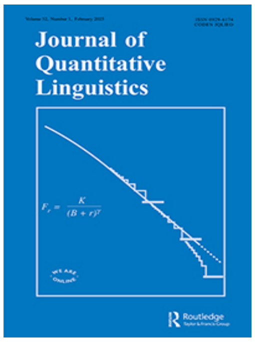

Journal of Quantitative Linguistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/njql20

Linguistic Compression and Cognitive Load: A Quantitative Study of the Menzerath–Altmann Law in Interpreted, L2 and Native English Speech

Ziqiao Su, Han Xu & Kanglong Liu

To cite this article: Ziqiao Su, Han Xu & Kanglong Liu (23 Mar 2025): Linguistic Compression and Cognitive Load: A Quantitative Study of the Menzerath–Altmann Law in Interpreted, L2 and Native English Speech, Journal of Quantitative Linguistics, DOI: 10.1080/09296174.2025.2468573

To link to this article: https://doi.org/10.1080/09296174.2025.2468573

Published online: 23 Mar 2025.

JOURNAL OF QUANTITATIVE LINGUISTICS
https://doi.org/10.1080/09296174.2025.2468573

Linguistic Compression and Cognitive Load: A Quantitative Study of the Menzerath–Altmann Law in Interpreted, L2 and Native English Speech

Ziqiao Su, Han Xu and Kanglong Liu

Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong SAR, China

ABSTRACT

This study examines the applicability of the Menzerath–Altmann Law (MAL) across three varieties of English speech: interpreted speech (IE), second language (L2) speech and native English (NE) speech. By analysing the relationship between sentence length and clause length, the research evaluates the fitting parameters *a* and *b* of the MAL formula to investigate syntactic patterns. The results confirm MAL's validity for all varieties, supporting the principle that longer linguistic structures consist of shorter components. The study also identifies distinct and shared features of each speech type. Parameter *a* indicates that L2 speech aligns more closely with IE than NE, likely due to shared cognitive constraints, while IE and NE exhibit minimal differences, reflecting interpreters' high proficiency. Parameter *b* shows that IE has the steepest reduction in sentence length with increasing complexity, reflecting interpreters' automated processing strategies under high cognitive load. In contrast, L2 speakers manage complex structures more deliberately, indicating a more gradual adaptation to complexity and distinct cognitive approaches. These findings offer insights into the application of linguistic laws and the influence of interpreting on language production, contributing to the understanding of constrained versus unconstrained language use. This study opens avenues for further quantitative investigations into linguistic variation.

1. Introduction

Language is a complex system, consisting of interrelated, multi-layered subsystems (Jiang & Ma, 2020). The intricate nature of this structure not only exemplifies linguistic complexity but also presents significant challenges for thorough analysis. Quantitative linguistics addresses these challenges by employing mathematical methods to investigate linguistic phenomena (Schusterová et al., 2017). Notably, the Menzerath–Altmann Law (MAL) offers a novel quantitative framework that elucidates the connections between linguistics and mathematics, thus providing a means with which

CONTACT Kanglong Liu kl.liu@polyu.edu.hk

to examine the relationships within the structural layers of language (Altmann, 1980).

The German linguist and phonetician Paul Menzerath, a principal founder of the MAL, formulated this law after examining the relationship between the quantity and length of syllables within words. His observation, often summarized as 'the bigger the whole, the smaller its parts' (Menzerath, 1954, p. 101), remained largely unnoticed by linguists for some time (Andres et al., 2012). It was not until 1980 that Gabriel Altmann, regarded as 'the founder of quantitative linguistics in Germany and the Nestor of modern quantitative linguistics' (Köhler et al., 2021, p. 189), expanded Menzerath's work. Altmann introduced the terms 'language construct' and 'constituent' to define the MAL, generalizing the law across all linguistic levels. He concluded that 'the longer a language construct, the shorter its components' (Altmann, 1980, p. 1), thereby demonstrating an inverse relationship between the length of language structures and their components across languages and linguistic levels. The MAL can be expressed mathematically as the following functions:

$$y = ax^{b}e^{-cx}(b<0) \tag{1}$$

$$y = ae^{-cx} \tag{2}$$

$$y = ax^{b}(b<0) \tag{3}$$

In equations (1)–(3), x represents the length of the structure, measured by the total number of its components, while y is the average length of its components. This average length corresponds to the language unit at the nearest lower linguistic level. The parameters a, b, and c are empirically determined constants, with e representing Euler's number – approximately equal to 2.718. The length of a linguistic structure is measured by the total number of its components, which is always expressed as an integer. In contrast, the component lengths are typically represented as averages and are usually expressed in decimals. According to the equations outlined, the MAL follows a power law, which indicates that, as the length of the components decreases, the total length of the structure generally increases (Schusterová et al., 2017).

Given the significance of selecting the appropriate computational formula for accurate analysis, it is crucial to understand how these different formulas can influence data interpretation and the conclusions drawn (Andres et al., 2012, 2014). In general linguistic inquiries, the simplified version of the MAL is often considered sufficient. Because previous research has frequently relied on formula (3) to examine the relationship between sentences and clauses (Jiang & Jiang, 2022; Jiang & Ma, 2020; Li et al., 2023; Y. Wang & Chen, 2022;

Xu & He, 2020), parameter *c* 'can be set to 0' (Köhler, 2012, p. 148). Therefore, following this precedent, we have utilized only parameters *a* and *b*, which enhances clarity and enables a more straightforward interpretation of the results.

The MAL reflects the idea that the limitations of human perceptual processing necessitate a balance between the continuous flow of linguistic information and the efficient use of cognitive resources (Xu & He, 2020). This balance aligns with the principle of cognitive economy (Fenk & Fenk-Oczlon, 1993), which suggests that linguistic structures are optimized to minimize cognitive load. In the context of sentence and clause length relationships, parameter *a* represents the baseline of the fitting curve, corresponding to the mean length of a one-clause sentence ($x = 1$) as per the MAL. Parameter *b* defines the slope of the curve, illustrating how mean clause length changes as the number of constituent clauses increases (Jiang & Ma, 2020). Therefore, as the length of a higher-level linguistic unit, such as a sentence, increases, the length of its lower-level components, such as clauses, tends to decrease. Otherwise, the structural complexity of the higher-level unit would grow proportionally with its length, ultimately exceeding human cognitive capacity (Jin & Liu, 2017). Due to the sensitivity of parameters *a* and *b* to linguistic structure, they may serve as useful predictors in various linguistic applications. For example, these parameters can potentially serve as good predictors for register differentiation, typological classification of language, authorship attribution and similar fields (Mačutek & Mikros, 2015).

The parameters of the MAL exhibit a structurally systematic quantitative relationship with syntactic complexity and research indicates that these parameters can effectively capture the hierarchical organizational characteristics of linguistic units, reflecting how different levels of structure are organized within language. Jiang and Jiang (2022), through the studying modes of language interpretations, found that the MAL parameters *a* and *b* can reveal differences in language representation under different interpreting modes, with those differences being influenced both by the limitations of human cognitive capacity and the modes of interpreting. In particular, the value of parameter *b* in simultaneous interpreting is significantly lower than it is in consecutive interpreting, thus suggesting that sentence components decrease more rapidly in simultaneous interpreting, perhaps in response to a higher cognitive load. Furthermore, Jiang and Ma's (2020) study supports the hypothesis of translation language as a 'third code' and proposes that MAL parameters *a* and *b* can be used to differentiate translation language from original language, thereby revealing changes in syntactic complexity during the translation process. In that light, the extant research results indicate that the MAL parameters not only reflect the economical processing of language but also demonstrate how syntactic complexity changes during translation and interpreting activities, relationships that are influenced by cognitive load and language modes. Furthermore, these insights validate the application of the MAL in syntactic research and offer a new perspective for understanding the cognitive mechanisms involved in linguistic compression.

In this study, we have sought to extend the understanding of the MAL by applying it to the syntactic structures of three varieties of English speech: interpreted English speech, English as a second language speech, and native English speech, specifically within interpretation scenarios. By building upon the existing research, our work highlights the practical significance of the MAL, illustrating its relevance in the field of interpreting studies and contributing to the broader understanding of linguistic compression.

2. Related Work

Although the term ‘linguistic compression’ has not yet reached consensus, extensive research has highlighted its significance. The concept can be traced back to Zipf’s (1949) seminal research on the relationship between word length and frequency, which observed that high-frequency words tend to be shorter than low-frequency ones, reflecting what he termed the principle of least effort in human language use. This principle suggests that speakers tend to employ shorter linguistic forms to express frequently occurring concepts, thereby minimizing the overall communicative effort. Building on this foundation, linguistic compression has emerged as a fundamental mechanism in human communication, referring to speakers’ systematic tendency to reduce linguistic forms while maintaining information content to optimize cognitive resources. Specifically, Yin and van Schijndel (2023) defined linguistic compression as ‘significant cognitive efforts to compress information’ (p. 7922). This process encompasses a series of complex linguistic operations including removing irrelevant information, summarizing concepts, and constructing fluent expressions. Several studies have explored related compression mechanisms from various perspectives, including information theory and pragmatics. In information theory, Shannon’s (1948) compression principle demonstrated how more information can be encoded using fewer bits. In linguistics, researchers have discovered that speakers tend to compress linguistic elements to enhance their communication efficiency (Jaeger, 2010). In the context of this study, linguistic compression plays a critical role, as it reflects speakers’ ability to compress and adapt syntactic components to varying cognitive demands during real-time language production. This adaptive mechanism underscores the intricate balance between efficiency and expressiveness in human communication.

The MAL illustrates a fundamental pattern of information compression in language systems. This compression manifests in human language through systematic relationships: longer words are composed of shorter syllables (Altmann, 1980), and longer sentences consist of shorter clauses (Teupenhayn & Altmann, 1984). The compression acts as an adaptive mechanism for managing cognitive resources during language production (Gustison et al., 2016; Jiang & Jiang, 2022), thereby allowing speakers to maintain an efficient flow of information while also optimizing their cognitive load (Fenk & Fenk-Oczlon, 1993). The validity of the MAL has been corroborated by multiple studies across various languages and linguistic units (Mačutek et al., 2019), which have consistently demonstrated that the MAL applies not only to different languages but also to different levels of language structure. Interestingly, the MAL has also proven valid in non-linguistic contexts. For instance, vocal sequences in non-primate species, such as penguins, exhibit patterns that are consistent with the MAL (Favaro et al., 2020). In linguistic research, the most commonly analysed units are sentences, clauses, words, and phrases (Chen & Liu, 2022; Jiang & Ma, 2020; Köhler, 1982; Li et al., 2023; Mačutek et al., 2017). Köhler (1982) was among the first to validate the MAL, correlating sentence lengths with constituent clause lengths in German texts. Similarly, Mačutek et al. (2017) applied the MAL to syntactic dependency structures in Czech, finding that larger clauses contained shorter phrases. Xu and He (2020) further validated the MAL by examining English language sentence lengths (measured by the number of clauses) and the length of their constituent clauses (measured by the number of words) across both academic spoken and written registers in academic settings. Chen and Liu (2022) employed the MAL to investigate the hierarchical relationships between linguistic units in written Chinese, focusing on its implications for register classification. Their analysis revealed a predictable mathematical relationship between the lengths of higher-level linguistic units, such as sentences and clauses, and their lower-level components, such as words, characters, and even strokes. These findings are consistent with the earlier work by Köhler (1982), thus further validating the MAL across different languages and linguistic systems. Indeed, in addition to its relevance to hierarchical linguistic structures, research on the MAL has extended the law's application to the distribution of distinct words within large texts, uncovering statistical patterns that reflect the regularities of word usage. Eroglu (2013, 2014) explored these implications in a broader context, demonstrating how the MAL can also be applied to non-linguistic organizations and showing that the principle of cognitive economy observed in language can similarly manifest in other complex language systems.

Despite being tested across various linguistic levels and continuing to be developed, the MAL was initially formulated without fully considering the boundary conditions inherent in all languages (Altmann, 2014; Eroglu, 2014). Studies such as Mačutek et al. (2019) have demonstrated the MAL's applicability to spoken Czech, suggesting its potential for analysing spoken as well as written language. However, this applicability does not extend to random texts. Dębowski (2007) emphasized this by comparing English and Polish texts with their random character sequences arranged by frequency. His findings revealed a significant structural divergence between random strings and natural language, underscoring that the MAL cannot be applied indiscriminately and requires further refinement. Whereas the MAL shows strong validity in written language, its effectiveness in spoken language remains debated. Although some studies, like that of Mačutek et al. (2019), support its applicability to spoken registers, other researchers argue that more evidence is needed to confirm its reliability in the spoken domain (Jiang & Jiang, 2022). This ongoing controversy highlights the need for continued research to clarify the MAL's limitations and potential across different linguistic contexts.

The application of the MAL to translation studies, particularly in the field of interpreting, has been relatively rare (Jiang & Jiang, 2022; Jiang & Ma, 2020). Interpreting, which involves the spoken form of translation, differs fundamentally from written translation and presents unique cognitive challenges. As translation has been described as 'the most extreme end of the bilingual activation continuum', it requires rapid bidirectional switching between source and target texts, thus imposing significant cognitive constraints on language processing (Kruger & van Rooy, 2016b, p. 121). These cognitive restrictions could have implications for the MAL, especially in the real-time demands of interpreting. Recent research in interpreting has increasingly focused on translated English and non-native varieties of English, commonly referred to as constrained languages (Chen et al., 2024; Liu et al., 2022; Liu et al., 2023). However, few studies have specifically investigated spoken constrained varieties, with most focusing instead on lexical (Kajzer-Wietrzny, 2022; Kajzer-Wietrzny & Ivaska, 2020) and syntactic features (Chen et al., 2024; Liu et al., 2023). Given the distinct linguistic patterns that interpreting may exhibit compared with written translation, it is essential to incorporate interpreting into the constrained language research paradigm (Kajzer-Wietrzny, 2015). Doing so would provide a more comprehensive understanding of how spoken translation operates under different cognitive and linguistic pressures, potentially offering new insights into the role of the MAL in these contexts.

Within the context of constrained languages, simplification has become a key focus in research on interpreting. This phenomenon refers to interpreters' tendency to use simpler language than that of the target language, aligning with Baker's (1993) proposal to study translation universals by comparing translated language with non-translated native language. It should be noted that simplification is a complex, multidimensional process

(Kruger & van Rooy, 2016a), with research yielding varying results due to the selective use of linguistic indicators. This inconsistency underscores the need for more robust quantitative frameworks to holistically approach simplification (Z. Wang et al., 2024). Importantly, simplification should be seen as part of a broader strategy of adjusting to linguistic structures, rather than merely reducing syntactic complexity. This adjustment involves balancing the need to convey information clearly and efficiently within the cognitive constraints of the interpreter, which does not necessarily entail simplifying sentence structures in the traditional sense. Another critical factor is the influence of different varieties of English on how language structures are adjusted during interpretation. Kruger and van Rooy (2016b) analysed syntactic and pragmatic transfer effects in reported-speech constructions across three contact varieties of English influenced by Afrikaans, demonstrating that these varieties are distinct forms of English emerging from language contact situations. These varieties reflect diverse socio-cognitive conditions and bilingual activation strengths, differing not only in structural features but also in how they interact with cognitive constraints and language processing. Drawing on models like the MAL could provide deeper insights into this intricate behaviour by offering a structured way to analyse how interpreters balance cognitive constraints and language processing. Interpreting, as a sophisticated cognitive activity, reflects the underlying linguistic structures, showcasing the interplay between cognitive limitations and language representations (Jia & Liang, 2020). The MAL modelling may help illuminate the different interpretation outcomes that result from varying modalities and cognitive pressures. In a notable study, Jiang and Jiang (2022) explored whether the principles of MAL, traditionally applied to written language, could be extended to the dynamic environment of oral interpretation. Although their findings were promising, the study's dataset – consisting of only 11 texts for consecutive interpreting and 22 for simultaneous interpreting – was relatively small to draw definitive conclusions. The limited sample size makes it difficult to assess whether the observed patterns would hold across a broader range of interpreting scenarios. Given the diversity of topics, styles, and linguistic complexities in real-world interpreting, a larger dataset would provide a more accurate and comprehensive understanding of how the MAL applies to interpreting contexts.

Despite these limitations, however, previous research has suggested that the MAL may offer significant understanding of certain complexities of spoken language. Fitting parameters have been shown to differentiate between various language levels, types of registers, and text types, or they may be constrained by other factors (Cramer, 2005; Hou et al., 2017; Jiang & Ma, 2020; Wang & Chen, 2022; Xu & He, 2020). This implies that the MAL could serve as a promising tool for analysing the unique features of interpreted language, particularly when considering the inherent constraints of spoken language in interpreting contexts. However, a gap remains in our understanding of how mathematical model parameters should be interpreted in these scenarios. Most studies have focused on verifying the MAL within certain boundaries or under specific conditions, exploring how parameter values reflect the relationship between linguistic structures and their components. Yet, there is a notable lack of research on applying the MAL to interpreted language specifically. This gap underscores the need for further exploration to fully understand how the MAL can inform the analysis of interpreted speech in comparison to other forms of constrained speech, such as L2 and unconstrained English speech.

3. Materials and Methods

3.1. Research Questions

Given the relative lack of research on interpreted language from the perspective of constrained languages and recognizing interpretation as ‘an extreme case of translation’ (Shlesinger & Ordan, 2012, p. 54), this study seeks to bridge that gap by exploring the applicability of the MAL in modelling the inter-relation between sentence length and clause length in interpreted speech, second language speech, and native English speech from the constrained language perspective. Specifically, this study aims to investigate the complex parameter relationships of these three varieties of English speech and address the following research questions:

RQ1: Is the MAL valid for L1–L2 interpreted speech and L2 speech, which are considered spoken constrained languages, as well as for native English speech, which is considered an unconstrained language?

RQ2: If the MAL proves to be valid, what differences exist among these varieties of English speech in terms of syntactic relations at the sentence-clause level? Specifically, can the fitting parameters *a* and *b* of the MAL formula reliably measure these differences?

RQ3: Will L1–L2 interpreted speech and L2 speech, as constrained languages, exhibit similar or different patterns in terms of the MAL when compared to native English speech? If differences are observed, what might explain these variations?

3.2. Corpus

This study utilizes the Political Debate English Comparable Corpus (PDECC) from Liu et al. (2023), which provides a well-structured dataset across three distinct varieties of English speech: Native English (NE), Second Language English (L2), and interpreted English (IE). The PDECC is particularly suited for this research due to its comparability in text size, time span, and genre. Each sub-corpus comprises 50 texts of approximately 2,000 words, covering key political, social, and economic topics, including budget control, social welfare, and human rights, as shown in Table 1.

Table 1. Overview of the corpus.

Sub-corpus	Source	Texts	Tokens	No. of sentences	No. of clauses	Producer
NE	UK Parliament debates (2016–2020)	50	98,854	4,424	10,224	L1 speakers (English)
IE	HK LegCo debates (2016–2020)	50	100,637	5,916	10,364	L2 speakers (English)
L2	HK TV interviews (2016–2020)	50	104,922	5,996	12,400	L2 speakers (English)

The IE sub-corpus comprises Cantonese-to-English simultaneous interpreting (L1–L2 interpreted speech) transcripts from Legislative Council debates in Hong Kong between 2016 and 2020. The NE sub-corpus consists of native English (L1) speeches from debates in the UK House of Commons during the same period. The L2 sub-corpus was compiled using interviews from two prominent Hong Kong television shows, *All About Money* and *Talk and Walk*, where native Cantonese (L2) speakers conducted interviews in English on political, financial and social topics from 2016 to 2020. All the debates and interviews included in the PDECC are publicly accessible. By incorporating three distinct varieties of English speech – native, non-native, and interpreted English – the PDECC provides a strong foundation for cross-linguistic analysis using the MAL. This diversity allows for the exploration of linguistic patterns and structural relationships across different speech types, making it possible to assess how unit size affects larger linguistic constructs. The controlled thematic content and consistent discourse context further validate the use of the MAL, ensuring that any observed variations are due to language differences rather than contextual influences (Liu et al., 2023; Xu & Liu, 2023). Consequently, the PDECC offers both reliability and depth in examining linguistic variations in political discourse, making it an ideal resource for applying the MAL.

3.3. Methods

Since the segmentation of English clauses is well-established, we adopted the method outlined by Köhler (1982), which determines the number of clauses in a sentence based on the count of finite verbs. Identifying the number of finite verbs is crucial for accurately determining the number of clauses. This approach has been widely adopted and validated in academic research (Jiang & Ma, 2020; Li et al., 2023; Xu & He, 2020). In this study, sentence length is defined by the number of clauses, while clause length is calculated by dividing the total number of words in a sentence by the number of clauses.

After transcribing and cleaning all corpora, we applied TagAnt 2.0.5 (Anthony, 2022) for part-of-speech (POS) tagging, which uses two tags to mark sentence boundaries: the _SENT tag, which identifies sentences ending with punctuation marks such as ?, ! or ., and the _: tag, which captures sentences ending with punctuation like ;, -, – or :. These tags were used to accurately determine sentence boundaries. To ensure comprehensive and accurate extraction of finite verbs, we identified them using ten specific POS tags: VBD (verb be, past tense), VBP (verb be, present, non-third person), VBZ (verb be, present, third person singular), VHD (verb have, past tense), VHP (verb have, present, non-third person), VHZ (verb have, present, third person singular), VVD (verb, past tense), VVP (verb, present, non-third person), VVZ (verb, present, third person singular), and MD (verb, modal). Human verification was also performed to further refine the identification and measurement of finite verbs. While no segmentation framework can account for every linguistic nuance, our current rules are both fine-grained and objective, aiming to balance linguistic accuracy with practical application.

All statistical analyses in this study were conducted using Python with its scientific computing libraries. To begin our analysis, we first examined and compared the cumulative relative frequency distribution of sentence length across NE, IE, and L2. Following this, we tested the validity of the MAL in these three varieties of English speech. For the MAL fitting process, we used the curve_fit function from SciPy to fit the data to the MAL formula (3), where x represents sentence length (measured in clauses) and y represents the mean clause length (measured in words) for sentences of length x. The parameters a and b characterize the quantitative relationship between these variables as the fitting coefficients. For each text, we first calculated the average clause length for each sentence length category, then fitted these data points to the MAL formula using non-linear least squares regression. The fitting results were assessed based on the coefficient of determination (R^2), with $R^2 > 0.7$ indicating an acceptable fit, $R^2 > 0.85$ reflecting a relatively good fit and $R^2 > 0.9$ considered a good fit (Jin & Liu, 2017; Li et al., 2023).

Once the MAL was validated, we explored whether the fitting parameters could effectively differentiate between the outputs of NE, IE, and L2. We conducted the Kruskal-Wallis test to determine whether significant differences existed in the parameters across the three varieties of English speech. To ensure the reliability of the results, we excluded sentence lengths with very low frequency (fewer than five occurrences). For instance, sentences consisting of 11 clauses appeared only twice in the NE corpus and were omitted to avoid potential bias.

4. Results

4.1. Frequency Distribution of Sentence Length Measured in Clauses

The length of English sentences in each text collection was determined by two factors: the overall word count and the total number of clauses. As shown in Figure 1, sentence frequencies in all three corpora tend to decrease as sentence length increases. Interestingly, the IE corpus not only starts at a higher frequency but also exhibits a steeper decline, with around 40% of its sentences consisting of just one clause, compared to approximately 30% in both the NE and L2 corpora. In addition, sentence lengths in the NE corpus range from one to eleven clauses, while IE texts range from one to eight clauses, and L2 texts range from one to thirteen clauses.

In terms of cumulative frequency distribution, Figure 2 shows that sentences exceeding seven clauses are rare in all three corpora. This observation aligns with Miller's (1956) theory on the limitations of short-

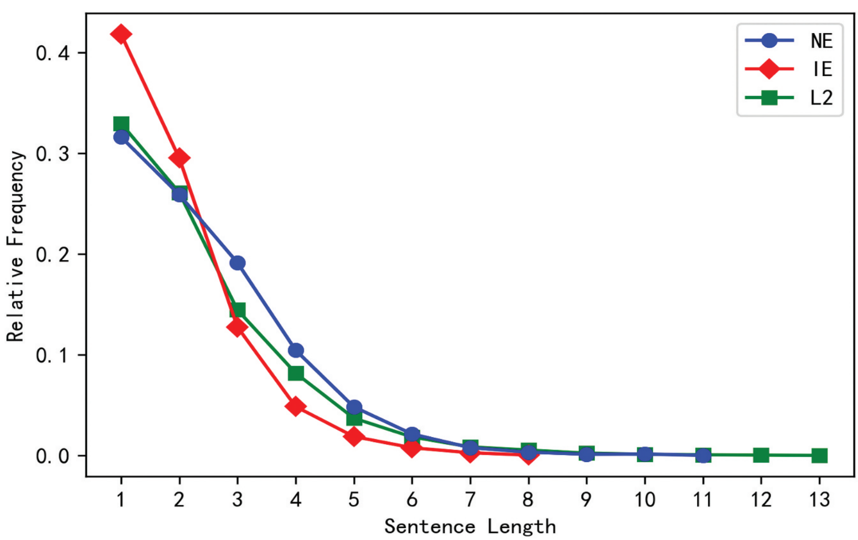

Figure 1. Frequency distribution of sentence length (in clauses).

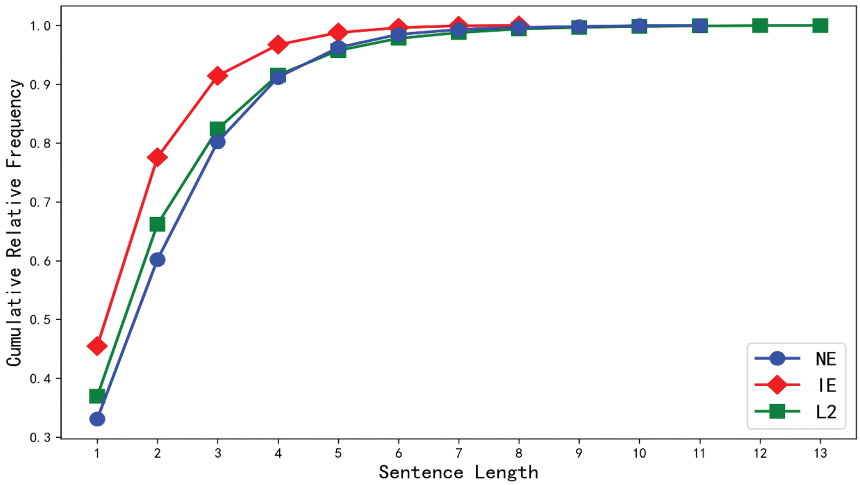

Figure 2. Cumulative relative frequency distribution of sentence length (in clauses).

term memory, which posits that humans can hold a limited number of items at once. Cowan (2001) later refined this threshold to four items, further emphasizing the cognitive constraints that influence sentence structure. These findings are consistent with those of Li et al. (2023) and also support Zipf's (1949) principle of least effort, which suggests that language is shaped by the need to optimize cognitive efficiency. The comparison of sentence length distributions across the corpora highlights how cognitive processing limits influence sentence construction in different varieties of English.

4.2. MAL Testing for NE, IE and L2 Corpora Separately

We analysed the relationship between sentence and clause lengths across the three corpora to determine their fit with the MAL. For each sentence length, the average clause length was calculated to simulate this relationship. Figure 3 shows a clear decline in average clause length as sentence length increases, as predicted by the MAL. While the trend is generally consistent, minor fluctuations, especially in the NE and L2 corpora, can be observed at longer sentence lengths.

The fitting results in Figure 4 demonstrate a strong consistency between the observed data and the MAL model, with R^2 values of 0.939 for NE, 0.917 for IE, and 0.947 for L2, as indicated in Table 2. These values suggest a better overall fit for L2, though all three varieties of English speech conform well to the MAL model. The results affirm the validity of the MAL across both constrained (IE

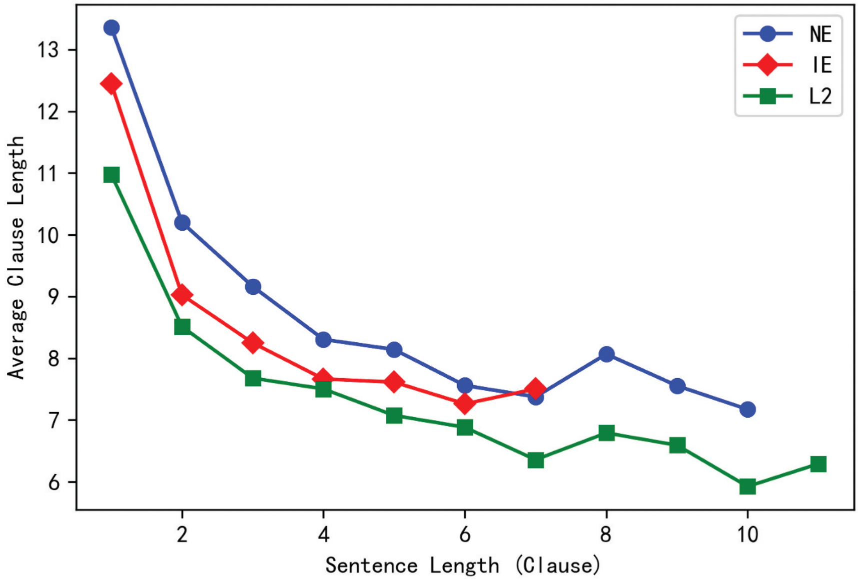

Figure 3. Relation between average clause length and sentence length in NE and IE sub-corpora.

and L2) and non-constrained (NE) varieties of English speech, indicating that English sentence length follows the expected patterns. While the fit is not without minor irregularities, the results provide strong evidence of MAL's applicability across different English speech varieties.

4.3. Differences in the Fitting Parameters of NE, IE and L2

Having validated the MAL for English at the sentence level across NE, IE, and L2 corpora, we examined the structural parameters distinguishing these English varieties. Table 3 and Figure 5 present the statistical information for parameters *a* (the length of one element construct) and *b* (the gradient, representing the rate of clause shortening as sentence length increases). Parameter *a* represents the average sentence length (in words) for a single-clause sentence. It functions as a scaling factor in the MAL equation, determining the initial magnitude of *y* and shaping the overall curve. While *a* is a fitted parameter and may not correspond directly to observed values, it serves as a theoretical baseline that characterizes the general trend in the data. The results show that NE has a slightly higher mean value for parameter *a* (13.073) compared to IE (12.803) and L2 (10.750), reflecting how different speaker groups initiate their discourse based on linguistic proficiency and cognitive strategies. Parameter *b* quantifies the rate at

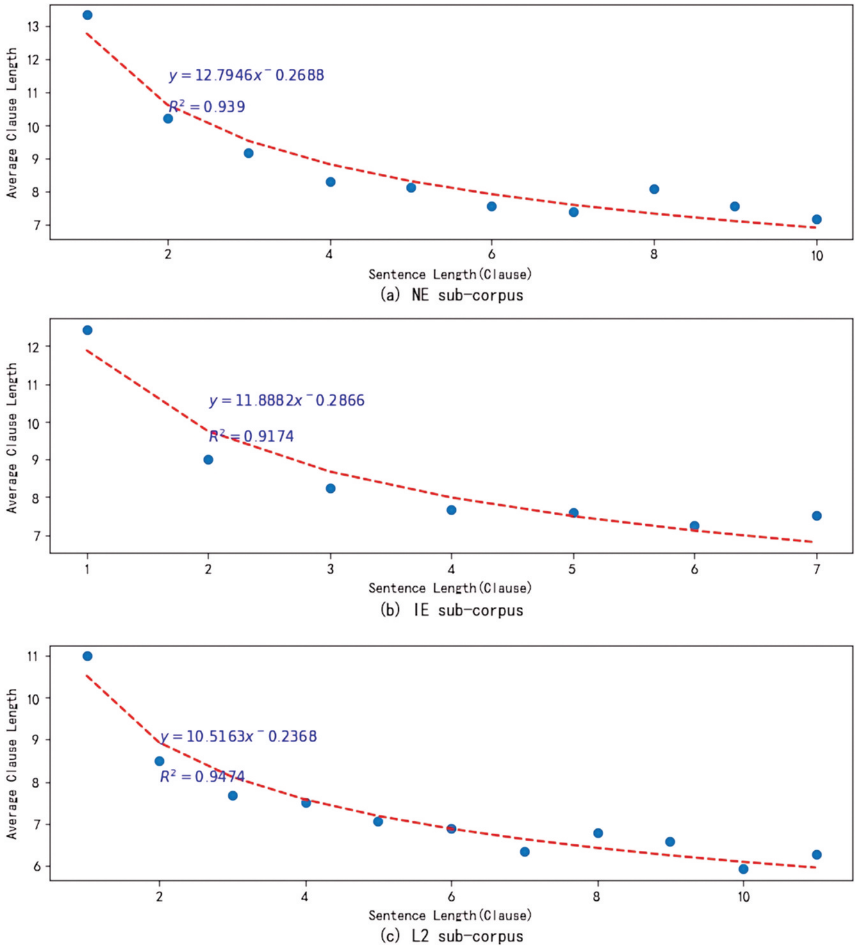

Figure 4. Fitting results of the MAL for NE and IE sub-corpora, respectively.

Table 2. MAL fitting results for the three corpora.

Corpus	NE	IE	L2
Number of points	10	7	11
Parameter of *a*	12.795	11.888	10.516
Parameter of *b*	−0.269	−0.287	−0.237
R^2	0.939	0.917	0.947
F value	1798.572	892.553	2789.244
p	<.001	<.001	<.001

which clause length decreases as additional clauses are added, thus reflecting how sentence complexity evolves with increasing sentence length. This decreasing trend in clause length can indicate how speakers manage

Table 3. Overview of statistical information of the parameter *a* and *b* of the three sub-corpora.

Sub-corpus	Parameter	Mean	Median	SD	Minimum	Maximum
IE	a	12.803	12.347	2.929	8.983	20.346
NE	a	13.073	12.906	1.980	9.306	19.886
L2	a	10.750	10.566	1.482	8.024	14.750
IE	b	−0.351	−0.346	0.125	−0.609	−0.081
NE	b	−0.292	−0.285	0.099	−0.536	−0.054
L2	b	−0.249	−0.263	0.092	−0.466	−0.030

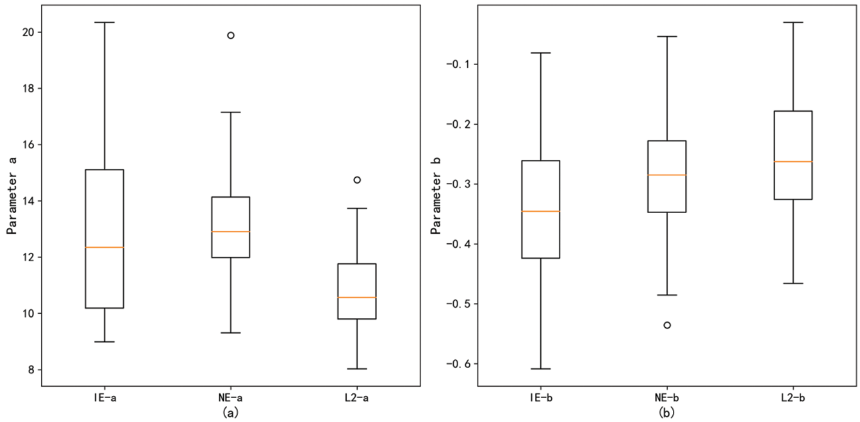

Figure 5. Distribution of NE, IE and L2 parameters.

cognitive load and information structuring in longer discourses. For parameter *b*, NE has a less negative value (−0.292), indicating a more gradual decline in clause length as sentences lengthen, while IE shows the most negative value (−0.351), reflecting a steeper rate of decline in clause length. L2, with the least negative value (−0.249), shows the slowest rate of decline among the three corpora. This confirms that IE responds more strongly to sentence lengthening, leading to faster reductions in clause length, while NE and L2 exhibit a more gradual decrease. Overall, the results indicate that NE starts with longer clauses but maintains a more moderate pace in reducing complexity, whereas IE demonstrates heightened syntactic compression and L2 employs a more deliberate complexity management approach. This analysis provides valuable insights into the adaptive strategies used by each group to manage cognitive and linguistic demands during speech.

