Hou, R., Huang, C. R., Ahrens, K., & Lee, Y. S. (2019). Linguistic characteristics of Chinese register based on the Menzerath—Altmann law and text clustering. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqz005
Pre-published version provided to meet funding guidelines. Refer to published version for final version.

Linguistic Characteristics of Chinese Register Based on the Menzerath – Altmann Law and Text Clustering

Hou, Renkui[1, 2], Chu-Ren Huang[2], Kathleen Ahrens[3], Yat-Mei Sophia Lee[2]

[1]School of Humanities, Guangzhou University, Guangzhou, China;

[2]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Kowloon, Hong Kong;

[3]Department of English, The Hong Kong Polytechnic University, Kowloon, Hong Kong

Hou, R., Huang, C. R., Ahrens, K., & Lee, Y. S. (2019). Linguistic characteristics of Chinese register based on the Menzerath—Altmann law and text clustering. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqz005
Pre-published version provided to meet funding guidelines. Refer to published version for final version.

Linguistic Characteristics of Chinese Register Based on the Menzerath – Altmann Law and Text Clustering

Abstract: This article explores the linguistic features of different registers in Chinese through text clustering driven by the Menzerath–Altmann (MA) law. We propose to calculate the average word length distribution according to clause length. The MA law predicts that texts from different registers will show differences in terms of average word length distribution in texts. As predicted by the MA law, analysis result 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 $y = ax^b e^{-cx}$. We hypothesize that it is the situation type, i.e. whether the text is dialectic or monologue, that is the linguistic characteristic behind the dichotomy of word length distribution. To confirm these register-distinguishing linguistic features, texts were represented by the average word length distribution and the fitted parameters using the vector space model and clustered according to their register categories. Good clustering results show that average word length distribution in certain length clauses and their fitted parameters can be used as the distinctive characteristics of these three registers.

Keywords: Register, Linguistic characteristics, Average word length distribution, Text clustering

1 Introduction

Register is often considered one of the most important contributing factors of text varieties (Biber & Conrad 2009). Generally speaking, a register is associated with a particular situation of use. The register perspective attributes the linguistic characteristics of a text variety with analysis to the situation of use of the variety. In this paper, we propose to explore the linguistic characteristics of selected registers based on the Menzerath–Altmann law and text clustering. The Menzerath–Altmann law (MA law hereafter) predicts the distributional patterns of self-organizing complex systems and is one of the foundational laws of synergetic linguistics (Khöler 1993). In this view, the MA law is used to predict the distributional correlation between a linguistic unit and its constituents.

Biber (2012) argued strongly that register differences should be considered in the description of all linguistics levels, i.e., lexical, grammatical, and lexico-grammatical. The significance of comparing different registers in studies of Chinese grammar was introduced by Lv (1992). Zhang (2012) has shown that there are a wide range of linguistic variations among written Chinese registers. They argue that linguistic studies may come up with over-simplified or even wrong conclusions if registers are not considered. By taking registers into consideration, linguistic studies can provide descriptions of significant and meaningful usage variations. Based on this, we propose that we can use the methodology for register classification to discover the linguistic characteristics of selected registers.

The categories of register in a language is extremely limited due to the generality of the topology of context, the systematic differentiation of language materials and the technicality of the special ways of expressing objects. There is no universally agreed upon taxonomy of register in Chinese language. Following Biber and Conrad (2009), who regard the register differences as a continuum of variation, Feng (2010) also considers the Chinese register as a continuum with binary opposition. His continuum ranges from conversational informal to written formal registers, in which formality

is the primary element. Yuan and Li (2005) took a discrete approach and proposed seven registers: conversational, officialese, scientific, news, literary and art, lectures, and advertisements. This current study does not compare different taxonomies of registers in the Chinese language but instead focuses on the quantitative register characteristics.

Quantitative linguistics is one of the main streams of mathematical linguistics. The first attempt to bring quantitative methods into linguistics started to rise at the end of nineteenth and the beginning of the twentieth century. Mathematics started to be used within other scientific disciplines during that time; likewise it found its way into the linguistics as well. The English mathematician De Morgan firstly applied quantitative analysis into linguistic research in 1851. There are theoretical and applicational significance to this kind of research. For example, Cramer (2005) proposed that the investigation of the statistical aspects of language would advance natural language processing research as well as basic linguistic research. Mathematical linguistic studies on Chinese, on the other hand, came much later. Huang et al. (1998) showed that mathematical properties in natural language have both theoretical and applicational implications. Liu and Huang (2012) and Feng (2012) introduced the theory and methodology of quantitative linguistics including the frequency distribution law and the MA law, among others.

Register can also be studied using quantitative approaches. 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. Hou and Jiang (2016), with a text mining approach, validated that parts of speeches could differentiate registers in Chinese. Linguistic devices can lead to both inherent and idiosyncratic behaviors, and these behaviors can be considered as the quantitative characteristics of the corresponding registers. 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.

