of the text become strongly pronouncement. In addition, in the context where a text/speech is planned with the audience in mind, it requires time for a listener/reader to think and respond. This is not possible for *News Broadcasting* as the news broadcasting is continuous. Hence it is

strictly a one-way communication with minimal influence of the addressee on the planning. This dialogic interpretation is also consistent with Biber's (1986) study showing that *Fiction* is closer to conversation than to either academic prose or planned speeches. It is also important to note that the degree of formality of register does not correspond to word length or clause/sentence averages reported earlier in this paper.

In LCMC, the number of texts in each register differs. This may affect the linear regression analysis between parameters *a* and *b*. In future studies, this factor should be considered and the number of texts from each register should be as similar as possible.

4 Conclusion

Quantitative linguistics treats languages as self-organizing and self-regulating systems. Synergetic linguistics holds that there are interrelated relationships among the various language levels (Köhler 1984, 2005). As an important law, the MA law explores the relationship between a language construct and its immediate components. This paper examined degrees of formality of register and the distance between two registers based on the MA law from the perspective of quantitative linguistics and regression analysis.

News Broadcasting, *Sitcom Conversation*, and *TV Conversation* texts were selected to form a corpus for this preliminary study. The results show that, as predicted by MA law, average word length decreases as the increase of clause length for most clauses. The logarithm of average word length distributions can be fitted by the Formula (1a-1). The fitting results shown that, for the texts from each register, the relationship between clauses and their constituent words abides by the MA law.

All the texts were represented by their corresponding fitted parameters, *a* and *b*, obtained from Formula (1a-1). There were obvious boundaries between the texts from various registers. The functional correlation between these two parameters, *a* and *b*, was fitted by linear regression in each register. Analysis indicates that the *a*-intercept can be used as an index to represent the formality degree of the register and to quantify the distances between two registers. The *News Broadcasting* register is more formal than both the *TV Conversation* and *Sitcom Conversation* registers. The same experiments were carried out on texts from 6 additional registers from LCMC, and confirmed the validity of using *a*-intercept to represent the formality degrees of registers and to quantify the distance between two registers.

In addition, by combing the results of two studies, we show that the *a*-intercept values of the 8 registers can be group into three clusters corresponding to *informal*, *semi-formal*, and *high-formal* registers. We further show that the three clusters correspond to three different modes of communication: dialogic (and informal), informative/persuasive (with targeted audience and semi-formal), and pronouncement (and high-formal). This is consistent with Hou et al.'s (under review) result showing that the average word length differences in different genres can be explained by cost of planning, where more interactive genres require more planning and hence shorter units.

In sum, we propose *a*-intercept as an effective index to represent the degrees of formality of a register and to quantify the distances between various registers based on the MA law and regression analysis. In addition, we show that the range of the *a*-intercept can be attribute to the

modes of communication typical of each register. Thus our study further developed and formally realized Biber's (1994) claim that registers are varieties in a continuum which may still be analytically identified as different categories.

REFERENCES

Altmann, G. (1980). Prolegomena to Menzerath´s law. *Glottometrika 2, 1-10.*

Bakhtin, M.M. (1981). Discourse in the Novel. In: *The Dialogic Imagination: Four Essays* (Vol. 1). 262-349. Austin: University of Texas Press.

Benešová, M., & Čech, R. (2015). Menzerath-Altmann Law Versus Random Model. In: G.K. Mikros & J. Mačutek (Eds), *Sequences in Language and Text (pp. 57-69).* Berlin/Boston: de Gruyter.

Benešová, M. (2016). *Text segmentation for Menzerath-Altmann law testing*. Palacký University, Faculty of Arts.

Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language 62:384–414.*

Biber, D. (1988). *Variation across Speech and Writing.* Cambridge: Cambridge University Press.

Biber, D. (1994). An analytical framework for register studies. *Sociolinguistic perspectives on register*, 31-56.

Biber, D. (1995). On the role of computational, statistical, and interpretive techniques in multi-dimensional analyses of register variation: A reply to Watson. *Text – Interdisciplinary Journal for the Study of Discourse*, *15*(3), 341-370

Biber, D. & Conrad, S. (2009). *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory, 8(1), 9-37.*

Cacoullos, R. T. (1999). Construction frequency and reductive change: Diachronic and register variation in Spanish clitic climbing. *Language variation and change, 11(2), 143-170.*

Chao, Y. R. (1968). *A Grammar of Spoken Chinese*. Berkeley and Los Angeles: University of California Press.