To investigate whether the variation in parameters *a* and *b* of the MAL among the three sub-corpora follows a similar pattern from a statistical perspective, we applied the Kruskal-Wallis test, given the non-normal distribution of the data. The results show significant differences among the three corpora for both parameter *a* ($H = 28.84$, p

< .001) and parameter b ($H = 17.76$, $p < .001$), indicating that these varieties differ significantly in terms of syntactic relations at the sentence-clause level.

Dunn's post-hoc test was conducted to identify which specific groups exhibited significant differences. As shown in Table 4, for parameter a, there is no significant difference between NE and IE ($p = .196$). However, there are highly significant differences between NE and L2 ($p < .001$) and between IE and L2 ($p < .001$). The results indicate that sentences with a single clause are, on average, shorter in L2 compared to both IE and NE, as suggested by the lower values of parameter a. Notably, the difference between L2 and NE is larger than that between L2 and IE. For parameter b, significant differences exist between NE and IE ($p < .05$) and IE and L2 ($p < .001$), while the difference between NE and L2 is marginally significant ($p = .052$). The smallest difference is between NE and L2, followed by NE and IE, and the largest difference is between IE and L2. These findings indicate that syntactic compression processing differs more strongly between IE and L2, while NE and L2 exhibit more similar patterns in the rate of clause shortening as sentence length increases, highlighting the distinct strategies these varieties employ to handle syntactic complexity.

The findings suggest that IE exhibits simpler syntactic structures compared to NE. Additionally, the analysis shows that, as constrained languages, IE and L2 share some common characteristics. While there is no significant difference in parameter a between IE and NE, which could be attributed to variations in interpreter proficiency, it indicates that interpreters maintain near-native performance in terms of basic clause length. In contrast, the greater difference in parameter b between L2 and IE, compared to L2 and NE, suggests that L2 speakers perform more similarly to NE speakers when adjusting clause length to accommodate increasing sentence complexity. This demonstrates that constrained languages are not a homogeneous category, but rather encompass various types of language output, each with its own distinct features and strategies.

Figure 6 shows the combined scatterplot of NE (blue), L2 (green), and IE (red) texts according to parameters a and b of the MAL model. The different varieties of English speech form distinct clusters, with minimal overlapping,

Table 4. Dunn's test results of the parameter a and b among the three sub-corpora.

Sub-corpus	Parameter	Difference	p-value	Sample Size (NE, IE, L2)	Degrees of Freedom
NE-IE	a	0.270	.196	50, 50	98
NE-L2	a	2.323	<.001**	50, 50	98
IE-L2	a	−2.053	<.001**	50, 50	98
NE-IE	b	0.059	<.05*	50, 50	98
NE-L2	b	−0.043	.052	50, 50	98
IE-L2	b	0.102	<.001**	50, 50	98

*Statistically significant at $p < .05$; **Statistically significant at $p < .001$.

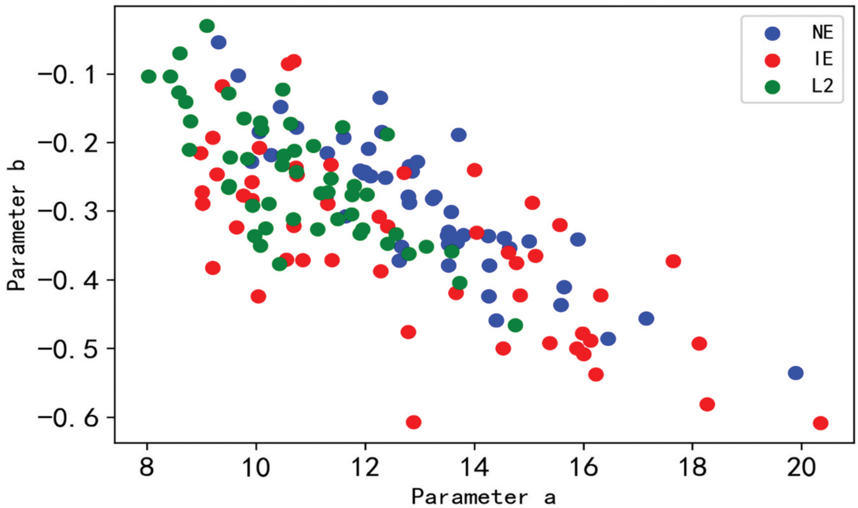

Figure 6. Scatterplot of NE, IE and, L2 texts' distribution through the *a* and *b* parameters of the MAL model.

indicating that parameters *a* and *b* effectively differentiate these varieties of English speech. The scatterplot also illustrates a clear negative correlation between parameters *a* and *b*, which aligns with Cramer's (2005) hypothesis of a systematic inverse relationship between these parameters. NE data points cluster towards higher values of parameter *a* and less negative values of parameter *b*. In contrast, IE data points show a broader distribution, with generally lower values of *b*. L2 data points tend to cluster around lower values of parameter *a* and higher (less negative) values of *b*. This pattern demonstrates that, while both parameters contribute to differentiating the varieties of English speech, parameter *b* exhibits greater sensitivity to group differences, further supporting the findings of Jiang and Ma (2020).

Building on the distribution patterns observed in Figure 6, we further conducted K-Means clustering to quantitatively assess the grouping of NE, IE, and L2 varieties. Figure 7 presents the results of the K-Means clustering analysis based on the structural parameters *a* and *b* of the MAL model, alongside a comparison between the K-Means cluster memberships and the original speech categories (NE, IE and, L2) in Figure 8. For the clustering analysis, the number of predefined clusters was set to three (k = 3), determined by the elbow method, which indicated that three clusters best captured the natural structure in the data. The clustering algorithm grouped the data points based on the values of parameters *a* and *b*, and the mean values for each cluster are reported in Table 5. Each cluster shows distinct characteristics. Cluster 0, with a mean *a* of 10.07 and mean *b* of −0.23, primarily aligns with L2 speech patterns, as its simplified syntactic structures (lower

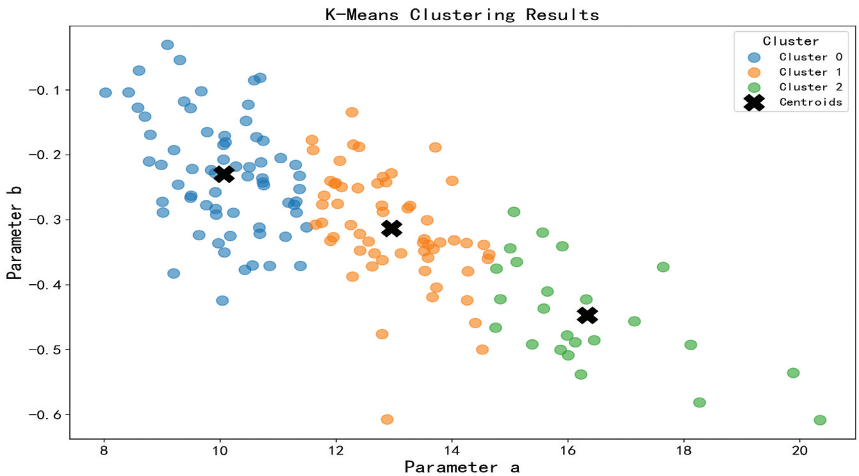

Figure 7. K-Means clustering results for parameters *a* and *b*.

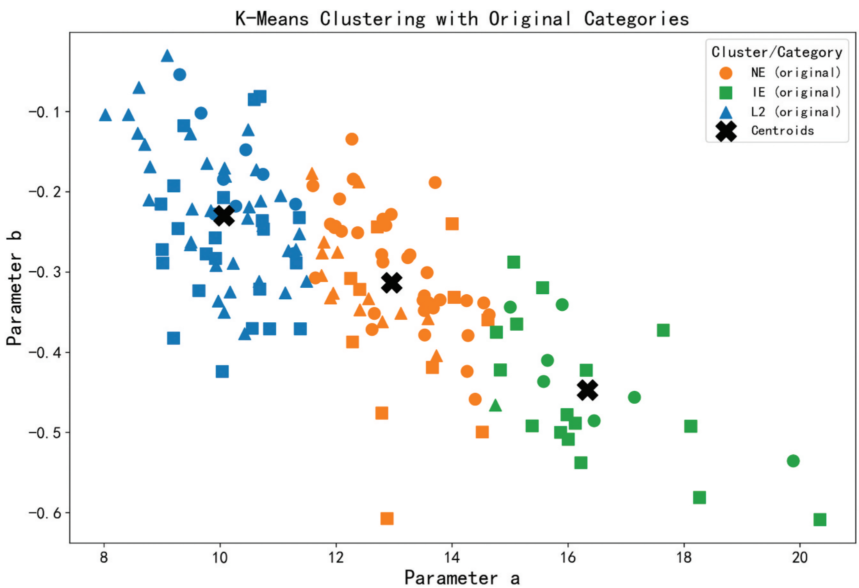

Figure 8. Comparison between K-Means clusters and original speech categories.

a values) and reduced compression rate (lower *b* values) are characteristic of second language production. Cluster 1, with a mean *a* of 12.96 and mean *b* of –0.31, exhibits characteristics more closely related to NE. Cluster 2, with a mean *a* of 16.33 and mean *b* of –0.45, is dominated by IE samples. The right panel in Figure 8 compares the clustering results with the original data categories (NE, IE, and L2). While there is some degree of overlap between

Table 5. Cluster statistics.

Cluster	Mean Parameter *a*	Mean Parameter *b*	NE Number of Members	IE Number of Members	L2 Number of Members
0	10.07	−0.23	9	23	35
1	12.96	−0.31	34	11	14
2	16.33	−0.45	7	16	1

the clustering results and original categories, the differences highlight the nuanced role of the parameters in distinguishing speech varieties. These results reinforce the effectiveness of parameter *b* in identifying distinct varieties of English speech, extending its applicability beyond register and authorship attribution (Mačutek et al., 2017).

5. Discussion

This study compares corpora of interpreted speech, second language speech, and native English speech to examine how both constrained and non-constrained varieties of English speech adhere to the MAL. By analysing the relationship between sentence length measured by the number of clauses and by the number of words, we aimed to investigate how these varieties conform to the MAL model. Our findings offer insights into pragmatic translation and demonstrate how cognitive load and translation features influence language structure.

5.1. Exploring the Validity of the MAL in NE, IE, and L2

The analytical results from this study demonstrate that the relationship between sentence length and clause length in all corpora adheres to the MAL. Specifically, as sentence length (measured by the number of clauses) increases, the average length of the constituent clauses (measured by the number of words) decreases, aligning with MAL predictions at the syntactic level. The sentences in Table 6 illustrates this phenomenon with five sentences containing one to five clauses, showing how the average clause length decreases as the number of clauses increases.

Table 6 shows the average clause length decreases from 14 to 6.4 as the number of clauses increases from one to five. These findings, consistent with those of Jiang and Jiang (2022) and Jiang and Ma (2020), confirm that the sentence-clause relationship follows the MAL. According to Köhler (1989), shortening the constituents of longer constructs is essential to balance the continuous flow of linguistic information with the efficient use of cognitive resources. Expanding both a linguistic unit and its components simultaneously could overwhelm the cognitive system (Seeber, 2013). The observed

Table 6. Decreasing clause length with increasing sentence complexity (taken from IE sub-corpus).

Sentence	Number of Clauses	Word Tokens	Average Clause Length (words)
1a) Now we have got some national leaders making a point focusing on this business.	1	14	14.0
1b) If the original lease was granted on policy considerations for promoting certain objectives, for example, development of an individual industry, whether the policy consideration is still valid.	2	27	13.5
1c) It is just because when the complaints were made to the police, they never set up for any investigation into these cases.	3	22	7.33
1d) So if they get more here and the parent company gets less, then we are going to suffer because you are extending the arrangement to onshore aircraft leasing activities.	4	29	7.25
1e) We feel very sad because the administration is sitting on hundreds of billions of dollars but there are grass-roots people who have no bargaining power and they are living in subdivided units.	5	32	6.4

sentence–clause interactions, particularly in the IE and L2 corpora, help prevent the formation of overly complex structures that could exceed human cognitive capacity (Li et al., 2023). Moreover, the phenomenon of increasing clauses and decreasing average clause length may reflect strategies to manage memory limitations and optimize linguistic processing. As Köhler (1989) suggests, the shortening mechanism arises from memory constraints: the longer the construct, the more space must be reserved for structural information between the constituents. Therefore, the size of the constituents is to be lowered.

Our results indicate that the coefficients of determination *(R^2)* for all three corpora – NE, IE, and L2 – are greater than 0.9, demonstrating a strong correlation between the observed data points and the theoretical values predicted by the MAL. This confirms that, across both constrained (IE, L2) and non-constrained (NE) varieties of English speech, as the number of clauses in a sentence increases, the average length of each clause shortens. This finding reflects the principle of cognitive economy (Fenk & Fenk-Oczlon, 1993). The IE corpus, in particular, shows a steeper decrease in clause length with increasing sentence length, indicated by a smaller value of parameter b. This suggests that clauses in IE are more strongly adjusted and reorganized to accommodate sentence complexity compared to NE and L2. In all three corpora, parameter b is negative, reinforcing that, on average, clause length decreases as sentence length increases. Overall, the validity of the MAL is confirmed across interpreted speech, second language speech, and native English speech at the sentence–clause level. Despite being considered 'an extreme case of translation' (Shlesinger & Ordan, 2012, p. 54) or a constrained language variety (Liu et al., 2023), interpreted speech, like native English speech, adheres to the principles of the MAL. This supports the findings of Jiang and Ma (2020) and Jiang and Jiang (2022), which show that, in natural languages, as sentences grow longer or contain more clauses, the average clause length shortens. These results address the first research question, confirming that the MAL applies universally across different varieties of English speech, including interpreted and native speech.

5.2. Differentiating NE, IE, and L2 through the MAL Parameters

The MAL offers a valuable framework for analysing how NE, IE, and L2 manage sentence and clause complexity. By examining parameters a and b, clear distinctions emerge between these varieties, reflecting differences in syntactic complexity and strategies for handling linguistic challenges. Among the three, NE stands out as the most syntactically complex, with the highest value of parameter a, indicating that native speakers tend to begin sentences with more elaborate and complex clauses. This reflects their full command of the language, enabling them to construct detailed and nuanced sentences. This finding aligns with Liu and Afzaal's (2021) research on the syntactic complexity of translated versus non-translated texts, revealing that translations are generally less complex than their non-translated counterparts, underscoring key differences in syntactic patterns between native speakers and L2 speakers. Native speakers tend to produce more structurally varied and intricate sentences, often employing longer clauses and a wider range of syntactic constructions. In contrast, L2 speakers may rely on shorter, simpler sentence structures due to the cognitive demands of managing two languages simultaneously (Xu & Liu, 2023) as reflected by the lower value of parameter a. Interpreters, despite their cognitive constraints, maintain near-native performance in terms of basic clause length, likely due to their high proficiency in English. Furthermore, the moderate rate of clause shortening, as indicated by parameter b, shows that native speakers preserve syntactic richness even as sentence complexity increases, adjusting clause length with ease. This ability highlights their cognitive flexibility, allowing them to construct more complex sentences without strain. On the other hand, L2 speakers often divide their attention between managing syntax, vocabulary, and other linguistic tasks, which limits their ability to produce the same level of syntactic complexity (Vercellotti, 2019). While their rate of clause shortening is closer to that of NE than to IE, L2 speech relies on simpler syntactic structures overall. The larger difference in parameter b between L2 and IE, compared to L2 and NE, suggests that L2 speakers are less able to compress clauses as effectively as interpreters when sentence complexity increases. IE shows the steepest rate of clause shortening, reflecting the high cognitive demands of real-time interpreting, which require interpreters to compress information quickly.

Both L2 and IE can be understood as constrained languages, a term used to describe varieties of English speech that operate under significant cognitive, social, or communicative limitations. As defined by Kruger and van Rooy (2016a), constrained language refers to the language produced in contexts characterized by particularly conspicuous constraints, such as bilingualism, translation, or proficiency limitations. Lanstyák and Heltai (2012) argue that these communicative settings impose similar cognitive and social constraints, often leading to reduced syntactic complexity and lower lexical diversity. Traditionally, this concept has been widely applied to translation studies (Chen et al., 2024; Kruger & van Rooy, 2016a; Kwok et al., 2023), but recent work by Liu et al. (2023) has expanded the framework to include interpreting. Their study highlights how cognitive constraints, such as the real-time processing demands of interpreting, also shape syntactic complexity in ways similar to those observed in translated and EFL varieties. Given their categorization as constrained languages, both L2 and IE exhibit common trends of structural adjustment compared to NE. Studies like those of Kajzer-Wietrzny and Ivaska (2020) highlight shared linguistic features between these varieties, such as reduced syntactic complexity and a tendency towards more frequent use of certain grammatical structures, including the optional complementizer ‘that’. These similarities are a result of the cognitive and linguistic challenges both L2 speakers and interpreters face (Kroll et al., 2018), whether from proficiency limitations in L2 or the real-time processing demands in IE. In both cases, speakers are required to simplify their sentence structures to ensure clarity and manage cognitive load, leading to more streamlined, less complex language compared to NE.

While L2 and IE are both categorized as constrained languages, the MAL analysis reveals deeper insights into how these varieties manage sentence complexity. Both L2 and IE tend to exhibit shorter clauses compared to NE. However, their approaches to managing complexity diverge considerably. In L2, sentence construction is relatively stable and consistent, primarily shaped by the constraints of language proficiency. L2 speakers tend to use shorter clauses from the start, as indicated by the lower value of parameter *a*, and maintain this level of simplification throughout, as reflected in the less negative value of parameter *b*. In contrast, IE shows a more dynamic response to cognitive pressures, with interpreters starting sentences with relatively complex clauses (similar to NE), but then rapidly shortening them as sentence complexity increases, as reflected by the steepest negative value of parameter *b*. This suggests that the approach to complexity is more reactive to the immediate cognitive demands of real-time speech processing, where interpreters need to compress and condense information quickly (Xu & Li, 2022). This rapid linguistic compression reflects interpreters’ automated strategies for managing high cognitive load, enabling them to balance comprehension and production in real time.

The differences in complexity management between L2 and IE can be attributed to the distinct cognitive constraints each variety faces. Kotze (2022) identifies five dimensions of constraints that affect language production: language activation, modality and register, text production, proficiency, and task expertise. In L2, complexity management is primarily influenced by proficiency and language activation constraints, where speakers often lack the full linguistic resources to handle complex syntactic structures. This leads to more consistent management of complexity across different sentence types. On the other hand, in IE, the modality and task expertise involved in interpreting play a larger role. Interpreters must process speech in real time while producing output, which imposes unique cognitive demands, leading to more aggressive clause shortening as a strategy to manage this cognitive load. This phenomenon aligns with the cognitive load approach proposed by Seeber (2011), who attributes the high cognitive load in simultaneous interpreting to the real-time combination of comprehension and production. These differences highlight the importance of understanding the specific constraints that shape each language variety rather than treating them both as similar examples of constrained language. The oral characteristics of simultaneous interpretation reflect the increasing cognitive load on translators or the competition for cognitive resources (Song & Li, 2021). The application of the MAL framework to both L2 and IE allows for a deeper understanding of the cognitive load underlying these constrained languages. Without the insights provided by the MAL, the differences in how L2 and IE handle sentence complexity might remain hidden. In particular, the rapid clause shortening in IE reveals the heightened cognitive effort required for real-time interpretation, compared to the steadier simplification seen in L2. This analysis suggests that simplification in constrained languages is not a monolithic process but rather a multidimensional one, shaped by the specific demands of each communicative context (Kruger & van Rooy, 2016a). The idea that language use operates along a continuum, as proposed by Grosjean (2001), further supports this analysis, positioning L2 and IE at different points based on their communicative contexts. The MAL analysis helps uncover the distinct ways in which each language variety adapts to its specific constraints, offering new insights into how cognitive pressure influences language production.

6. Conclusions

This study confirmed the applicability of the MAL to both constrained languages (interpreted speech and L2 speech) and non-constrained native English speech, demonstrating that longer linguistic structures tend to have shorter components. By analysing the fitting parameters *a* and *b*, the study revealed critical differences and similarities in how these varieties of English speech manage syntactic complexity. Both L2 and interpreted speech simplify linguistic features under cognitive load, but the distinct nature of their constraints results in differing patterns. In terms of sentence construction, interpreters, who possess high proficiency in English, typically begin with complex sentence structures but rapidly reduce complexity as sentences progress, reflecting the high cognitive demands of interpreting. In contrast, L2 speakers generally employ simpler sentence structures from the outset, which likely stem from limitations in their language proficiency. The use of the MAL offers new perspectives on language processing that might remain hidden with traditional linguistic measures. The rapid decline in clause length in interpreted speech, for example, reflects interpreters' automated strategies for managing complex sentences, which other indicators might not reveal as clearly. In contrast, L2 speakers show a more gradual reduction in sentence length, highlighting their more deliberate approach to processing complexity.

This study is not without limitations. It focuses on a specific Cantonese-speaking context, which may limit the generalizability of the findings to other linguistic environments. Additionally, individual differences in proficiency and experience among speakers and interpreters were not considered, which may influence the variability of parameters *a* and *b*. Future research should explore the application of the MAL to a wider range of language pairs and interpreting contexts to determine whether these findings hold across different linguistic and cultural settings. Additionally, examining individual differences in interpreter expertise and L2 proficiency could shed light on how personal factors influence the syntactic patterns revealed by the MAL. Another promising avenue for research would be to investigate the correlations between lexical and syntactic complexity measures and MAL parameters. Such analyses would deepen our understanding of how different linguistic features interact with memory constraints during simultaneous interpretation.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

This study was supported by two grants from The Hong Kong Polytechnic University [Project Nos: P0051009; P0043847].

ORCID

Ziqiao Su http://orcid.org/0009-0003-6959-3115
Han Xu http://orcid.org/0000-0001-5493-6857
Kanglong Liu http://orcid.org/0000-0003-3962-2563

Data Availability Statement

Corpus data concerning the study are publicly available on Open Science Framework (https://osf.io/4qwvb/).

References

Altmann, G. (1980). Prolegomena to Menzerath's law. In R. Grotjahn (Ed.), *Glottometrica* (Vol. 2, pp.1–10). Brockmeyer.

Altmann, G. (2014). Bibliography: Menzerath's law. *Glottotheory*, *5*(1), 121–123. https://doi.org/10.1515/glot-2014-0008

Andres, J., Benešová, M., Chvosteková, M., & Fišerová, E. (2014). Optimization of parameters in the Menzerath-Altmann law, II. *Acta Universitatis Palackianae Olomucensis, Facultas Rerum Naturalium, Mathematica*, *53*(2), 5–28.

Andres, J., Benešová, M., Kubáček, L., & Vrbková, J. (2012). Methodological note on the fractal analysis of texts. *Journal of Quantitative Linguistics*, *19*(1), 1–31. https://doi.org/10.1080/09296174.2011.608604

Anthony, L. (2022). *TagAnt (Version 2.0.5)* [Computer software]. Waseda University. https://www.laurenceanthony.net/software/TagAnt

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honor of John Sinclair* (pp. 233–250). John Benjamins.

Chen, H., & Liu, H. (2022). Approaching language levels and registers in written Chinese with the Menzerath-Altmann law. *Digital Scholarship in the Humanities*, *37*(4), 934–948. https://doi.org/10.1093/llc/fqab110

Chen, J., Li, D., & Liu, K. (2024). Unraveling cognitive constraints in constrained languages: A comparative study of syntactic complexity in translated, EFL, and native varieties. *Language Sciences*, *102*, 101612. https://doi.org/10.1016/j.langsci.2024.101612

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114. https://doi.org/10.1017/S0140525X01003922

Cramer, I. (2005). The parameters of the Altmann-Menzerath law. *Journal of Quantitative Linguistics*, *12*(1), 41–52. https://doi.org/10.1080/09296170500055301

Dębowski, Ł. (2007). Menzerath's law for the smallest grammars. In P. Grzybek & R. Köhler (Eds.), *Exact methods in the study of language and text* (pp. 77–86). De Gruyter Mouton.

Eroglu, S. (2013). Menzerath-Altmann law for distinct word distribution analysis in a large text. *Physica A: Statistical Mechanics and Its Applications*, *392*(12), 2775–2780. https://doi.org/10.1016/j.physa.2013.02.012

Eroglu, S. (2014). Menzerath-Altmann law: Statistical mechanical interpretation as applied to a linguistic organization. *Journal of Statistical Physics*, *157*(2), 392–405. https://doi.org/10.1007/s10955-014-1078-8

Favaro, L., Gamba, M., Cresta, E., Fumagalli, E., Bandoli, F., Pilenga, C., Isaja, V., Mathevon, N., & Reby, D. (2020). Do penguins' vocal sequences conform to linguistic laws? *Biology Letters*, *16*(2), 20190589. https://doi.org/10.1098/rsbl.2019.0589

Fenk, A., & Fenk-Oczlon, G. (1993). Menzerath's law and the constant flow of linguistic information. In R. Köhler & B. B. Rieger (Eds.), *Contributions to quantitative linguistics* (pp. 11–31). Kluwer Academic Publishers.

Grosjean, F. (2001). The bilingual's language modes. In J. Nichol (Ed.), *One mind, two languages: Bilingual language processing* (pp. 1–22). Blackwell Publishers.

Gustison, M. L., Semple, S., Ferrer-I-Cancho, R., & Bergman, T. J. (2016). Gelada vocal sequences follow Menzerath's linguistic law. *Proceedings of the National Academy of Sciences*, *113*(19), E2750–E2758. https://doi.org/10.1073/pnas.1522072113

Hou, R., Huang, C. R., Do, H. S., & Liu, H. (2017). A study on correlation between Chinese sentence and constituting clauses based on the Menzerath-Altmann law. *Journal of Quantitative Linguistics*, *24*(4), 350–366. https://doi.org/10.1080/09296174.2017.1314411

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62. https://doi.org/10.1016/j.cogpsych.2010.02.002

Jia, H., & Liang, J. (2020). Lexical category bias across interpreting types: Implications for synergy between cognitive constraints and language representations. *Lingua*, *239*, 102809. https://doi.org/10.1016/j.lingua.2020.102809

Jiang, X., & Jiang, Y. (2022). Menzerath-Altmann law in consecutive and simultaneous interpreting: Insights into varied cognitive processes and load. *Journal of Quantitative Linguistics*, *29*(4), 541–559. https://doi.org/10.1080/09296174.2022.2027657

Jiang, Y., & Ma, R. (2020). Does Menzerath-Altmann law hold true for translational language: Evidence from translated English literary texts. *Journal of Quantitative Linguistics*, *29*(1), 37–61. https://doi.org/10.1080/09296174.2020.1766335

Jin, H., & Liu, H. (2017). How will text size influence the length of its linguistic constituents? *Poznan Studies in Contemporary Linguistics*, *53*(2), 197–225. https://doi.org/10.1515/psicl-2017-0008

Kajzer-Wietrzny, M. (2015). Simplification in interpreting and translation. *Across Languages and Cultures*, *16*(2), 233–255. https://doi.org/10.1556/084.2015.16.2.5

Kajzer-Wietrzny, M. (2022). An intermodal approach to cohesion in constrained and unconstrained language. *Target*, *34*(1), 130–162. https://doi.org/10.1075/target.19186.kaj

Kajzer-Wietrzny, M., & Ivaska, I. (2020). A multivariate approach to lexical diversity in constrained language. *Across Languages and Cultures*, *21*(2), 169–194. https://doi.org/10.1556/084.2020.00011

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

Köhler, R. (1989). Das Menzerathsche Gesetz als Resultat des Sprachverarbeitungsmechanismus. In G. Altmann & H. Schwibbe (Eds.), *Das Menzerathsche gesetz in informationsverarbeitenden systemen* (pp. 108–116). Georg Olms Verlag.

Köhler, R. (2012). *Quantitative syntax analysis*. De Gruyter Mouton. https://doi.org/10.1515/9783110272925

Köhler, R., Kelih, E., & Goebl, H. (2021). Gabriel Altmann (1931–2020). *Journal of Quantitative Linguistics*, *28*(2), 187–193. https://doi.org/10.1080/09296174.2021.1902057

Kotze, H. (2022). Translation as constrained communication: Principles, concepts, and methods. In S. Granger & M. Lefer (Eds.), *Extending the scope of corpus-based translation studies* (pp. 67–97). Bloomsbury Academic.

Kroll, J. F., Dussias, P. E., & Bajo, M. T. (2018). Language use across international contexts: Shaping the minds of L2 speakers. *Annual Review of Applied Linguistics*, *38*, 60–79. https://doi.org/10.1017/S0267190518000119

Kruger, H., & van Rooy, B. (2016a). Constrained language: A multidimensional analysis of translated English and a non-native indigenised variety of English. *English World-Wide*, *37*(1), 26–57. https://doi.org/10.1075/eww.37.1.02kru

Kruger, H., & van Rooy, B. (2016b). Syntactic and pragmatic transfer effects in reported-speech constructions in three contact varieties of English influenced by Afrikaans. *Language Sciences*, *56*, 118–131. https://doi.org/10.1016/j.langsci.2016.04.003

Kwok, H. L., Laviosa, S., & Liu, K. (2023). Lexical simplification in learner translation: A corpus-based approach. *Research in Corpus Linguistics*, *11*(2), 103–124. https://doi.org/10.32714/ricl.11.02.06

Lanstyák, I., & Heltai, P. (2012). Universals in language contact and translation. *Across Languages and Cultures*, *13*(1), 99–121. https://doi.org/10.1556/Acr.13.2012.1.6

Li, Y., Gao, Y., & Lu, X. (2023). Effects of word limit on sentence length and clause length in academic journal article abstracts: A synergetic linguistic perspective. *Journal of Quantitative Linguistics*, *30*(3–4), 322–342. https://doi.org/10.1080/09296174.2023.2263249

Liu, K., & Afzaal, M. (2021). Syntactic complexity in translated and non-translated texts: A corpus-based study of simplification. *PLOS ONE*, *16*(6), e0253454. https://doi.org/10.1371/journal.pone.0253454

Liu, K., Liu, Z., & Lei, L. (2022). Simplification in translated Chinese: An entropy-based approach. *Lingua*, *275*, 103364. https://doi.org/10.1016/j.lingua.2022.103364

Liu, Y., Cheung, A. K. F., & Liu, K. (2023). Syntactic complexity of interpreted, L2 and L1 speech: A constrained language perspective. *Lingua*, *286*, 103509. https://doi.org/10.1016/j.lingua.2023.103509

Mačutek, J., Čech, R., & Milička, J. (2017). Menzerath-Altmann law in syntactic dependency structure. In S. Montemagni & J. Nivre (Eds.). *Proceedings of the Fourth International Conference on Dependency Linguistics* (pp. 100–107). Linköping University Electronic Press.

Mačutek, J., Chromý, J., & Koščová, M. (2019). Menzerath-Altmann law and prothetic/v/in spoken Czech. *Journal of Quantitative Linguistics*, *26*(1), 66–80. https://doi.org/10.1080/09296174.2018.1424493

Mačutek, J., & Mikros, G. K. (2015). Menzerath-Altmann law for word length motifs. In G. Mikros & J. Mačutek (Eds.), *Sequences in language and text* (pp. 125–132). De Gruyter Mouton.

Menzerath, P. (1954). *Die Architektonik des deutschen Wortschatzes* [The architecture of German vocabulary]. Dümmler.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81. https://doi.org/10.1037/h0043158

Schusterová, D., Motalová, T., & Wang, F. (2017). Menzerath-Altmann dìng lǜ: Lǐ lùn gòng xiàn yǔ jú xiàn [Menzerath-Altmann law: Contributions and limitations of the theory]. *Linguistic Research*, *1*, 45–58.
Seeber, K. G. (2011). Cognitive load in simultaneous interpreting: Existing theories-new models. *Interpreting*, *13*(2), 176–204. https://doi.org/10.1075/intp.13.2.02see
Seeber, K. G. (2013). Cognitive load in simultaneous interpreting: Measures and methods. *Target*, *25*(1), 18–32. https://doi.org/10.1075/target.25.1.03see
Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
Shlesinger, M., & Ordan, N. (2012). More spoken or more translated? Exploring a known unknown of simultaneous interpreting. *Target*, *24*(1), 43–60. https://doi.org/10.1075/target.24.1.04shl
Song, S., & Li, D. (2021). Orality features of simultaneous interpreting and its implications for interpreting training: A case of TV interpreting. *Foreign Language Education*, *42*(2), 86–91. https://doi.org/10.16362/j.cnki.cn61-1023/h.2021.02.017
Teupenhayn, R., & Altmann, G. (1984). Clause length and Menzerath's law. In J. Boy & R. Köhler (Eds.), *Glottometrika* (Vol. 6, pp. 127–138). Brockmeyer.
Vercellotti, M. L. (2019). Finding variation: Assessing the development of syntactic complexity in ESL speech. *International Journal of Applied Linguistics*, *29*(2), 233–247. https://doi.org/10.1111/ijal.12225
Wang, Y., & Chen, H. (2022). The Menzerath-Altmann law on the clause level in English texts. *Linguistics Vanguard*, *8*(1), 331–346. https://doi.org/10.1515/lingvan-2022-0048
Wang, Z., Cheung, A. K., & Liu, K. (2024). Entropy-based syntactic tree analysis for text classification: A novel approach to distinguishing between original and translated Chinese texts. *Digital Scholarship in the Humanities*, *39*(3), 984–1000. https://doi.org/10.1093/llc/fqae030
Xu, C., & Li, D. (2022). Exploring genre variation and simplification in interpreted language from comparable and intermodal perspectives. *Babel*, *68*(5), 742–770. https://doi.org/10.1075/babel.00289.cui
Xu, H., & Liu, K. (2023). Syntactic simplification in interpreted English: Dependency distance and direction measures. *Lingua*, *294*, 103607. https://doi.org/10.1016/j.lingua.2023.103607
Xu, L., & He, L. (2020). Is the Menzerath-Altmann law specific to certain languages in certain registers? *Journal of Quantitative Linguistics*, *27*(3), 187–203. https://doi.org/10.1080/09296174.2018.1532158
Yin, F., & van Schijndel, M. (2023). Linguistic compression in single-sentence human-written summaries. In H. Bouamor, J. Pino & K. Bali (Eds.), *Conference on Empirical Methods in Natural Language Processing, Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 7922–7935). Association for Computational Linguistics.
Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley.