2024; 118(3): 51–69
https://doi.org/10.58221/mosp.v118i3.15574

Moderna Språk

Research Article

Distribution of sentence length of English complex sentences

Jinlu Liu
Zhejiang University of Finance and Economics

Nan Yang
Zhejiang University of Finance and Economics

Haitao Liu*
Fudan University

Received August, 2023; accepted July, 2024;
published online December, 2024

Abstract: In previous studies based on many languages, the distributions of sentence length fit several distribution models. Moreover, those research findings are based on a mixture of all kinds of sentences, which constitute the most complex syntactic units. How is the distribution of sentence length of English complex sentences manifested individually? To answer this question, with the aid of Altmann-Fitter software (2013), we analyzed and compared the distribution of sentence length of English complex sentences comprehensively, judging by Brown and LOB corpus, the three research findings were obtained. Firstly, the frequency distributions of sentence length of English complex sentences well follow the Extended Positive Negative Binomial distribution; secondly, text type or genre could have a significant effect on the distribution of sentence length of English complex sentences; thirdly, there are no any significant differences in the distributions of sentence length of complex sentences between British and American English. The above research findings suggest that human language is a probabilistic system by nature.

Keywords: English complex sentence, sentence length, distribution, Brown corpus, LOB corpus

1 Introduction

There exist many grammatical units in human languages, of which the biggest grammatical unit is sentence (Quirk et al., 1985, p. 47). Consequently, sentence length, which is defined by the number of words included in the sentence, has been attracting much interest in linguistic studies (Köhler, 2012). Longer sentences, which will take longer time and more effort to pronounce, write and read, are more inconvenient to process and less economical than short sentences (Sigurd et al., 2004, p. 47). This perhaps means that sentence length is a crucial factor influencing language understanding. Perkins et al. (1986, p. 139) used elicited imitation tasks to show that sentence length tends to be positively correlated with the difficulty of the sentence repetition task. Yan et al. (2016, p. 522) also claimed that as the sentence becomes longer, the level of cognitive pressure for elicited imitation tasks increases. In other words, adding words to a sentence could increase the load on immediate memory, increasing the difficulty in an imitation task. As sentence length increases, correct imitation will decrease (Miller, 1973, p. 1–

***Corresponding author:** Haitao Liu, E-mail: htliu@163.com

2). It has been shown that sentence length was the strongest predictor of difficulty of elicited imitation tasks (Miller, 1956b, p. 133; Perkins et al., 1986; Yan et al., 2016).

Liu (2018, p. 149) claimed that language is a human-driven complex adaptive system, the length of a sentence is not arbitrary. Its use will be restricted strictly by human nature. That is to say, there will be a certain relationship between the use of sentence and the patterns of human behavior and cognition. For instance, in some written works, the distribution of sentence lengths is known to depend on the style of an author (Sichel, 1974, p. 25; Grzybek, 2002; Wu & Li, 2022; Haverals et al., 2022). In addition, according to Goldsmith, sentence length is also considered a reliable stylistic marker (Mannion & Dixon, 2004, p. 497). Some studies have shown that different lengths of sentences can serve as a basis for classifying different genres of registers (Chen & Liu, 2022). This evidence may inform the regularity of sentence length. Best (2002, p. 136) argued if sentence length is a variable, its different values will exist in texts in certain proportions and a very frequently observed model is the Hyper-poisson distribution. Popescu et al. (2014) also found that the length distribution of many linguistic units well fit the same model, that is, Zipf-Alekseev function, which is consistent with the principle of least effort (Zipf 1949, p. 1). In addition, Yu et al. (2021) also stated that the frequency distribution of sentences follows a general pattern, which is formed by basic cognitive mechanisms. Fenk-Oczlon and Pilz (2021) found that larger phoneme inventories correlate with shorter words and clauses, and languages with more speakers have more phonemes per syllable, shorter words, more monosyllabic words, and more words per clause, consistent with Zipf's law. Similarly, dependency distance involving syntactic complexity and understanding difficulty will also be affected by sentence length (Ferrer-i-Cancho & Liu, 2014, p. 143; Jiang & Liu, 2015, p. 94).

Previous studies on sentence length have encompassed a variety of languages, most of which focus on the relationship between sentences and words or sentences and texts, underscoring the exploration of language universals. To delve into these universals, a single language study in linguistics is never enough. Extensive testing is the only possibility of finding a common and more stable background (Popescu et al., 2014, p. 111). For example, 398 Slovenian texts from different genres, 152 Slovenian texts, and 333 Slovenian texts are analyzed respectively with regard to their sentence length and word length to show that both factors play an important role in text classification (Grzybek et al., 2005, p. 53; Antić et al., 2006, p. 117; Kelih et al., 2006, p. 382). Besides, 117 samples of German literary prose texts written by 52 authors were analyzed to conclude that an increase in sentence length goes along with an increase in word length (Grzybek & Stadlober, 2007, p. 205). On the basis of 199 Russian texts, there seems to be no strong relationship between sentence length and word length, which is related to the inter-textual perspective (Grzybek et al., 2007, p. 617). Moreover, the sentence lengths of 77 Hindi texts (Pande & Dhami, 2015, p. 338) and 113 Japanese texts were also investigated (Ishida & Ishida, 2007, p. 28). The two studies revealed that the sentence length distribution in Hindi aligns more closely with the Extended Positive Negative Binomial (EPNB) model, whereas the distribution in Japanese is more consistent with the Hyper-pascal model. As for sentence length in English, previous research found an almost perfect fit of a variant of gama distribution for a corpus consisting of different text genres (Sigurd et al., 2004, p. 37). In addition, their research data showed that the formula can be used to distinguish between different kinds of text genres.

A study by Miller (1956a, p. 96) showed that, from a cognitive perspective, sentence length should be within a certain range, and he ever stated that the number of chunks (such as letters, words, numbers, etc.) that one can hold in short-term memory is 7 ± 2. In other words, cognitive limitations prevent sentences from being expanded indefinitely in length. Parenthetically, Sperling (1960, p. 6) claimed that the limit of immediate verbal (auditory) memory is 4 ± 1. Perkins et al. (1986) suggested that the length of the sentences be set at seven to eight syllables.

All the same, Naiman (1974) chose sentences of 15 syllables for first- and second-grade second language learners and considered the length appropriate. However, these studies were all based on a mix of several sentence types. Deng et al. (2021, p. 1) pointed out that the inconsistencies in many syntactic related research results may be due to the insufficient accuracy of sentence classification, suggesting the possibility of differences between different sentence types.