Chen, H.H. (1994). The contextual analysis of Chinese sentences with punctuation marks. *Literary and linguistic computing, 9(4): 281-289.*

Chen, H & H Liu (2016) How to Measure Word Length in Spoken and Written Chinese, *Journal of Quantitative Linguistics, 23:1, 5-29.*

Chen, Keh-jiann, Chu-Ren Huang, Li-ping Chang, and Hui-Li Hsu. 1996. Sinica Corpus: Design Methodology for Balanced Corpora. In: B.-S. Park and J.B. Kim. (Eds.) *Proceedings of the 11th Pacific Asia Conference on Language, Information and Computation. Seoul:Kyung Hee University. pp. 167-176.*

Chen, Keh-Jiann, Chi-Ching Luo, Ming-Chung Chang, Feng-Yi Chen, Chao-Jan Chen, Chu-Ren Huang, and Zhao-Ming Gao. (2003). Sinica Treebank: Design Criteria, Representational Issues and Implementation. In: Anne Abeillé (Ed.), *Treebanks: Building*

and Using Parsed Corpora (pp. 231-248). Dordrecht; Boston: Kluwer Academic Publishers.

Conway, D., & White, J. (2013). *Machine learning for hackers*. (Chen, Kaijiang, Yizhe Liu & Xiaonan, Meng, Trans). Beijing. China: China Machine Press.

Cramer, I. (2005). The parameters of the Altmann-Menzerath law. *Journal of Quantitative Linguistics, 12, 41–52.*

Eroglu, S. (2014). Menzerath-Altmann law: Statistical mechanical interpretation as applied to a linguistic organization. *Journal of Statistical Physics, 157(2) 392-405.*

Feng, S. (2010). On mechanisms of register system and its grammatical property. *Studies of the Chinese Language, 5, 400–412.*

Grzybek, P. (2007). Do we have problems with Arens' law? A new look at the sentence-word relation. In: P. Grzybek and E. Stadlober (Eds.), *Exact Methods in the Study of Language and Text: Dedicated to Gabriel Altmann on the Occasion of His 75th Birthday, 62, 205.*

Hammerl, R., & Sambor, J. (1993). *O statystycznych prawach jezykowych*. Warszawa: Zakład Semiotyki Logicznej Uniwersytetu Warszawskiego.

Hou, R., Chu-Ren Huang and Yat-mei Lee. (2018). Linguistic Characteristics of Chinese Register Based on the Menzerath – Altmann Law and Text Clustering. (Under review).

Hou, R., Yang, J., & Jiang, M. (2014). A Study on Chinese Quantitative Stylistic Features and Relation Among Different Styles Based on Text Clustering. *Journal of Quantitative Linguistics, 21(3), 246-280.*

Hou, R, Chu-Ren Huang, Hue San Do & Hongchao Liu (2017): A Study on Correlation between Chinese Sentence and Constituting Clauses Based on the Menzerath-Altmann Law, *Journal of Quantitative Linguistics. 24(4): 350-366.*

Hou, R., Huang, C., & Liu, H. (2017). A study on Chinese register characteristics based on regression analysis and text clustering. *Corpus Linguistics and Linguistic Theory*, (Online). doi:10.1515/cllt-2016-006

Huang, B. & Liao, X. (2002). *Modern Chinese*. Beijing: High Education Press.

Huang, Chu-Ren and Shi, D. (2016). *A Reference Grammar of Chinese*. Cambridge: Cambridge University Press.

Huang, C.-R. & K.-J. Chen. (2017). Sinica Treebank. In: N. Ide and J. Pustejovsky (eds), *Handbook of Linguistic* Annotation. Berlin & Heidelberg: Springer.

Hřebíček, L. (1992). *Text in communication: Supra-sentence structure*. Bochum, Brockmeyer.

Hřebíček, L. (1995). *Text levels: Language constructs, constituents and Menzerath-Altmann law.* Trier: WVT.

Hřebíček, L. (1997). *Lectures on text theory.* Prague: Academy of Sciences of the Czech Republic, Oriental Institute.

Kelih, E. (2010). Parameter interpretation of Menzerath's Law: Evidence from Serbian. In P. Grzybek, E. Kelih & J. Mačutek (Eds.): *Text and Language, Structures, Functions, Interrelations, Quantitative Perspectives (pp. 71–78).* Wien: Praesens.