The MA law is one of the best known quantitative linguistic laws and 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).

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

$$\frac{y'}{y} = -c + \frac{b}{x}$$

The solution to this differential equation is the function:

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

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 (quoted by Köhler 2012).

Hou, R., Huang, C. R., Ahrens, K., & Lee, Y. S. (2019). Linguistic characteristics of Chinese register based on the Menzerath—Altmann law and text clustering. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqz005
Pre-published version provided to meet funding guidelines. Refer to published version for final version.

The MA law is one of the best-known laws of quantitative linguistics. By predicting the correlations between successive hierarchical levels of language, it demonstrates that language is a self-organizing and self-regulating system. Previous research has validated the MA law at different linguistic levels. 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. 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. Hou et al. (2017) showed that the relationship between sentences and their constituent clauses abide by the MA law in written formal register texts, but not in *TV Sitcom* and *TV talkshow*. 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. Jin and Liu (2017) discussed the interdependence between text size and its linguistic constituents at text level. Xu and He (2018) explored the relationship between sentence length and clause length in English based on the MA law taking the factor of register into consideration. This paper will examine the relationship between clause and its immediate component based on the MA law and demonstrate that these parameters are affected by the registers in Chinese.

1.1 Research question and methodology

Hou, Yang and Jiang (2014) showed that the word length distribution can be used as the linguistic characteristics of Chinese registers using text clustering. Synergetic linguistics sees language as an open, dynamic, self-organizing, and self-adaptive system with multiple levels, each of which can be defined as a sub-system and interacts. It is clear that word length is not isolated linguistic phenomenon given one accepts the distinction of linguistic levels, as (1) phoneme, (2) syllable/morpheme, (3) word, (4) clause, and (5) sentence (Levels may be a little different in Chinese). The units of all five levels are characterized by length, again mutually influencing each other, resulting in specific frequency length distributions. The Menzerath-Altmann law, as a general linguistic law in synergetic linguistics, describes the interrelation between language entities in mathematical terms. Under the context of synergetic linguistics, the average word length in clauses could reflect the relatively internal stable features of language entities in the external dynamic, ever-changing language system. Thus we take a different approach based on the earlier result and calculate the average word lengths for clauses with different lengths in this paper. Our assumption is the average word length reflects the distributional differences. Based on this assumption, we further hypothesize that the average word length distribution according to clause lengths can differentiate various registers and be used as the linguistic characteristics of Chinese registers.

Given the average word length distribution information, we fitted the data using Formula (1) and explored the relationship between clause length and word length based on the MA law. The texts can be represented by the fitted parameters using vector space models and clustered. We can validate whether these three parameters can differentiate various registers and can be used as linguistic characteristics of Chinese registers through clustering results.

Effective register analysis is always comparative. It is virtually impossible to know what is distinctive about a particular register without comparing it to other registers. So we selected more than one register text to establish the corpus.

Different from Indo-European languages, it is difficult to define the sentence and the clause in

Chinese language. Chinese sentences can be delineated clearly with speech cues, but not with texts cues only (Huang & Shi, 2016; Lu, 1993). However, the 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. 2003; Huang and Chen 2017 for Sinica TreeBank) is to mark all segments between punctuation marking pauses in utterances as sentences. Such punctuation marks include commas, semicolons, colon, periods, exclamation marks and question marks. Although such segmentation method yields very short sentences, we can in fact borrow this approach for a reliable operational definition of Chinese clauses. That is, a clause is the minimal unit between two punctuation marks. Some studies of quantitative linguistics also define the clauses using such methods, for example, Hou, Huang and Liu (2017), Hou, et al. (2017) and Chen (1994).

Words are the immediate constituents of the clauses (and not sentences), hence the clause length can be defined as the number of words. Note that the definition of "word" has generated some controversies in linguistics and is certainly the center of on-going debate in Chinese corpus linguistics (Huang and Xue 2012). We take the segments delineated by blank spaces in the texts, segmented by the Chinese lexical analysis system, as operationally defined words. Yet, even with words delineated, there are still different approaches to calculate word length. For example, from the perspective of speech, actual duration of articulation could be used to measure word length; and in alphabetic writing systems, number of letters is often taken as the default measure. Taking into consideration of the Chinese orthography, we defined word length as the number of Chinese characters (*Hanzi*, *汉字*) included (Hou et al. 2014; Chen and Liu 2016). This is not only provides an easy and non-ambiguous measure, it is also linguistically felicitous as each character stands for one syllable and typically (though with rare exceptions) also one morpheme (Huang and Shi 2016). The average word lengths in clauses with certain length were calculated as the value of the total number of Chinese characters divided by the number of words in these clauses. The texts are represented by the corresponding average word length and hierarchical clustered. We can estimate whether average word length distribution in certain length clauses can be used as linguistic characteristics of the selected registers.

Formula (1) was selected to fit the average word length distribution and explored the relationship between the clause and the constituting word based on the MA law.

The determination coefficient (R^2) was used to validate the fitted results; 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. The determination coefficient can be calculated using Equation (2), in which *obs* and *pred* refer to the observed and predicted values of average word lengths respectively, and *mean*(*obs*) refers to the mean observed value:

$$R^2 = \frac{\sum(obs-pred)^2}{\sum\left(obs-mean(obs)\right)^2} \qquad (2)$$

The function between average word length and clause length were fitted by Formula (1) in every text. Then the texts from various registers were represented by the fitted parameters, *a*, *b* and *c,*

using a vector space model.

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.

2 Corpus establishment and preprocessing

The texts from "*News Co-Broadcasting*", the situation comedy "*I Love My Family*", and "*Behind the Headlines with Wentao*" were selected to represent the *New Broadcasting*, *Sitcom conversation* and *TV Talk Show* registers respectively. These texts and registers were selected for several reasons. Most crucially, we needed to ensure that the texts we studied came from the same sources and were created during the same period of time, yet differ significantly in the contexts of use. After surveying other possibilities, including available balanced corpora, none met the required criteria. These three types of texts that we chose, however, were all from the same source (i.e. TV subtitles) and from roughly same period of time, hence allowing us to complete comparative studies of the registers without compounding factors of other variations.

The selection of these registers is an integral and crucial part of the design of our study. These three registers can be grouped into three different two versus one contrasting pairs. That is, we identified three potential criteria for defining situation types and each of them is shared by two registers but not the other one. In this way, the effect of these register defining situation types can be confirmed by distributional similarity of the two sharing registers as well as distinctive distributional patterns of the third. First, for modes of delivery, both *TV talk show* and *Sitcom conservation* are dialectic, with *News Broadcas*t is not. That is, both *TV talk show* and *Sitcom conservation* involve multiple participants in dialogue, while *News Broadcast* involves only one-way delivery. Secondly, for topicality, both *News Broadcas*t and *TV talk show* are topical and focus on current issues, while *Sitcom conservation* is not. Third, in terms of preparedness content, *Sitcom conservation* and *News Broadcas*t are scripted (written to be spoken and written to be read respectively); while *TV talk show* is free-flow and non-scripted. This design of studying registers with critical situational contrasts facilitates possible linguistic accounts for any contrastive distributional features among the three registers.

The Central China TV (CCTV) program, "*News Co-Broadcasting*", mainly consists of brief introductions of important state policies and both domestic and international events. It is characterized by formal use of serious language. We treat this as the representative of the *News Broadcasting* register. The texts of *News Co-broadcasting* were obtained from the National Broadcast Language Resources Monitoring and Research Centre.

"Behind the Headlines with Wentao" is a program of Phoenix Satellite TV in which the host discusses current hot issues with invited guests. They chat freely, rather than reading scripts that are prepared ahead of time, aiming to generate audience interest while teasing apart truth from falsehood yet without presumed "right" answers. The language use is representative of the *TV Talk show* register. Textual materials of *Behind the Headlines with Wentao* were collected from the website of Phoenix Satellite TV.

The situational comedy, "*I Love My Family*", tells the story of a family via dialogues, and is representative of the *Sitcom Conversation* register. The texts of *My Love My Family* were downloaded from the Internet. Sitcom conversations are scripted but delivered in an informal style and on informal conversational issues rather than news events.

Hou, R., Huang, C. R., Ahrens, K., & Lee, Y. S. (2019). Linguistic characteristics of Chinese register based on the Menzerath—Altmann law and text clustering. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqz005
Pre-published version provided to meet funding guidelines. Refer to published version for final version.

Note that these three registers can be characterized and differentiated by three different features: *New Broadcasting* is + Scripted, and +Topical, -Dialogue; *TV conversation* is -Scripted, and +Topical, +Dialogue; and *Sitcom Conversation* is + Scripted, and -Topical, +Dialogue. As these three registers differ minimally by one feature of either stylistic or topical nature, we can also explore the correlation between textual distance and textual features.

The Chinese lexical analysis system created by Institute of Computing Technology of Chinese Academy of Science (ICTCLAS) was used for word segmentation and part-of-speech tagging. ICTCLAS has been acknowledged with 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 approximate 98% for the recognition of Chinese names[1].

The size of the sub-corpora in terms of text, word types, and word tokens are shown in Table 1.

Table 1: Size of sub-corpora from the different register

	Text Number	Type	Token
News Co-Broadcasting	50	24, 812	41, 8943
Behind the Headlines with Wentao	50	16, 372	35, 7663
I Love My Family	60	14, 107	31, 7661

3 Experiments

3.1 Frequency distribution of clause length in terms of words

Firstly, we calculated the relative frequencies of clauses with certain lengths in terms of words and established the frequency distributions of clause length in each register, as shown in Figure 1.

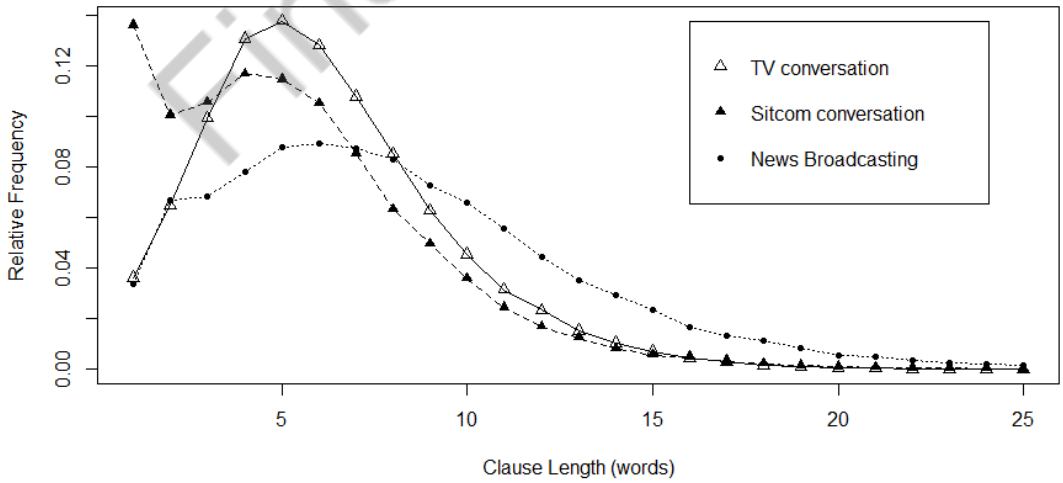


Figure 1: Frequency distributions of clause length in terms of words

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

Hou, R., Huang, C. R., Ahrens, K., & Lee, Y. S. (2019). Linguistic characteristics of Chinese register based on the Menzerath—Altmann law and text clustering. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqz005
Pre-published version provided to meet funding guidelines. Refer to published version for final version.

Figure 1 demonstrated that there are similar changing tendencies of relative frequencies of clauses with the increase of length from various registers. In *Sitcom conversation* text, the one-word clauses are more than that with other lengths. Based on the observation of *Sitcom conversation* text, more frequent interactions in daily conversation and the characteristics of comedy lead to this phenomenon. The frequency of clauses in texts from the other two registers, *News Broadcasting* and *TV talk show,* first increase and then decrease with clause length in terms of words. From Figure 1, we also see that most of clauses with high occurrence frequencies are short.

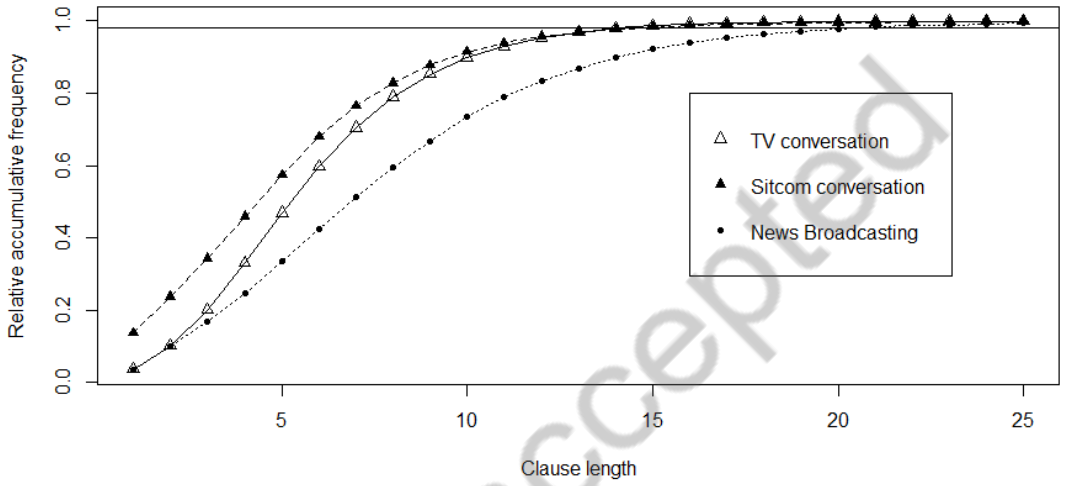


Figure 2: Relative accumulative frequency of clause length in terms of words

The cumulative relative frequency distributions of clauses in the different register texts were shown in the Figure 2. From Figure 2 we can observe that most of clauses are composed of a few words in the texts from different registers. More than 98% of clauses in *TV talk show* and *Sitcom conversation* are composed of 1-15 words. About 99% of clauses in *News broadcasting* composed of less than 20 words. Figure 1 showed that the short clauses appear more frequently and longer clauses appear less frequently. Figure 2 also shows that most clauses are short.