Sentence can be categorized into three types: simple sentence, compound sentence, and complex sentence. From a linear perspective, the syntactic complexity of these three sentence types increases sequentially, with complex sentences often being regarded as the most syntactically complex type (Diessel, 2004). Complex sentences are also the most important means of expressing conditionality, opposition, comparison, simultaneity, sequence and other syntactic relations (Tskhovrebov & Shamonina, 2023). A complex sentence refers to a sentence form in which a main clause is followed by one or more subordinate clauses (Quirk et al., 1985, p. 987; Diessel, 2004, p. 1; Burton-Roberts, 2011, p. 171; Owens, 2016, p. 397; Lastres-López, 2020, p. 50). In this regard, complex sentences not only contain more intricate syntactic elements but also tend to be longer compared to other types of sentences. From the perspective of maximum threshold of sentence length, research on sentence length of complex sentences will become particularly necessary. Based on the above points, our present study focuses on exploring the sentence length of English complex sentences from both a macroscopic and microcosmic perspective. We will answer three questions discussed in the following.

Previous studies indicate that the sentence length distribution across various languages adheres to either the Extended Positive Negative Binomial (EPNB) distribution or the Hyper-pascal distribution, with English sentence length distribution aligning with the EPNB. However, there is a gap in research regarding the distribution of sentence lengths in English complex sentences. Thus, our research question (RQ1) is: How is the sentence length of English complex sentences distributed in the use of language?

From a more detailed point of view, different genres may differ in language use. Complex sentences, with their intricate structure and capacity to convey more information than simple and compound sentences, what are the differences and similarities between the length of complex sentences in different genres of texts? Additionally, American and British English represent the two most recognized English varieties all over the world, despite the many similarities between British English and American English, which can be attributed to their distinct language environments, there are also notable differences (Davies, 2005; Baker, 2017: 236). The points discussed above give rise to our RQ2 and RQ3 for this study: Is the distribution of sentence length of English complex sentences influenced by different genres? Considering the distribution of sentence length of English complex sentences, does British English differ from American English?

2 Research materials and method

2.1 Language materials

The use of a corpus often brings to light surprises and usages that would likely be overlooked if the investigator were relying solely or chiefly on introspection (Rudanko, 2011, p. 2). Therefore, in light of our research objectives, we chose the Brown corpus representing American English (Francis, 1965, p. 267) and the LOB corpus standing for British English (Johansson et al., 1978). The Brown corpus refers to the standard corpus of edited present-day American English, compiled by W. Nelson Francis and Henry Kučera of Brown University in

the 1960s. Like its American counterpart, the LOB corpus, namely the Lancaster-Oslo/Bergen Corpus of British English was directed by Geoffrey N. Leech and Stig Johansson at the University of Lancaster (1970-1976) and the University of Oslo (1977-1978) and designed to match the Brown corpus in size closely, text category and composition in the 1970s. As for their size, each corpus is composed of 500 text samples of about 2,000 words each, including roughly a million words per corpus. Although the data from these two corpora may be out-of-date, they still have significant research value. Firstly, these two corpora can help researchers gain a deeper understanding of language development and evolution trends of. In addition, due to the extensive use and research of the Brown and LOB corpora for many years, they have always been the standard works of English balanced corpora (Feng, 2002). Therefore, their research results have high stability and reliability and irreplaceably important value in linguistic and academic research. Therefore, despite their small scale, these two corpora are still important reference resources for studying language learning, language change, grammar structure, and other aspects. Therefore, in response to the research questions of this study, the Brown corpus and LOB corpus can still be effectively utilized for research.[1]

Traditionally, English complex sentences are divided into two basic types: the first includes the sentences containing coordinate clauses, the second includes the sentences containing subordinate clauses. In fact, this classification method is consistent with the defined multiple sentences, that is, in addition to simple sentences, combine compound sentences with complex sentences into one category (Quirk et al., 1985, p. 719). The first type consists of two or more clauses functionally equivalent and symmetrical, while the other one consists of two or more clauses constituting an asymmetrical relationship (Diessel, 2004, p. 43). Specifically, it should be pointed out that in the process of our annotating research materials in two corpora, the complex sentences are only limited to those sentences containing subordinate clauses and all types of subordinate clauses are finite. In short, if a sentence contains a finite subordinate clause, we will list it as a complex sentence.

In current research, the number of subordinate clauses contained in a complex sentence directly determines the density of subordinate clauses in that complex sentence. For example, the complex sentence “Dr Fortran says if I exercise my leg more, maybe I can use a cane when I’m big” includes a noun clause with an omitted conjunction and two adverbial clauses, therefore, the clause density is 3. In addition, in sentences with direct quotations, such as complex sentences like “The way you were careful?”, he snorted.”, the direct quotation part is the object of the verb “snorted”. Therefore, we consider it as an object clause, so the overall clause density of this sentence is 2.

In addition, according to the criteria proposed by Karlsson (2007, p. 110), several clauses in one complex sentence sometimes have a situation where one clause contains another, and there is a dependency relationship between the clauses, forming an embedding chain (embedding chain or abbreviated as e-chain) in complex sentences, also known as embedding depth. The minimum value of this embedding depth is 1, which means when there is only one clause in a

1 The Brown corpus and the LOB corpus have a total of 105180 single sentences, each consisting of 15 text files. Each text file represents a language style and is represented by a letter between A and R, namely A, B, C, D, E, F, G, H, J, K, L, M, N, P, R. One, for details, please refer to the following two official website links:
Brown corpus: http://icame.uib.no/brown/bcm.html
LOB corpus: http://www.helsinki.fi/varieng/CoRD/corpora/LOB/basic.html
In addition, all text files have been automatically coded using the "CLAWS5" part of speech annotation set, and the label list for this annotation set is: http://ucrel.lancs.ac.uk/claws5tags.html.
The end of the sentence has already been syntactically annotated using the “_SENT” method, so it is directly used as the “feature substring for automatic sentence segmentation”. To avoid garbled characters, the entire set of text files has been batch converted to the UTF-8 characters encoding format.

complex sentence, or when there are two or more clauses that are not dependent on each other. For example, here are two examples from the Brown corpus. The sentence "That's what he said" only contains 1 noun clause, and the embedding depth of the clause in this complex sentence is 1; "The route which he had traveled and which he believed had developed into a trade route was followed by his setters earlier than he had expected." This complex sentence contains three subordinate clauses, namely two relative clauses and a comparative adverbial clause, but these three subordinate clauses do not form interdependence, so the embedding depth of the subordinate clause in this complex sentence is still 1.

