

Critique on *Investigating the Hierarchical Relationship Between Clause and Phrase Using the Menzerath-Altmann Law: Evidence from Academic Research Articles*

Name: Li Li

ID: 2407513

Introduction

Investigating the Hierarchical Relationship Between Clause and Phrase Using the Menzerath-Altmann Law: Evidence from Academic Research Articles is written by Mengge Wang and Jingyang Jiang in July 2024 in *Journal of Quantitative Linguistics*. As stated in the title, this article mainly explores the hierarchical relationship between sentence, clause, phrase and word in academic research articles by using the Menzerath-Altmann Law. There has been limited research on the phrases as the direct constituent of a clause. The study of this article will concentrate on the academic research articles, which is considered as one genre with high degree formality.

Theory

In this article, two theory foundation can be traced. The first is syntactic analysis, which was traditionally considered as dominant element in modern English grammar (David, 2018, p. 226). It is widely acknowledged that hierarchical structure of syntax attracted more attention than its linear order. In this article, the authors follow the rules of syntactic hierarchical structure of sentence, clause, phrase, and word as shown in Figure 1.

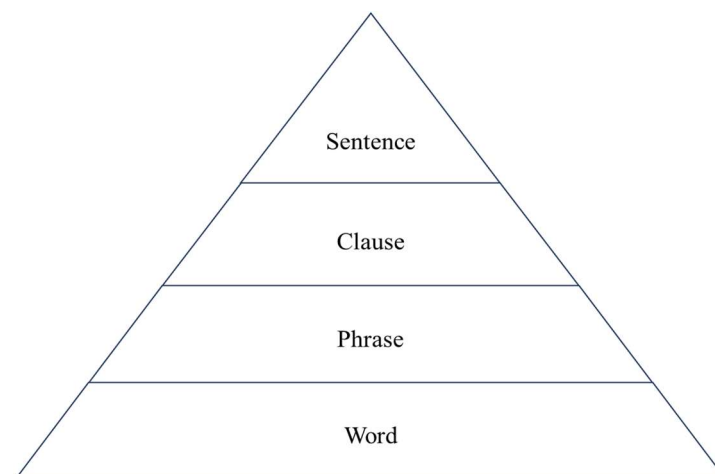


Figure 1: The hierarchical structure

There are various approaches to syntax, each offering unique insights into syntactic structures and relationships. In the structural approach to syntax, one of the most commonly used methods for analyzing English syntax is Immediate Constituent Analysis (ICA). This technique breaks down sentences into their fundamental components or constituents, revealing the underlying structure and relationships within the sentence. Consider the example, “The boys shyly touched the puppy”, to illustrate and better understand the structure of a sentence.

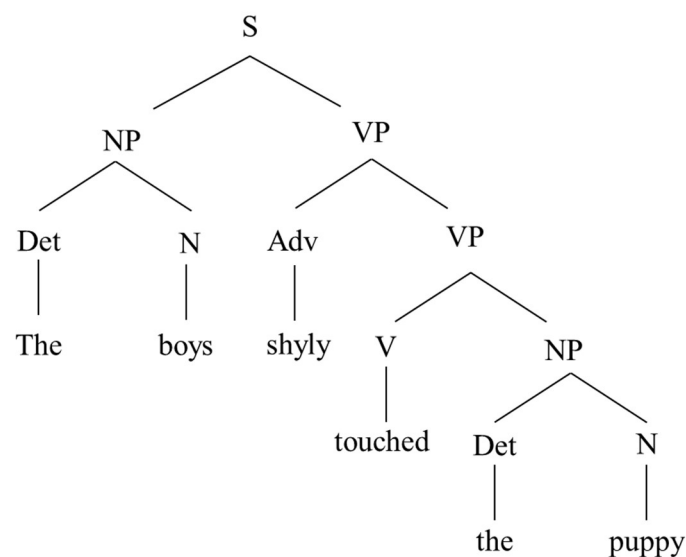


Figure 2: Immediate Constituent Analysis

As shown in Figure 2, by using Immediate Constituent Analysis, linguists can systematically examine the hierarchy and relationships between different parts of a sentence. This approach not only clarifies how words come together to form meaningful phrases but also sheds light on the syntactic rules governing English grammar.

In this article, sentences are analyzed using dependency grammar, which is widely considered a practical method for English syntax analysis. Dependency grammar focuses on the relationships (dependencies) between words in a sentence, where each word is linked to a “head” word, typically forming a hierarchical dependency tree. This approach simplifies the representation of sentence structure while enhancing the clarity of grammatical relationships, making it an invaluable tool for both linguistic analysis and natural language processing.

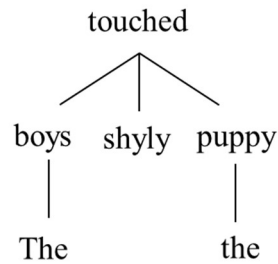


Figure 3: Dependency Grammar Analysis

As shown in Figure 3, *touched* serves as the main verb, governing both its subject (*boys*) and its object (*puppy*). The determiners (*The*) modify the respective nouns (*boys and puppy*). Additionally, *shyly* provides information about the manner in which the action (*touched*) is performed.

Compared to Immediate Constituent Analysis (ICA), dependency grammar provides a more direct representation of the subjects, predicates, and other components within a sentence. This approach allows for a straightforward depiction of grammatical relations, making it particularly effective for understanding sentence syntax.

Second, the Menzerath-Altmann Law (MAL) is a linguistic principle that states the size of a linguistic construct and the size of its direct constituents (Köhler, 2012, p. 84). This law was primarily to describe the syllables are pronounced longer in short words than in long words. Then, it was developed in the relationships between sentences and clauses, between clauses and words, which demonstrates the relationship of the construct and its constituents.

The article aims to explore the relationships between the clauses and phrases. The author considers phrases as the direct constituents of the clauses instead of words. In addition, the authors aim to test whether the correlations between sentences and clauses, and between clauses and phrases abided with the MAL. And how the correlation between each level fitted the MAL still need to be tested and further explored.

Method & Results

The analysis of the methods and results will be validated by the research questions.

RQ1: Is the MAL valid at the sentence-clause-phrase level regardless of the RA part-genres? If yes, does the clause length decrease with the increase of sentence length? (Wang and Jiang, 2024, p. 301)

RQ2: Is the MAL valid at the clause-phrase-word level regardless of the RA part-genres? If yes, does the phrase length decrease with the increase of clause length? (Wang and Jiang, 2024, p. 301)

RQ1 and RQ2 are to test the validity of MAL at the sentence-clause-phrase and at the clause-phrase-word. Based on the two research questions, this article takes quantitative research. Because the MAL is a dominant rule in the field of quantitative linguistics which provides a mathematical approach to language (Grzybek, 2006, p. 1).

RQ1 & RQ2 Method

The study used 18 academic research articles from high-impact journals in applied linguistics and education, with authors confirmed to be native English speakers. The articles were converted from PDF to plain text to facilitate automated parsing and analysis. The processed text was cleaned to retain only the actual written content of the articles, ensuring the data was uniform and ready for analysis.

- **Linguistic Unit Identification**

In this article, the authors segment the corpus into meaningful linguistic units (sentences, clauses, phrases) for analysis. Sentence boundaries were identified using punctuation marks like periods, question marks, and exclamation points to ensure that sentences are the primary constructs for subsequent hierarchical analysis. Clauses were defined as linguistic units containing a subject and a finite verb. The Stanford Dependency Parser, developed by Stanford University, was used to identify clauses syntactically and ensure consistency in the extraction process. Phrases were identified as smaller syntactic units dependent on the clause's predicate. This segmentation lays the groundwork for testing the hierarchical relationships between sentences, clauses, and phrases.

- **MAL Application**

The MAL is applied to test its validity across two linguistic levels.

Sentence-Clause-Phrase (SL-CL): The relationship between sentence length (in clauses) and clause length (in phrases).

Clause-Phrase-Word (CL-PL): The relationship between clause length (in phrases) and phrase length (in words).

The MAL formula was applied to the extracted data.

$$y = ax^b e^{-cx}$$

x : Construct length (number of clauses).

y : Mean length of constituents (phrases or words).

a, b, c : Parameters calculated during fitting.

Only constructs lengths appearing at least five times in the corpus are analyzed.

NLREG Software fitted the MAL model by estimating parameters a , b , and c and provided R^2 values to evaluate the goodness-of-fit.

$R^2 > 0.75$: Acceptable.

$R^2 > 0.8$: Good fitting

$R^2 > 0.9$: Very good fit.

RQ1 & RQ2 Results

The study examined the validity of the MAL at two levels:

- **Sentence-Clause-Phrase (SL-CL)**

At sentence-clause-phrase level, it finds that phrases are valid constituents of clauses across different part-genres, with strong fits in most sections ($R^2 > 0.9$), except for the Results section ($R^2 = 0.6448$). Generally, clause length decreases as sentence length increases, but in the Abstract section, clause length increases with shorter sentences, indicating potential interplay between the Menzerathian effect and other syntactic influences.

- **Clause-Phrase-Word (CL-PL)**

At clause-phrase-word level, the results reveal a low overall fit ($R^2 = 0.5826$) with only three of seven part-genres showing acceptable fits. Excluding the shortest clause length improves the fit significantly, indicating that phrases effectively serve as intermediate units between clauses and words.

- **Explanation**

In the data analysis, three seventh part-genre do not align with the fitting models. The results reveal a negative correlation between sentence length and clause length, as well as between clause length and phrase length. The authors use the dependency analysis to explain these findings and suggest that the effects of the dependency distance measure (DDM) against the Menzerathian effect. Dependency distance measures the linear distance between a word (dependent) and its governing word (head) in a sentence. The average DD for all dependencies in a construct was calculated using:

$$MDD(a\ sentence) = \frac{1}{n-1} \sum_{i=1}^{n-1} |DD_i|$$

where n is the number of dependencies.

