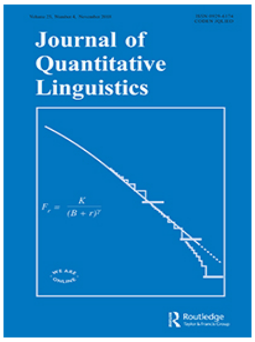

Journal of Quantitative Linguistics

ISSN: 0929-6174 (Print) 1744-5035 (Online) Journal homepage: https://www.tandfonline.com/loi/njql20

Can the Probability Distribution of Dependency Distance Measure Language Proficiency of Second Language Learners?

Jinghui Ouyang & Jingyang Jiang

JOURNAL OF QUANTITATIVE LINGUISTICS
2018, VOL. 25, NO. 4, 295–313
https://doi.org/10.1080/09296174.2017.1373991

Routledge
Taylor & Francis Group

Can the Probability Distribution of Dependency Distance Measure Language Proficiency of Second Language Learners?

Jinghui Ouyang and Jingyang Jiang

Department of Linguistics, School of International Studies, Zhejiang University, Hangzhou, China

ABSTRACT

It has been found that the length distribution of many linguistic units fits well the same model, the Zipf-Alekseev function. In this article, we aimed to find out whether this holds for English learners' interlanguage and whether the parameters in probability distribution of dependency distance can measure the language proficiency of second language learners. We selected 367 participants of English learners of nine consecutive grades and fitted different probability distribution models to dependency distances of their writings in English and of self-built contrastive dependency treebanks based on Wall Street Journal Corpus. It was found that: (1) the Zipf-Alekseev distribution well captures the probability distribution of dependency distance of each grade and native speakers; (2) the probability distribution of dependency distance well measures second language learners' language proficiency at different learning stages; (3) high-level learners don't present exactly the same parameters in the probability distribution of dependency distance as those of native speakers, which means learners' language proficiency is not as high as that of English native speakers and second language learners' syntactic acquisition process is always constrained by the tendency of dependency distance minimization. This study corroborates that quantitative linguistic methods can be well utilized in second language acquisition researches.

1. Introduction

Human natural languages may follow certain universal laws. As early as the 1930s, Zipf (1936) proposed the law of distribution for word frequencies, stating that the frequency of any word in natural languages is inversely proportional to its rank in the frequency table. In addition, the distribution of morpheme length (Pustet & Altmann, 2005), word length (e.g. Pande & Dhami, 2012; Eger, 2013; Narisong et al., 2014; Kalimeri et al., 2015; Chen & Liu, 2016), sentence length

CONTACT Jingyang Jiang ✉ jjy203@163.com, jy-jiang@zju.edu.cn

(e.g. Sigurd, Eeg-Olofsson, & van de Weijer, 2004; Pande & Dhami, 2015) and dependency relations (e.g. Liu, 2009a) has all been extensively examined in quantitative investigations of human languages and has been found to follow regular patterns. What's more, Popescu et al. (2014) found that the length distribution of many linguistic units fits well the same model, the Zipf-Alekseev function.

Moreover, numerous previous studies (Liu, 2008; Jiang & Liu, 2015; Lu & Liu, 2016; Liu, Xu, & Liang, 2017) have shown that the distribution of dependency distances has presented certain regularity. By investigating probability distributions of dependency distances in six texts extracted from a Chinese dependency treebank, Liu (2007) found that the right truncated zeta distribution can well capture the investigated distribution. Jiang and Liu's (2015) further study on the effects of sentence length on dependency distance based on a parallel English–Chinese dependency treebank also found that the right truncated zeta can fit well the distribution of dependency distance of 420 sentences (210 sentences in English and 210 in Chinese). By analysing dependency treebanks of 30 languages, Lu and Liu (2016) found that for the majority of 30 languages, the distribution of dependency distance conforms to certain models, namely, Stretched Exponential Distribution for 'short sentences' and Truncated Power Law Distribution for 'long sentences'.

Dependency distance refers to the linear distance between two linguistic units having a syntactic relationship within a sentence (Heringer, Bruno, & Rainer, 1980; Hudson, 1995). The linear distance between two words with a syntactic relationship, restrained by human working memory, can reflect the comprehension difficulty of syntactic structure (Liu, 2008). If dependency distance can indicate how human working memory constrains language comprehension and production, and if most people have a similar working memory capacity, it is probable that the distribution of dependency distance of human languages fits well certain distribution patterns (Jiang & Liu, 2015). The distribution of dependency distances is a kind of length distribution. Popescu et al. (2014) found that the Zipf-Alekseev function can well capture the distribution of many linguistic units of physical length, including syllable length, morpheme length, word length, sentence length, etc. In addition, previous studies found that the probability distribution of dependency distances well fits Zipf-like laws, including Zipf-Alekseev (Jiang & Liu, 2015). Moreover, Popescu et al. (2014) found that parameters in the Zipf-Alekseev function can reflect the peculiarities of human languages to some extent. For example, the parameter *a* of a language of an older stage is greater than that of a younger one.

As a structurally intermediate status between the native and target languages (Selinker, 1969; Brown, 1994), the second language learners' language system, defined as 'interlanguage', is also a kind of human natural language. In second language acquisition (SLA), the evaluation of learners' language proficiency at different learning stages is always the research focus. Syntactic complexity, along with accuracy and fluency, has been proposed as an essential construct in the description of second language proficiency (Nearysundquist, 2016). As a kind of natural language, theoretically, the syntactic complexity of a second language learners' interlanguage can also be measured by dependency distance (Jiang & Ouyang, 2017). If parameters in probability distribution can reflect the universalities as well as peculiarities of languages as we mentioned above, can we measure second language learners' language proficiency through the variations of parameters in the Zipf-Alekseev distribution of dependency distance at different learning phases of SLA?

Such research on SLA by applying the methods and models of quantitative linguistics has its significance and value not only to quantitative linguistics, but also to SLA. Quantitative linguistics investigates languages using statistical methods, aiming to find out universal laws of languages, which can be used to solve practical problems in various fields of applied linguistics, including language teaching and language learning. Our research demonstrates how to apply methods, models or findings from quantitative linguistics to solve problems in applied linguistics, more precisely, in SLA. As for SLA, our research shows how to integrate quantitative linguistics into the study of language learners' interlanguage, so as to reveal linguistic phenomenon in a more scientific and universal manner.

Therefore, we chose to study how the parameters in the probability distribution of dependency distances in second language learners' writings vary to observe their developmental language proficiency during their English learning processes. To study SLA in the QL framework in general, and from the probability distribution in particular, we raised three specific research questions as follows:

Question 1: Does the probability distribution of the dependency distance of second language learners' interlanguage fit well the Zipf-Alekseev distribution?

Question 2: Can the parameters in the Zipf-Alekseev distribution well reflect second language learners' language proficiency at different learning stages?

Question 3: Will high-level second language learners present the same parameters in the probability distribution of dependency distance as those of English native speakers?

2. Materials and methods

This study tries to answer the above-mentioned three research questions. Detailed information about the study's method and materials, including participants, materials, procedures and data analysis are described as follows.

2.1. Participants

The participants were 367 Chinese students from two high schools and one university in Zhejiang Province, China. The participants' basic information, including the number and the age of each grade is presented in Table 1. The

Table 1. A brief profile of participants.

Group	Number	Age	Years of English Learning
First Grade of Junior High School (J1)	75	12–13	3–4
Second Grade of Junior High School (J2)	61	13–14	4–5
Third Grade of Junior High School (J2)	69	14–15	5–6
First Grade of Senior High School (S1)	78	15–16	6–7
Second Grade of Senior High School (S2)	74	16–17	7–8
Third Grade of Senior High School (S3)	79	17–18	8–9
First Grade of University (U1)	40	18–19	9–10
Second Grade of University (U2)	28	19–20	10–11
First Grade Postgraduate of English Major (P1)	26	22–23	13–14

subjects range from first graders of junior high school to first grade postgraduates of English major, which spans 9 grades. In China, most students start learning English from fourth grade of elementary school. After their second year at university, Chinese students usually do not need to take any English lessons. Therefore, choosing Chinese junior high school students, senior high school students, undergraduates of first and second grade (non-English majors) and postgraduates as participants can help us observe almost the whole process of learning English. To discover whether Chinese EFL learners will present the same parameters in the probability distribution as English native speakers, we also compared English native speakers with Chinese first grade postgraduates of English major, who, in the current study, are deemed the representatives of high-level EFL learners in China.

2.2. Materials

Our self-built dependency treebank contains 367 English compositions written by the above-mentioned participants within the prescribed time limit in the class, with a total of 58,583 words, with about 6500 words from each grade. The compositions collected are narrative ones. The topics of the compositions are basically about their own experiences, such as 'my weekend', 'an embarrassing experience', 'an unforgettable experience', 'an annoying experience', and so on. We controlled the genre and the subject matter by assigning compositions of similar topics to make a better longitudinal comparison. The contrastive dependency treebank of native English was extracted from the Wall Street Journal (WSJ) Corpus. We selected linguistic data randomly from the WSJ Corpus and built four sub-corpora with about 6500 words of each corpus as contrastive dependency treebanks.

2.3. Procedure

After the students finished the writing tasks, we carefully selected the data and inputted them into a computer. All 369 compositions were kept in a TXT format. Each composition was labelled with a unique code indicating the student's school, grade, number and the topic. Students' compositions were faithfully keyboarded into the computer in exactly the same way as they were presented, including capitalization, punctuation, spelling and grammatical mistakes. The POS (Part-of-Speech) annotation and dependency relation tagging were automatically done by Stanford Parser 3.6.0, a tagging software developed by Stanford University. To meet the requirements of our research, we modified some of the Stanford typed dependencies and established a new syntactic relation system. Moreover, we made an error tagging system and labelled L2 writing errors, including lexical errors and grammatical errors.

Although Stanford Parser can provide an effective version of annotation of all the raw data, there still exist quite a few mistakes because, on the one hand, the accuracy of the programme does not reach 100%; on the other hand, those participants with relatively low language proficiency will be very likely to make language mistakes, increasing the inaccuracy of the programme. So after preliminary annotation done by Stanford Parser, we did the manual check and modification. Moreover, applying the exactly same tagging systems to annotate the four contrastive corpora makes our research more accurate and scientific.

2.4. Data analysis

The concept of (dependency) distance is often used in the syntactic analysis framework, with phrases or dependency relations as its basic constituents. The current paper uses the syntactic analysis framework of dependency grammar in which sentence structure is analysed using the dependency relations between words in a sentence (Tesnière, 1959; Hudson, 2007, 2010; Nivre, 2006; Liu, 2009b). A dependency relation has three core properties: binary, asymmetry and labeledness.

Based on these three properties, we can build a syntactic dependency tree or directed dependency graph as the representation of a sentence. In this paper, we use directed acyclic graphs to present dependency structure. Figure 1 is a directed acyclic graph that shows a dependency analysis of the sentence 'He must have good ideas'.

In Figure 1, all the words in a sentence are connected by grammatical relations. For example, the subject and the object depend on the main verb; prepositions (not exemplified in Figure 1) depend on the nouns or verbs that they

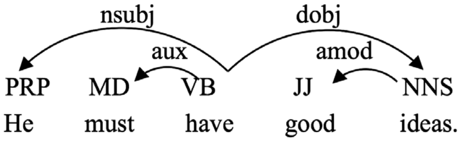

Figure 1. Dependency structure of 'He must have good ideas.'

modify, and so on. In each pair of connected words, one is called the dependent and the other is called the governor. The labelled arc is directed from the governor to the dependent.

The linear distance between governor and dependent is defined as 'dependency distance'. The concept was first used in Heringer et al. (1980), who extracted the idea from the depth hypothesis of Yngve on phrase structure grammar (Yngve, 1960, 1996). The term 'dependency distance' was introduced in Hudson (1995) and defined as 'the distance between words and their parents, measured in terms of intervening words'.

A method (Liu, Hudson, & Feng, 2009) was proposed for measuring the mean dependency distance of a sentence (*MDD*), of a sample of a treebank (a corpus with syntactic annotation) or of a particular dependency type in a treebank. Formally, let $W_1 \ldots W_i \ldots W_n$ be a word string. For any dependency relation between the words W_a and W_b, if W_a is a governor and W_b is its dependent, then the *DD* between them can be defined as the difference 'a–b'; by this measure, adjacent words have a *DD* of 1 (rather than 0 as is the case when *DD* is measured in terms of intervening words). When 'a' is greater than 'b', the *DD* is a positive number, which means that the governor follows the dependent; when 'a' is smaller than 'b', the *DD* is a negative number and the governor precedes the dependent. However, in measuring *DD* the relevant measure is the absolute value of *DD*.

The mean dependency distance of an entire sentence can be defined as:

$$MDD(dependency\ type) = \frac{1}{n} \sum_{i=1}^{n} |DD_i| \tag{1}$$

Here 'n' is the number of words in the sentence and 'DD_i' is the dependency distance of the ith syntactic link of the sentence. In a sentence, there is generally one word (the root verb) without a governor, whose *DD* is therefore defined as zero.

For instance, a series of *DD*s can be obtained from the sentence in Figure 1 as follows: 2 1 0 1 2. In other words, the example has two dependencies with *DD* = 1 and two dependencies with *DD* = 2. Using equation (1), the *MDD* of this sentence is 6/4 = 1.5.

This formula can also be used to calculate the *MDD* of a larger collection of sentences, such as a treebank:

$$MDD(the\ sample) = \frac{1}{n-s} \sum_{i=1}^{n-s} |DD_i| \tag{2}$$

In this case, 'n' is the total number of words in the sample, and 's' is the total number of sentences in the sample. DD_i is the *DD* of the ith syntactic link of the sample.

In the previous studies of the distribution of dependency distances, the Zipf-like laws (Liu, 2007; Jiang & Liu, 2015; Lu & Liu, 2016) were found to well fit the distribution model of dependency distances. The distribution of dependency distances is a kind of length distribution. Popescu, Best, and Altmann (2014) found that the Zipf-Alekseev function is adequate for the distribution of any linguistic units of physical length. Based on the distribution models in the researches of length distribution for linguistic units and in previous quantitative linguistic studies of distribution of dependency distances (Jiang & Liu, 2015), we assume that the investigated distributions obey the Zipf-Alekseev model (Hřebíček, 1996; cited from Strauss & Altmann, 2006). Two assumptions were adopted by Hřebíček:

(1) the logarithm of the ratio of the probabilities P_1 and P_x is proportional to the logarithm of the class size, i.e.

$$\ln\left(P_1/P_x\right) \propto \ln x$$

(2) the proportionality function is given by the logarithm of Menzerath's law (Hierarchy), i.e.

$$\ln\left(P_1/P_x\right) = \ln\left(Axe^b\right)\ln x$$

yielding the solution

$$P_x = P_1 x^{-(a+b\ln x)},\ x\ =\ 1, 2, 3, \ldots \tag{3}$$

If equation (1) is considered a probability distribution, then P_1 is the norming constant, otherwise it is estimated as the size of the first class, $x = 1$. Very often, diversification distributions display a diverging frequency in the first class while the rest of the distribution behave regularly. In these cases, one usually ascribes the first class a special value α, modifying equation (1) as

$$P_x = \begin{cases} \alpha, x = 1 \\ \frac{(1-a)x^{(a+b\ln x)}}{T}, x = 2, 3, \ldots, (n) \end{cases} \tag{4}$$

where

$$T = \sum_{j=2}^{n} j^{-(a+b\ln j)}, a, b\varepsilon\mathfrak{R}, 0 < \alpha < 1.$$

Distributions (3) and (4) are called Zipf-Alekseev distributions. If n is finite, equation (4) is called a Right truncated modified Zipf-Alekseev distribution. In our project, we used the Altmann-Fitter software to fit the model to the investigated data (Altmann-Fitter, 2013). In addition, NLREG 6.3 was used to examine the relation between parameters in probability distribution models.

3. Results and discussions

The results of previous corpus-based researches and psychological experiments have indicated that human languages have a tendency toward dependency distance minimization (Liu et al., 2017). This tendency suggests that, although human languages differ in pronunciation, vocabulary, grammar, etc., their syntax may be constrained by universal mechanisms, and their evolution may have a universal model (Lu & Liu, 2016). Dependency distance, which is defined as the linear distance between two words that are syntactically related, can reflect the comprehension difficulty of syntactic structure (Liu, 2008). Therefore, the *DDM* is considered as a result of the constraint of the cognitive mechanism and the effect of 'the principle of least effort' on syntactic structure. Moreover, humans tend to avoid the use of long-distance dependency distances to reduce cognitive cost. As a result, dependency distance distribution may present a certain pattern.

Figure 2 shows the distributions of dependency distances of nine grades and in four sub-corpora from WSJ. It can be seen that the 13 distribution curves are all concave down. Numerous previous studies have corroborated that the distribution of dependency distances of natural languages can fit well the exponential distribution or power law distribution (Ferrer-i-Cancho, 2004; Liu, 2007; Liu, Jiang & Liu, 2015; Lu & Liu, 2016). Therefore, we fitted different exponential distribution and power law distribution models to the dependency distances of each grade and of WSJ. The quantities of various dependency relations related to dependency distances at each grade and in WSJ were computed with the quantitative linguistic software of Altmann-Fitter to determine the probability distribution models suitable for dependency distances of each grade and WSJ. The only problem is the chi-square goodness-of-fit test, whose reliability is increasingly doubted by linguists (Wang, 2012; Mačutek & Wimmer, 2013). It is inadequate if the sample size is very large, but no criterion as to what 'large' means has been given so far. Therefore, we opt for the determination coefficient R^2, regarding it as more effective and reliable.

From only the values of the coefficient of determination R^2 among 13 groups (nine grades and four WSJ groups) of dependency distances, the dependency distances of these groups reveal the following probability distribution: Waring (b, n); extended logarithmic (θ, α); mixed geometric (q_1, q_2, α); right truncated Waring (b, n); and right truncated modified Zipf-Alekseev (a, b; $n = x$-max, α fixed). The mean values of R^2 of the 13 groups with suitable models were calculated and tabulated as shown in Table 2.

It can be seen that although Chinese EFL learners' English proficiency is not as high as that of native English speakers, the distributions of dependency distances in their English writings are similar to those of English native speakers, as

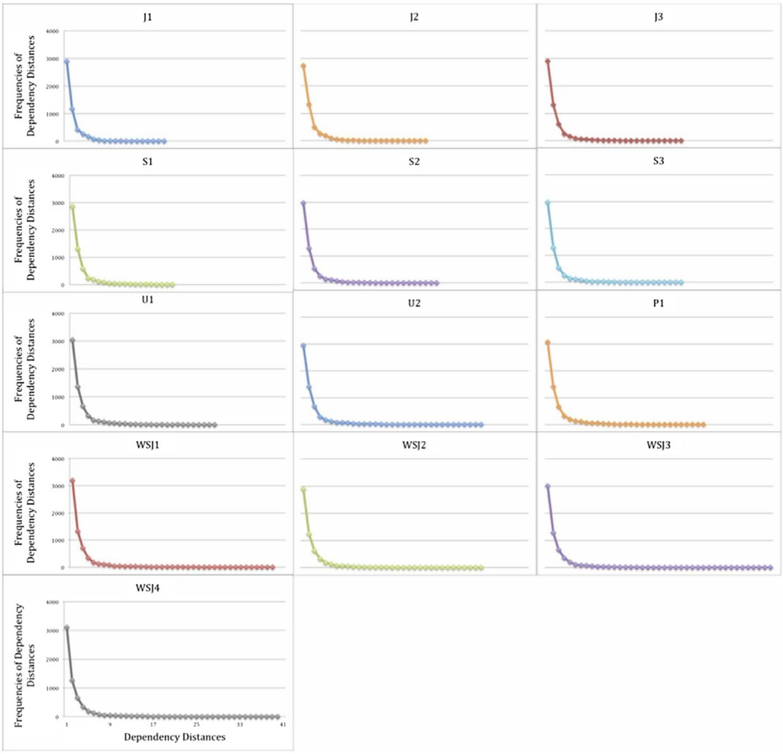

Figure 2. The distributions of dependency distances of each grade and of WSJ.

Table 2. The determination coefficient R^2 of fitting of different models to the dependency distances of each grade and WSJ.

Group	Right truncated modified Zipf-Alekseev	Waring	Right truncated Waring	Extended logarithmic	Mixed geometric
J1	0.9988	0.9990	0.9990	0.9994	0.9991
J2	0.9975	0.9976	0.9975	0.9997	0.9972
J3	0.9991	0.9993	0.9993	0.9947	0.9996
S1	0.9986	0.9981	0.9980	0.9991	0.9989
S2	0.9970	0.9982	0.9980	0.9994	0.9983
S3	0.9979	0.9990	0.9989	0.9999	0.9993
U1	0.9982	0.9991	0.9988	0.9991	0.9997
U2	0.9963	0.9976	0.9974	0.9988	0.9988
P1	0.9973	0.9989	0.9987	0.9994	0.9995
WSJ1	0.9990	0.9994	0.9993	0.9985	0.9992
WSJ2	0.9991	0.9997	0.9997	0.9991	0.9997
WSJ3	0.9990	0.9997	0.9995	0.9988	0.9991
WSJ4	0.9995	0.9996	0.9996	0.9987	0.9987

the dependency distances in learners' English writings fit well some of the same probability distribution models as native English speakers. The fitting results to the right truncated modified Zipf-Alekseev give a positive answer to Question 1.

These results indicate that although Chinese EFL leaners' English proficiency is not as high as that of native English speakers, the distributions of dependency distances in their English writings can fit certain probability distributions that the writings of native English speakers fit, which indicates the universality of human natural languages. Furthermore, it can be inferred that second language learners' syntactic acquisition process is always constrained by the tendency of *DDM*, which makes the probability distribution of dependency distances in Chinese EFL learners' English writings follow a specific regularity in different language acquisition periods. And this can be attributed to the effect of the constraint of working memory capacity and 'the principle of least effort' on syntactic structure.

However, the fitting results of these distributions to different grades and to WSJ are not the same. The different fitting results of the same probability distribution model may be the manifestation of different language proficiency. Therefore, we choose the right truncated modified Zipf-Alekseev distribution model to observe whether the parameters in the probability distribution model can reflect second language learners' and native speakers' English language proficiency.

There are, altogether, four parameters in the right truncated modified Zipf-Alekseev distribution: a, b, n and α. The statistic results of the parameters are presented in Table 3. To obtain more accurate statistics of WSJ, we calculated the mean values of parameters of the four random sub-corpora of WSJ to be: $a = 0.7918$; $b = 0.4532$; $n = 37$; $\alpha = 0.5045$.

For a clearer illustration of the variations of parameters (a, b, α), see Figure 3. It is clearly shown that a increases along with the increase of grades and gradually approaches 0.7918 of WSJ. b falls along with the increase of grades and gradually reaches 0.4232 of WSJ. The value of α, ranging between 0.49 and 0.58, doesn't present an increasing or decreasing tendency.

Table 3. Fitting the right truncated modified Zipf-Alekseev to the dependency distances to different grades and to WSJ.

	Parameters								
Group	a	b	α	X^2	$P(X^2)$	DF	N	C	R^2
J1	0.3372	0.9325	0.5755	38.0044	0.0001	11	19	0.0076	0.9988
J2	0.1164	0.9183	0.5185	43.8182	0.0000	13	23	0.0083	0.9975
J3	0.4027	0.7734	0.5328	23.7847	0.0486	14	21	0.0043	0.9991
S1	0.5027	0.6783	0.5193	43.8244	0.0004	17	25	0.0078	0.9986
S2	0.3878	0.6891	0.5162	68.1798	0.0000	15	20	0.0123	0.997
S3	0.3784	0.7323	0.5364	50.9597	0.0000	16	25	0.0092	0.9979
U1	0.5495	0.5683	0.4999	64.4543	0.0000	21	28	0.0106	0.9982
U2	0.4194	0.6150	0.4913	109.8569	0.0000	22	33	0.0183	0.9963
P1	0.4578	0.5853	0.4952	78.4422	0.0000	22	29	0.0127	0.9973
WSJ1	0.8803	0.4235	0.5079	63.6525	0.0001	27	34	0.0101	0.999
WSJ2	0.6724	0.5030	0.5016	43.2867	0.0130	25	33	0.0075	0.9991
WSJ3	0.7255	0.4621	0.4981	54.3986	0.0020	28	41	0.009	0.999
WSJ4	0.8889	0.4242	0.5105	62.5308	0.0002	28	40	0.0102	0.9995

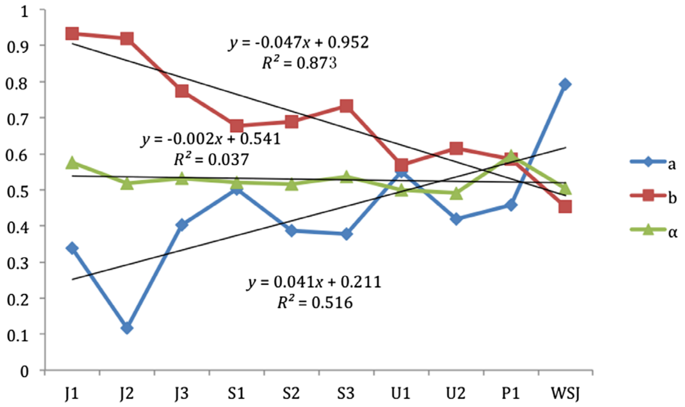

Figure 3. The variations of parameters (*a*, *b*, *a*) of the right truncated modified Zipf-Alekseev, fitting the dependency distances of different grades and of WSJ.

To explore whether there is a correlation between Chinese EFL learners' English proficiency and the parameters *a* and *b*, we performed the correlation analysis between the grades and the parameters: *a*, *b* and *α*. The results of the data analysis show that parameter *b* is highly correlated with the grades, which means that along with the increase of grades or learners' English proficiency, parameter *b* decreases significantly. The regression equation is $y = -0.047x + 0.952$, $F(1, 8) = 54.969$, $p < 0.01$, $R^2 = 0.873$. And parameter *a* is moderately correlated with the grades. The regression equation is $y = 0.041x + 0.211$, $F(1, 8) = 8.528$, $p = 0.019$, $R^2 = 0.516$. However, there is no correlation between the parameter *α* and the grades ($R^2 = 0.037$, $p > 0.05$). Through the correlation analysis, we found that along with the increase of learners' grades and the improvement of their English proficiency, the parameter *a* presents a significantly increasing tendency, while *b* decreases significantly. It can be concluded that the values of parameters *a* and *b* in the right truncated modified Zipf-Alekseev can well reflect the English proficiency of Chinese EFL learners. The *a* of Chinese EFL learners of high English language proficiency is bigger than that of low language proficiency; on the contrary, the *b* of Chinese EFL learners of high English language proficiency is smaller than that of low language proficiency, which helps answer Question 2, that the parameters in the Zipf-Alekseev distribution well reflect second language learners' language proficiency at different learning stages. The values of *a* and *b* can be good indicators of second language learners' language proficiency. Moreover, the variations of these parameters of the right truncated modified Zipf-Alekseev distribution indicate that, along with the increase of Chinese EFL learners' grade, or English proficiency, the probability distribution of dependency distances in their English writings gradually gets closer to that of native English speakers.

Furthermore, we adopted the analytical method used in Popescu's (2014) research to investigate the relationships between parameters *a* and *b*.

Figure 4 shows that the relationship of parameters *a* and *b* fits well the power function $y = 1.0500–0.7754*x^{0.9297}$ ($R^2 = 0.7482$), which agrees well with Popescu's (2014) finding that capturing the relationship between parameters *a* and *b* in the Zipf-Alekseev formula always yields a very regular function that can be expressed by the power function $y = k+mx^b$ or linearly, i.e. with $b = 1$. The statistical analytical results further verify Popescu's (2014) conclusions that in the unified model there are merely differences in the parameters, and the parameters themselves are part of a dynamic system displaying self-regulation, further illustrating that the length of any unit in language abides by the same regularity which can be considered a universality of human language.

To observe more clearly how the distribution of dependency distances in Chinese EFL learners' English writings gets closer to that of native English speakers, along with the increase of grades, we took the logarithm of the frequencies of different dependency distances at each grade and in WSJ to reduce the number of parameters in the right truncated modified Zipf-Alekseev and obtained Figure 5.

From Figure 5, we can see that the distribution of the logarithm of the dependency distances in Chinese EFL learners' English writings has a tendency to approach that of native English speakers. Then, we did the linear fitting of the logarithm of frequencies of dependency distances at each grade and in WSJ and obtained the variations of parameters of linear fitting results (Figures 6, 7 and 8). The formula of linear fitting is:

$$f(x) = ax + b \quad x = 1, 2, 3 \cdots, n$$

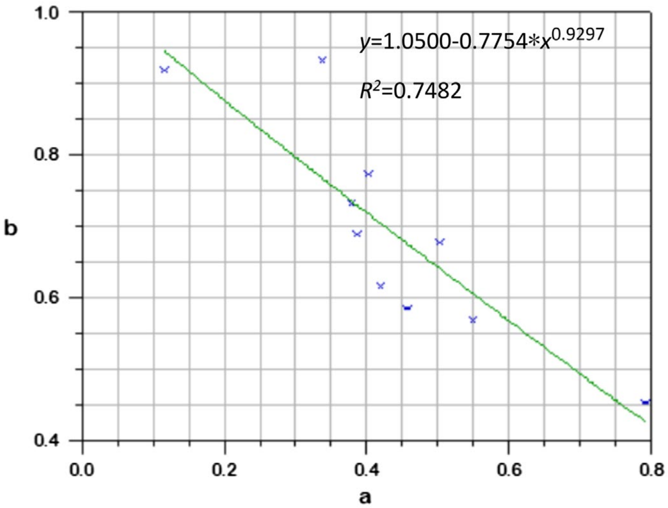

Figure 4. The relation of parameters *a* and *b*.

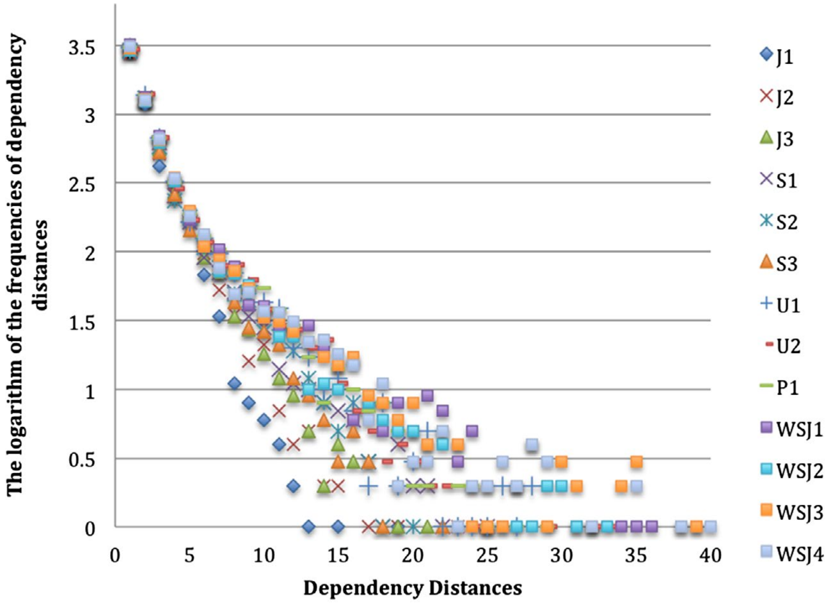

Figure 5. The variations of the logarithm of the different dependency distances of each grade and of WSJ.

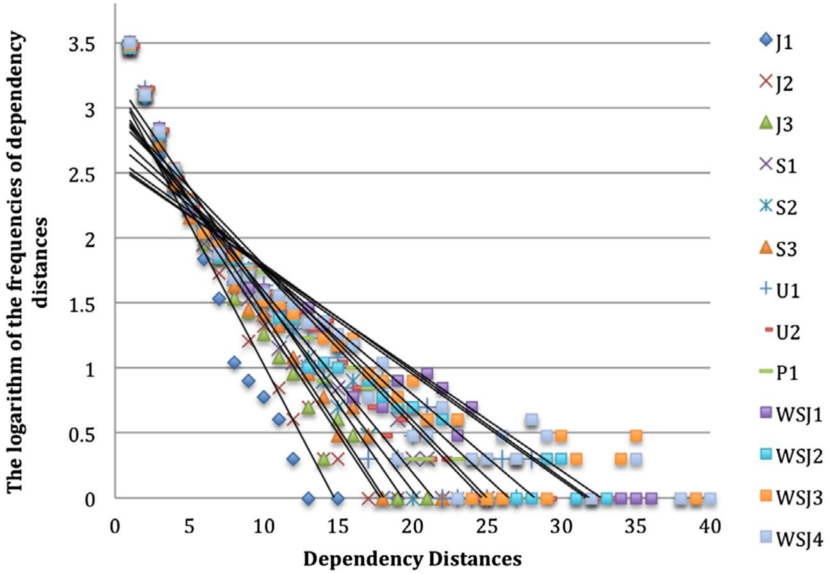

Figure 6. The linear fitting results of the logarithm of frequencies of dependency distances of each grade and of WSJ.

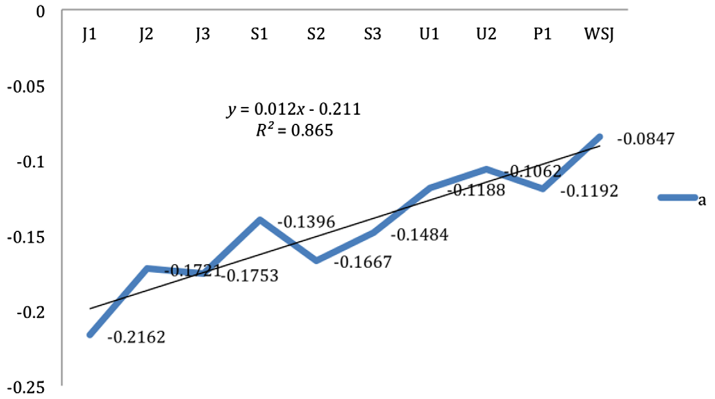

Figure 7. The variations of parameter *a* in linear fitting to each grade and to WSJ.

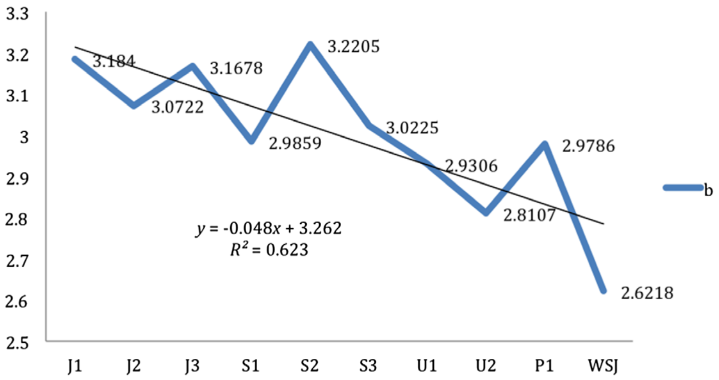

Figure 8. The variations of parameter *b* in linear fitting to each grade and to WSJ.

where *a* and *b* are two parameters. To obtain more accurate statistics of WSJ, we calculated the mean values of parameters of the four random sub-corpora of WSJ.

As presented in Figures 7 and 8, parameters *a* and *b* show, respectively, increasing and decreasing tendencies. Parameter *a* gradually increases from –0.2162 at J_1 to –0.1192 at P_1, with several falls and finally approaches –0.0847 in WSJ. And, parameter *b* gradually falls from 3.184 at J_1 to 2.9786 at P_1, with several increases and finally approaches 2.6218 in WSJ. We further did the correlation analysis between the grades and the parameters *a* and *b*. The results of the data analysis show that parameter *a* is highly correlated with the grades, which means that along with the increase of grades or learners' English proficiency, parameter *a* decreases significantly. The regression equation is $y = 0.012x - 0.211$, $F(1, 8) = 51.474$, $p < 0.01$, $R^2 = 0.865$. And, the parameter *b* is moderately correlated with the grades. The regression equation is $y = -0.048x + 3.262$, $F(1, 8) = 13.237$, $p = 0.007$, $R^2 = 0.623$.

The correlation analysis shows that along with the increase of learners' grades and the improvement of their English proficiency, the parameter *a* presents a significant decreasing tendency, while *b* increases significantly, proving our hypothesis that the values of parameters *a* and *b* in the linear fitting results of the logarithm of frequencies of dependency distances can also reflect well the English proficiency of Chinese EFL learners at different learning stages.

The variations of parameters of the fitting results of the right truncated modified Zipf-Alekseev and the linear fitting results of the logarithm of frequencies of dependency distances at each grade and in WSJ have suggested that along with the increase of grades, the probability distribution of dependency distances in Chinese EFL learners' English writings gets closer to that of native English speakers.

To explore whether high-level Chinese EFL learners present the same parameters in the probability distribution of dependency distance as those of English native speakers (Question 3), we need to compare the parameters of the right truncated modified Zipf-Alekseev fitting and linear fitting at P_1 (the representations of the high-level English learners in China for the current study) and in WSJ. The parameters (a, b, n, α) of the right truncated modified Zipf-Alekseev fitting at P_1 and in WSJ are respectively 0.4578, 0.5853, 29, 0.5952 and 0.7918, 0.4232, 37, 0.5045. In addition, the parameters of linear fitting (a, b) of the logarithm of frequencies of dependency distances at P_1 and in WSJ are respectively −0.1192, 2.9786 and −0.0847, 2.6218. In terms of the values of parameters, the probability distribution of the dependency distance of high-level Chinese EFL learners is not the same as that of native English speakers. Therefore, through comparing the parameters of the right truncated modified Zipf-Alekseev fitting and linear fitting to P_1 and to WSJ, the high-level Chinese EFL learners don't present the exact same probability distribution of dependency distance as that of native English speakers.

Through observing the variations of parameters, we also found that the parameters of the writings of university students don't present a steadily increasing tendency towards native English speakers, but present fluctuation changes, which indicates that the probability distribution of the dependency distances of Chinese EFL leaners' English writings does not continue developing towards that of native English speakers after entering universities. The stagnation of the probability distribution of dependency distances at university level can be attributed to the interlanguage fossilization of Chinese high-level EFL learners. Interlanguage fossilization is a phenomenon of SLA in which second language learners develop and retain a linguistic system, or interlanguage, that is self-contained and different from both the learner's first language and the target language. Second language learners' language system, called 'interlanguage' (Selinker, 1972), is a dynamic language system between the native and target languages. According to Corder (1978), this temporary and changing grammatical system, interlanguage, which is constructed by the learner, approximates the grammatical system of the target language. After Chinese EFL learners enter universities and continue their English learning, due to the learning environment, teaching approaches (Yi & Rui, 1998), the lack of motivation, and so on, their English proficiency does not keep improving and enters a plateau period. In this plateau period of English learning, or in the case of interlanguage fossilization, Chinese EFL learners' syntactic structural system does not seem to develop anymore. Moreover, under the pressure of *DDM*, or the constraint of working memory capacity and the effect of 'the principle of least effort' on syntactic structure, the development of syntactic structures is restricted. Therefore, we hold the view that Chinese EFL learners' interlanguage fossilization after entering universities and the pressure of *DDM* cause the stagnant development of the *MDD* and the distribution of dependency distances.

4. Conclusions and implications

Our data, derived from the cross-sectional dependency treebank of Chinese EFL learners' English writings from nine grades, suggest that the parameters of probability distribution of dependency distance can measure well the language proficiency of second language learners. The probability distribution of the dependency distance of second language learners' interlanguage fits the Zipf-Alekseev distribution well, which helps show that length distribution of many linguistic units follows certain regularities, in this case, the Zipf-Alekseev distribution. Second language learners' language system also follows certain distribution patterns that other human natural languages follow, which is a linguistic universality caused by human working memory constraint.

The fitting results of the right truncated modified Zipf-Alekseev and the linear fitting results of the logarithm of frequencies of dependency distances to each grade and to WSJ have suggested that the values of parameters can well reflect second language learners' proficiency. In the Zipf-Alekseev distribution, the parameters *a* and *b* can be good indicators of second language learners' language proficiency. In addition, along with the increase of second language learners' language proficiency, the parameters in the probability distribution of dependency distances in their writings get closer to those of native English speakers. However, the high-level second language learners don't present the exact same parameters in the probability distribution of dependency distance as those of native speakers, which means learners' language proficiency is not as high as that of native English speakers.

What's more, it is found that the parameters in the probability distribution of dependency distances of second language learners' writings are stabilized after learners entered universities. This may be the result of both their interlanguage fossilization after entering universities and the pressure of *DDM*.

The current study corroborates that quantitative linguistic research methods can be well utilized in language research, including research in SLA. It demonstrates the possibility of adopting quantitative linguistics and dependency grammar to discover the relationship between language and human cognitive mechanism by probing into the process of SLA. In addition, it verifies that second language learners' interlanguage at different learning stages can be described by different parameters under a unified theoretical framework. Undoubtedly, more evidence is needed to test whether our findings are applicable to other second language learners with different native languages and target languages. Moreover, by doing this, it can be shown that, through mathematical models and formulas, quantitative linguistic research methods can well reflect the universalities and peculiarities of human languages.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work is supported by the National Social Science Foundation of China [grant number 17AYY021]; the Fundamental Research Funds for the Central Universities.

ORCID

Jinghui Ouyang http://orcid.org/0000-0002-2980-9511
Jingyang Jiang http://orcid.org/0000-0003-3019-3918

References

Altmann-Fitter. (2013). Altmann-Fitter user guide. The third version. Retrieved August 29, 2016, from http://www.ram-verlag.eu/wp-content/uploads/2013/10/Fitter-User-Guide.pdf

Brown, H. D. (1994). *Principles of language learning and teaching* (3rd ed.). New Jersey: Prentice Hall Regents.

Chen, H., & Liu, H. (2016). How to measure word length in spoken and written Chinese. *Journal of Quantitative Linguistics, 23*(1), 5–29.

Corder, S. P. (1978). Language-learner language. In J. C. Richards (Ed.), *Understanding second and foreign language learning* (pp. 71–92). Rowley, MA: Newbury House.

Eger, S. (2013). A contribution to the theory of word length distribution based on a stochastic word length distribution model. *Journal of Quantitative Linguistics, 20*(3), 252–265.

Ferrer-i-Cancho, R. (2004). Euclidean distance between syntactically linked words. *Physical Review E, 70*, 056135.

Heringer, H., Bruno, S., & Rainer, W. (1980). *Syntax: Fragen, Lösungen, Alternativen* [Syntax: Issues, Solutions, Alternatives]. Munich: Wilhelm Fink Verlag.

Hřebíček, L. (1996). Word associations and text. In P. Schmidt (Ed.), *Glottometrika 15* (pp. 12–17). Trier: Wissenschaftlicher Verlag Trier.

Hudson, R. (1995). Measuring syntactic difficulty. Unpublished paper. Retrieved October 4, 2016 from http://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf

Hudson, R. (2007). *Language networks: The new word grammar*. Oxford: Oxford University Press.

Hudson, R. (2010). *An Introduction to Word Grammar*. Cambridge: Cambridge University Press.

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.

Jiang, J., & Ouyang, J. (2017). Dependency distance: A new perspective on the syntactic development in second language acquisition. *Phys Life Rev, 21*, 209–210. doi:10.1016/j.plrev.2017.06.018

Kalimeri, M., Constantoudis, V., Papadimitriou, C., Karamanos, K., Diakonos, F. K., & Papageorgiou, H. (2015). Word-length entropies and correlations of natural language written texts. *Journal of Quantitative Linguistics, 22*(2), 101–118.

Liu, H. (2007). Probability distribution of dependency distance. *Glottometrics, 15*, 1–12.

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

Liu, H. (2009a). Probability distribution of dependencies based on Chinese dependency treebank. *Journal of Quantitative Linguistics, 16*(3), 256–273.

Liu, H. (2009b). *Dependency grammar: From theory to practice*. Beijing: Science Press.

Liu, H., Hudson, R., & Feng, Z. (2009). Using a Chinese treebank to measure dependency distance. *Corpus Linguistics and Linguistic Theory, 5*(2), 161–174.

Liu, H., Xu, C., & Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Phys Life Rev, 21*, 171–193. doi:10.1016/j.plrev.2017.03.002

Lu, Q., & Liu, H. (2016). Does dependency distance distribute regularly? *Journal of Zhejiang University (Humanities and Social Sciences Online Edition), 2016*(4), 1–14.

Mačutek, J., & Wimmer, G. (2013). Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics, 20*, 227–240.

Narisong, Jiang, J., & Liu, H. (2014). Word length distribution in Mongolian. *Journal of Quantitative Linguistics, 21*(2), 123–152.

Nearysundquist, C. A. (2016). Syntactic complexity at multiple proficiency levels of L2 German speech. *International Journal of Applied Linguistics, 27*(1), 242–262.

Nivre, J. (2006). *Inductive dependency parsing*. Dordrecht: Springer.

Pande, H., & Dhami, H. S. (2012). Model generation for word length frequencies in texts with the application of Zipf's order approach. *Journal of Quantitative Linguistics, 19*(4), 249–261.

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.

Popescu, I.-I., Best, K.-H., & Altmann, G. (2014). *Unified modeling of length in language (= studies in quantitative linguistics 16).* Lüdenscheid: RAM-Verlag . ISBN 978-3-942303-26-2.

Pustet, R., & Altmann, G. (2005). Morpheme length distribution in Lakota. *Journal of Quantitative Linguistics, 12*(1), 53–63.

Selinker, L. (1969). Language transfer. *General Linguistics, 9*, 67–92.

Selinker, L. (1972). Interlanguage. *IRAL, 10*(3), 209–231.

Sigurd, B., Eeg-Olofsson, M., & van de Weijer, J. (2004). Word length, sentence length and frequency - Zipf revisited. *Studia Linguistica, 58*(1), 37–52.

Strauss, U., & Altmann, G. (2006). *Diversification - laws in quantitative linguistics*. Retrieved March 12, 2007, from http://www.uni-trier.de/uni/fb2/ldv/lql_wiki/index.php/Diversification

Tesnière, L. (1959). *Eléments de la syntaxe structurale* [Elements of Structural Syntax]. Paris: Klincksieck.

Wang, L. (2012). Word length in Chinese. *Issues in Quantitative Linguistics, 3*, 39–53. Lu̇dendscheid: Ram Verlag.

Yi, F., & Rui, Y. (1998). *Toward the 21st century EFL teaching and learning*. Chongqing: Chongqing Press.

Yngve, V. (1960). A model and hypothesis for language structure. *Proceedings of the American Philosophical Society, 104*(5), 444–466.

Yngve, V. (1996). *From grammar to science: New foundations for general linguistics*. Amsterdam & Philadelphia: John Benjamins.

Zipf, K. G. (1936). *The psychobiology of language*. London: Routledge.