Linguistic Variation and Standardization of English as a Global Lingua Franca in a Multicultural Context

Yalin Wang *
School of International Relations, National University of Defense Technology, Nanjing, 210039, Jiangsu, China
wangyalin19990529@163.com

Abstract: Under the multicultural context, there is always the phenomenon that languages use multiple forms to express the same or similar meanings, among which there are even some usages that deviate from the traditional linguistic norms, and this phenomenon of covariation is the basis and prerequisite of ephemeral change. This paper mainly adopts the research methods of the Modern Chinese Corpus and the AmE Brown Family Corpora Corpus, and explores the features of language variation and standardization based on dependent syntactic analysis in a multicultural context. The intrinsic motivation of language variation and the constraint mechanism of variation are explored, the intrinsic law of language variation is discovered, and the cognitive sociolinguistic msodel of language variation is revealed. It is found that the process of language variation can be described by an S-shaped curve, with variation starting very slowly and increasing faster after it has increased to a certain level (around 20%). After the proportion of new forms reaches about 80%, the growth rate slows down significantly until it reaches the norm.
Keywords: Multiculturalism; Corpus; Dependent Syntax; Language Variation; S-curve

1. INTRODUCTION

English is one of the major international languages in the world today, not only a widely used communication tool in the society, but also a popular and popular international teaching language. Language is a communication tool needed in daily life and has social attributes, so social factors will inevitably affect the development of language and make language variation. Language variation refers to the linguistic differences produced by language users when using spoken or written language due to the different language communication contexts, which are mainly manifested in the aspects of phonetics, words, sentence structure, and stylistics (Eckert, 2017; Kidd & Donnelly, 2020). Phonology is the external form of language, which is prone to ambiguity or misunderstanding caused by pronunciation in daily communication (Fishman, 2020; García & Otheguy, 2017; Kim et al., 2018). Vocabulary is the basis of language, and lexical compilation is to transform and change the original meaning of English words before putting them back into use (Lohndal et al., 2019;

Rosa & Flores, 2017). And grammatical variation refers to violating grammatical rules and utilizing ungrammatical structures, including omitted structures, collocational variation, punctuation variation, and order variation, and other methods, in order to achieve the effect of focus (Bomer, 2017; Gleason & Ratner, 2022). Semantic variation mainly refers to the words and word collocation does not conform to the semantic content and logical laws, so that the meaning expressed by the utterance deviates from the conventional, or even violates the conventional is a variant method of thinking mode (De Villiers & De Villiers, 2017; Kornexl, 2017; Milroy & Milroy, 2017). As an international language, its reliability and feasibility must be fully maintained. The standardization of English is not only a reliable pillar of English as an international language for teaching and learning, but also the soul of all English written expressions and creative literary terms (Jenkins & Leung, 2019; Lee, 2017). Violation of this linguistic system and standards will inevitably lead to confusion in language communication (Akujobi & Ebere, 2022). New standards will certainly emerge, but the already existing scientific rules of the English language are bound to continue. Literature (Mauranen et al., 2020) shows that academic English is showing a multimodal way of combining texts due to the use of Internet technology, which is characterized by linguistic variation beyond lexico-grammatical correctness in different writing cultures. Literature (Mair & Leech, 2020) compares linguistic data from digital corpora in order to carry out research on syntactic variation in English, and argues that the spread of linguistic innovations is the result of a combination of internal and external factors. Literature (Dragojevic, 2017) proposes the concept of language attitudes, including standard and non-standard language variants, and analyzes the new reflection of the evaluation of language attitudes in terms of two dimensions: status and solidarity, aiming to study the impact of language change on the social environment in different contexts. Literature (Kidd et al., 2018) argues that there are significant differences between individuals in the process of language acquisition and language processing, which is the result of the joint influence of self-cognitive abilities and complex environments, and finds that the experience of individuals plays an important role in the variation of the language system. Literature (Greenhill et al., 2017) investigated the dynamic change process of the language system and found that grammatical structures contain more signals from the language system than basic vocabulary and this form deepens over time, but the rate of change of grammatical structures is faster. Literature (Utami, 2022) compared the differences between Vietnamese and American English accents in terms of characterization of

suprasegmental segments, acoustic changes, and correspondences, and found differences in the number of consonants between the two. For the first time, the study makes a more comprehensive and systematic research on the variability and standardization of English, and makes a comprehensive discussion on the two classifications of linguistic variation, the rules of variation, the motives of variation, and the constraining mechanisms. In the multicultural context, the nature and characteristics of English prototypes are analyzed with the help of Corpus Word Parser corpus annotation tool. With the help of the Modern Chinese Corpus and the actual corpus of AmE Brown Family Corpora, various aspects of English variation are elaborated in detail. The structural laws of English lexical and syntactic relations are also elucidated based on the dependent syntax theory and the syntactically annotated corpus, providing a new perspective for the study of English variation. The internal characteristics and cognitive motives of English variation are revealed at the micro level.

2. STUDY DESIGN

Based on dependent syntax, this paper examines the linguistic variation and standardization features of English in a multicultural context, and explores the effects of clause type and dependency position on the dependency bias and processing difficulty of Chinese-English relational clauses. In this research context, this paper will present the research questions for the current research status, followed by the introduction of the research corpus, and finally the introduction of the measurement indexes, research tools and research steps.

2.1 Research Theory

In order to explore the motivation behind the variation of English as a global lingua franca in a multicultural context, the first step is to reveal how idiomatic variation is generated and analyze what factors play a driving role in the process of generation, so as to find out the motivation of language variation (Fatima et al., 2023).

2.1.1 The Process Theory of Language Production

Linguistic variation is a creative use of language by a cognitive agent to achieve a target conceptualization, and its process can be explained by the process of language generation, which is the process of decoding conceptual content from the generation of the target concept in the brain

to the final formation of grammatical constructions (Benítez Fernández, 2023). Language generation is the decoding of complex concepts by elaborating grammatical constructions, and this creative process of linguistic conceptualization causes the simultaneous excitation of a series of linguistic units. The brain constructions of a linguistically excited group are constructed based on the complex interaction between the target concepts and the cognitive syntactic network, and the conversion from conceptual content to symbolic structure is constrained by the linear nature of language. Thus, the problem of decoding can be reduced to that of reducing complexity in multidimensional conceptualization through cognitive control in order to generate maximally relevant grammatical constructions. According to the theory of billiard ball pattern, it can be considered that the language prototype is like a billiard ball. In order to make the "billiard ball" enter the target net directly, efficiently and accurately, the player always tries to find out the most suitable and effective angle from multiple angles of force, and use the most appropriate force in order to realize the final goal. Linguistic variation is the process by which a cognizer tries to find the best way of association from the cognitive syntactic network that reduces complexity in order to realize a new target concept. It is just like the language prototype stored in people's brain, which is subject to the most suitable and effective "external force" in the specific communication situation, and then undergoes dynamic transmission to finally realize the target conceptualization. The dynamic force in this process (which may be cognitive or socio-cultural, or both) is highly variable because of the specific communicative situation, and there may be different angles and strengths of force, which makes the billiard ball of the prototype roll towards the target. Thus, the creative use of language needs to take into account the various types of external forces.

2.1.2 Creating a process theory of use

Realizing that the process of language variation is a use event and recognizing the complexity of the use event is necessary when analyzing the dynamics of language variation. Language variation is a creative use based on usage events with specific cognitive processes. The categorization of language variation use events is shown in Figure 1, where idiomatic structures are defined as being complex linguistic units that can be activated as brain criteria. Thereby encoding or decoding a specific use event, that is, an idiomatic activation group can instantiate idiomatic types into specific idiomatic symbols in discourse.

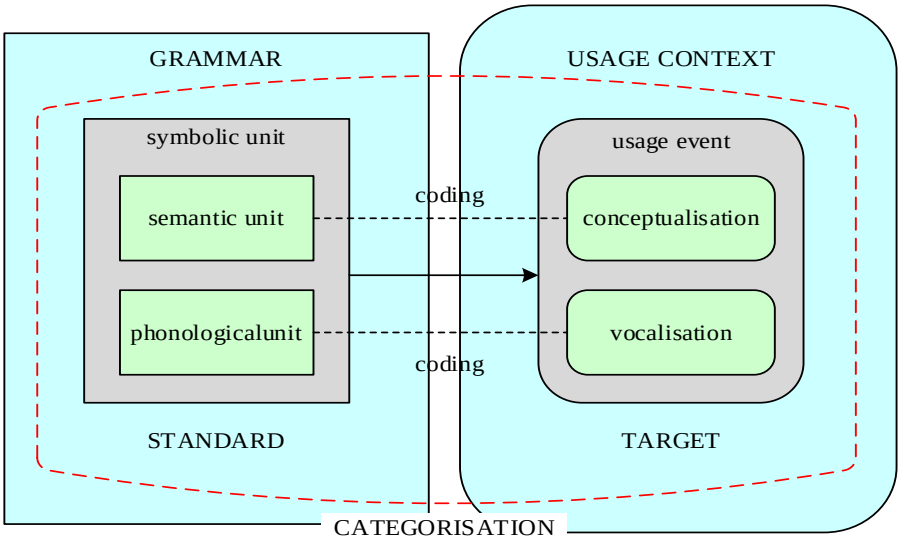

Figure 1: Categorization of linguistic variation using events

An adaptation of the use event model is shown in Figure 2, which further applies this process of categorization to linguistic variation. As can be seen, any idiomatic activation group is a semiotic standard used to encode a target conceptualization of a concrete situation (or to decode a target vocalization during comprehension). However, since idioms involve metaphorical expressions, abstract target meanings do not come to be written literally, but are based on non-literal idiomatic structures. According to the cognitive linguistic definition, idioms convey complex scenarios. Therefore, semantics is always encoded through a pre-coded structure that contains a literal scene. This literal scene reaches a certain level of consistency and serves as a semantic scaffold or context for mutating, shaping, reconceptualizing, or reconstructing the target meaning.

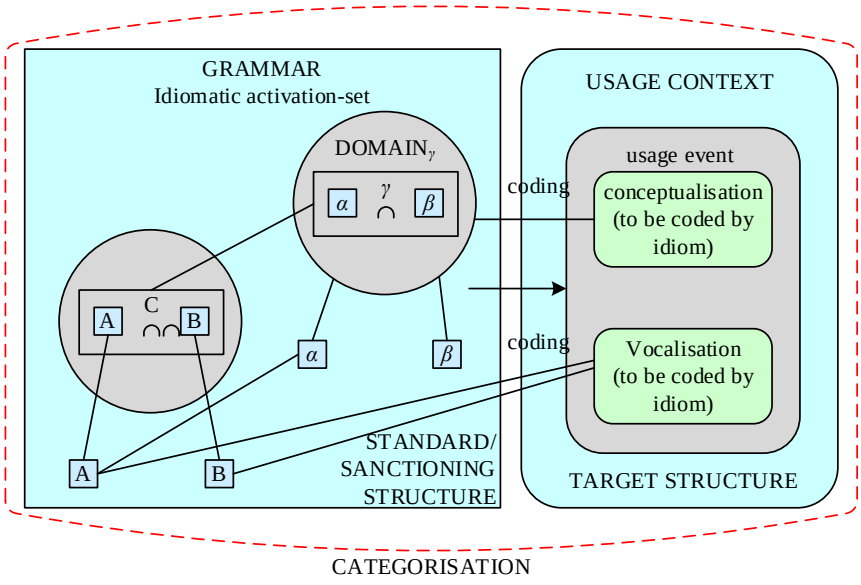

Figure 2: Usage of idioms Events

To characterize linguistic variation and standards as a form of linguistic creativity, it is necessary to provide cognitively driven explanations for the encoding of creative idioms. And it is important to consider that there are important socio-cultural factors behind these explanations. As is evident in linguistic and psycholinguistic discussions of the study of language variation, idiom generation has not been reduced to a straightforward lexical rule process, and lexico-grammatical changes in idiomatic variants must be explained in terms of creative processing. Creative idiomatic coding is based on brain computation, which involves brain manipulation of idiomatic activation groups and integration with other predefined symbolic units and schemas.