We invited 12 high school English teachers to annotate the corpora manually, and they were from Tai'an No.19 Senior High School, Wucheng County No.2 Senior High School in Dezhou City, Meiqu County Senior High School in Meizhou City, and Zhejiang Wuxing High School. All of them have certain teaching experience and good knowledge of English grammar. The corpus consists of approximately 2 million words of text, with a total of 42,655 complex sentences and approximately 160,000 words annotated by each teacher. To ensure the validity of the annotation, we require all teachers involved in artificial syntactic annotation to uniformly adhere to a consistent set of syntactic annotation guidelines (Quirk et al., 1985, p. 985-1044; Hudson, 1998, p. 61-85). Meanwhile, the teachers responsible for syntactic annotation will receive corresponding compensation for each complex sentence annotated artificially. Considering the reliability of the syntactic annotation, during the process of automatic sentence segmentation, we executed programming codes to randomly duplicate a portion of the entire set of English complex sentences. For example, "*This is the book my teacher bought for me.*" is a complex sentence introduced by an omitted relative pronoun. When we segment the sentence, we duplicate it twice, but the duplicated sentences are not placed next to each other. As a result, the complex sentence will be annotated twice by the teachers responsible for syntactic annotation. After we receive the annotated sentences, we will compare the duplicated annotated sentences during the review process to see if the annotations are consistent. If they are consistent, we will select one to include in the corpus; if they are inconsistent, we will choose the correct one to include in the corpus. The examples will be shown in Table 1.

Table 1

Examples of English sentence types

Number	Types of clauses contained	Is it a complex sentence	English Sentences
1	nominal clause	Yes	You see what I mean.
2	relative clause	Yes	That was all she said.
3	adverbial clause	Yes	It ended when he tumbled.
4	non-finite clauses	No	It is necessary to do it like this.

In the process of annotating complex sentences, relative clauses, noun clauses, and adverbial clauses are sequentially marked as "a", "b", and "c". The specific format for annotating complex sentences is shown in Table 2 below.

Table 2

Examples of English complex sentence annotation

Number	Clause Density	Embedding Depth	Clause Identification	Complex Sentences
1	1	1	b	You see what I mean.
2	2	1	ab	All I know is that they are gone.
3	2	2	bb	You thought I did not mean what I said.
4	3	2	bbc	I replied that I hoped it would not, unless he ran the way of trouble.
5	4	3	cbcb	While I was drinking it, I wondered what Peter Rakosi would say when I told him I wanted to marry his daughter.
6	7	4	cbacbac	While they were told that there were some normal people who reacted differently than they had, they were also informed that there were other normals who reacted as they had.

2.2 Research method

As for the distribution model of sentence length, Wimmer & Altmann (1999) found that the Extended Positive Negative Binomial (EPNB) and Hyper-pascal are two distributions which give a good fit (in one displaced form). In addition, Pande & Dhami (2015, p. 346) pointed out that the Extended Positive Negative Binomial distribution (k, p; α fixed) in one displaced form could be considered as an appropriate distribution for the grouped frequency distribution of sentences of different lengths. EPNB and Hyper-pascal distributions are variants of the negative binomial distribution, designed for modeling data that exhibit clustering and long-tail characteristics. The EPNB enhances model flexibility with additional parameters, while the Hyper-pascal offers alternative parameterizations. Both distribution models effectively address over-dispersion in data and are well-suited for analyzing complex data distributions such as language unit lengths.

For this reason, we make all research samples tested or fitted by Altmann-Fitter (2013), a quantitative linguistics software package. By referring to coefficient of determination, R^2 and two parameters of EPNB, k and p to decide whether the degree of fit is good or not and which distribution model the sentence length of complex sentence will follow well.

3 Results and discussion

By analyzing the distribution of sentence length of English complex sentences in the Brown and LOB corpora, this study reveals the statistical characteristics of the distribution of sentence length of English complex sentences. The following sections discuss our key findings in detail

and explore the reasons behind these findings. In Section 3.1, we analyze RQ1, find a model suitable for the distribution of sentence length of English complex sentences, and further explore whether the sentence length distribution conforms to Zipf's law. In Section 3.2, we address RQ2 by analyzing the distribution of sentence length of English complex sentences across different genres and find some commonalities and characteristics. In Section 3.3, we compare the distribution of sentence length of complex sentences between British English and American English for RQ3 and find some common rules in the language.

3.1 Distribution of sentence length of English complex sentences

In previous studies, Best (2002, p. 136) argued that those different values of sentence length would exist in texts in certain proportions. A very frequently observed model is the Hyper-pascal distribution. However, the other studies claimed that the Extended Positive Negative Binomial Distribution (EPNB Distribution) in one displaced form could be considered as an appropriate distribution for the grouped frequency distribution of sentences of different lengths (Wimmer & Altmann, 1999; Pande & Dhami, 2015, p. 346). These two distribution models are known for their 'unimodal' shape, which reaches a peak at a small value and then declines as the value increases. These mentioned findings suggest that most sentences are concentrated in some medium length, with sentences that are very short or very long being less common. In contrast to the Hyper-pascal model, the EPNB model demonstrates superior performance when dealing with data exhibiting long tail characteristics.

In order to figure out whether sentence length of English complex sentences also follows these distribution models, we fitted our data to Altmann-Fitter (2013). The fitting results are presented in Table 3 and the concrete distribution is shown in Figure 1.

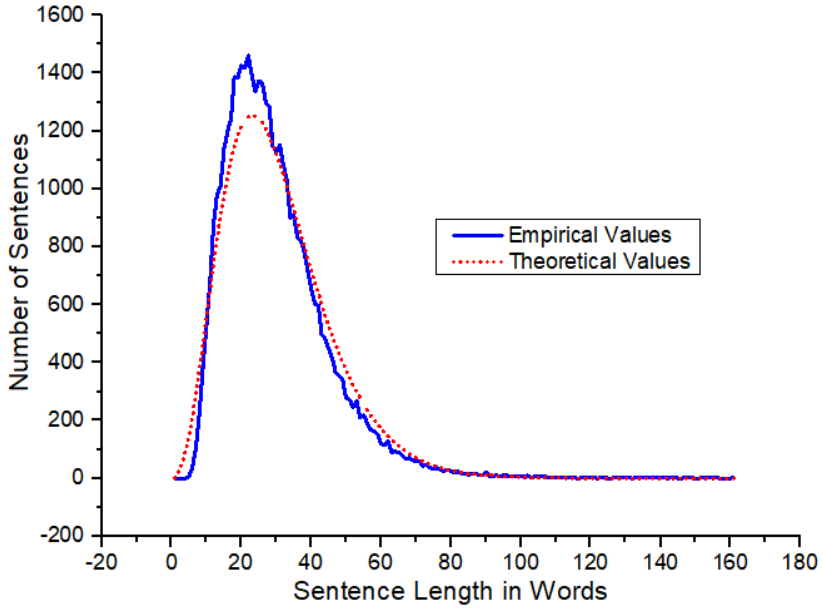
Table 3

Fitting of sentence length of complex sentences to EPNB

Distribution Model	R^2	k	p	α
EPNB Distribution	0.9781	4.3775	0.1295	1.0000

Figure 1

Frequency distribution of sentence length.

