Open Access Library Journal
2018, Volume 5, e4747
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Testing the Menzerath-Altmann Law in the Sentence Level of Written Chinese

Heng Chen

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

How to cite this paper: Chen, H. (2018) Testing the Menzerath-Altmann Law in the Sentence Level of Written Chinese. *Open Access Library Journal*, **5**: e4747.
https://doi.org/10.4236/oalib.1104747

Received: July 1, 2018
Accepted: August 5, 2018
Published: August 8, 2018

Abstract

Language unit is a fundamental conception in modern linguistics, but the boundaries are not clear between language levels. 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 in the sentence level.

Subject Areas

Linguistics

Keywords

Menzerath-Altmann Law, Sentence, Language Level, 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 used [1] [2]. Generally, five language units are recognized by grammarians: morpheme, word, phrase, clause and sentence [3]. 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. These language units include sound, word, phrase, sentence, phone, phoneme, morph, morpheme, syllable, affix, word-group, etc.

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 well knit and coherent whole. However, linguistics is traditionally preoccupied with the fine detail of language structure, or in other words, the language

phenomena at the microscopic scale rather than at the system level [4]. 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 [3].

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. The whole is composed of its parts, which interact with each other. Language units in the same levels are relatively homogeneous. Therefore, the relation between two adjacent language levels is a "whole-part" relationship.

Actually, in quantitative linguistics, the relationship between "whole-part" has been extensively investigated [5]. This relation was investigated and tested on many linguistic levels and in many languages and even on some non-linguistic data [6]. [7] 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. The law has also been used to study phenomena on the supra-sentencial level and fractal structures of text [8] [9]. 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 in the sentence level, *i.e.*, "sentence-clause-word". The rest 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 "sentence-clause-word" levels. 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**.

The language units we will test in this paper are *word*, *clause* and *sentence*. The reason why we do not include *phrase* here is that, a complete sentence or

Table 1. Basic information of LCMC.

Language units	Scale
Character (tokens)	1,314,058
Character (types)	4705
Clauses (types)	126,455
Sentence (types)	45,969
Word (types)	847,521

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 firstly define the other language units, and then give our methods of defining *phrase*.

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 difficulty distinguishing *sentence*.

Clause is not tagged in LCMC, nor in any other corpus available. Generally speaking, *clause* is the smallest independent grammatical unit of expression. But this definition can hardly be used to obtain the clauses in LCMC. [10] 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 be used 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 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

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

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 examine in this paper are "word > clause > sentence" (here we use ">" to direct to a higher-rank unit in written Chinese). **Table 2** shows the Menzerathian data of this group.

As can be seen in **Table 2**, the sentence length is measured in clause, and the clause length is measured in word.

The Menzerathian function is used to fit the data, and the fitting results are displayed in **Figure 1**.

As can be seen from the fitting results in **Figure 1**, 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.

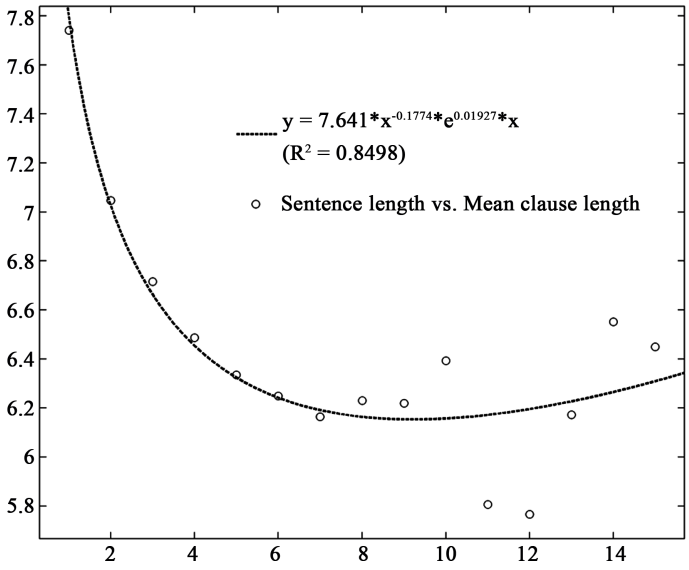

Figure 1. Fitting Menzerath-Altmann law to the hierarchical data of "word > clause > sentence".

Table 2. Hierarchical data of "word > clause > sentence".

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		

4. Conclusions

In Section 3, we tested the Menzerath-Altmann law in the "word > clause > sentence" levels. The results show that they are in line with the Menzrathian law.

Language is a system. This view has been put forward for about 100 years, however, it has never been realized until quantification is introduced into linguistics. In this paper, we show that Menzerath-Altmann law can be an efficient way of finding the basic language units in a language in the sentence level. Since language is a complex adaptive system, in the future, we will investigate into this question from a diachronic perspective to see if these Menzerathian levels will change over time.

Acknowledgements

This work is supported by the Education Department of Guangdong Province

"Innovative Strong School Project" Youth Innovation Talents Project (Humanities and Social Sciences) (Project Number: 2017WQNCX046).

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Altmann, G. (1996) The Nature of Linguistic Units. *Journal of Quantitative Linguistics*, **3**, 1-7. https://doi.org/10.1080/09296179608590059

[2] Chen, H. and Liu, H. (2016) How to Measure Word Length in Spoken and Written Chinese. *Journal of Quantitative Linguistics*, **23**, 5-29. https://doi.org/10.1080/09296174.2015.1071147

[3] Lyons, J. (1968) Introduction to Theoretical Linguistics. Cambridge University Press, London.

[4] Liu, H. and Cong, J. (2014) Empirical Characterization of Modern Chinese as a Multi-Level System from the Complex Network Approach. *Journal of Chinese Linguistics*, **42**, 1-38.

[5] Menzerath, P. (1954) Die Architektonik des deutschen Wortschatzes. Dümmler, Bonn.

[6] Altmann, G. and Schwibbe, H. (Eds.) (1989) Das Menzerathsche Gesetz in informationsverarbeitenden Systemen. Georg OlmsVerlag, Hildesheim.

[7] Köhler, R. (2012) Quantitative Syntax Analysis. Walter de Gruyter, Berlin/Boston. https://doi.org/10.1515/9783110272925

[8] Hřebíček, L. (1995) Text Levels: Language Constructs, Constituents and the Menzerath-Altmann Law. WVT, Wiss. Verlag Trier.

[9] Andres, J. (2010) On a Conjecture about the Fractal Structure of Language. *Journal of Quantitative Linguistics*, **17**, 101-122. https://doi.org/10.1080/09296171003643189

[10] Luke, K. (2006) On the Status of the Clause in Chinese Grammar. *Chinese Linguistics*, **15**, 2-14.