Köhler, R. (1982). Das Menzerathsche Gesetz auf Satzebene. In: W. Lehfeldt & U. Strauss (Eds.), *Glottometrika 4 (pp. 103 – 113*). Bochum: Brockmeyer.

Köhler, R. (1984). Zur Interpretation des Menzerathschen Gesetzes. In: W. Lehfeldt & U. Straus (Eds.), *Glottometrika 6, 177-183*. Bochum: Brockmeyer.

Köhler, R. (1989). Das Menzerathschen Gesetz als Resultat des Sprachverarbeitungs-mechanismus. In: Altmann, Schwibbe (1989): *108-112*.

Köhler, R. (2005). Synergetic Linguistics. In: R. Köhler, G. Altmann & R.G. Piotrowski (eds.). *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook*. Berlin, New York: Walter de Gruyter, 760-775.

Köhler, R. (2012). *Quantitative syntax anal*ysis (Vol. 65). Berlin: Walter de Gruyter.

Kułacka, A. (2010). The coefficients in the formula for the Menzerath-Altmann law. *Journal of Quantitative Linguistics*, *17*(4), 257-268.

Lv, S. (1992). Studies on Chinese grammar through comparison. *Foreign Language Teaching and Research*. (2).

McEnery, A. & R. Xiao. (2004). The Lancaster corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study. In_ M. Lino, M. Xavier, F. Ferreire, R. Costa, R. Silva (eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC) 2004, pp. 1175–1178.* Lisbon, May 24 –30, 2004.

Menzerath, P. (1954). *Die Architektonik des deutschen Wortschatzes* (Vol. 3). F. Dümmler.

Motalová, T., Spáčilová, L., Benešová, B., Kučera, O. (2014). *An application of Menzerath-Altmann law to contemporary written Chinese*. Křížkovského, Olomouc: Univerzita Palackého v Olomouci.

Popescu, I.-I., Mačutek, J., & Altmann, G. (2009). *Aspects of word frequencies*. Lüdenscheid: RAM-Verlag.

Ščigulinská, J. & Schusterová, D. (2014). *An Application of the Menzerath-Altmann Law to Contemporary Spoken Chinese.* Palacký University in Olomouc.

Tuldava, J. (1995). Informational measures of causality. *Journal of Quantitative Linguistics, 2(1), 11-14.*

Wang, K., & Qin, H. (2014). What is peculiar to translational Mandarin Chinese? A corpus-based study of Chinese constructions' load capacity. *Corpus Linguistics and Linguistic Theory, 10(1), 57-77.*

Wilson, A. (2017) Units and Constituency in Prosodic Analysis: A Quantitative Assessment, *Journal of Quantitative Linguistics, 24:2-3, 163-177.*

Yuan, H. and Li, X. (2005*). Outline of Chinese Register*. China, Beijing: The Commercial Press.

Zhang, Z. S. (2012). A corpus study of variation in written Chinese. *Corpus Linguistics and Linguistic Theory, 8(1), 209-240.*

Zhu, D. (1982). *Lectures on Grammar.* Beijing, China: Commercial Press.

Appendix

Appendix 1:
The occurrence frequencies of clauses with certain lengths
(raw numbers)

Clause length	*TV Conversation*	*Sitcom Conversation*	*News Broadcasting*
1	*2068*	*7963*	1743
2	*3687*	*5884*	3446
3	5652	6177	3514
4	7445	6843	4020
5	7843	6704	4507
6	7294	6160	4588
7	6138	4997	4492
8	4851	3707	4260
9	3583	2907	3735
10	2593	2105	3378
11	1800	1443	2854
12	1340	993	2279
13	874	739	1821
14	594	494	1516
15	405	337	1207
16	263	281	865
17	182	183	693
18	102	137	579
19	68	90	432
20	48	65	295
21	34	52	258
22	19	38	192
23	15	37	132
24	16	25	111
25	6	21	83

Appendix 2

he relative frequency distributions of clause length (for Figure 1)

Clause length	*TV Conversation*	*Sitcom Conversation*	*News Broadcasting*
1	0.036328	0.136194	0.033945
2	0.064768	0.100636	0.067111
3	0.099287	0.105648	0.068435
4	0.130784	0.117038	0.078289
5	0.137775	0.114661	0.087774
6	0.128131	0.105357	0.089351
7	0.107824	0.085466	0.087481
8	0.085216	0.063402	0.082963
9	0.062941	0.049720	0.072739
10	0.045550	0.036003	0.065786
11	0.031620	0.024680	0.055582
12	0.023539	0.016984	0.044383
13	0.015353	0.012639	0.035464
14	0.010435	0.008449	0.029524
15	0.007114	0.005764	0.023506
16	0.004620	0.004806	0.016846
17	0.003197	0.003130	0.013496
18	0.001792	0.002343	0.011276
19	0.001195	0.001539	0.008413
20	0.000843	0.001112	0.005745
21	0.000597	0.000889	0.005025
22	0.000334	0.00065	0.003739
23	0.000263	0.000633	0.002571
24	0.000281	0.000428	0.002162
25	0.000105	0.000359	0.001616

Appendix 3

: Average word length distribution in clauses (for Figure 3)

	TV Conversation	*Sitcom Conversation*	*News Broadcasting*	*Whole*
1	1.957447	1.489263	2.530694	1.725667
2	1.635476	1.502039	2.151045	1.711646
3	1.55585	1.39728	1.916809	1.574681
4	1.493519	1.357555	1.885137	1.52869
5	1.476756	1.335561	1.850189	1.515409

6	1.460356	1.324378	1.826431	1.507021
7	1.453102	1.310958	1.810234	1.510307
8	1.452098	1.310898	1.792165	1.524282
9	1.442243	1.311318	1.782032	1.529139
10	1.446317	1.310309	1.773475	1.547709
11	1.44298	1.308574	1.767185	1.56293
12	1.451244	1.305975	1.758556	1.571824
13	1.44288	1.310086	1.759811	1.582366
14	1.441799	1.307981	1.758575	1.600834
15	1.424033	1.309397	1.764154	1.614845
16	1.44249	1.301601	1.759176	1.608809
17	1.446671	1.303439	1.769544	1.633382
18	1.412854	1.281833	1.780944	1.651453
19	1.452012	1.319883	1.790205	1.679483
20	1.439583	1.296923	1.785254	1.666789
21	1.439776	1.320513	1.777224	1.674834
22	1.425837	1.327751	1.812973	1.709383
23	1.457971	1.347826	1.816535	1.693053
24	1.484375	1.266667	1.850601	1.716009
25	1.473333	1.367619	1.883855	1.762909

Appendix 4

Fitted parameters of average word length distribution in clauses (for Figure 5, 6 and 7. “qq”, “wj”, and “xw” refer to *TV Conversation*, *Sitcom Conversation*, and *News Broadcasting*, respectively)

Files	*a*	*B*
qq01.txt	2.067773	−0.18753
qq02.txt	2.174008	−0.20846
qq03.txt	1.947793	−0.14294
qq04.txt	1.807163	−0.10454
qq05.txt	1.751832	−0.10506
qq06.txt	1.764547	−0.09116
qq07.txt	1.779792	−0.10791
qq08.txt	1.858832	−0.13347
qq09.txt	1.753004	−0.10043
qq10.txt	1.892101	−0.14266
qq11.txt	1.893699	−0.14804
qq12.txt	2.05125	−0.16539
qq13.txt	1.931095	−0.14453
qq14.txt	2.217134	−0.24779

qq15.txt	2.10442	-0.20811
qq16.txt	1.990768	-0.19279
qq17.txt	1.995214	-0.1665
qq18.txt	1.727310	-0.08338
qq19.txt	1.759958	-0.09278
qq20.txt	2.132648	-0.20043
qq21.txt	1.802140	-0.10913
qq22.txt	1.831594	-0.12511
qq23.txt	1.615169	-0.05312
qq24.txt	1.788015	-0.10808
qq25.txt	1.831414	-0.11988
qq26.txt	1.872961	-0.13428
qq27.txt	1.929761	-0.15053
qq28.txt	1.803591	-0.11334
qq29.txt	1.91029	-0.14825
qq30.txt	1.698243	-0.09146
qq31.txt	1.986195	-0.17809
qq32.txt	1.764805	-0.10245
qq33.txt	2.314057	-0.25443
qq34.txt	2.011485	-0.1705
qq35.txt	1.774028	-0.10153
qq36.txt	2.253452	-0.22888
qq37.txt	1.800376	-0.11194
qq38.txt	1.965715	-0.16464
qq39.txt	1.867041	-0.12519
qq40.txt	1.716586	-0.08507
qq41.txt	1.834335	-0.13484
qq42.txt	1.750414	-0.10254
qq43.txt	1.777919	-0.11176
qq44.txt	1.667633	-0.07651
qq45.txt	1.711596	-0.08762
qq46.txt	1.701851	-0.08325
qq47.txt	1.76669	-0.10654
qq48.txt	1.563749	-0.03122
qq49.txt	1.893359	-0.1482
qq50.txt	1.892036	-0.16296
wj01.txt	1.701541	-0.10547
wj02.txt	1.579630	-0.0974
wj03.txt	1.567663	-0.11157
wj04.txt	1.487902	-0.07347
wj05.txt	1.466492	-0.03893

wj06.txt	1.373107	-0.01848
wj07.txt	1.574886	-0.06957
wj08.txt	1.464686	-0.05789
wj09.txt	1.459430	-0.04956
wj10.txt	1.526858	-0.0552
wj11.txt	1.657220	-0.09628
wj12.txt	1.584129	-0.09103
wj13.txt	1.685830	-0.11615
wj14.txt	1.477337	-0.03961
wj15.txt	1.584944	-0.08975
wj16.txt	1.587453	-0.08484
wj17.txt	1.479296	-0.04599
wj18.txt	1.489070	-0.06742
wj19.txt	1.581871	-0.10483
wj20.txt	1.810669	-0.17034
wj21.txt	1.594398	-0.10822
wj22.txt	1.434462	-0.04705
wj23.txt	1.562341	-0.08129
wj24.txt	1.55812	-0.09029
wj25.txt	1.577619	-0.08739
wj26.txt	1.527094	-0.06899
wj27.txt	1.519326	-0.07362
wj28.txt	1.510108	-0.08433
wj29.txt	1.597706	-0.10607
wj30.txt	1.398341	-0.01865
wj31.txt	1.486941	-0.0775
wj32.txt	1.64755	-0.09942
wj33.txt	1.54406	-0.07909
wj34.txt	1.507677	-0.0689
wj35.txt	1.585655	-0.10438
wj36.txt	1.550824	-0.08769
wj37.txt	1.479014	-0.07302
wj38.txt	1.480225	-0.04912
wj39.txt	1.443864	-0.03998
wj40.txt	1.534121	-0.07684
wj41.txt	1.462054	-0.05437
wj42.txt	1.523679	-0.06365
wj43.txt	1.510244	-0.08121
wj44.txt	1.400162	-0.05061
wj45.txt	1.478317	-0.06013
wj46.txt	1.406906	-0.0327

wj47.txt	1.495283	-0.07339
wj48.txt	1.47248	-0.0704
wj49.txt	1.432348	-0.05111
wj50.txt	1.551323	-0.09809
wj51.txt	1.559035	-0.08117
wj52.txt	1.547542	-0.07581
wj53.txt	1.469425	-0.05357
wj54.txt	1.44971	-0.04541
wj55.txt	1.643353	-0.11486
wj56.txt	1.421602	-0.04071
wj57.txt	1.411729	-0.04461
wj58.txt	1.475764	-0.06114
wj59.txt	1.466146	-0.07185
wj60.txt	1.472642	-0.05403
xw01.txt	2.262991	-0.12554
xw02.txt	2.198158	-0.10987
xw03.txt	2.24177	-0.12304
xw04.txt	2.282072	-0.11802
xw05.txt	2.387058	-0.13759
xw06.txt	2.324207	-0.13598
xw07.txt	2.269689	-0.10807
xw08.txt	2.285678	-0.10362
xw09.txt	2.425591	-0.11979
xw10.txt	2.475266	-0.14716
xw11.txt	2.539164	-0.15114
xw12.txt	2.513899	-0.11853
xw13.txt	2.355283	-0.11542
xw14.txt	2.379863	-0.13813
xw15.txt	2.302483	-0.10163
xw16.txt	2.196296	-0.11534
xw17.txt	2.259619	-0.10839
xw18.txt	2.29023	-0.10474
xw19.txt	2.312217	-0.10316
xw20.txt	2.093065	-0.0775
xw21.txt	2.328397	-0.12352
xw22.txt	2.212437	-0.09836
xw23.txt	2.32851	-0.11559
xw24.txt	2.38001	-0.13449
xw25.txt	2.285232	-0.09528
xw26.txt	2.331219	-0.10743
xw27.txt	2.500296	-0.15373