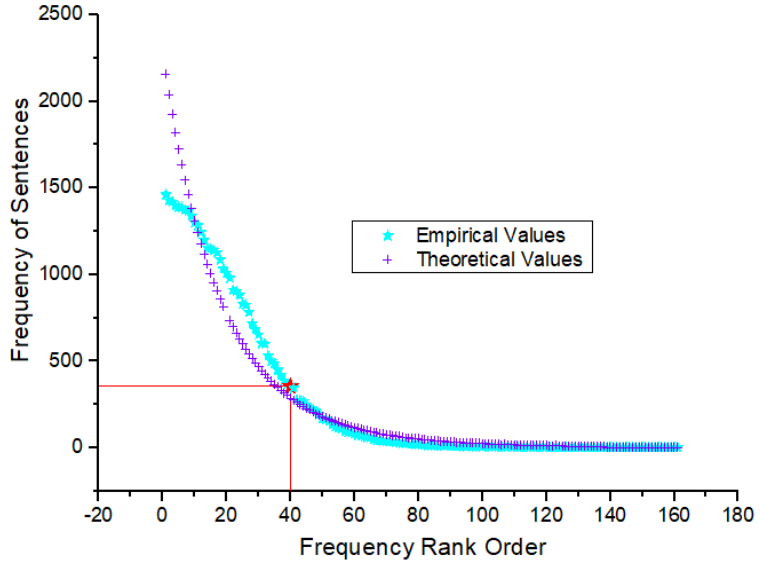


This result in Table 3 means that the frequency distribution of sentence lengths of English complex sentences follows EPNB Distribution well (R^2=0.9781＞0.9), while the Hyper-pascal distribution is an inappropriate distribution model. From Figure 1, it can be seen that the fitting curves for sentence length and number of sentences have a high degree of overlap, and the trend is also similar. When the sentence length is 22, the number of sentences is the highest. This implies that during language use, people have a propensity to employ complex sentences that are around 22 words. The long tail in Figure 1 also illustrates the potential for longer sentences in actual language use, while the occurrence of extremely short sentences is quite rare. It is worth mentioning that although the empirical value at the peak is different from the theoretical value, it indicates that the number of sentences in the corpus exceeds the theoretical value. This may be due to the specific field, style, or writing habits of the corpus, and does not affect the degree of fitting.

In addition, Popescu et al. (2014) found that the length distribution of many linguistic units fit the Zipf-Alekseev function, which is consistent with the principle of least effort (Zipf, 1949, p. 1). As for the frequency of sentences in a text, is it a power-law function of frequency rank order of its length? Our present study shows that it is quite true and it fits Zipf-Mandelbrot function (a = 12.000, b = 212.1910, R^2 = 0.9139), as shown in Figure 2.

Figure 2

Rank distributions of sentence length.



We can see much more clearly with the help of the auxiliary lines that the proportion of sentences listed in the top 40 (161 in all) in sentence length is 89.30%, that is, the ratio of 38,091 to 42,655. Of the top 40, almost all sentences are under 40 in length, as shown in Figure 3 and the length of the most used sentences is 22, whose total number is 1,460.

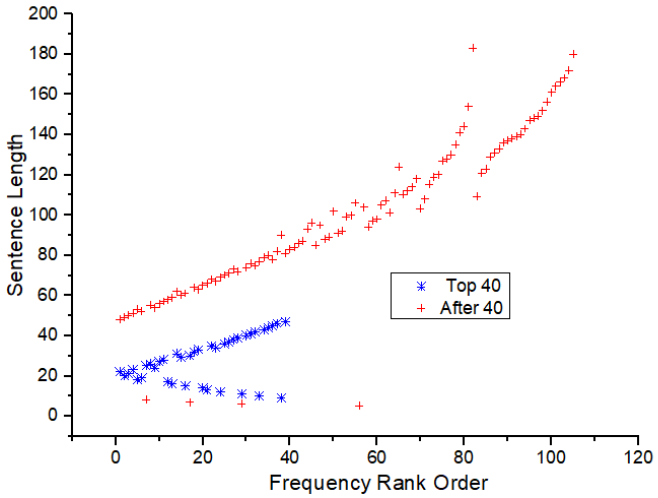
Through the above data analysis, we should emphasize that this kind of distribution of sentence frequency follows a linguistic universal, shaped by fundamental common cognitive mechanisms and reflecting overall tendency in language use (Yu et al., 2021).

Understanding the syntactic difficulty of a sentence can be measured by a lot of metrics, of which dependency distance involving syntactic complexity and understanding difficulty will also be affected by sentence length (Ferrer-i-Cancho & Liu, 2014, p. 143; Jiang & Liu, 2015, p. 94). Generally speaking, the longer a sentence is, the greater its dependency distance will be. After all, human language is a human-driven complex adaptive system (Liu, 2018, p. 149). Previously, scholars have found in their research on the Dependency Distance Minimization (DDM) a problem in natural language that human cognitive mechanisms lead people to tend to use sentences with smaller dependency distances language, making sentences less complex (Liu, 2008; Futrell et al., 2015; Ferrer-i-Cancho et al., 2022). Therefore, the universality of human language is determined by the universality of human cognition to some extent. Consequently, it is extremely natural that we often use those shorter complex sentences more.

Due to the proportion of sentences listed in the top 40 of sentence length being as high as 89.30%, in order to further discover, we divided the rank into below 40 and above 40, and made the following Figure 3.

Figure 3

Distribution of sentence length ranking top 40 and the rest.



From Figure 3, it can be seen that when the sentence length is 0-20, the longer the sentence length, the more sentences there are; when the sentence length is 20-40 and longer than 40, the longer the sentence length, the fewer sentences there are. This may be due to people constantly making dynamic adjustments while using language, following the principle of least effort.

The frequency of words in a text is a power-law function of its frequency rank order with an exponent around -1 (Zipf, 1932, 1949). From the fitting results and the degree of curve overlap, in terms of our present study, a complex sentence consisting of many words is also a power-law function of its frequency rank order, with an exponent around -1. The smaller the length of complex sentences, the higher their frequency of occurrence. The longer the length of a complex sentence, the lower its frequency of occurrence. That is to say, the longer the length of a complex sentence, the fewer times people use this construction.Similar to the distribution of word frequency, dependency distance, and the whole sentence length, the distribution of complex sentence lengths also adheres to Zipf's law, which indicates that humans neither construct sentences entirely at random nor strictly by rule, but rather show a preference for sentences of certain lengths and structures. These distributions reflect the probabilistic nature of language at the syntactic level, suggesting that human language is likely to be a probabilistic system in essence.