3.2 Average word length distribution

In this paper, the average word length according to clauses length was calculated as the number of Chinese characters of the given clauses divided by the number of words included in the corresponding clauses.

The correspondence between average word length and clause length was established, as shown in Figure 3. The clause length is the independent variable and the average word length is the dependent variable. From Figure 3, we can observe that there are differences in the average word length distribution in different registers. Interestingly, of the three features that differentiate these three registers, it seems that one has the most significant influence on the word length. The unique feature of *News Broadcasting* being read and non-dialectic corresponds to its longest average word length. This could be explained in terms of the limit on working memory of human brain (Köhler

1989). Interactive dialogue requires both planning and real time reaction to other interlocutors, hence the working memory cannot be fully devoted to production of the utterance. A non-dialogue register such as *News Broadcasting* does not have such additional processing requirements and the speaker can devote all memory to the delivery of longer linguistic units.

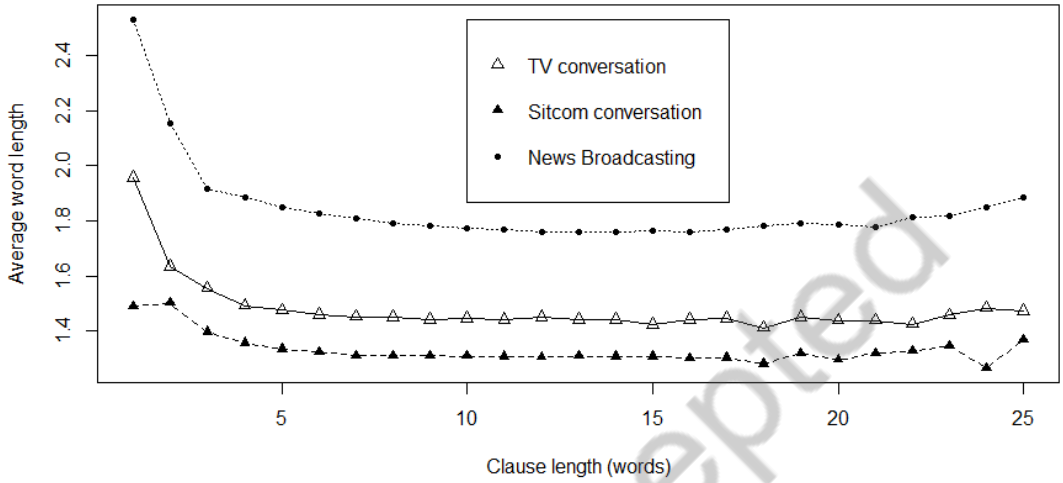


Figure 3: Average word length distributions in the clauses

3.3 Regression analysis

Figure 3 shows that there is a reverse relationship between average word length and clause length in each register when clauses are short. The average word length in *News broadcasting* and *TV talk show* texts decreases with the increase of length in most clauses. In *Sitcom conversation*, since one-word clauses are dominated by one-character words, the average word length in one-word clauses is smaller than that in the clauses with 2 words. Most of these words are interjection words and often as conversational turn-fillers and not content sentences. In the clauses with more than 1 word, the average word length decreases with the increase of clause length. These short conversational utterances should in general be treated as exception to the MA law.

This reverse relationship between constructs, clauses, and their immediate constituents words follows the prediction of the MA law. Formula (1) was selected to fit this relationship between clause and its immediate constitute. The fitted results are shown in Table 2 and Figure 4.

Table 2: Fitted results of link between average word lengths and clause length

	a	*b*	*c*	R^2	*p*-value
TV talk show	1.844	−0.168	-0.013	91.75%	4.632×10^{-13}
Sitcom Conversation	1.513	−0.094	-0.007	81.19%	4.007×10^{-9}
News Broadcasting	2.415	−0.214	-0.017	96.42%	2.2×10^{-16}

In Table 2, the values of R^2 show that the relationship between average word length and clause length can be fitted by Formula (1) for each of the three registers: *News Broadcasting*, *TV talk show*,

Hou, R., Huang, C. R., Ahrens, K., & Lee, Y. S. (2019). Linguistic characteristics of Chinese register based on the Menzerath—Altmann law and text clustering. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqz005
Pre-published version provided to meet funding guidelines. Refer to published version for final version.

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), *X*1 (the logarithm of clause length distribution) and *X*2 (clause length distribution). For each register, the value of parameter *b* is negative, which indicates that average word length decreases with clause length firstly. Thus, as can be seen from Table 2 and Figure 4, the relationship between clauses and their constituent words abide by the MA law in each register.

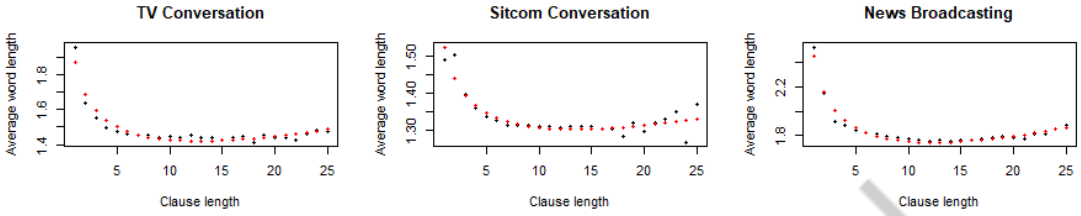
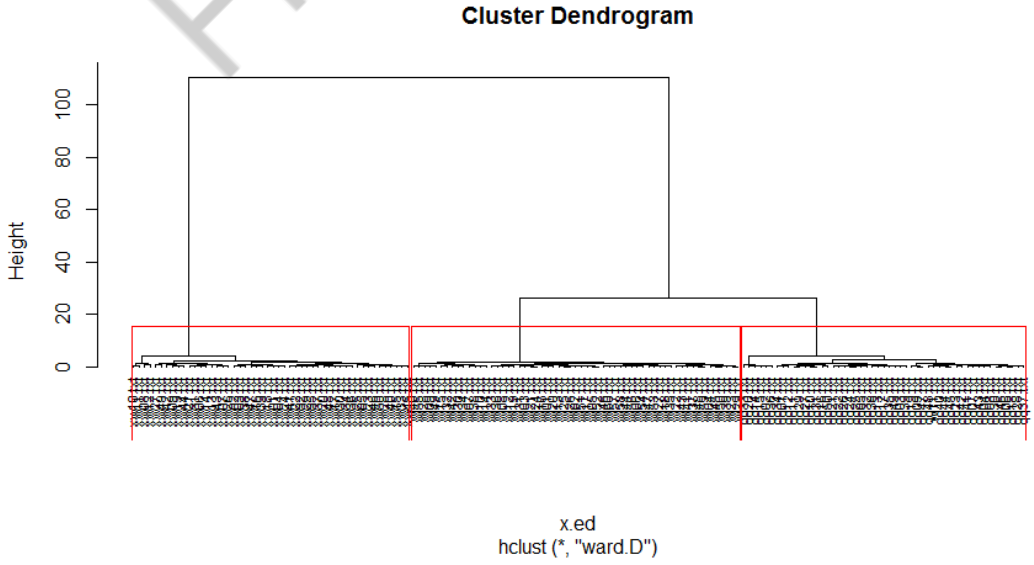


Figure 4: Fitted results of relationship between average word length and clause length (black dots represent the observed values of average word length; red dots represent the fitted values of average word length)

3.4 Text clustering

From Figure 3, there are differences between the average word length distributions in clauses with certain length in the different registers. Clustering analysis was used to determine whether this difference is accidental or inherent.

Agglomerative hierarchical clustering does not require an a priori decision of the number of clusters, hence it was selected in this research. In this algorithm, each document is first put into its own cluster. The two nearest clusters are then combined recursively. The result of this algorithm is a tree of clusters, called a dendrogram, which shows how the clusters are related. By cutting the dendrogram at a desired level, a clustering of the data items into disjoint groups is obtained (Halkidi el at., 2001).



Hou, R., Huang, C. R., Ahrens, K., & Lee, Y. S. (2019). Linguistic characteristics of Chinese register based on the Menzerath—Altmann law and text clustering. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqz005
Pre-published version provided to meet funding guidelines. Refer to published version for final version.

Figure 5: Clustering dendrogram of texts represented by average word length distribution

The texts are represented by average word length distribution using vector space model and clustered using the agglomerative hierarchical clustering algorithm. In this process, the Euclidian distance was used to represent the distances between different texts. The sum of squared deviations ("Ward") was used to calculate the dissimilarity between clusters. The number of clusters need not be determined in advance before the texts were clustered. The clustering results are shown in Figure 5 and Table 3.

Table 3: Clustering results of texts represented by average word length in the clauses (1:25)

	Cluster 1	Cluster 2	Cluster 3
TV talkshow	50	0	0
Sitcom conversation	1	59	0
News broadcasting	0	0	50
Entropy	0.139	0	0

In the hierarchical dendrogram, as shown in Figure 5, the leaf and non-leaf nodes represent the texts and the text clusters respectively, and the distances between any pair of texts can be measured by the height of their common ancestor. The clusters which are shown in Figure 5 are *News broadcasting*, *Sitcom conversation* and *TV talk show* from left to right. In addition, the distances between texts from different registers are also shown in Figure 5. But we can only see *News broadcasting* are far from other two registers and cannot see which one of the two registers, *Sitcom conversation* or *TV talk show,* are farther away from *News broadcasting*. In the meantime, we also can see *Sitcom conversation* and *TV talk show* are close.

Cophenetic Correlation Coefficient (CPCC) was adopted here to internally validate the result of the hierarchical clustering (Halkidi et al., 2001). The CPCC measures how well a dendrogram represents the pairwise distances among the points of a data set. The external criterion—supervised validation—evaluates the results of the clustering algorithm based on a pre-specified structure, which is imposed on a data set and reflects our intuition about the clustering structure of the data set.

The CPCC and the entropy were calculated to validate the clustering result. The entropy of each cluster is shown in Table 3 — the weighted entropy of all the clusters is 0.044, and the CPCC is 0.90. These validation values show that the clustering result is fairly good. The good clustering result shows that the average word length distribution can be used to differentiate the different register texts as shown in Figure 5.

Table 4: Clustering results of texts represented by the fitted parameters of average word length distribution in the clauses (1:25)

	Cluster 1	Cluster 2	Cluster 3
TV talkshow	44	4	2
Sitcom conversation	5	0	55
News broadcasting	0	50	0
Entropy	0.475	0.381	0.219

The average word length distribution in each text is fitted using the Formula (1). Then the texts were represented by the fitted parameters – *a*, *b* and *c* using vector space models. The agglomerative hierarchical clustering algorithm was used to cluster these texts from these three registers. In this clustering process, the Euclidian distance was used to represent the distances between different texts. The sum of squared deviations ("Ward") was used to calculate the dissimilarity between clusters. The clustering result is shown in Table 4 and Figure 6.

The entropy of each cluster is shown in Table 4. The weighted entropy of all the clusters is 0.352, and the CPCC is 0.822. From this, we can see that the clustering result using these three parameters to represent the texts is not as good as using the average word length distribution to represent the texts. Figure 6 also shows that the distances between different clusters are less than the above clustering result. These parameters are the deep linguistic characteristics and are not controlled consciously by the speakers.

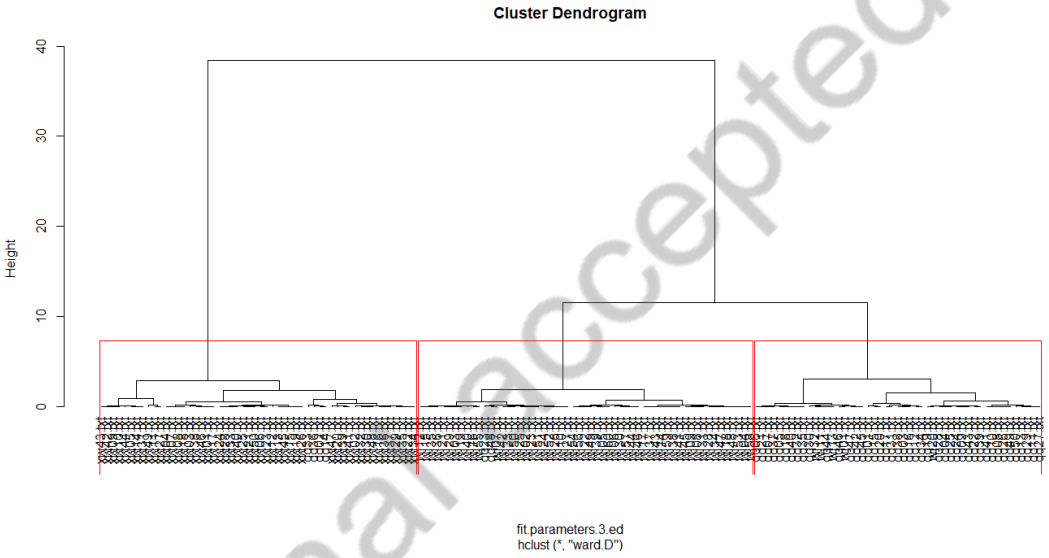


Figure 6: Clustering dendrogram of texts represented by the fitted parameters of the average word length distribution

Then we established the classification model using these three parameters to represent the texts. The support vector machine (SVM) was used to establish the classification model. Sample validation was used to validate the corresponding classification model. In this process, the ratio between training data and testing data is 3:1. The classification result is shown in Table 5. This classification result also demonstrates that these three fitted parameters can be used as the characteristics of the selected registers.

Table 5: Classification result of texts represented by the fitted parameters of average word length distribution in the clauses (1:25)

	Cluster 1	Cluster 2	Cluster 3
TV talk show	11	0	0
Sitcom conversation	0	16	0
News broadcasting	0	0	13

Hou, R., Huang, C. R., Ahrens, K., & Lee, Y. S. (2019). Linguistic characteristics of Chinese register based on the Menzerath—Altmann law and text clustering. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqz005
Pre-published version provided to meet funding guidelines. Refer to published version for final version.

3 Conclusion

This paper presents an innovative approach to register analysis and classification. Instead of using the attested and descriptive word length distribution as linguistic characteristics of the registers based on text clustering (Hou, Yang and Jiang 2014), we propose to directly utilize the essence of MA law and the self-organizing nature of the linguistic system. That is, we propose to use instead the distribution of average word lengths according to different clause lengths. In order to ensure that the observed differences are inherent to the registers and not just accidental, we perform three different clustering analyses to confirm the distributional features. The good clustering results demonstrate that these differences are not accidental and are indeed inherent linguistic characteristics of the selected registers. The results also confirm that the average word length distribution in clauses with certain length can be used as linguistic characteristics in selected registers.

In addition, in order to clarify the contribution of different situation types to the register differences, we design our study with three registers that can be differentiated by three different situation types. Each situation type feature is shared by two registers and different from the third. Our closer analysis showed that the average word length differences correspond most closely with whether the text is dialectic in nature or not (i.e. whether it involves dialogue). We hypothesize that dialogues require both planning and listening, hence requiring some working memory to be set aside. Non-dialectic register, however, can devote working memory to the preparation and production of the utterances hence having distinctively longer average word lengths across different clause lengths. Interestingly, neither the topicality nor whether the text is scripted has strong impact on average length. Although the lack of correlation with topicality and processing can be expected, the lack of correlation with being scripted or not is somewhat surprising as scripted texts can be committed to memory and should take less planning. A possible explanation is that *Sitcom conversation* (vs. the non-scripted *TV talk show*) is scripted to be like natural dialogue and hence is artificially short; while the fact that *News Broadcast* is scripted in fact is an additional feature to facilitate longer units. This conjecture should be tested in future studies.

It is important to note that there is a negative relationship between the average word length and the clause length when clauses are short. This relationship can be fitted by the Formula (1). The fitted result showed that this relationship abides by the MA law. The texts are represented by these three fitted parameters of the average word length distribution. The clustering and classification results show that these fitted parameters can differentiate these three registers.

Some linguistic characteristics of registers are overt, and can be manipulated by speakers, and some are covert. The two characteristics proposed in this paper, average word length in clauses with certain length and the fitted parameters of them, are covert and difficult to consciously manipulate by the speakers. We believe that these are reliable, objective and scientific characteristics of registers that can be easily obtained for all registers.

Acknowledgements: We would like to thank the anonymous DSH reviewers for their insightful and helpful comments.

Funding: Research on this paper was funded by National Social Science Fund in China (Grant

Award Number: 16BYY110)

Reference

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

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. England Cambridge: Cambridge University Press.

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.

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.

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

Feng Z. (2012). 用计量方法研究语言. Foreign Language Teaching and Research, 44(2), 256-269.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. Journal of Intelligent Information Systems, 17, 107–145.

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., & Jiang, M. (2014). Analysis on Chinese quantitative stylistic features based on text

mining. *Digital Scholarship in the Humanities*, 31(2): 357-367.

Hou, R., Chu-Ren Huang, Hue San Do & H. Liu. (2017). A Study on Correlation between Chinese Sentence and Constituting Clauses Based on the Menzerath-Altmann Law, Journal of Quantitative Linguistics, DOI: 10.1080/09296174.2017.1314411

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, C. R., Chen, K. J., & Gao, Z. M. (1998). Noun class extraction from a corpus-based collocation dictionary: An integration of computational and qualitative approaches. Quantitative and Computational Studies of Chinese Linguistics, 339-352.

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

Huang, Chu-Ren & Shi, D. 2016. A Reference Grammar of Chinese. Cambridge: Cambridge University Press.

Huang, Chu-Ren, and N. Xue. 2012. Words without boundaries: computational approaches to Chinese word segmentation. *Language and Linguistics Compass*. *6*(8). 494-505.

Jin, H and H. Liu. (2017). How will text size influence the length of its linguistic constituents? Poznań Studies in Contemporary Linguistics. 53(2):197:225.

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

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

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

Köhler, R. (1993). Synergetic linguistics. In *Contributions to quantitative linguistics*, pp. 41-51. Springer, Dordrecht.

Liu, H. & Huang, W. (2012). Quantitative Linguistics: State of the Art, Theories and Methods. Journal of Zhejiang University (Humanities and Social Sciences). 42(2). 178-192.

Lu J. (1993). The features of Chinese sentences. Chinese Language Learning. No.1, 1-6.

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

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

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

R Core Team. (2016). R: A language and environment for statistical computing. Vienna. Austria: R Foundation for Statistical Computing. https://www.R-project.org

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

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

Xu, L., & He, L. (2018). Is the Menzerath-Altmann Law Specific to Certain Languages in Certain Registers?. Journal of Quantitative Linguistics. DOI:10.1080/09296174.2018.1532158.

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

Hou, R., Huang, C. R., Ahrens, K., & Lee, Y. S. (2019). Linguistic characteristics of Chinese register based on the Menzerath—Altmann law and text clustering. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqz005
Pre-published version provided to meet funding guidelines. Refer to published version for final version.

Linguistic Theory, 8(1), 209-240.