xw28.txt	2.374066	-0.12564
xw29.txt	2.210489	-0.08788
xw30.txt	2.229068	-0.09742
xw31.txt	2.39812	-0.13752
xw32.txt	2.241518	-0.09986
xw33.txt	2.375414	-0.11892
xw34.txt	2.228828	-0.09917
xw35.txt	2.233510	-0.09978
xw36.txt	2.186077	-0.09676
xw37.txt	2.202082	-0.10072
xw38.txt	2.235197	-0.11707
xw39.txt	2.170009	-0.09935
xw40.txt	2.386215	-0.11978
xw41.txt	2.163245	-0.08660
xw42.txt	2.448241	-0.13281
xw43.txt	2.462103	-0.14008
xw44.txt	2.387655	-0.11001
xw45.txt	2.349125	-0.11095
xw46.txt	2.278891	-0.10759
xw47.txt	2.069112	-0.06881
xw48.txt	2.234222	-0.09467
xw49.txt	2.69523	-0.18178
xw50.txt	2.339337	-0.12004

Appendix 5

Average word length distribution in clauses (LCMC, for Figure 9, the average word length distributions in clauses whose range is 1:12 words were fitted.)

	Officialese	*Science*	*Fiction*	*Press Editorials*	*Press Reportage*
1	3.062147	2.517738	2.041815	2.520772	2.387789
2	2.291935	2.022654	1.65941	1.99269	1.834146
3	1.973881	1.851996	1.545702	1.837147	1.713834
4	1.991392	1.76871	1.46331	1.759375	1.678852
5	1.934568	1.717284	1.442179	1.74123	1.661885
6	1.878258	1.707954	1.424236	1.699459	1.61875
7	1.853913	1.703171	1.414539	1.697101	1.628486
8	1.841814	1.687843	1.395501	1.697993	1.614583
9	1.838235	1.671431	1.399111	1.678824	1.611985
10	1.810336	1.677444	1.408616	1.71008	1.627921
11	1.796671	1.666633	1.399324	1.699655	1.608276
12	1.821721	1.663522	1.408932	1.696912	1.624351

13	1.820926	1.650267	1.42096	1.705882	1.605604
14	1.838724	1.673993	1.40803	1.722084	1.602814
15	1.816798	1.679961	1.463043	1.680417	1.617687
16	1.794444	1.674213	1.407095	1.698138	1.623326
17	1.821238	1.646278	1.410256	1.700073	1.620098
18	1.838574	1.668022	1.412698	1.673127	1.60463
19	1.810729	1.660254	1.440000	1.730884	1.723977
20	1.815476	1.640761	1.504545	1.714706	1.677381
21	1.772109	1.660588	1.47619	1.690476	1.70000
22	1.758117	1.682497	1.563636	1.73445	1.693182
23	1.849275	1.678261	1.434783	1.68530	1.601449
24	1.783333	1.69086	1.583333	1.777778	1.666667

Appendix 6

The fitted parameters of average word length distribution in clauses (LCMC, “h” represents *Officialese*, “j” represents *Science*, ”k” represents *Fiction*, “c”represent *Press Editorials*, “r” represent *Press Reportage*)

	a	*B*
h01.txt	3.894010	−0.33009
h02.txt	3.931898	−0.33874
h03.txt	2.057789	−0.02896
h04.txt	2.011199	−0.04385
h05.txt	2.04932	−0.07152
h06.txt	2.478661	−0.16447
h07.txt	2.462683	−0.1686
h08.txt	2.256338	−0.1177
h09.txt	2.343932	−0.12074
h10.txt	2.011827	−0.03794
h11.txt	2.177471	−0.09293
h12.txt	2.185661	−0.09616
h13.txt	2.473934	−0.14672
h14.txt	3.203523	−0.26464
h15.txt	4.438227	−0.42259
j01.txt	2.256356	−0.12903
j02.txt	2.211942	−0.12504
j03.txt	2.21564	−0.13305
j04.txt	2.108875	−0.12448
j05.txt	2.191512	−0.14079
j06.txt	2.533191	−0.21083

j07.txt	2.392433	-0.17568
j08.txt	2.31026	-0.15318
j09.txt	2.373629	-0.17504
j10.txt	2.356452	-0.16436
j11.txt	2.151169	-0.12766
j12.txt	2.460924	-0.19736
j13.txt	2.764089	-0.22633
j14.txt	2.482294	-0.16497
j15.txt	2.390709	-0.1706
j16.txt	2.378474	-0.14962
j17.txt	2.2828	-0.12115
j18.txt	2.372927	-0.17545
j19.txt	2.360044	-0.18185
j20.txt	2.264953	-0.16002
j21.txt	2.099943	-0.12058
j22.txt	2.00819	-0.09356
j23.txt	2.169798	-0.11982
j24.txt	2.133982	-0.11114
j25.txt	2.183392	-0.12744
j26.txt	2.070529	-0.10486
j27.txt	2.146686	-0.08837
j28.txt	2.647627	-0.19237
j29.txt	2.335038	-0.17791
j30.txt	2.310879	-0.17349
j31.txt	2.294596	-0.16076
j32.txt	2.172026	-0.12477
j33.txt	2.733758	-0.2399
j34.txt	2.687748	-0.23046
j35.txt	2.285372	-0.16717
j36.txt	2.107397	-0.12615
j37.txt	2.219698	-0.13427
j38.txt	2.29143	-0.15376
j39.txt	2.214308	-0.13661
j40.txt	2.247157	-0.14454
k01.txt	1.657834	-0.06306
k02.txt	1.86132	-0.11084
k03.txt	1.851445	-0.12097
k04.txt	1.685644	-0.11368
k05.txt	1.731522	-0.11959
k06.txt	1.862992	-0.1279
k07.txt	1.880562	-0.15794

k08.txt	1.870938	-0.13575
k09.txt	1.88173	-0.14079
k10.txt	1.644948	-0.08455
k11.txt	1.557019	-0.0398
k12.txt	1.712411	-0.10302
k13.txt	1.885497	-0.14816
k14.txt	1.967994	-0.15197
k15.txt	1.94156	-0.12746
k16.txt	1.968872	-0.10686
k17.txt	1.984294	-0.10546
k18.txt	2.11039	-0.17604
k19.txt	2.131986	-0.22162
k20.txt	2.113132	-0.22887
k21.txt	1.664552	-0.08383
k22.txt	1.579961	-0.05674
k23.txt	1.908861	-0.16672
k24.txt	2.214675	-0.23127
k25.txt	1.939691	-0.17327
k26.txt	1.844373	-0.14724
k27.txt	1.991633	-0.16936
k28.txt	1.886768	-0.13761
k29.txt	1.718277	-0.10842
k30.txt	1.710238	-0.1129
k31.txt	2.043509	-0.18767
k32.txt	2.002826	-0.17315
nc01.txt	2.18501	-0.09482
nc02.txt	2.104087	-0.09234
nc03.txt	2.099482	-0.09271
nc04.txt	2.172759	-0.1178
nc05.txt	2.350765	-0.16227
nc06.txt	2.326585	-0.16922
nc07.txt	2.244471	-0.16514
nc08.txt	2.372331	-0.20013
nc09.txt	2.39383	-0.20885
nc10.txt	2.335363	-0.16448
nc11.txt	2.350964	-0.15258
nc12.txt	2.402742	-0.16292
nc13.txt	2.346136	-0.15682
nc14.txt	2.16682	-0.11142
nc15.txt	2.017009	-0.04722
nc16.txt	2.478327	-0.15619

nc17.txt	2.477328	-0.16893
nc18.txt	2.43489	-0.16924
nc19.txt	2.22708	-0.1273
nc20.txt	2.326956	-0.14678
nc21.txt	2.042853	-0.09147
nc22.txt	1.835245	-0.05065
nr01.txt	1.955519	-0.13577
nr02.txt	2.428588	-0.20546
nr03.txt	2.488269	-0.18487
nr04.txt	2.082228	-0.12232
nr05.txt	2.009029	-0.10886
nr06.txt	1.791718	-0.05526
nr07.txt	1.715746	-0.06794
nr08.txt	1.884986	-0.11167
nr09.txt	2.078765	-0.10581
nr10.txt	2.091409	-0.08449
nr11.txt	2.084471	-0.08436
nr12.txt	2.118468	-0.10367
nr13.txt	2.077391	-0.12496
nr14.txt	2.042974	-0.11603
nr15.txt	2.045316	-0.11082
nr16.txt	2.065749	-0.12572
nr17.txt	2.110759	-0.12921
nr18.txt	1.844938	-0.05597
nr19.txt	1.86599	-0.05585
nr20.txt	2.35474	-0.15559
nr21.txt	2.478857	-0.19649
nr22.txt	2.470704	-0.21101