A quantitative probe into the hierarchical structure of written Chinese

Heng Chen
Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou, China
chenheng@gdufs.edu.cn

Haitao Liu
Department of Linguistics,
Zhejiang University, Hangzhou, China
lhtzju@yeah.net

Abstract

Language unit is a fundamental conception in modern linguistics, but the boundaries are not clear between language levels both in the past and present. As language is a multi-level system, quantification rather than microscopic grammatical analysis should be used to investigate into this question. In this paper, Menzerath-Altmann law is used to make out the basic language units in written Chinese. The results show that "stroke > component > word > clause > sentence" is the hierarchical structure of written Chinese.

1 Introduction

Language levels and language units are critical conceptions in a language system, and they are highly related with the entities in a language, as well as the methods in linguistics. The conception of language unit is definitely put forward by Saussure in the first half of the 20th century. In his seminal book representing the birth of modern linguistics, Saussure puts forward the conceptions of language entities or language units and analyzes the methods as well as difficulties of dividing the spoken chain into language units. Moreover, Saussure distinguished the concept of language units from speech units. Language unit becomes the fundamental problem in modern linguistics. The conception of language level is introduced by American descriptive linguistics. Gleason (1956) distinguishes three types of language levels: language levels of structure, analysis and speech. Later, a number of other linguistic theories (Halliday, 1985; Hudson, 2010; Miyagawa et al., 2013; Nordström, 2014) treat language as a multi-level system.

Generally, five language units are commonly recognized by grammarians: morpheme, word, phrase, clause and sentence (Lyons, 1968). However, different linguistic schools have different opinions upon the systematicness of language, therefore, the methods and standards they use to divide language levels and units are different. Mackey (1967) lists seven sets of language levels from different linguistic schools, and the maximum is Bulundaer's 14 levels, and the minimum is Harris's 2 levels. These language units include sound, word, phrase, sentence, phone, phoneme, morph, morpheme, syllable, affix, word-group, and so on. The boundaries between language levels are not clear in the past for the lack of a common standard, however, it is the same after the introduction of the conception of language level.

The most characteristic feature of modern linguistics is structuralism. Briefly, this means that language is not a haphazard conglomerate of words and sounds but a tightly knit and coherent whole. However, linguistics is traditionally preoccupied with the fine detail of language structure (Hudson, 2010:104), or in other words, the language phenomena at the microscopic scale rather than at the system level (Liu and Cong, 2014). Therefore, it is not ordinarily feasible to analyze each language level separately, and the work must be carried on simultaneously on all levels. Moreover, the results should be stated in terms of an orderly hierarchy of levels (Lyons, 1968).

Quantification is necessary in the inquiry into the structure of the language system (Altmann, 1987, 1996). Without quantification, it would be extremely difficult to investigate language as a multi-level system empirically. Unfortunately, systems thinking in linguistics is generally unaffected by quantitative methods. Liu & Cong (2014) characterize modern Chinese as a multi-level system from the complex

network. However, their emphasis is on the levels of grammatical analysis, for example, syntax and semantics, but not language levels. In this paper, we try to analyze the language levels of written Chinese as a multi-level system using Menzerath-Altmann law.

Menzerath-Altmann law is a general statement about the natural language constructions which says: the longer is a construction, the shorter are its constituents. Language is a whole complex system, and it is a set of relations. The language units correlate with each other in different levels and in complex ways through the relations. The whole is composed by its parts, and they restrains mutually. Language units of the same levels are relatively homogeneous. Therefore, the relation between two adjacent language levels is "whole-part".

Actually, in quantitative linguistics, the relationship between "whole-part" has been extensively investigated (Menzerath, 1954; Krott, 1996; Uhlírová, 1997; Mikros and Milička, 2014; Milička, 2014). The relation was investigated and tested on many linguistic levels and in many languages and even on some non-linguistic data (Baixeries et al., 2013). Köhler (1984) conducted the first empirical test of the Menzerath-Altmann law on "sentence > clause > word", analyzing German and English short stories and philosophical texts. The tests on the data confirmed the validity of the law with high significance. Heups (1983) evaluates 10,668 sentences from 13 texts separated with respect to text genre and her results also confirm the Menzeraht-Altmann law with high significance. The law has also been used to study phenomena on the supra-sentencial level (Hřebíček, 1990, 1992) and fractal structures of text (Hřebíček, 1994; Andres, 2010). This is why this law is considered one of the most frequently corroborated laws in linguistics. The law is a good example of the importance of the quantitative linguistic methodology since it clearly shows that the "independent language subsystems" are in fact interconnected by relationships which are hard to detect by a qualitative research.