3.2 Effect of text types on sentence length of English complex sentences

In language, genres are differentiated and identifiable text types (Purcell-Gates et al., 2007, p. 11) and different genres may mean that the use of language varies. According to Wang's (2020) research, most syntactic features exhibit significant differences in different language genres between English and Chinese, such as dependency direction and dependency genre, making the ability to identify these features an effective criterion for classifying language genres in different languages. However, some other indicators exhibits same or similar rules in certain language laws, reflecting the universality of human language. For example, the sentence length rank frequency distribution of all language genres follows the same probability distribution, and the distribution of dependency distance follows a long tail distribution, which conforms to

the trend of "dependency distance minimization" (Liu, 2008; Futrell et al., 2015; Ferrer-i-Cancho et al., 2022). And in the previous section we proved that the distribution of complex sentence lengths, like other language units, conforms to certain general laws, yet these language units may vary across different text genres. Therefore, we are curious whether the distribution of complex sentences will also exhibit the same pattern in different text genres. Or can it serve as an indicator to distinguish different text genres? That is to say, from the perspective of the difference in text types, what general and specific characteristics does the sentence length of English complex sentences reflect? To this end, we draw the distributions of sentence length of 15 different text types in all, as shown in Figure 4. Through the Altman Fitter, the distributions of sentence length of English complex sentences were fitted by the Extended Positive Negative Binomial distribution, and the fitting results are shown in Table 4.

Figure 4

Distributions of sentence length of 15 text types.

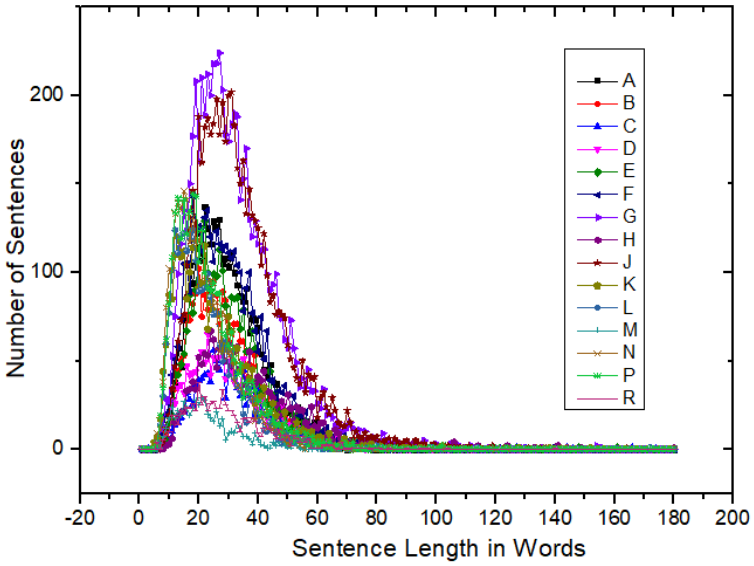


Table 4

Fitting of sentence length of complex sentences

Text types	k	p	c	R^2
A-Press: reportage	6.1646	0.1805	0.0387	0.9792
B-Press: editorial	5.4298	0.1654	0.0493	0.9705
C-Press: reviews	4.1181	0.1326	0.2135	0.8304
D-Religion	4.0284	0.1253	0.0804	0.9403
E-Skills, trades and hobbies	5.2525	0.1561	0.0969	0.9429
F-Popular lore	5.0735	0.1519	0.0495	0.9716
G-Belles lettres, biography, essays	4.8755	0.1344	0.0416	0.9760

H-Miscellaneous	3.9978	0.1049	0.1052	0.9201
J-Learned and scientific writings	5.0431	0.1326	0.0532	0.9664
K-General fiction	3.6972	0.1362	0.0854	0.9453
L-Mystery and detective fiction	4.9784	0.1919	0.0945	0.9511
M-Science fiction	3.5186	0.1401	0.2156	0.8794
N-Adventure and western fiction	4.9000	0.1937	0.0723	0.9522
P-Romance and love story	4.4190	0.1663	0.0782	0.9389
R-Humor	3.0814	0.1131	0.2032	0.8209

From Figure 4, it can be seen that the overall sentence length distribution of the 15 text types is very similar, with peaks appearing between 20-45. This indicates that regardless of the type of text, people use language in accordance with the principle of least effort and tend to use fewer sentences that are too short or too long. This is due to the limitations of working memory. However, it can also be seen from the figure that the peak values of texts of different genres are different, which may be related to the reading difficulty of different text types. We will discuss this later.

As Table 4 indicates, the all the mean values of R^2 show that the distributions of sentence length of English complex sentences of all genres are well captured by the distribution ($R^2 > 0.82$). These findings possibly also show that language is a self-regulating system, which features invariant entities, namely certain language laws or regularity (Köhler & Altmann, 1986, p. 254).

The following Figure 5 (The X-axis represents sentence length and the Y-axis represents frequency of sentences) displays the fitting results of sentence length distribution of each genre.

Figure 5

Fitting Extended Positive Negative Binomial distribution to sentence length of different text types.

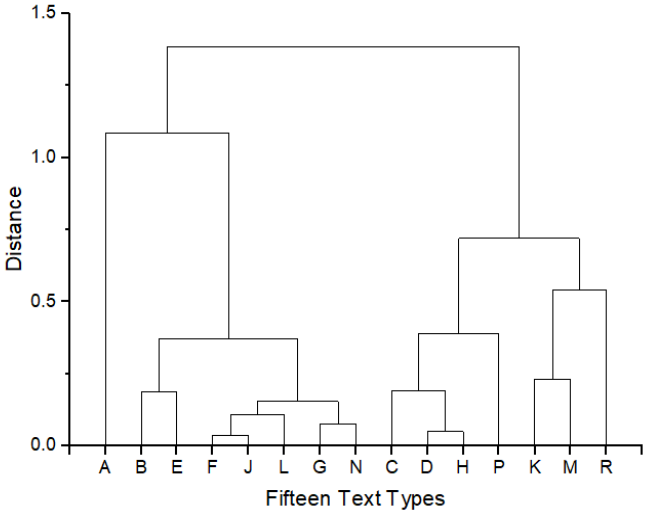
In combination with the Table 4 analysis, it is evident that there is an extremely similar distribution trend of sentence length of complex sentences across all text genres, but there are differences in the peaks and the degree of the curve. In order to figure out the differences among

the fifteen genres, a one-way analysis of variance (ANOVA) test was conducted on the mean values of frequency of sentence length for different genres. The result of the ANOVA test is significant ($F(14, 2225) = 14.486$, $P = .000$, $\eta^2 = 0.084$). This test result shows that the variation of sentence length distribution reaches a significant level, that is, the differences among different genres are revealed.