2.2 Research corpus

The corpus of Chinese-English relational clauses is selected from the Modern Chinese Corpus and AmE Brown Family Corpora respectively, which was created in 1995 and contains three major categories of natural sciences, humanities and social sciences, and more than 40 subcategories, covering a wide range and a large corpus size. The data volume of the whole corpus is about 100 million characters, and the corpus materials are diversified and representative, which are suitable for the research of language features and sentence patterns.AmE Brown Family Corpora is a collection of Brown1962, Frown1995, CROWN2008, CROWN2022, all of which are the corpus of the English native language, with diversified types of corpus, totaling 5 million characters, which can represent the English language. Totaling 5 million words, it can represent the linguistic expression level of English native speakers and is suitable for the study of English relational clauses in this research. Among them, there are 100 English subject and object relational clauses and 100 English subject and object relational clauses each, totaling 400 sentences. After the corpus was selected and collected, it was manually categorized and proofread, and the dependent syntactic treebank of Chinese main and object relative clauses and the dependent syntactic treebank of English main and object relative clauses were created independently.

2.3 Research methodology

2.3.1 Measurement indicators

This study uses dependency distance as a measure of linguistic variation in the syntactic complexity of Chinese-English relational clauses. Dependency distance refers to the linear distance from the dominant word to the subordinate word, and the difference between the ordinal number of the dominant word and that of the subordinate word can be positive or negative

due to the dominant word's forward or backward position in the sentence. Dependency distance is the difference between the number of the dominant word and the number of the subordinate word, and the average dependency distance of a sentence or a treebank is the average of the absolute values of the dependency distances of all the dependencies in the sentence or treebank. The specific measurement formula is as follows:

$$MDDs = \frac{1}{n-1}\sum_{i=1}^{n-1}\left|DDi\right| \quad (1)$$

In the above formula, MDDs represent the average dependency distances of sentences, n represents the number of words in a sentence, and DDi represents the i nd dependency in a sentence. Based on the above measurement formula, the dependency distances of the dependencies whose end-of-sentence punctuation governs the verb of the main clause are removed.

Based on the above formula of dependency distance and the scientific nature of dependency distance, this study quantifies the processing difficulty of English main and object relational clauses, which helps to better explore the influence of clause type and dependency position on the syntactic complexity of Chinese-English relational clauses, and then explore the deep cognitive motives affecting the variation of the English language.

2.3.2 Semantic Dependency Approach

Semantic dependencies include:

(1) Primary semantic roles, each of which corresponds to the existence of a nested relation and an inverse relation.

(2) Event relations, which describe the relationship between two events.

(3) Semantic dependency markers, which mark dependency information such as speaker tone.

It is because semantic dependency analysis contains such a wealth of information that how to utilize this useful information to calculate the similarity of sentences becomes the focus of research in our research work. When measuring the semantics of an entire sentence, factors such as punctuation, intonation, conjunctions, prepositions, etc. usually have less impact on the semantics. Such markers are ignored in the calculation and only semantically dependent roles with actual meaning are counted. The calculation process is as follows:

1) The number of identical semantic dependency roles:

$$SdpSim1(S1,S2) = \frac{2\times SameRole(S1,S2)}{Role(S1)+Role(S2)} \quad (2)$$

where $SameRole(S1,S2)$ denotes the number of identical semantic dependency roles contained in sentence $S1$ and sentence $S2$, and $Role(S1)$ and $Role(S2)$ denote the number of semantic dependency roles contained in each of sentences $S1$ and $S2$.

2) The position of the same semantic dependency roles in the sentence:

$$SdpSim2(S1,S2)=\frac{\sum_{i=1}^{n} dis(r_{1i},r_{2i})}{n} \quad (3)$$

$$dis(r_{1i},r_{2i})=1-\left|\frac{dis(r_{1i})-dis(r_{2i})}{dis(r_{1i})+dis(r_{2i})}\right| \quad (4)$$

where r_{1i} and r_{2i} denote the ith identical semantic role in sentences $S1$ and $S2$, and $dis(r_{1i})$ and $dis(r_{2i})$ denote the distance of the role from the root node of the semantic dependency syntactic tree; the closer the distance, the more similar the sentences are.

3) Sentence similarity based on semantic dependency analysis:

$$SdpSim(S1,S2)=\alpha SdpSim1(S1,S2)+\beta SdpSim2(S1,S2) \quad (5)$$

After experimentation, the calculations are most valid at $\alpha=0.2$, $\beta=0.8$.

2.3.3 Sentence similarity measures

The similarity between sentences is affected by many factors, such as word shape, sentence length, word order, word meaning, syntactic structure and so on. The degree of similarity between two sentences is not only related to the semantic information of the sentences, but also to the syntactic information of the sentences, and a single aspect of them cannot be considered solely to achieve the desired effect. Therefore, these factors mentioned above are regarded as different feature items of the sentences in the process, and these feature items are comprehensively utilized and assigned different weights according to their importance to get the final similarity degree of the sentences. The calculation process is as follows: Sentence similarity is measured based on the occurrence of the same words in the two sentences. The word similarity of sentences $S1$ and $S2$ is:

$$WordSim(S1,S2)=\frac{2\times SameWord}{Len(S1)+Len(S2)} \quad (6)$$

where $Len(S1)$ and $Len(S2)$ denote the number of words in the two sentences, respectively, and $SameWord$ is the number of identical words

contained in the two sentences.

(2) Sentence length similarity: Sentence length also reflects the degree of similarity between sentences to some extent, the closer the length of two sentences the greater the degree of similarity. Let $S1$ and $S2$ be two sentences, then the sentence length similarity of $S1$ and $S2$ is:

$$LenSim(S1,S2)=1-\left|\frac{Len(S1)-Len(S2)}{Len(S1)+Len(S2)}\right| \quad (7)$$

where and $Len(S2)$ denote the number of words in the two sentences, respectively.

(3) Semantic similarity: In this paper, we adopt Zhi.com as a semantic dictionary to measure the similarity between sentences. Let $S1$, $S2$ be two sentences, $S1$ contains the word $A_1, A_2, \cdots, A_m$, $S2$ contains the word $B_1, B_2, \cdots, B_n$, and the similarity between words $A_i(1 \le i \le m)$ and $B_j(1 \le j \le n)$ is calculated as $S(A_i, B_j)$. Then the semantic similarity between sentences $S1$ and $S2$ is:

$$SematicSim(S1,S2)=\left(\frac{\sum_{i=1}^{m} a_i}{m}+\frac{\sum_{j=1}^{n} b_j}{n}\right)\Big/2 \quad (8)$$

Among them:

$$a_i=\max\left(S(A_i,B_1),S(A_i,B_2),\cdots,S(A_i,B_n)\right) \quad (9)$$

$$b_i=\max\left(S(A_1,B_j),S(A_2,B_j),\cdots,S(A_m,B_j)\right) \quad (10)$$

(4) Sentence similarity

The formula for the multi-factor based sentence similarity measure model is as follows:

$$SenSim(S1,S2)=\gamma_1 WordSim(S1,S2)+\gamma_2 LenSim(S1,S2)+\gamma_3 SematicSim(S1,S2)+\gamma_4 SdpSim(S1,S2) \quad (11)$$

Among them:

$$\gamma_1+\gamma_2+\gamma_3+\gamma_4=1 \quad (12)$$

2.4 Research tools

The study of language variation needs to be guided by mature theories, scientific research methods and effective research tools. So as to facilitate the efficient and scientific processing of the real corpus of language and help researchers to be able to obtain effective research data, and then to discover the characteristics and laws of language variation and enrich the research in the

field of language. The research tools used in this study are Excel, Corpus Word Parser and IBM SPSS Statistics 26.

(1) Corpus Word Parser is a free corpus annotation tool with a simple interface and easy operation. The software can recognize and process different character encodings, and it can perform word segmentation and lexical annotation on Chinese text, which is fast and accurate. Therefore, in this study, the Corpus Word Parser software is used to perform lexical segmentation on the corpus of Chinese relational clauses, and then after manual proofreading, the lexical segmentation of Chinese relational clauses is completed.

(2) IBM SPSS Statistics 26 is a commonly used statistical analysis tool at present. This software integrates data organization and analysis functions, which can help users to make effective statistical analysis and data inference on data. Its basic functions are diverse, and it can perform data analysis such as descriptive statistics, comparing means, general linear models, generalized linear models, correlation analysis, regression analysis, non-parametric tests, and time series forecasting. In this study, in order to ensure the reliability of the data, the average dependency distance of Chinese-English relational clauses was analyzed and tested with the help of SPSS to finally arrive at the results of the study.

2.5 Steps of the study

This study includes the following four steps:

(1) Divide the target corpus, label the corpus and complete the construction of the library. Based on the above search terms, this study selects Chinese and English relational clauses from the Modern Chinese Corpus of the State Language Commission and the AmE Brown Family Corpora, respectively, and then divides Chinese main and object relational clauses into English main and object relational clauses according to the syntactic roles of the center word in the relational clauses.

(2) Subsequently, the Chinese relative clauses were divided by Corpus Word Parser, and the final results were manually proofread.

(3) Then, based on the collected Chinese-English relational clauses corpus and the syntactic analysis theory of dependent syntax, we created the dependent syntactic treebank of Chinese main and object relational clauses and the dependent syntactic treebank of English main and object relational clauses by manually annotating them in Excel.

(4) Finally, data extraction and interpretation. Based on the annotation of Chinese and English relative clauses in the Dependency Syntax Treebank, we extracted and calculated the average dependency distance of Chinese and English relative clauses with the help of tools such as Excel and SPSS, tested

the significance of the obtained data, and made linguistic explanations for the linguistic variations that appeared in the study.

3. FINDINGS

3.1 Variation trajectories in English lexicon and syntax

The study of vocabulary variation through two subjects in this paper begins with plotting the trajectories of the development of vocabulary variables (TTR, D-value, unique word proportion and word frequency profile LFP) for each of the two groups of subjects using Excel scatter plots and polynomial trend lines (2nd degree). The two subjects were two college students (one male and one female) from a first-year non-English major at a university. Then the vocabulary variability of each of the two groups was described by maximum-minimum extreme value plots, followed by a probability level test of the vocabulary variability using a Monte Carlo simulation-based re-sampling technique to examine the vocabulary development patterns and the paths of improvement.

3.1.1 Trajectories of lexical variation

This section plots the vocabulary development trajectories of the 2 subjects in their oral English development using Excel scatter plots and polynomial trend lines (2nd degree), which include the following measures, word frequency profile (LFP) vs. proportion of unique words, TTR vs. D-value. The developmental trajectories of each measure of the vocabulary variables (word frequency profile LFP vs. proportion of unique words, TTR vs. D-value) for the male subjects are shown in Figure 3. It can be seen that the overall trend of the subjects' (male) high-frequency word proportions (LFP) was decreasing over the 16 recording epochs. However, it is clear that there is a great deal of variation in the developmental process of the subject (male), as shown by the scatterplot with data and straight lines. The LFP hovered at a low level in the predevelopmental period, rose to a peak of 82.75% in the mid-developmental period, i.e., at the 8th test, and then declined to a low to moderate level value. At the same time, the proportion of unique words was generally increasing and the developmental variation was prominent, as evidenced by the fact that the predevelopmental period was at a moderate level, but dropped to the trough in the middle of the period, and then came to a low-moderate horizontal value in the latter period, which is well in line with the later development of the LFP.

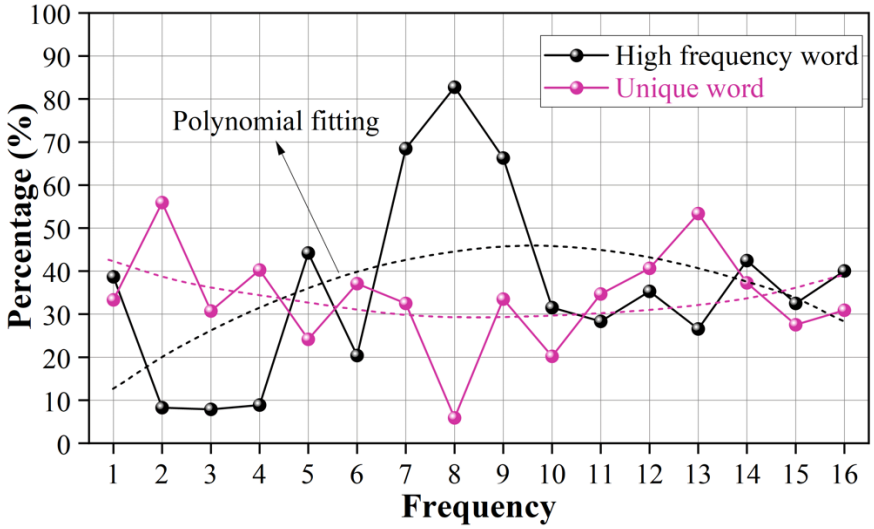

Figure 3: Trajectory of lexical variation (Subject male)

The polynomial trend line (degree 2) for female subjects is shown in Figure 4, and again the overall trend in the proportion of high-frequency words (LFP) for the subjects (female) was decreasing over the 16 recording time periods. However, it is clear, as shown in the scatter plot with data and straight lines, that there is a great deal of variation in the development of the subject (female), from the predevelopmental period when LFP hovered at a low level. Similarly it rises to a peak of 87.09% in the middle of the development, after which the downward trend is evident and finally stays at a low to medium level value. Meanwhile, the proportion of unique words is generally rising, and the developing variation, although not as prominent as LFP, can be clearly observed in the presence of linguistic variation, especially in the middle and late stages of development.

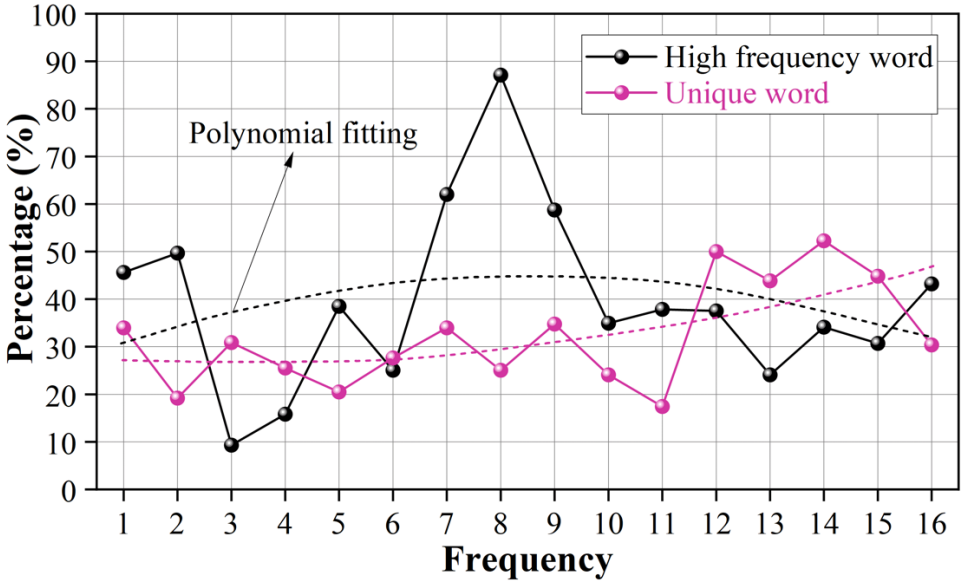

Figure 4: Trajectory of lexical variation (Subject female)

3.1.2 Syntactic variation trajectories

This section is to plot the syntactic development trajectories of the 2 subjects in their oral English development using Excel scatter plots and polynomial trend lines (2nd degree), which include the following measures, simple sentence proportion, complex sentence proportion, finite verb proportion W/FV, and sentence length MLT. The 16-variable trajectories of the male subjects' English simple and complex sentences, and the restricted verb ratio W/FV are shown in Figure 5. It can be seen that the overall trend of the subjects' (male) simple sentences, complex sentences, and finite verbs (W/FV) is upward. However, it is obvious that, as shown in the scatterplot with data and straight lines, there is a large amount of variation in the development of the subjects (male) in all three measures, which is manifested in the fact that the proportion of simple sentences is very high in the early and middle stages of the process, then begins to decline after the middle stage, and then rises again in the late stage of the process to 55.14%, forming a typical U-shaped developmental trajectory. The proportion of complex sentences is relatively low at the beginning (10.85%), and only begins to rise gradually in the middle and late stages, and is basically equal to the proportion of simple sentences in the final stage of development.W/FV starts at a low level, and begins to climb after the middle stage. Finally, it stabilizes at a low-to-moderate value.

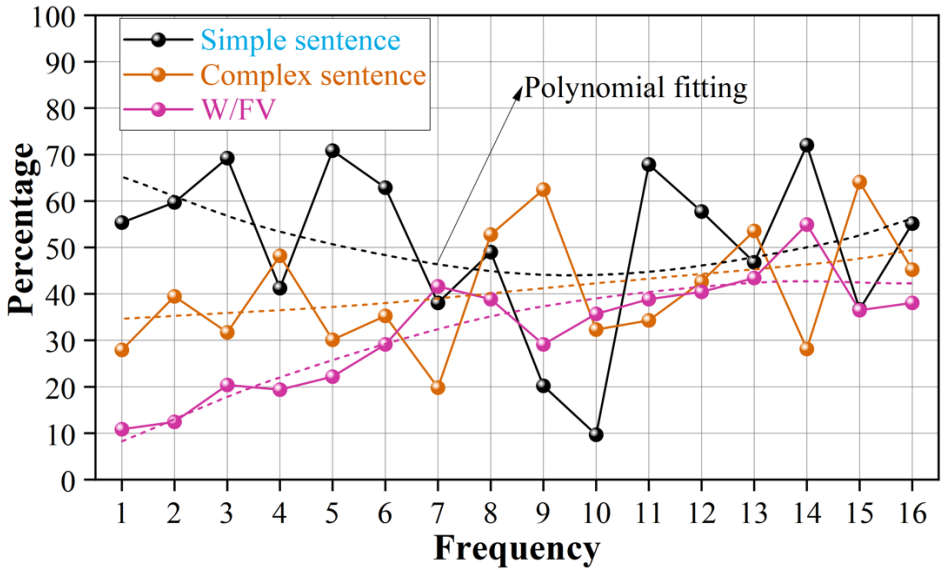

Figure 5: Syntactic variation trajectory (Subject male)

The trajectories of the 16 variations of simple and complex sentences, and the ratio W/FV of finite verbs in English of the female subjects are shown in Figure 6. It can be seen that the overall trend of simple sentences, complex sentences, and finite verbs (W/FV) of the subjects (female) is upward, and the overall upward trend of W/FV is very prominent. It is clear, however, that

there is a great deal of variation in the development of the subjects (female) on all 3 measures, as shown by the scatterplots with data and straight lines. This is reflected in the fact that the proportion of simple sentences is very high in the early stage (49.8%), starts to decrease after the middle stage, and then increases again in the late stage and stabilizes in the final stage, forming a typical U-shaped developmental trajectory. The proportion of complex sentences is relatively low at the beginning (27.4%), but climbs from the middle stage onwards and stabilizes at the end, and is basically the same as the proportion of simple sentences.W/FV is low at the beginning, but starts to rise significantly after the middle stage, and stabilizes at a high value at the end.

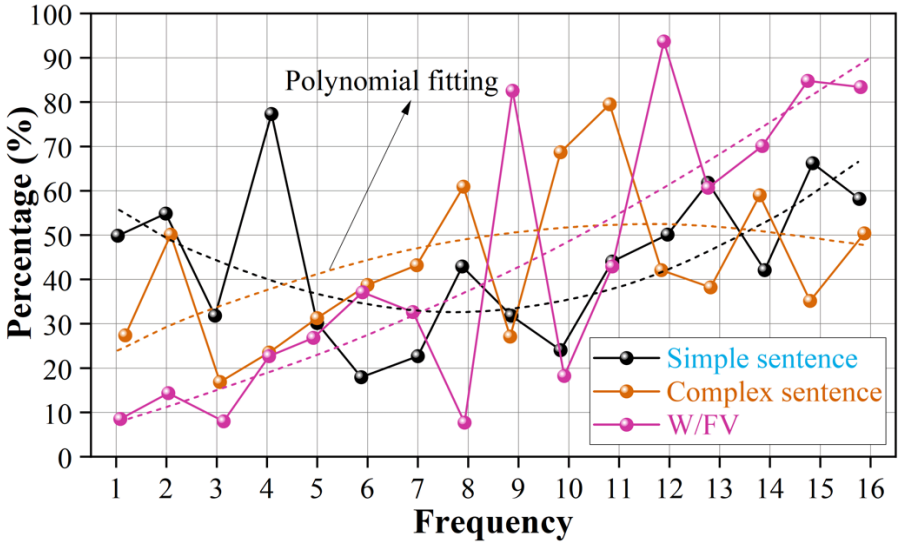

Figure 6: Syntactic variation trajectory (Subject female)

In summary, the overall trend of the subjects' English simple sentences, complex sentences and finite verbs (W/FV) is upward, with the variation in the development process of simple sentences showing a change from high to low to high, and the variation in the development process of complex sentences showing a change between lower and higher. Meanwhile, the variation in the development of W/FV varies between low, medium and high.The overall trend of MLT is slightly upward, with the variation in the development varying from very low to high to medium.

3.2 Typical Dependent Tree Forms

The results of the variant trajectories of English vocabulary and syntax show that the level at which a node is located is a fundamental linguistic property, and that homogeneity is shown among the subpools of the same language. Therefore, a typical representative graph of the dependency tree can be depicted according to the average value. Table 1 shows the average width of each level summarized by each of the six libraries of the English lexical and

syntactic language, reflecting the distribution of the levels in which each node is located. Based on this table, a typical dependency tree schema for English lexical and syntactic variants is depicted as shown in Figure 7. It can be seen that layer 3 of the English lexicon is wider, with an average layer width of 4.71, which implies that the average collocation (coefficient of variation) of predicates is larger than that of the English syntax, i.e., it side by side indicates that the English lexicon is more compact. In English language, the width of nodes located in tier 3 are all maximal, after which the number of nodes starts to decrease. However, the width of the English lexicon at all levels is consistently greater than the width of English syntax. Starting from level 6, the trend starts to flip and English syntax is a bit wider than the lexical dependency tree. And the average number of layers is more, which is a side effect of the fact that there is more nesting in English, which is also an important reason why the average dependency distance is smaller in English.

Table 1: Summary of dependency trees of English vocabulary and syntactic variation

Test	English Vocabulary	English Syntax	Test	English Vocabulary	English Syntax
1	1.00	1.00	9	0.27	0.78
2	3.55	3.16	10	0.11	0.45
3	4.71	3.92	11	0.05	0.27
4	4.18	3.78	12	0.01	0.15
5	3.03	2.95	13	0.00	0.09
6	1.92	2.34	14	0.00	0.05
7	1.12	1.72	15	0.00	0.02
8	0.54	1.12	16	0.00	0.01

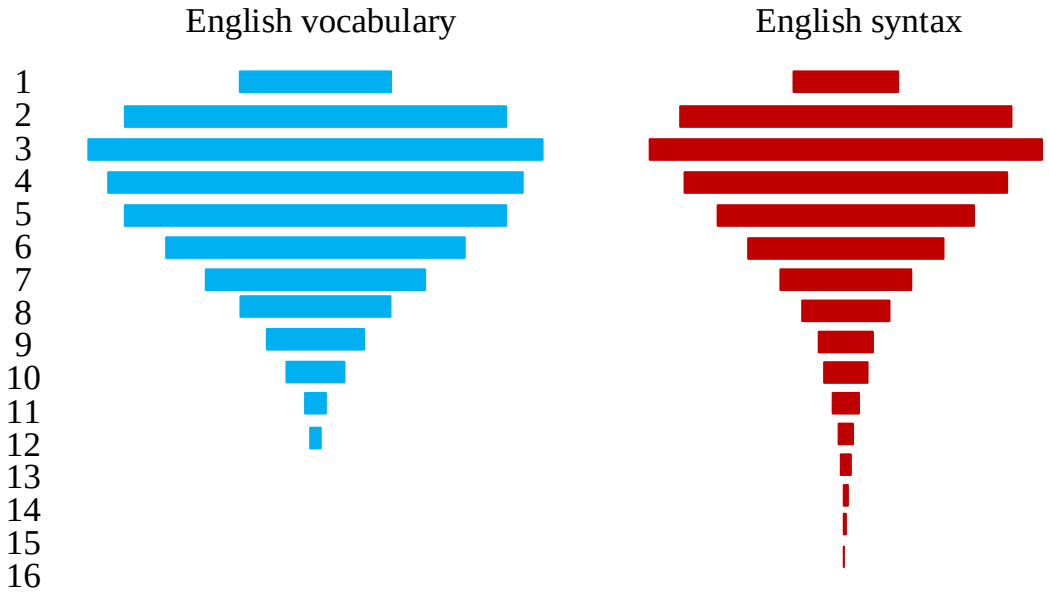

Figure 7: Diagram of a typical dependency tree

3.3 Analysis of the measurement patterns of language variation

Measurement Linguistics is a sub-discipline of Mathematical Linguistics, which takes real corpus as the object of study, uses quantitative methods to study human language, explains the phenomena on the basis of observing and describing the linguistic phenomena, and then makes it possible to predict the phenomena. Linguists have generalized and summarized some laws of metrological linguistics in their observation of linguistic phenomena, most of which are used to describe cooccurring linguistic phenomena and have been widely studied and applied. For example, the famous Zipf's law is used to describe the word frequency-frequency order relationship in text. Due to the scarcity of ephemeral linguistic data, there is a comparative lack of laws on linguistic variation. Currently, the most recognized is the Piotrowski-Altman law. The central idea of the Piotrowski-Altman law is that language change is the result of the interaction of old and new forms, and that the process of language change is an S-shaped curve about time. The proportion of the use of innovative forms of language and the process of propagation over time can be represented by an S-shaped curve as shown in Figure 8. It can be seen that the spread of linguistic variation is very slow at the beginning, and after slowly increasing to a certain level (at about 20%), the momentum increases and the growth rate accelerates. The growth rate reaches a maximum around the time when the momentum of the new forms exceeds that of the old forms, after which the growth rate gradually slows down. After the proportion of new forms reaches about 80%, the growth rate slows down significantly until it is finalized. This process is shown as an S-curve on the ratio-time diagram.

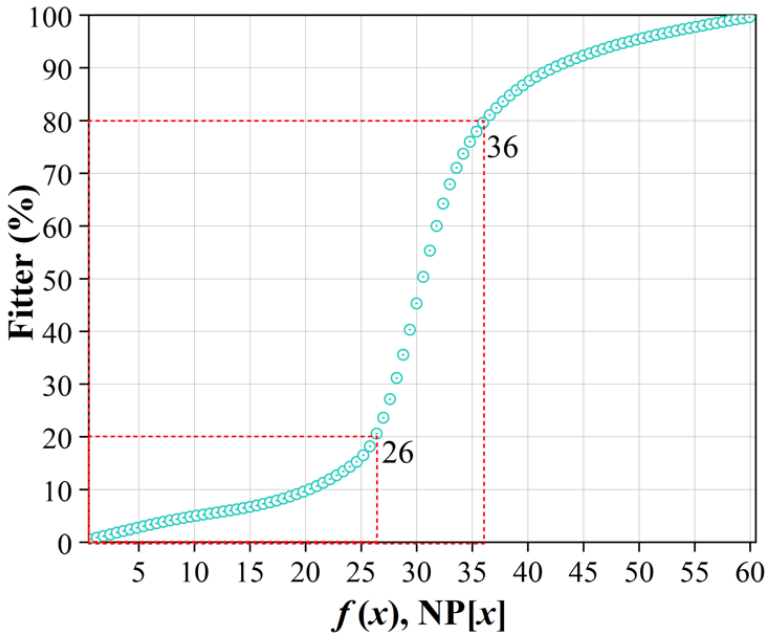

Figure 8: S-curve of language variation

The core idea of the S-curve model is that "all language changes are the result of the alternation of old and new forms", and the formula of the S-curve is obtained, i.e., the law of English language variation in a multicultural context is:

$$p = \frac{e^{k+st}}{1+e^{k+st}} \quad (13)$$

Where, p denotes the proportion of changing forms, t is a time variable, s and k are constants.

There are many variants of the formula of the S-curve, and if the form of the English language changes completely and the final proportion reaches 100%, the coefficient of variation of the English language reaches the criterion and will not change. The criterion of English language coefficient of variation is:

$$p = \frac{1}{1+ae^{-bt}} \quad (14)$$

Historical data on language variation are more difficult to obtain, and the S-curve model of language variation, although confirmed in some data on the historical variation of language forms, needs to be supported by more historical data. In fact, with the exception of the famous "roundabout do usage" material, there is little long-term temporal data available to fit S-curves. Currently, there are no studies that have fitted S-curve functions to data on changes in usage of English personal pronouns over time. The "constant rate hypothesis" has only been tested in a more limited number of cases than the S-curve model, and more data over time are needed.

4. CONCLUSION

Language is a complex adaptive system, and at the same time the development of language is a dynamic system with its own development law. The development of human language is a historical process, and the syntactic structure and output characteristics of language make different languages have holistic and localized features, which can be deeply explored with the help of dynamic and data-driven methods. Based on the theory of dependency syntax, this study reveals the dynamics of English variation with the help of dependency distance as a measurement index, and explores the facilitating and restricting mechanisms of English formation and variation under the tension between solidarity and variability. Finally, a cognitive sociolinguistic interpretation of English variability is made, and the measurement law of English variability is summarized. Experimental Results The experiment

proves that there are a variety of factors affecting language variation and change, some of which are of the language structure itself, which will lead to the diversity of the phenomena of variation and change in different languages. Some of the causes are social, which will affect one or more relevant language communities. Many more are functional, closely related to cognitive, psychological and other factors, which are common to all human beings and which explain the cross-language prevalence of certain variation and change phenomena.

References

Akujobi, O. S., & Ebere, P. E. (2022). STANDARDIZATION AND CODIFICATION OF NIGERIAN ENGLISH: BLUEPRINT FOR GRAMMATICALITY AND ACCEPTABILITY. *CACH Journal of Humanities and Cultural Studies*, *3*(1), 122-136.

Benítez Fernández, M. (2023). Linguistic variation, social meaning and covert prestige in a Northern Moroccan Arabic variety. *Languages*, *8*(1), 89.

Bomer, R. (2017). Leading the Call: What Would it Mean for English Language Arts to Become More Culturally Responsive and Sustaining? *Voices from the Middle*, *24*(3), 11-15.

De Villiers, J. G., & De Villiers, P. A. (2017). The acquisition of English. In *The crosslinguistic study of language acquisition* (pp. 27-139). Psychology Press.

Dragojevic, M. (2017). Language attitudes. In *Oxford research encyclopedia of communication*.

Eckert, P. (2017). Age as a sociolinguistic variable. *The handbook of sociolinguistics*, 151-167.

Fatima, M., Siddique, A. R., Ahmad, M., & Mahmood, M. A. (2023). Exploring linguistic variation in Pakistani English newspaper editorials through multidimensional analysis. *Newspaper Research Journal*, *44*(4), 425-451.

Fishman, J. A. (2020). Who speaks what language to whom and when? In *The bilingualism reader* (pp. 55-70). Routledge.

García, O., & Otheguy, R. (2017). Interrogating the language gap of young bilingual and bidialectal students. *International Multilingual Research Journal*, *11*(1), 52-65.

Gleason, J. B., & Ratner, N. B. (2022). *The development of language*. Plural Publishing.

Greenhill, S. J., Wu, C.-H., Hua, X., Dunn, M., Levinson, S. C., & Gray, R. D. (2017). Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences*, *114*(42), E8822-E8829.

Jenkins, J., & Leung, C. (2019). From mythical 'standard'to standard reality: The need for alternatives to standardized English language tests. *Language Teaching*, *52*(1), 86-110.

Kidd, E., & Donnelly, S. (2020). Individual differences in first language acquisition. *Annual Review of Linguistics*, *6*(1), 319-340.

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in cognitive sciences*, *22*(2), 154-169.

Kim, M., Crossley, S. A., & Skalicky, S. (2018). Effects of lexical features, textual properties, and individual differences on word processing times during second language reading comprehension. *Reading and Writing*, *31*, 1155-1180.

Kornexl, L. (2017). Standardization. *The History of English*, *2*, 220-235.

Lee, H.-K. (2017). Heading toward the global standardization of English education in

Korean universities: A case study of an English program in a Korean university. In *English education at the tertiary level in Asia* (pp. 84-108). Routledge.

Lohndal, T., Rothman, J., Kupisch, T., & Westergaard, M. (2019). Heritage language acquisition: What it reveals and why it is important for formal linguistic theories. *Language and Linguistics Compass*, *13*(12), e12357.

Mair, C., & Leech, G. N. (2020). Current changes in English syntax. *The handbook of English linguistics*, 249-276.

Mauranen, A., Pérez-Llantada, C., & Swales, J. M. (2020). Academic Englishes: A standardised knowledge? In *The Routledge handbook of world Englishes* (pp. 659-676). Routledge.

Milroy, J., & Milroy, L. (2017). Varieties and variation. *The handbook of sociolinguistics*, 45-64.

Rosa, J., & Flores, N. (2017). Unsettling race and language: Toward a raciolinguistic perspective. *Language in society*, *46*(5), 621-647.

Utami, T. K. P. (2022). Vietnamese English Accent Vs American English Accent: Locating the Phonological Variation. *Linguistics and Literature Journal*, *3*(2), 90-97.