It has been found that the MAL is valid when sentence lengths grow to 4 or more. In contrast, for shorter sentences of 1 to 3, the DDM effect exerts a stronger influence, allowing construct and constituent lengths to increase simultaneously. The opposing influence of DDM on the Menzerathian shortening effect does not simultaneously affect both linguistic levels (Wang and Jiang, 2024, p. 315). This interaction between the Menzerathian shortening effect and DDM illustrates a dynamic balance within language systems.

RQ3: Can the fitting parameters of the MAL in syntactic dependency structure differentiate the size-restricted RA Abstract from the RA bodies? (Wang and Jiang, 2024, p. 301)

RQ3 is to test whether the MAL parameters (a , b , c) can distinguish between different sections of academic articles, such as the Abstract vs. the main body. It is necessary to analyze how these parameters relate to the distinct characteristics of each section.

RQ3 Method

Parameters (a , b , c) derived from the MAL analysis were used as features for clustering. This analysis was conducted at two levels: Sentence-Clause-Phrase (SL-CL) and Clause-Phrase-Word (CL-PL). The analytical procedures were conducted

using R 4.3.0. Initially, the Euclidean distance between texts from different part-genres was calculated, followed by a clustering analysis employing Ward's method as the linkage criterion (Wang and Jiang, 2024, p. 305).

RQ3 Results

The comparison between results at the sentence-clause-phrase level and the results at the clause-phrase-word level indicates that Abstract stands as a distinct cluster, while the RA bodies tend to cluster together only at the clause-phrase-word level. This differentiation emphasizes the contextual variation in language use and demonstrates the effectiveness of the MAL parameters in distinguishing between part-genres. Furthermore, excluding the second MAL regime does not affect the differentiation of abstracts, suggesting that the predictive power of MAL parameters varies depending on the level of linguistic units analyzed.

Conclusion

Phrases, rather than individual words, serve as intermediate constituents between clauses and larger linguistic units. The validity of the MAL is assessed across various linguistic levels (sentence-clause-phrase and clause-phrase-word) and academic text genres, providing insights into the relationship between construct complexity and constituent size.

The results also revealed genre-specific patterns: Abstract, constrained by its size limitations, exhibited distinct syntactic behaviors, further supporting the self-regulating nature of language. Abstracts demonstrated unique trends, such as shorter dependency distances and higher proportions of adjacent dependency relations, which align with the strict word limits characteristic of this part-genre (Wang and Jiang, 2024, p. 319).

Contribution

The research uses corpus analysis, advanced dependency parsing, and statistical modeling, demonstrating the accuracy and efficiency of quantitative modeling and clustering in a scientific way. The logical flow of the research process is clear and enhances the Menzerath-Altmann Law (MAL) fitness at multiple linguistic levels in the academic articles, which is a unique contribution to the study of genres.

The study specifically examines part-genres such as Abstract, Introduction, Literature Review, Method, Results, and Conclusion. It demonstrates the features of each part while emphasizing size constraints and a high degree of compression in Abstracts, providing valuable insights for researchers. In addition, the study focuses on English academic writing, the framework can be extended to other languages and text types, suggesting a broader scope for future research.

Limitations

The study on linguistic identification is using dependency grammar. It is known that different approaches may have different definitions at linguistic levels. There are confusions about the distinction between clauses and phrases in this field. For instance, the phrase identified in this article are analyzed by dependency grammar. If the phrase identified by structural grammar, the results may vary. Thus, the identification of clause and phrase would be attaching great importance and impact the final conclusion.

The MAL formula varies much. There are a lot of variant formula in this field. Although the authors adopt the most universal one, the parameters a , b , c are not clear. Eroglu (2014, p. 393) argues that there is no convincing theoretical support for the substantiated interpretation of its parameters. Schusterovac et al. (2017, p. 55) also claims that there are multiple mathematical models used for the Menzerath–Altmann law, and the universal applicability of these models has not been widely verified.

References

- Crystal, D. (2019) *The Cambridge encyclopedia of the English language*. Cambridge: Cambridge University Press.
- Eroglu, S. (2014) ‘Menzerath–Altmann Law: Statistical Mechanical Interpretation as Applied to a Linguistic Organization’, *Journal of Statistical Physics*, 157(2), pp.392–405. DOI: <https://doi.org/10.1007/s10955-014-1078-8>.
- Grzybek, P., Stadlober, E. and Kelih, E (2007) ‘The Relationship of Word Length and Sentence Length: The Inter-Textual Perspective’, *Studies in classification, data analysis, and knowledge organization*, pp.611–618. DOI: https://doi.org/10.1007/978-3-540-70981-7_70.
- Köhler, R. (2012) *Quantitative Syntax Analysis*. Walter de Gruyter.
- Schusterova, D., Motalova T., and Wang B. F. (2017) ‘Menzerath-Altmann定律: 理论贡献与局限’, *Linguistic Studies*, (1), pp.45–58.
- Wang, M. and Jiang, J. (2024) ‘Investigating the Hierarchical Relationship Between Clause and Phrase Using the Menzerath-Altmann Law: Evidence from Academic Research Articles’, *Journal of Quantitative Linguistics*, 31(4), pp.298–322. DOI: <https://doi.org/10.1080/09296174.2024.2375998>.