In order to further explore the differences between single text type or genre, we conducted hierarchical clustering analysis. This method groups data points based on their similarities, creating a hierarchical structure that allows us to see how different texts are related. By doing so, we obtained the language cluster; it results from adopting parameter k, p, c, and R^2 of fitting of sentence length of complex sentences in 15 different text types, which is shown in Figure 6.

Figure 6

Cluster analysis based on parameter k, p, c and R^2.



Interestingly, the ANOVA and cluster analyses consistently display the difference between A (Press: reportage) and R (Humor) ($P = 0.003$). Obviously, the former belongs to a most formal or serious text style, while the latter the most informal or entertaining text style. These differences are enough to reflect a fact that text type or significantly affect the distribution of sentence length of English complex sentences significantly. In addition, it can be seen from the figure that the connection between the C (Press: reviews) and H (Miscellaneous) is relatively close. This may be due to the fact that news and religion belong to more formal genres, and the use of language is also more formal, resulting in longer sentence lengths. Although K (General fiction), L (Mystery and detective fiction), M (Science fiction), and N (Adventure and western fiction) all belong to the category of novels, they show certain differences in sentence length. The sentence lengths of M (Science fiction) and K (General fiction) are relatively similar, while those of N (Adventure and western fiction) and G (Belles lettres, biography, essays) are similar. Wang (2020) found that, as a whole, the sentence lengths of online and novel language styles are closer, and detective novels are the most difficult type of novel to read, so they are closer to J (Learned and scientific writings), which is also difficult to read. And the more relaxed and lively language style like P (Romance and love story) is much different from other types of novels. This not only proves that different genres may influence sentence length, but even in the same genre of "novel", there may be differences due to genres. Biber's research (1986, p.

407) suggests that novels differ from all other types of oral/writing production due to their strong preference for reporting events, rather than providing explanatory statements about actual events or information. Scott (1988, p. 59) once pointed out that adult language is woven by intricate semantic relationships, which may lead to people producing different styles of articles and sentences based on different types of novels.

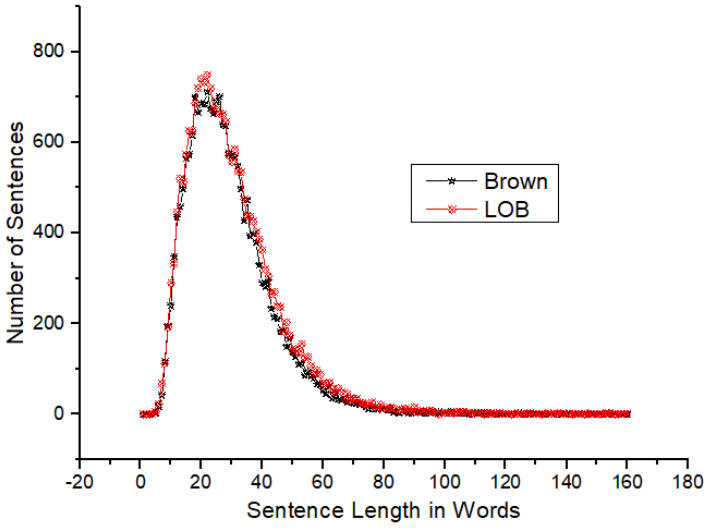
3.3 The comparison of sentence length of complex sentences between British English and American English

Generally speaking, there are both similarities and differences between British English and American English. The differences involve grammar, vocabulary (Bock et al., 2006, p. 64), spelling (Baker, 2017, p. 236), punctuation (Algeo, 2006, p. 2; Carrie & McKenzie, 2018, p. 313), and idioms, etc. However, the related research on the use of syntactic structure, e. g. the issues related to complex sentences between them, is relatively insufficient. Here, we draw the distributions of sentence length of British and American English to investigate their comparison of sentence length of complex sentences between the two varieties, as shown in Figure 7.

From Figure 7, which shows the distributions of sentence length of complex sentences in Brown and LOB corpus, it appears that the two bending lines almost overlap. To put it simply, there seem to be similar distribution trends in aspect of sentence length of complex sentences. Correspondingly, after the independent sample t-test of the sentence length of complex sentences in two corpora, the results show that there is no significant difference on the whole ($t(356) = 0.375$, $P = 0.708$, $d = 0.04$) between American English ($M = 114.94$, $SD = 208.31$) and British English ($M = 123.35$, $SD = 215.37$).

Figure 7

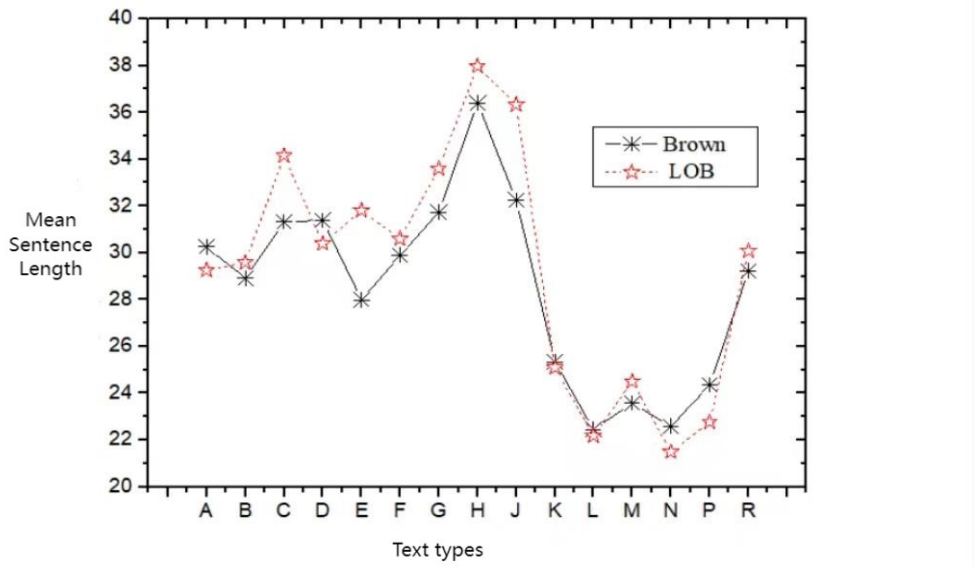
Distributions of sentence length of British and American English.



Meanwhile, from the perspective of average sentence length of complex sentences of fifteen genres in Brown and LOB corpus, we draw their line graphs, as shown in Figure 8.

Figure 8

Comparison of the distribution of mean sentence length between British English and American English in different text styles.