In this paper, we will test the construction units in written Chinese, which includes stroke, component, character, word, clause and sentence. We do not include phrase in our language units list because it is hard to divide a sentence into one or several independent phrases in written Chinese. Despite of this, it can be inferred from the Menzerathian results of "sentence-clause-word". That is to say, if "sentence-clause-word" fits well with Menzerath's law, then the unit phrase in written Chinese can be left out, or we should reconsider phrase as an indispensable language unit in written Chinese.

The remainder of this paper is organized as follows. Section 2 introduces the materials and methods of the present study. Section 3 presents the results of the tests for different hierarchical language units. Section 4 concludes the study and makes suggestions for further research.

2 Materials and Methods

We use the Lancaster Chinese corpus (LCMC) as the testing material. The corpus is segmented and part of speech (POS) tagged, and its basic information is in table 1.

Language units	scale
Character (tokens)	1,314,058
Character (types)	4,705
Clauses (types)	126,455
Sentence (types)	45,969
Word (types)	847,521

Table 1. Basic information of LCMC

The language units we will test in this paper are stroke, component, character, word, clause and sentence. The reason why we do not include phrase here is that a complete sentence or clause cannot be divided into several sequential phrases, both theoretically and practically.

All the language units are easy to get in LCMC by using some tools except clause. Therefore, in the following, we will first define the other language units, and then give our methods of defining phrase.

The stroke is a segment written with one uninterrupted movement. The component is the constructing units of characters which have more than one strokes. The character are logograms used in the writing of Chinese, which is called hanzi in Chinese. For example, the word "语言"("yǔ yán", which means "language") consists of two characters "语，言" ("yǔ, yán", which means "language, parole"), and the two characters have nine strokes "丶, 讠, 一, 丨, 冂, 一, 丨, 冂, 一" and seven strokes "丶, 一, 一, 一,

丨, 𠂉, 一" respectively, eleven in total. "语" "言" have five components "讠" "五" "口" and one component "言" (means "parole"), respectively. To measure the number of strokes and components of a word, we used a list consisting of 20,902 characters (CJK Unified Ideographs) with numbers of strokes and components of each character.

In written Chinese, sentences are separated from one another by using special marks of punctuation (full-stop, question-mark, exclamation-mark). As for our case, the sentences are tagged in LCMC, so here there is no difficulties distinguishing sentence.

Clause is not tagged in LCMC, nor in any other corpus available. Xing (1997:13) states that clause is the smallest independent grammatical unit of expression. But this definition can hardly be used to obtain the clauses in LCMC. Lu (2006) analyzes a long sentence from a literary book and claims that the constituents just between two punctuations (comma and period) can be defined as clauses roughly. We believe that although this method is not so exact in grammatical analyses, it can in large-scale-corpus studies. But we need to state that, since in LCMC sentences are tagged, we choose comma and semicolon as our marks of clause boundaries.

After obtaining all the statics with respect to language units in LCMC, we use the Menzerath-Altmann law to fit the hierarchical data.

Menzerath-Altmann law (short for Menzerathian function) describes the mathematical relation between two adjacent language units, and its model function is

$$y = ax^b e^{-cx} \tag{1}$$

.

In this function, y represents the length of the upper language unit, and x represent the mean length of the lower language unit; a, b, c are parameters which seem to depend mainly on the level of the language units under investigation: much more than on language, the kind of text, or author as previously expected, and e is natural constant, which equals 2.71828 approximately. The goodness of fit can be seen from determination coefficient R^2. We say the result is accepted for $R^2 > 0.75$, good for $R^2 > 0.80$, and very good for $R^2 > 0.90$.

3 Results

The language units we will examine in this paper are stroke > component > character > word > clause > sentence (here we use ">" to direct to a higher-rank unit in written Chinese). Since the Menzerath-Altmann law is only used to fit the data of two adjacent language units, we corroborate that the fitting results of these five groups, namely "component> character > word", "stroke > character > word", "stroke > component > word", "component > word > clause", "word > clause > sentence", can answer the question of the hierarchical structure in written Chinese. We will give the result of each group in the following. We begin with the word level since it is regarded as the most basic language unit in all languages.

3.1 Component > Character > Word

The Menzerathian data of "component > character > word" can be seen in Table 2. In this group, word length is measured in character, and character length is measured in component. Mean character length can be calculated with this function:

$$M_i = \frac{F_i}{F_i' * i} \tag{2}$$

In this function, i refers to word length class, i.e. the first column in Table 2; M_i represents mean character length of word length class i (if a word's length is 1, then it belongs to word length class 1, and the like), i.e. the second column in Table 2; F_i represents the sum length of all the characters (measured in component) in the words (based on tokens) of word length class i; F_i'represents the number of words (based on tokens) of word length class i.

Word length (in character)	Mean character length (in component)	Word length (in character)	Mean character length (in component)
1	2.4592	6	2.2054
2	2.5899	7	2.1860
3	2.5435	8	2.1354
4	2.5372	9	2.4222
5	2.1536	10	2.7000

Table 2. Hierarchical data of "component > character > word"

In table 2, we can see that there are ten word length classes and their corresponding mean character lengths. We fit the Menzerathian function introduced in section 2 to the two groups of variables. The goodness of fit indicator $R^2 = 0.1625$ means that the fitting result is unaccepted, which indicate that the hierarchical group "component > character > word" does not line with Menzerath-Altmann law. Therefore, next, we need to test two other possible groups, "stroke > character > word" and "stroke > component > word" to find out the hierarchy in word level.

3.2 Stroke > Character > Word

We replace component in "component > character > word" with stroke, and the Menzerathian data can be seen in Table 3. In this group, word length is measured in character, but character length is measured in stroke instead.

Word length (in character)	Mean character length (in stroke)	Word length (in character)	Mean character length (in stroke)
1	6.9359	6	6.1622
2	7.4136	7	6.2326
3	7.2189	8	6.2708
4	7.1969	9	6.5778
5	6.2356	10	6.4000

Table 3. Hierarchical data of "stroke > character > word"

We fit the Menzerathian function to the two groups of variables in Table 3, and the goodness of fit indicator $R^2 = 0.5009$ means that the fitting result is also unaccepted. This indicates that the group "stroke > character > word" does not line with Menzerath-Altmann law. Then the only possible group in the word level is "stroke > component > word".

3.3 Stroke > Component > Word

The Menzerathian data of "stroke > component > word" is displayed in Table 4. In this group, word length is measured in component, and component length is measured in stroke.

Word length (in component)	Mean component length (in stroke)	Word length (in component)	Mean component length (in stroke)
1	3.45959	13	1.72858
2	2.80834	14	1.62894
3	2.44086	15	1.71641
4	2.21272	16	1.62715
5	2.00806	17	1.55203
6	1.86860	18	1.66435
7	1.81350	19	1.90789
8	1.80166	20	1.350
9	1.80735	21	1.71428
10	1.78970	22	1.98484
11	1.80674	23	1.34782
12	1.74935	25	1.960

Table 4. Hierarchical data of "stroke > component > word"

Then we fit the Menzerathian function to the data in Table 4, and we have a good result this time. The fitting is displayed in Figure 1, and the fitting results, i.e. parameters (with 95% confidence bounds) and determination coefficient are shown in the bottom of Table 4. The value of the goodness of fit indicator R^2 is 0.8982, which means that the result is good, and the group "stroke > component > word" lines with Menzerath-Altmann law.

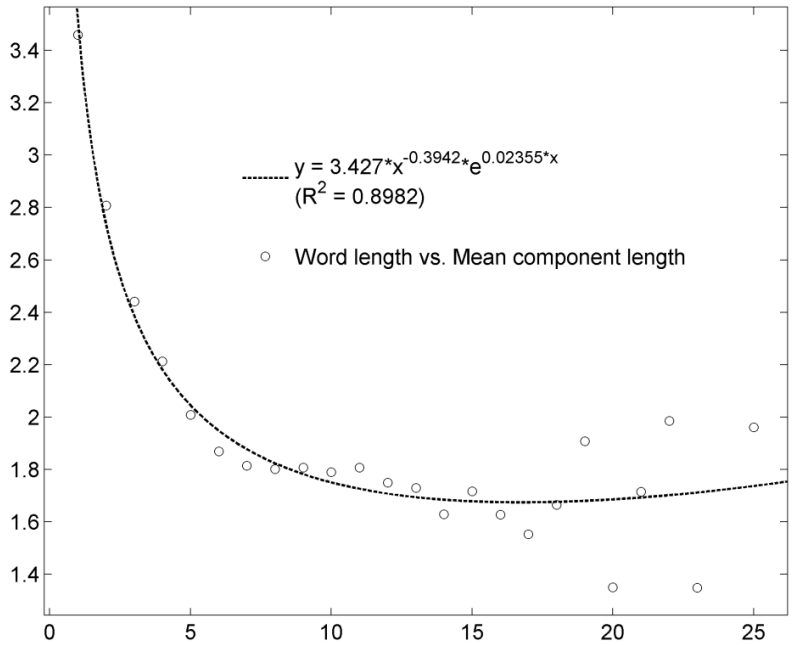

Figure 1. Fitting Menzerath-Altmann law to the hierarchical data of "stroke > component > word"

In sum, in word-level, component is its immediate lower basic language unit. Since above word level, there are two other language units, we first need to examine the group "component > word > clause" to determine if we need go into "component > word > sentence".

3.4 Component > Word > Clause

Table 5 shows the Menzerathian data of "component > word > clause". In this group, clause length is measured in word, and word length is measured in component.

Clause length (in word)	Mean word length (in component)	Clause length (in word)	Mean word length (in component)	Clause length (in word)	Mean word length (in component)
1	5.5445	12	3.9150	23	4.0552
2	4.5248	13	3.9402	24	4.1348
3	4.1405	14	3.9494	25	4.1948
4	3.9387	15	3.9944	26	4.1137
5	3.8897	16	3.9733	27	4.2187
6	3.8444	17	4.0052	28	3.8613
7	3.8383	18	4.0247	29	4.1614
8	3.8458	19	4.0453	30	4.2573
9	3.8657	20	4.0729	31	4.1608
10	3.8738	21	4.0674	32	4.0275
11	3.8966	22	4.1309	33	4.3384

Table 5. Hierarchical data of " component > word > clause"

We then fit the Menzerathian function to the two groups of variables in Table 5. The fitting is displayed in Figure 2, and the results is shown in the bottom of Table 5. As can be seen from the value of R^2 (0.7657) in Table 5, the fitting result is accepted.

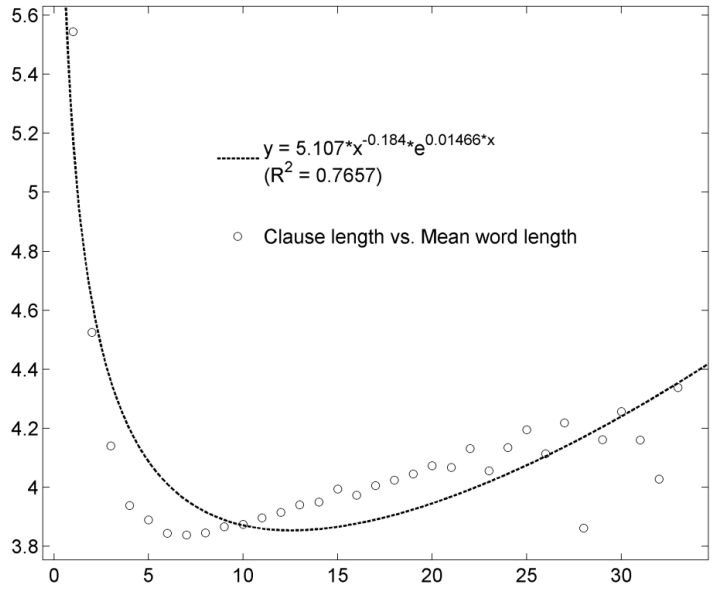

Figure 2. Fitting Menzerath-Altmann law to the hierarchical data of " component > word > clause"

Although the fitting result ($R^2 = 0.7657$) in this group is not as good as in "stroke > component > word" ($R^2 = 0.8982$), the group "component > word > clause" lines with Menzerath-Altmann law. This means that clause is the immediate higher language unit of word, thus we need not go into the group "component > word > sentence". Ultimately, we only have "word > clause > sentence" to be tested.

3.5 Word > Clause > Sentence

Table 6 shows the Menzerathian data of this group. As can be seen in Table 6, the sentence length is measured in clause, and the clause length is measured in word.

Sentence length (in clause)	Mean clause length (in word)	Sentence length (in clause)	Mean clause length (in word)
1	7.7407	9	6.2194
2	7.0465	10	6.3932
3	6.7162	11	5.8068
4	6.4866	12	5.7661
5	6.3357	13	6.1723
6	6.2485	14	6.5510
7	6.1646	15	6.4500
8	6.2296		

Table 6. Hierarchical data of "word > clause > sentence"

The Menzerathian function is again used, and the fitting is displayed in Figure 3. As can be seen from the fitting results in Table 6, the goodness of fit indicator R^2 (0.8498) indicates the result is good. This means that the group "word > clause > sentence" lines with Menzerath-Altmann law.

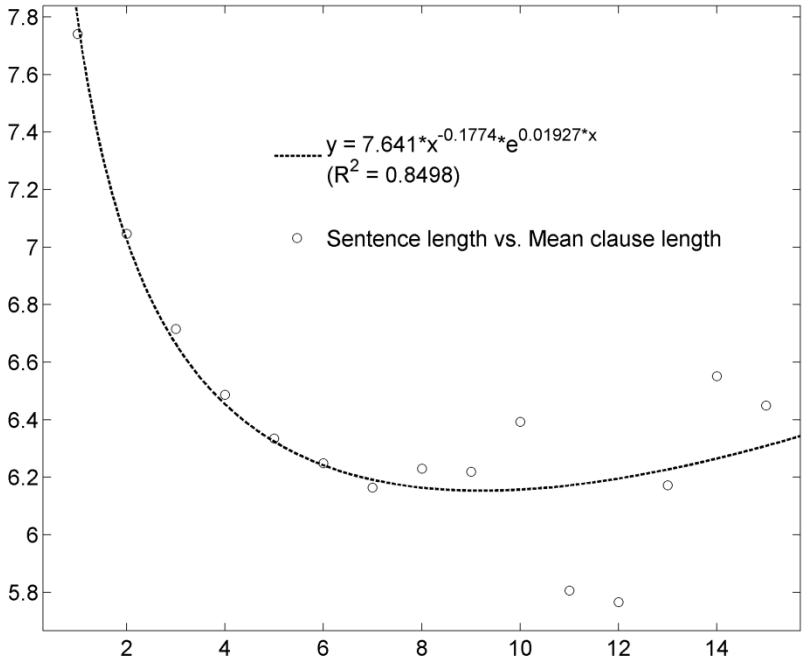

Figure 3. Fitting Menzerath-Altmann law to the hierarchical data of "word > clause > sentence"

4 Discussions and Conclusions

In section 3 we tested five Menzerathian groups, namely "component > character > word", "stroke > character > word", "stroke > component > word", "component > word > clause", "word > clause > sentence". The results shows that only "stroke > component > word", "component > word > clause" and "word > clause > sentence" line with Menzrath-Altmann law. However, the fitting results of "component > word > clause" and "word > clause > sentence" are not as good as that of "stroke > component > word". We think that there are two possible reasons. One reason is the data sparseness problem: clause length distribution is sparser than word length distribution because the length range of word is more fixed than that of clause. The other reason may be the rough way of segmenting clauses by means of punctuations: the clauses obtained in this way may be a little bigger or smaller than the practical situation. Generally, the results indicate that "stroke > component > word > clause > sentence" is a Menzrathian hierarchy in written Chinese.

Character is an easy-to-distinguish language unit in written Chinese; clause is commonly regarded as one level of language unit by grammarians. However, they are not included in the Menzerathian hierarchy, i.e. they are not basic language units. For character, the reason may be that although there are thousands of single-character words, they are not enough for communication. The combinations of characters into multi-character words makes ends meet. In classic Chinese, Character may be a basic language unit, however, it is replaced by word in modern Chinese, because the classic Chinese habitually uses monosyllable words while the modern Chinese prefers to choose multi-syllable words to express the same meaning. As for phrase, first, it is difficult to segment a sentence into several phrase sequences; secondly, from a quantitative perspective, the main reason may be that clause can directly be composed of words, but not via one level of phrase.

That language is a system has been put forward for about 100 years, however, it has never been realized until quantification is introduced into linguistics. In this paper, we shows that Menzerath-Altmann law can be an efficient way of finding the basic language units in a language. In the future, we will investigate into this question from a diachronic perspective to see if the basic language units have changed with time.

Acknowledgements

This work was supported by the National Social Science Fund of China (Grant No. 18CYY031) and the MOE Project of the Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies.

Reference

Andrea Krott. 1996. Some Remarks on the Relation between Word Length and Morpheme Length. Journal of Quantitative Linguistics, 3 (1): 29-37.

G. Heups (1983). Untersuchungen zum Verhaltnis von Satzlange zu Clauselange am Beispiel deutscher Texte verschiedener Textklassen. In R. Kohler & J. Boy (Eds.), Glottometrika 5 (pp. 113 – 133). Bochum: Brockmeyer.

Gabriel Altmann. 1987. The Levels of Linguistic Investigation. Theoretical Linguistics, 14(2-3): 227-240.

Gabriel Altmann. 1996. The Nature of Linguistic Units. Journal of Quantitative Linguistics, 3(1): 1-7.

Georgios Mikros and Jiří Milička. 2014. Distribution of the Menzerath's law on the syllable level in Greek texts. In Altmann, G., Čech, R., Mačutek, J., & Uhlířová, L. (Eds.), Empirical approaches to text and language analysis, pp. 180-189, RAM-Verlag.

Haitao Liu and Jin Cong. 2014. Empirical Characterization of Modern Chinese As A Multi-Level System From The Complex Network Approach. Journal of Chinese Linguistics, 42(1): 1-38.

Jackie Nordström. 2014. Language as a Discrete Combinatorial System, rather than a Recursive-Embedding One. The Linguistic Review, 31(1): 151-191.

Jan Andres. 2010. On a Conjecture about the Fractal Structure of Language. Journal of Quantitative Linguistics. 17: 101–122.

Jaume Baixeries , Antoni Hernández-Fernández , Núria Forns & Ramon Ferrer-i-Cancho. 2013. The parameters of the Menzerath-Altmann Law in genomes. Journal of Quantitative Linguistics, 20: 94-104.

Jiří Milička. 2014. Menzerath's Law: The Whole is Greater than the Sum of its Parts. Journal of Quantitative Linguistics, 21(2): 85-99.

John Lyons. 1968. Introduction to Theoretical Linguistics. London: Cambridge university press.

Jr Henry A. Gleason. 1955. An Introduction to descriptive linguistics. New York : Holt, Rinehart and Winston.

Ludk Hřebíček. 1990. The constants of Menzerath-Altmann's Law. In R. Hammerl (ed.), Glottometrika 12. Bochum: Brockmeyer.

Ludk Hřebíček. 1992. Text In Communication: Supra-Sentence Structures. Universitätsverlag Dr. N. Brockmeyer.

Ludk Hřebíček. 1994. Fractals in Language. Journal of Quantitative Linguistics.1: 82-86.

Ludmila Uhlírová. 1997. Length vs. Order: Word Length and Clause Length from the Perspective of Word Order. Journal of quantitative linguistics, 4(1-3): 266-275.

Michael A. K. Halliday. 1985. An Introduction to Functional Grammar, Edward Arnold Ltd.

Paul Menzerath. 1954. Die Architektonik des deutschen Wortschatzes. Bonn: Dümmler.

Reinhard Köhler. 1984. Zur Interpretation des Menzerathschen Gesetzes. In R. Köhler, J. Boy (Eds.), Glottometrika 6, pp. 177-183. Bochum: Brockmeyer.

Richard, Hudson. 2010. An introduction to word grammar. Cambridge University Press.

Shigeru Miyagawa, Robert C. Berwick, and Kazuo Okanoya. 2013. The emergence of hierarchical structure in human language. Frontiers in Psychology, 4:71.

William F. Mackey. 1967. Language Teaching Analysis. Longman.