It is apparent that their overall changing trend is consistent and there exists no significant difference in distribution of average sentence length on the whole (t (28) = 0.479, P = 0.636, d = 0.17) As previously discussed, it is acknowledged that sentence length can impact the cognitive load experienced by individuals during language processing. Our study reveals a high degree of similarity between British English and American English in terms of distribution of sentence length, indicating that the cognitive demands related to sentence length are consistent across these two varieties. Therefore, regardless of the language variety, the usage of language remains within the bounds of human cognitive capabilities.

Language information is only stored temporarily in short-term memory, so rote repetition is possible only if the sentences are short, or repetition will be labored or ineffective if the sentence length exceeds the capacity of short-term memory (Yan et al., 2016, p. 508). Although British English and American English display differences in vocabulary, grammar, and other linguistic aspects, which can be attributed to cultural, political, and other external influences, the remarkable consistency in the use of complex sentences between these two varieties highlights the pervasive cognitive constraints that shape language use across linguistic variants. This indicates that, despite surface differences, the underlying cognitive mechanisms governing sentence structure are universally applicable.

4 Conclusion

In order to explore the distribution of sentence length, previous studies based on many languages found the distributions of sentence length fitting several distribution models. As mentioned, those research findings were based on the mixture of all kinds of sentences. Individually, since complex sentences are the most complex syntactic units, it is essantial to explore the concrete distribution of sentence length in English.

Liu (2018, p. 149) claimed that language is a human-driven complex adaptive system. In this sense, the length of a sentence should not be arbitrary and will be restricted strictly by human

nature. Therefore, there will be a certain relationship between sentence length and human individual style. For example, in some written works, the distribution of sentence lengths depends on the author's characteristics (Sichel, 1974, p. 25). For Goldsmith, sentence length is also considered a reliable stylistic marker. These pieces of evidence may indicate the regularity of sentence length.

With the help of Altmann-Fitter (2013) and Brown and LOB corpus, we have analyzed and compared the distribution of sentence length of English complex sentences comprehensively. In response to RQ1, we found that the distributions of sentences length of English complex sentences fit the Extended Positive Negative Binomial distribution ($R^2 = 0.9781$), this finding supports the applicability of the EPNB model in capturing the statistical properties of different languages. In response to RQ2, our analysis revealed that text type or genre significantly influences the distribution of sentence lengths in English complex sentences. The use of complex sentences is adapted to the specific demands and conventions of each genre, which may relate to differences in reading difficulty and stylistic preferences. As for RQ3, there are no significant differences in the distributions of sentence length of complex sentences between British and American English. This similarity suggests that these two varieties of English adhere to similar cognitive constraints concerning the usage complex sentences, despite other lexical, grammatical, and spelling differences.

The distribution of sentence length of English complex sentences follows an EPNB distribution and a power law distribution, suggesting that the occurrence of sentences of different lengths is not random, but follows a predictable pattern, with some sentence lengths being very common and many others being rare. Concurrently, the constraints of human working memory imply that speakers prefer shorter sentences because they are less cognitively demanding. However, this preference is not absolute but probabilistic, simply indicating that a certain sentence length is more likely to occur. Our findings therefore suggest that human language is a probabilistic system by nature, which may also show the universal feature of human language to some extent. And such a pattern of sentence distribution of complex sentences is perhaps molded by the common human cognition mechanism under the restriction of principle of least effort.

Our study primarily focused on the distribution of sentence length in English complex sentences but did not differentiate between various types of subordinate clauses, representing a clear limitation of our research. According to Deng et al. (2021), different types of subordinate clauses may exert distinct influences on syntactic studies. Therefore, future research should account for the diversity of clause types to gain a more precise understanding of complex sentences. Additionally, our study was confined to the English language. To bolster the universality of our findings, we advocate for future studies to encompass other languages, allowing for a comparative analysis of the universality and specificity of sentence length distribution across different languages. Such cross-linguistic research will be instrumental in revealing universal patterns in sentence length distribution and offering novel insights into the commonalities and differences in linguistic structures.

Acknowledgement

Funding: This article is supported by Zhejiang Provincial Philosophy and Social Science Planning Project (22JCXK12YB).

References

Algeo, J. (2006). *British or American English?* Cambridge University Press.

Altmann-Fitter. (2013). *Altmann-Fitter user guide. The third version*. https://www.ram-verlag.eu/wp-content/uploads/2013/08/Fitter-User-Guide.pdf

Antić, G., Kelih, E., & Grzybek, P. (2006). Zero-syllable words in determining word length. In P. Grzybek (Ed.), *Contributions to the science of text and language. Word length studies and related issues* (pp. 117–156). Springer.

Baker, P. (2017). *American and British English: divided by a common language?* Cambridge University Press.

Best, K. (2002). The distribution of rhythmic units in German short prose. *Glottometrics, 3*, 136–142.

Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language, 62*(2), 384–414. https://doi.org/10.2307/414678

Bock, K., Cutler, A., Eberhard, K. M., Buttefield, S., & Humphreys, K. R. (2006). Number agreement in British and American English: Disagreeing to agree collectively. *Language, 82*(1), 64–113.

Burton-Roberts, N. (2011). *Analysing sentences: An introduction to English syntax*. Longman.

Carrie, E., & McKenzie, R. M. (2018). American or British? L2 speakers' recognition and evaluations of accent features in English. *Journal of Multilingual and Multicultural Development, 39*(4), 313–328.

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

Davies, C. (2005). *Divided by a common language: A guide to British and American English*. Houghton Mifflin Company.

Deng, Y., Lei, L., & Liu, D. (2021). Calling for more consistency, refinement, and critical consideration in the use of syntactic complexity measures for writing. *Applied Linguistics, 42*(5), 1021–1028. https://doi.org/10.1093/applin/amz069

Diessel, H. (2004). *The acquisition of complex sentences*. Cambridge University Press.

Feng, Z. (2002). Evolution and present situation of corpus research in China. *Journal of Chinese Language and Computing, 11*(2), 127–136.

Fenk-Oczlon, G., & Pilz, J. (2021). Linguistic complexity: Relationships between phoneme inventory size, syllable complexity, word and clause length, and population size. *Frontiers in Communication, 6,* 1–7.

Ferrer-i-Cancho, R., & Liu, H. (2014). The risks of mixing dependency lengths from sequences of different length. *Glottotheory, 5*(2), 143–155.

Ferrer-i-Cancho, R., Gómez-Rodríguez, C., Esteban, J. L., & Alemany-Puig, L. (2022). Optimality of syntactic dependency distances. *Physical Review. E, 105*(1), 014308. https://doi.org/10.1103/PhysRevE.105.014308

Francis, N. (1965). A standard corpus of edited present-day American English. *College English*, *26*(4), 267-273.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences, 112*(33), 10336–10341. https://doi.org/10.1073/pnas.1502134112

Grzybek, P. (2002) Quantitative aspekte slawischer texte (am Beispiel von Pukins Evgenij Onegin). *Wiener Slavistisches Jahrbuch, 48,* 21–36.

Grzybek, P., & Stadlober, E. (2007). Do we have problems with Arens' law? A new look at the sentence-word relation. In P. Grzybek & R. Köhler (Eds.), *Exact methods in the study of language and text* (pp. 205–218). Mouton de Gruyter.

Grzybek, P., Stadlober, E., & Kelih, E. (2007). The relationship of word length and sentence length. The inter-textual perspective. In R. Decker & H. J. Lenz (Eds.), *Advances in data analysis. Studies in classification, data analysis, and knowledge organization* (pp. 611–618). Springer.

Grzybek, P., Stadlober, E., Kelih, E., & Antić, G. (2005). Quantitative text typology: The impact of word length. In C. Weihs & W. Gaul (Eds.), *Classification - the ubiquitous challenge* (pp. 53–64). Springer.

Haverals, W., Geybels, L., & Joosen, V. (2022). A style for every age: A stylometric inquiry into crosswriters for children, adolescents and adults. *Language and Literature: International Journal of Stylistics, 31*(1), 62–84.

Hudson, R. (1998). *English Grammar*. Routledge.

Ishida, M., & Ishida, K. (2007). On distributions of sentence lengths in Japanese writing. *Glottometrics, 15*, 28-44.

Jiang, J., & Liu, H. (2015). The effects of sentence length on dependency distance, dependency direction and the implications - Based on a parallel English-Chinese dependency treebank. *Language Sciences, 50*, 93–104.

Johansson, S., Leech, G. N., & Goodluck, H. (1978). *Manual of Information to Accompany the Lancaster-Oslo/Bergen corpus of British English*. University of Oslo.

Karlsson, F. (2007). Constraints on multiple initial embedding of clauses. *International Journal of Corpus Linguistics, 12*(1), 107–118. https://doi.org/10.1075/ijcl.12.1.07kar.

Kelih, E., Grzybek, P., Antić, G., & Stadlober, E. (2006). Quantitative text typology: The impact of sentence length. In M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger & W. Gaul (Eds.), *From data and information analysis to knowledge engineering* (pp. 382–389). Springer.

Köhler, R. (2012). *Quantitative syntax analysis*. Walter de Gruyter.

Köhler, R., & Altmann, G. (1986). Synergetische aspekte der linguistik. *Zeitschrift Für Sprachwissenschaft, 5*(2), 253–265.

Lastres-López, C. (2020). Subordination and insubordination in contemporary spoken English: *If*-clauses as a case in point. *English Today, 36*(2), 48–52. https://doi.org/10.1017/S026607841900021X

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science, 9*(2), 159–191.

Liu, H. (2018). Language as a human-driven complex adaptive system. *Physics of Life Reviews,26–27*, 149–151. https://doi.org/10.1016/j.plrev.2018.06.006

Mannion, D., & Dixon, P. (2004). Sentence-length and authorship attribution: The case of Oliver Goldsmith. *Literary and Linguistic Computing, 19*(4), 497–508. https://doi.org/10.1093/llc/19.4.497

Miller, G. A. (1956a). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81–97. https://doi.org/10.1037/h0043158

Miller, G. A. (1956b). Human memory and the storage of information. *IRE Transactions on Information Theory, 2*(3), 129–137.

Miller, J. F. (1973). Sentence imitation in pre-school children. *Language and Speech, 16*(1), 1–14. https://doi.org/10.1177/002383097301600101

Naiman, N. (1974). The use of elicited imitation in second language acquisition research. *Working Papers in Bilingualism, 3*, 1–37.

Owens, R. (2016). *Language development: An introduction (9th Edition)*. Pearson Education Limited.

Pande, H., & Dhami, H. S. (2015). Determination of the distribution of sentence length frequencies for Hindi language texts and utilization of sentence length frequency profiles for authorship attribution. *Journal of Quantitative Linguistics, 22*(4), 338–348. https://doi.org/10.1080/09296174.2015.1106269

Perkins, K., Brutten, S. R., & Angelis, P. J. (1986). Derivational complexity and item difficulty in a sentence repetition task. *Language Learning, 36*(2), 125–141. https://doi.org/10.1111/j.1467-1770.1986.tb00375.x

Popescu, I., Best, K. H., & Altmann, G. (2014). *Unified modeling of length in language*. RAM-Verlag.

Purcell-Gates, V., Duke, N. K., & Martineau, J. A. (2007). Learning to read and write genre-specific text: Roles of authentic experience and explicit teaching. *Reading Research Quarterly, 42*(1), 8–45. https://doi.org/10.1598/RRQ.42.1.1

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman.

Rudanko, J. (2011). *Changes in complementation in British and American English: Corpus-Based studies on non-finite complements in recent English*. Palgrave Macmillan UK.

Scott, C. M. (1988). Producing complex sentences. *Topics in Language Disorders, 8*(2), 44–62. https://doi.org/10.1097/00011363-198803000-00006

Sichel. H. S. (1974). On a distribution representing sentence-length in written prose. *Journal of the Royal Statistical Society Series A: Statistics in Society, 137*(1), 25–34. https://doi.org/10.2307/2345142

Sigurd, B., Eeg-Olofsson, M., & Van de Weijer, J. (2004). Word length, sentence length and frequency-Zipf revisited. *Studia Linguistica, 58*(1), 37–52. https://doi.org/10.1111/j.0039-3193.2004.00109.x

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs, 74*(11), 1–29. https://doi.org/10.1037/h0093759

Tskhovrebov, A. S., & Shamonina, G. N. (2023). Syntactic features of Russian speech of two generations of bilinguals and monolinguals: A complex sentence. *Russian Language Studies, 21*(3), 293–305.

Wang, Y. (2020). Quantitative syntactic features of genres from multi-perspectives. [Doctoral thesis, Zhejiang University]. https://doi.org/10.27461/d.cnki.gzjdx.2020.000254

Wimmer, G. & Altmann, G. (1999). *Thesaurus of univariate discrete probability distributions.* Stamm Verlag.

Wu, K., & Li, D. (2022). Are translated Chinese Wuxia fiction and western heroic literature similar? A stylometric analysis based on stylistic panoramas. *Digital Scholarship in the Humanities, 37*(4), 1376–1393.

Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing, 33*(4), 497–528. https://doi.org/10.1177/0265532215594643

Yu, S., Xu C., & Liu, H. (2021). Statistical patterns of word frequency suggesting the probabilistic nature of human languages. http://arxiv.org/abs/2012.00187

Zipf, G. K. (1932). *Selected studies of the principle of relative frequency in language.* Harvard University Press.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley.