Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-022-01447-3 OPEN

Dependency distance minimization: a diachronic exploration of the effects of sentence length and dependency types

Xueying Liu[1], Haoran Zhu[1] & Lei Lei [2]✉

Dependency distance is regarded as an index of memory load and a measure of syntactic difficulty. Previous research has found that dependency distance tends to minimize both synchronically and diachronically due to the limited resource of working memory. However, little is known concerning the effects of different dependency types on the dependency distance minimization. In addition, previous studies showed inconsistent results on the anti-minimization of dependency distance in shorter sentences. Hence, a more fine-grained investigation is needed on the diachronic change of dependency distance with shorter sentences such as those of three or four words. To address these issues, this study intends to explore the diachronic change of dependency distance in terms of two variables, i.e., dependency types and sentence length. Results show that anti-minimization does exist in short sentences diachronically, and sentence length has an effect on diachronic dependency distance minimization of dependency types. More importantly, not all dependency types present a decreasing trend, while only nine types of dependency relations are responsible for the dependency distance minimization. Possible explanations for the findings are offered.

[1] Huazhong University of Science and Technology, Wuhan, China. [2] Shanghai International Studies University, Shanghai, China. ✉email: leileibama@outlook.com

Introduction

Dependency distance (Heringer et al., 1980; Liu, 2007) or dependency length (Temperley, 2007, 2008; Futrell et al., 2015) is defined as the linear distance between two syntactically related words. It is considered as a predictor of syntactic difficulty (Liu et al., 2017). More importantly, it is hypothesized to serve as an index of the working memory load imposed on language processing (Liang et al., 2017; Liu et al., 2017). Dependency distance has attracted considerable attention and has been widely explored in areas such as typological analysis of languages (Liu, 2010; Liu and Xu, 2011), second language studies (Jiang and Ouyang, 2017; Ouyang and Jiang, 2018), and code-switching and interpreting research (Wang and Liu, 2013, 2016; Jiang and Jiang, 2020).

Owing to the limited nature of human working memory (Miyake and Shah, 1999), it is hypothesized that a shorter dependency distance is preferred to reduce the cognitive burden and secure the efficient processing of language information (Yngve, 1960; Liu, 2008). The foregoing hypothesis seems supported by the tendency of dependency distance minimization (DDM) found in natural languages (Liu, 2008; Futrell et al., 2015). In addition, the DDM is hypothesized as a human language universal (Liu et al., 2016) in order to adapt to the limited resource of human memory (Lei and Jockers, 2020). For example, Liu (2007) found that, compared with the mean dependency distance (MDD) of two artificial languages, the overall MDD of Chinese is much shorter. Similarly, Gildea and Temperley (2010) found that the average dependency length (ADL) of German and English is significantly shorter than the ADL of the randomly generated samples of the two languages. More compelling evidence has also been obtained in large-scale cross-language studies such as Liu (2008) and Futrell et al. (2015). Liu (2008) investigated the MDDs of 20 languages, and found that all the examined languages exhibit shorter MDD compared to that of their randomly generated counterparts. A much similar finding was also observed across 37 languages in Futrell et al. (2015).

While all the aforementioned studies are synchronic ones, Lei and Wen (2020) extended the exploration of the DDM tendency more from a diachronic perspective. They found a general downward trend of dependency distance in the *State of the Union Addresses* across the past 227 years. More importantly, they also found that shorter sentences with ten or fewer words presented a tendency toward DDM. This finding seemed contradictory to that of Ferrer-i-Cancho and Gómez-Rodríguez (2021), which found the phenomenon of anti-dependency distance minimization (anti-DDM) in short sequences. More specifically, Ferrer-i-Cancho and Gómez-Rodríguez (2021) employed a simple binomial test to analyze the principles of DDM from the perspective of languages of different families. They found that the DDM might not work in shorter sentences with three or four words. The reason may be that in short sentences, DDM is more likely to be challenged by its competitor, i.e., the principle of surprisal minimization or predictability maximization (Levy, 2008; Ferrer-i-Cancho, 2017). The foregoing two principles of word order are in conflict when the placement of a governor and its dependents is considered. Based on the principle of surprisal minimization, the governor should be placed at either the beginning of a sentence (when the dependent is the target of predictability maximization) or the end of a sentence (when the governor is the target), while it should be placed at the center based on the principle of DDM (Ferrer-i-Cancho, 2017, see Fig. 1). Accordingly, the former placement of word order under the principle of surprisal minimization would maximize the dependency distance (Ferrer-i-Cancho, 2014). As the example in Fig. 1 shows, the MDD of the sentence according to surprisal minimization is 1.5 and that according to DDM is 1. However, a decreasing trend of dependency distance in shorter sentences was found in Lei and Wen (2020), which indicated that DDM also plays a major role in short sequences. Owing to the inconsistency of results in Lei and Wen (2020) and Ferrer-i-Cancho and Gómez-Rodríguez (2021), more diachronic investigation into the DDM, particularly on a more fine-grained scale of sentence length, is needed.

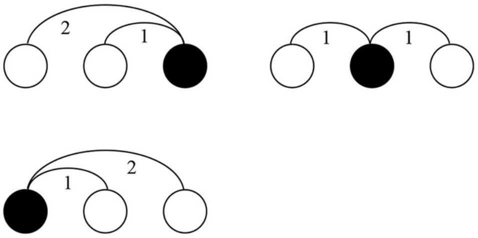

Fig. 1 Optimal sequential placement of a governor and its dependents when the length of a sentence is three. The figures on top left and bottom left indicate the optimal placements based on surprisal minimization and the one on top right shows the optimal placements based on DDM.

Another factor that may affect the MDD of a language is dependency types or syntactic structures, since different dependency types require different cognitive resources in language processing (Jiang and Jiang, 2020). For example, Liu et al. (2009a) manually specified the dependency types in a Chinese news treebank (*xinwen lianbo*) and measured the MDDs of distinct dependency types in this treebank. The results showed that the MDDs of some dependency types (e.g., clausal relation, sentential object) are larger than that of others (e.g., aspect adjunct, classifier complement). This finding showed an interaction between dependency distance and dependency types. In a more recent research, differences in the MDD of individual dependency types were also observed in spoken and written French (Poiret and Liu, 2020). Based on two treebanks, Poiret and Liu (2020) found that the MDDs of major dependency types (e.g., subject, oblique object) are greater in written French than in spoken French. As a result, further investigation into the diachronic change of dependency types is needed to examine their effects on DDM. One point to be noted here is that the dependency type *root*, i.e., the main verb of a sentence, is often ignored in previous studies (Liu, 2008; Ouyang and Jiang, 2018). However, the position of *root* has been argued to play an important role in sentence comprehension and production (Lei and Jockers, 2020). Therefore, it should be of interest to explore whether root position has experienced any diachronic change or any process of minimization in light of its role in syntactic complexity.

To summarize, previous studies may be limited from at least two perspectives as follows. First, existing studies showed inconsistency on the anti-minimization of dependency distance in shorter sentences. Hence, a more fine-grained analysis is needed on the diachronic change of MDD in shorter sentences such as those of three or four words. Second, little is known concerning the effects of different dependency types on the dependency distance minimization. To address the foregoing issues, the present study aims to explore the possible diachronic minimization of dependency distance with the consideration of two variables, i.e., dependency types and sentence length. The following three research questions are to be addressed in the study.

Question 1: How is the mean dependency distance diachronically distributed in short sentences? Particularly, does anti-DDM exist in the diachronic change of dependency distance in short sentences?

Question 2: What are the diachronic trends of dependency distance of different dependency types in sentences with different lengths?

Question 3: Does dependency distance minimization exist in all types of dependency, or is the minimization confined to only certain dependency types?

Methods

Metrics. As discussed earlier, dependency distance (DD) refers to the linear distance between two syntactically related words, one serving as the governor and the other the dependent (Heringer et al., 1980; Hudson, 2010). Dependency analysis is thus based on these dependencies in a sentence or a longer text. Consider a string of words represented as $W_1 \ldots W_i \ldots W_n$. For W_a and W_b linked by a certain type of dependency relation, if W_a acts as the governor and W_b the dependent, the dependency distance between them is calculated by the subtraction of their positions (i.e., $a - b$) (Liu et al., 2009a). For instance, Example (1) is a sentence adapted from *The Annual Message to the Congress on the State of the Union* in 1950 by President Harry S. Truman. Table 1 summarizes the dependency types of this sentence. As is shown in the table, the dependent *race* and its governor *reached* are linked by the dependency relation *nsubj*, which refers to nominal subject. The position of *race* is 3 and that of *reached* is 5. Hence, the dependency distance between them is $5 - 3 = 2$.

(1) *The human race has reached a turning point.*

It should be noted that the dependency type *root* differs from other types, for the reason that its dependency distance could be defined in two ways. Previous studies usually defined the dependency distance of *root* as zero in the calculation of MDD (Liu, 2008; Ouyang and Jiang, 2018), since it only indicates the position of the main verb or predicate in a sentence and has no governor (Liu, 2008). In contrast, more recent research such as Lei and Jockers (2020) argued that root position plays an important role in processing sentences and accordingly should be considered in the calculation of dependency distance. The present study adopted both algorithms of MDD calculation for two reasons. First, it intends to investigate the trend of MDD of sentences particularly on a more fine-grained scale of sentence length based on the study of Lei and Wen (2020). Hence, the former was adopted to calculate the MDD of sentences for comparison purposes. That is, we excluded root position in the calculation of MDD at the sentential level. Second, the present study also aims to explore the diachronic change of various types of dependency relations. In light of the important role of root position as aforementioned, its diachronic trend was also examined in the present study.

The present study adopted the methods proposed by Liu et al. (2009a) to calculate the MDDs of sentences and different dependency types. The MDD at the sentential level can be defined as in Formula 1. As for a certain dependency type, the MDD is measured with Formula 2.

Formula 1

$$\mathrm{MDD(sentence)} = \frac{\sum_{i=1}^{m} \left|\mathrm{DD}_i\right|}{m}$$

Formula 2

$$\mathrm{MDD}\left(\text{dependency type}\right) = \frac{\sum_{i=1}^{n} \left|\mathrm{DD}_i\right|}{n}$$

In Formula 1, m denotes the total number of dependency types in a sentence and DD_i stands for the dependency distance value of the i-th dependency type. For comparison purposes, we excluded the dependency types, *root* and *punct* (punctuation) in the calculation of MDD of sentences (Lei and Wen, 2020). Regarding the sentence in Example (1), the DD of *det* (*race-3, the-1*) is $3 - 1 = 2$ and the DDs of *amod*, *nsubj*, *nsubj*, *aux*, *det*, *amod*, *dobj* are 1, 2, 1, 2, 1, and 3, respectively. Therefore, its mean dependency distance is calculated as follows:

$$\mathrm{MDD}\left(\text{Example } 1\right) = \frac{2+1+2+1+2+1+3}{7} = \frac{12}{7} \approx 1.714$$

In Formula 2, n stands for the number of specific dependency types in a text. Similar to Formula 1, DD_i refers to the dependency distance of the i-th one in the collection of dependency types in the text. Assume Example (1) as a text (i.e., a set of sentences). The DD of its first *amod* (*race-3, human-2*) in the text is $3 - 2 = 1$, and the second one (point-8, turning-7) is $8 - 7 = 1$. Hence, the MDD of the dependency type *amod* of the text is 1.

$$\mathrm{MDD}(amod)\frac{1+1}{2} = 1$$

Data processing. Following Lei and Wen (2020), the present study employed the *State of the Union Addresses* delivered by 43 presidents of the United States of America as the dataset. The dataset was used for its features such as availability, long time-span, and comparability in genre (Savoy, 2015). First, all the texts of the addresses are freely available from the American Presidency Project homepage (http://presidency.proxied.lsit.ucsb.edu). Second, the span of 227 years (from 1790 to 2017) is long enough to trace possible diachronic changes in MDDs in terms of sentences and different dependency types. Last, all the texts are of the same genre, i.e., political texts, which could minimize possible impact of genre issues. The dataset contains 2,012,440 words and 71,155 sentences, and was processed as follows.

First, the Stanford CoreNLP (3.9.2) parser (Manning et al., 2014) was utilized to analyze the syntactic dependencies of these texts. It should be noted that, although the annotations of the Stanford CoreNLP (3.9.2) parser are not completely accurate, it is,

Table 1 Dependency relations of Example (1).

Dependent id	Token	Governor id	Governor	Dependency type	Dependency distance
1	The	3	race	det	2
2	human	3	race	amod	1
3	race	5	reached	nsubj	2
4	has	5	reached	aux	1
5	reached	0	ROOT	root	0/5
6	a	8	point	det	2
7	turning	8	point	amod	1
8	point	5	reached	dobj	3
9	.	5	reached	punct	/

as shown in Lei and Wen (2020), reliable and the parsing errors will not significantly affect the results.

Second, we classified the annotated texts into groups of different sentence lengths with the consideration of two criteria. Firstly, following previous studies (Jiang and Liu, 2015; Lei and Jockers, 2020; Lei and Wen, 2020), we divided all sentences in the corpus into categories including sentences with 0–10, 11–20, 21–30, and over 31 words. Such a design makes the experiments comparable with forgoing research. Secondly, we further classified the 0–10 group into two levels, i.e., 0–4 and 5–10, in order to explore whether anti-DDM exists in diachronic change of dependency distance. Lei and Wen (2020) found that shorter sentences with 0–10 words presented a tendency toward DDM, which seems contradictory to the anti-minimization of dependency distance in short sentences found by Ferrer-i-Cancho and Gómez-Rodríguez (2021). However, there is disparity of classification in these two studies. In Ferrer-i-Cancho and Gómez-Rodríguez (2021), short sentences were defined as sentences with three or four words, while Lei and Wen (2020) used the interval of 0–10 words, which may in part mask some underlying tendencies. Thus, a more fine-grained classification is needed. In brief, the data were categorized into groups of sentences with 0–4, 5–10, 11–20, 21–30, and 31+ words.

Third, we coded two Python scripts to calculate the MDDs of sentences and different dependency types. As for the calculation of the MDDs of different dependency types, a total of 39 dependency types occurred in our data, and the MDDs of 38 types were calculated, with the dependency type *punct* excluded since it was not considered in the calculation of the MDD in most previous studies (e.g., Liu, 2008; Lei and Wen, 2020).

Last, trend analyses were performed to examine the time series data, namely, the MDD values at the sentential level and that of different dependency types from 1790 to 2017. Since the MDD values were not normally distributed (e.g., $p = 0.000$ for MDDs in sentences with 5–10 words), we followed Zhu and Lei (2022) and employed the Mann-Kendall trend test, a commonly used non-parametric test, to detect significant trends in time series, and Theil-Sen's slope estimator to calculate the corresponding rate of change. Both tests were implemented with the Python package pyMannKendall (Hussain and Mahmud, 2019; https://github.com/Coder2cdb/pyMannKendall).

Results

Trend of MDDs of sentences with different lengths. The general trends of the MDDs of sentences with different lengths are plotted in Fig. 2 and the results of the Mann-Kendall trend tests are reported in Table 2. Two findings of interest are to be summarized as follows.

First, it was found that longer sentences had larger values of MDD. This corroborates the point that sentence length affects the MDD of a sentence (Ferrer-i-Cancho, 2013; Jiang and Liu, 2015). Second, our fine-grained analyses of shorter sentences showed an uptrend for sentences of 0–4 words and a downtrend for sentences of 5–10 words. The finding partially confirmed those in previous studies such as Ferrer-i-Cancho and Gómez-Rodríguez (2021). That is, anti-DDM may exist in shorter sentences, at least from a diachronic perspective of dependency distance.

Trend of MDDs of different dependency types. The trends of MDDs of different dependency relations at various sentence lengths were summarized in Table 3, which presented some interesting findings.

First, the number of dependency types with decreasing MDDs in sentences of 0–4 words is one, at sentence levels of 5–10, 11–20, 21–30, the numbers are 9, 10, 11, respectively, and in sentences with 31 and more words, the number is 18. This result showed that longer sentences have more dependency types with a decreasing tendency of dependency distance.

Second, the MDDs of various dependency types in shorter sentences such as of 0–4 words and 5–10 words showed different tendencies across the examined years. While the position of *root* in sentences of 0–4 words showed an increasing tendency, that in sentences with 5–10 words presented a decreasing trend. In addition, in sentences of 0–4 words, only the type *case* showed a decreasing trend (no significant change was found for most dependency types). However, in sentences with 5–10 words, there were nine dependency types showing a downtrend. The downward trend was also found for other dependency types in longer sentences with 11 or more words (see Table 3). To summarize, the phenomenon of DDM was found for many types of dependency relations for longer sentences, particularly for those with five or more words. In fact, nine types of dependency relations were found consistently decreasing in sentences of five or more words across the examined years (i.e., *acl: relcl*, *aux*,

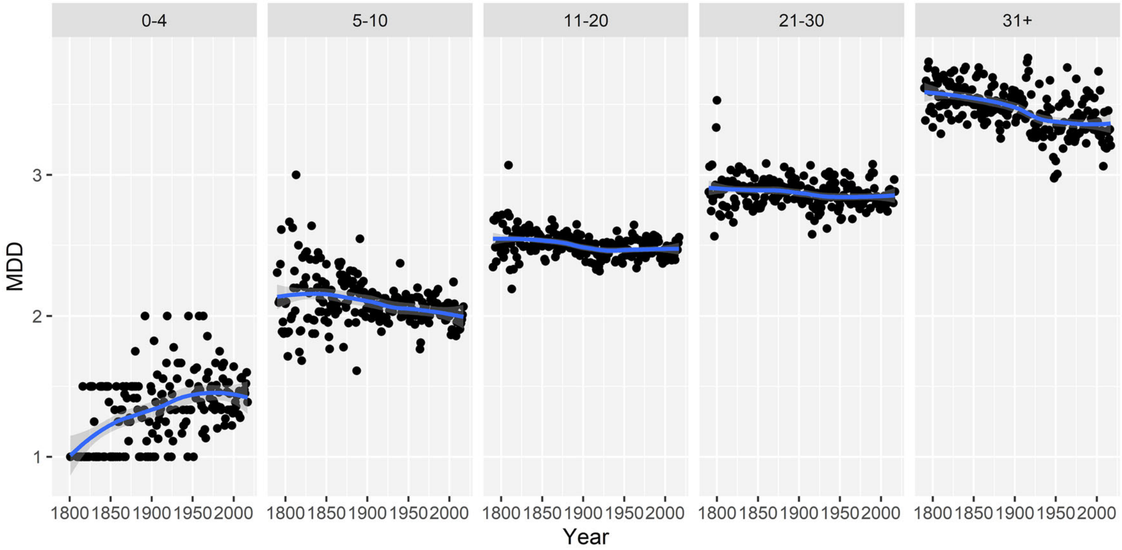

Fig. 2 Time series plots of MDDs at sentential level. The five columns show the trend of MDDs of sentences with different lengths. The blue lines represent fitting curves.

auxpass, *ccomp*, *mark*, *neg*, *nsubj*, *nsubjpass*, and *root*, see Fig. 3). However, the phenomenon of DDM was not found in shorter sentences such as those with four or fewer words. This finding provided evidence to the anti-DDM hypothesis proposed by Ferrer-i-Cancho and Gómez-Rodríguez (2021).

Last, there were also a number of dependency types that showed an increasing trend in each group. At sentence-length level 0–4, a significant uptrend of the MDD values were found for four dependency types (i.e., *root*, *appos*, *advmod*, and *cc*), which contributed to the increase of the general MDD in sentences of 0–4 words. In sentences of 5–10 words, ten types showed an increasing trend. At sentence levels of 11–20, 21–30, and 31+, the numbers of dependency types with increasing MDDs were 15, 16, and 8. Six types of dependency relations were found consistently increasing in sentences of five or more words across the examined years (i.e., *compound: prt*, *compound*, *amod*, *det*, *nmod:poss*, and *advmod*).

Table 2 Results of the Mann-Kendall trend tests on the MDDs of sentences with different lengths.

Length	Trend	*P*	Slope
0–4	Increasing	0.000	0.0018
5–10	Decreasing	0.000	−0.0007
11–20	Decreasing	0.000	−0.0004
21–30	Decreasing	0.004	−0.0003
31+	Decreasing	0.000	−0.0013

Discussion

The present study investigated the effects of sentence length and dependency types on dependency distance minimization over a long period of time. Below, we discuss the major findings of this study and its implications.

Roles of sentence length in diachronic minimization of dependency distance. This study showed that sentence length has an effect on diachronic DDM. Two points are worth discussing.

First, our results showed that longer sentences have a larger number of dependency types that showed a decreasing tendency of dependency distance. The reason for the finding was probably that the dependency types in longer sentences have longer

Table 3 Mann-Kendall trend tests results on MDDs of different dependency relations.

Type	0–4		5–10		11–20		21–30		31+	
	Trend	Slope	Trend	Slope	Trend	Slope	Trend	Slope	Trend	Slope
ROOT	↑	0.002	↓	−0.009	↓	−0.012	↓	−0.018	↓	−0.043
ccomp	—	0.000	↓	−0.008	↓	−0.009	↓	−0.014	↓	−0.025
nsubj	—	0.000	↓	−0.006	↓	−0.009	↓	−0.012	↓	−0.015
nsubjpass	—	0.000	↓	−0.003	↓	−0.009	↓	−0.009	↓	−0.014
aux	—	0.000	↓	−0.004	↓	−0.004	↓	−0.005	↓	−0.004
case	↓	0.000	↓	−0.001	↓	0.000	—	0.000	—	0.000
neg	—	0.000	↓	0.000[a]	↓	−0.001	↓	−0.001	↓	−0.002
auxpass	—	0.000	↓	0.000	↓	0.000	↓	0.000	↓	0.000
expl	—	0.000	↓	0.000	—	0.000	—	0.000	↓	−0.002
mark	—	0.000	—	0.000	↓	−0.003	↓	−0.004	↓	−0.010
acl:relcl	/	/	—	0.000	↓	−0.006	↓	−0.008	↓	−0.013
mwe	/	/	—	0.000	—	0.000	↓	0.000	↓	−0.001
dobj	—	0.000	—	0.000	—	0.000	↓	−0.001	↓	−0.002
csubjpass	/	/	—	0.750	—	0.043	—	−0.045	↓	−0.078
advcl	/	/	—	−0.003	—	−0.002	—	−0.001	↓	−0.020
xcomp	—	0.000	—	0.000	—	0.000	—	0.000	↓	−0.003
cop	—	0.000	—	0.000	—	0.000	—	0.000	—	0.000
det:predet	/	/	—	0.000	—	0.000	—	0.000	—	0.000
discourse	—	0.000	—	0.000	—	0.000	—	0.000	—	0.036
nmod.npmod	/	/	—	0.000	—	0.000	—	0.000	—	0.000
csubj	—	−1.000	—	0.000	—	−0.008	—	0.004	—	−0.021
appos	↑	0.113	—	0.000	↑	0.005	↑	0.004	↑	0.006
advmod	↑	0.000	↑	0.003	↑	0.004	↑	0.003	↑	0.002
cc	↑	0.045	↑	0.008	↑	0.005	↑	0.004	↓	−0.006
cc:preconj	/	/	—	0.000	—	0.000	↑	0.003	—	0.000
nmod:tmod	—	0.000	—	0.000	↑	0.014	↑	0.006	—	0.003
nmod	—	0.000	—	0.000	↑	0.001	↑	0.001	—	0.000
nummod	—	0.000	—	0.000	↑	0.000	↑	0.001	↑	0.001
parataxis	—	0.167	—	0.000	↑	0.019	↑	0.022	—	−0.011
iobj	—	0.000	↑	0.000	—	0.000	↑	0.000	—	0.000
acl	—	0.000	↑	0.000	↑	0.001	↑	0.001	↓	−0.001
compound:prt	—	0.000	↑	0.000	↑	0.000	↑	0.000	↑	0.000
compound	—	0.000	↑	0.000	↑	0.001	↑	0.000	↑	0.000
amod	—	0.000	↑	0.001	↑	0.001	↑	0.001	↑	0.001
det	—	0.000	↑	0.001	↑	0.001	↑	0.001	↑	0.001
nmod:poss	—	0.000	↑	0.001	↑	0.001	↑	0.001	↑	0.001
dep	—	0.007	—	0.002	↑	0.004	↑	0.003	—	0.000
conj	—	0.036	↑	0.005	↑	0.002	—	0.000	↓	−0.013

[a] For the types showing increasing/decreasing trend, the slope value 0.000 is close to zero, but not equal to zero.

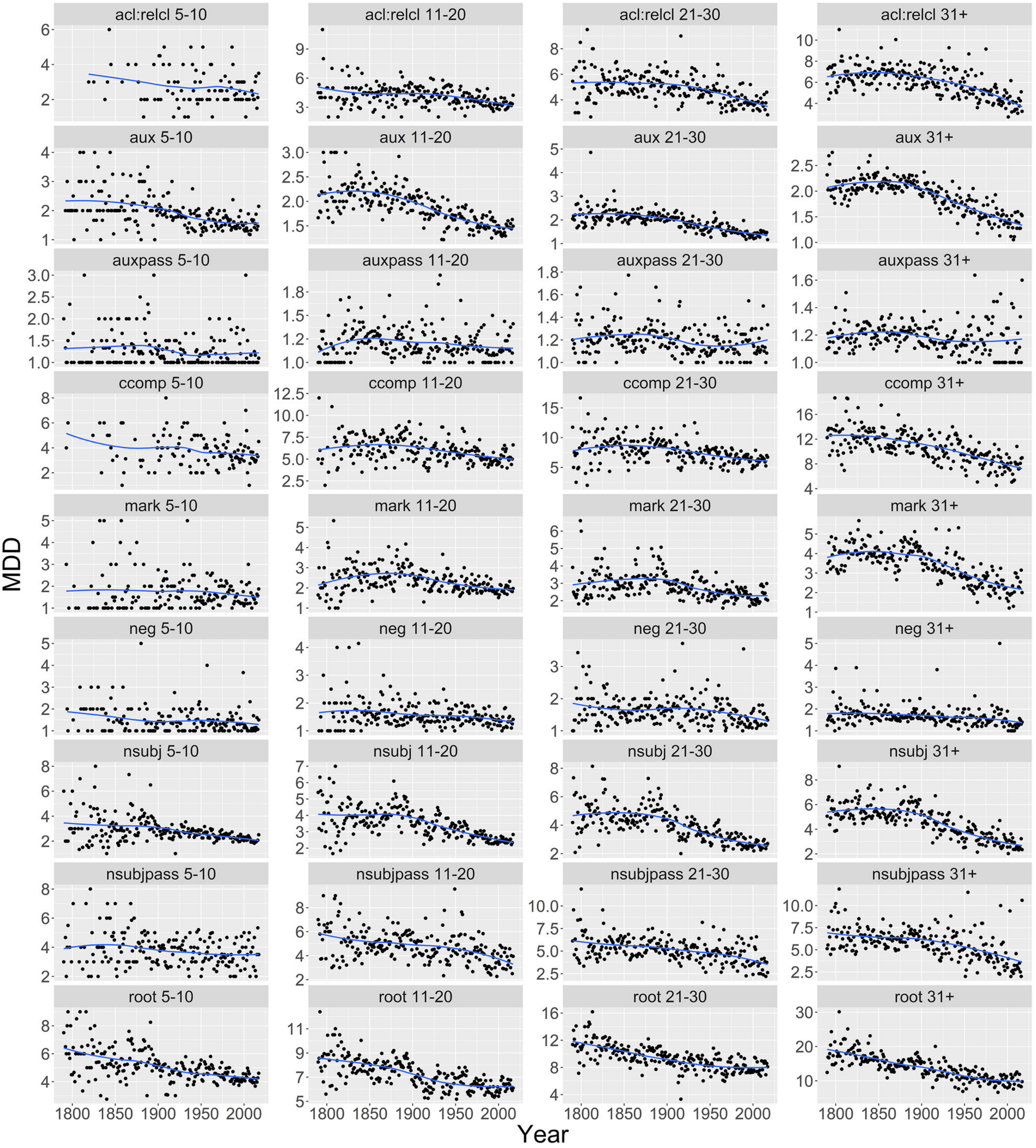

Fig. 3 Diachronic distribution of MDDs of the nine dependency relations presenting a downtrend in longer sentences. The columns from left to right displayed the MDDs of each relation at sentence levels of 5–10, 11–20, 21–30, and 31+.

dependency distances than those in short sentences. As dependency distances lengthen, more cognitive cost is required to process the language information. Hence, the pressure to reduce dependency distance is increasing. It should be noted that these decreasing trends are seemingly modest. One possible reason for it is the small value of mean dependency distance. As reported in Liu (2008), the MDDs of most languages are approximately 2 or 3 (2.543 for English). The minimum value of the MDD is one, and most values of the MDD fall between 2 and 4. Therefore, for such a small range of the MDD values, it seems difficult to have a change of large numbers. Hence, a significant change was detected in our study though the change of the values seemed modest. The other reason may be related to the fact that dependency distance is affected by the syntactic structure, particularly the word order, of a language (English in our case). A modest change of the word order is expected in a language over the past 200 years.

Second, the MDDs of sentences and root position showed an uptrend in shorter sentences of 0–4 words and a downtrend in sentences of five or more words across the examined years. In other words, the phenomenon of DDM did not apply to shorter sentences. This finding provided diachronic evidence to the anti-DDM hypothesis proposed by Ferrer-i-Cancho and Gómez-Rodríguez (2021). That is, in short sentences, DDM is not activated and its competitor, the principle of surprisal minimization is more likely to surface (Ferrer-i-Cancho and Gómez-Rodríguez, 2021). As mentioned earlier, DDM and surprisal minimization are two word order principles that have been found to be in conflict (Ferrer-i-Cancho, 2017). For example, according to surprisal minimization, if the verb appears last, more preverbal dependents could give more information to predict the verb's identity and location. However, DDM makes the opposite prediction, that is, the verb will be more difficult to process because it has more dependents, which require higher cost of working memory (Levy, 2008; Ferrer-i-Cancho, 2017). However, in short sentences of 0–4 words, there is no such conflict, since these sentences are so short that the memory cost can be neglected (Ferrer-i-Cancho, 2014). It has been found that the working memory has a limited capacity of 4 (i.e., the magical number 4) (Cowan, 2001). That is, the limitation in the human capacity to store and process information is believed to be four chunks (Cowan, 2001). Hence, the number of words in short sentences of 0–4 words would not exhaust this capacity limit in short-term memory. Accordingly, surprisal minimization will play a major role in short sentences, and hence results in increasing MDD.

Roles of dependency types in diachronic minimization of dependency distance. An important finding in our study is that not all dependency types in longer sentences showed a downtrend, but the minimization is confined to nine types, including *root*, *nsubj*, *nsubjpass*, *acl: relcl*, *ccomp*, *aux*, *auxpass*, *mark*, and *neg*. The decrease in dependency distance of these types may be explained as follows.

First, the decreasing tendency of root position is probably associated with its important role in sentence comprehension and production (Healy and Miller, 1970; Raeburn, 1979; Lei and Jockers, 2020). It has been argued that during sentence processing, an individual may expend a good amount of working memory to search for the main verb before she/he completely comprehends the sentence (Lei and Jockers, 2020). Accordingly, the later the verb appears, the heavier the cognitive load placed on working memory and the more difficult a sentence is. It is thus reasonable to assume that the earlier emergence of the main verb may be preferred to reduce cognitive burden and achieve efficient communication, which would contribute to the decrease of root position.

Second, the decreasing trends of the MDDs of the dependency type *nsubj* (nominal subject), *aux* (auxiliary), and their passive counterparts *nsubjpass* (passive nominal subject) and *auxpass* (passive auxiliary) may be closely related to or attributed to the downward trend of root position. The earlier emergence of the main verb leads to a shorter distance between this main verb and its subject or auxiliary verb. In (2a), for example, the subject *They* need to be kept in working memory until the verb *abandoned* is integrated (Ivanova and Ferreira, 2019). If *abandoned* shows up earlier as in (2b), the distance of the subject-verb relation (i.e., *nsubj*) will be shorter. The same is true for the dependency relation *aux* (between *have* and *abandoned*) as in (2a) and (2b). The distance of the former is shorter that of the latter.

(2) (a) ***They have*** for the most part ***abandoned*** *the hunter state and turned their attention to agricultural pursuits.*

(2) (b) ***They have abandoned*** *the hunter state and turned their attention to agricultural pursuits.*

Third, the dependency relation *acl: relcl* refers to a relative clause modifying the noun; *ccomp* is a dependent clause that serves as a complement; and *mark* is defined as a subordinating conjunction that marks a subordinate clause. A possible explanation for the decreasing tendencies in dependency distance of these three types is that they mark the subordinate clauses. A subordinate clause is embedded within a main clause and functions as a constituent of it (Cristofaro, 2003; Chen et al., 2021). Such a structure has been found to be one of the most important types of grammatical complexity and incur higher cognitive demand on language users (Carter and McCarthy, 2006; Chen et al., 2021). As a result, language users may be more pressurized and motivated to simplify such structures in order to reduce the higher processing cost resulted from it. One method is to reduce the words retained in working memory as fewer as possible, that is, to reduce the dependency distance of the relation that links them. For example, the reading time will decrease if the distance of dependency relation *acl: relcl* between *administrator* and *supervised* is shortened as shown in (3a) and (3b) (Grodner and Gibson, 2005; Bartek et al., 2011; Futrell et al., 2020). The same is true for the relation *mark* that links *who* and *supervised* as well as the relation *ccomp* between *proposed* and *create* as in Example (4).

(3) (a) The ***administrator who*** the nurse ***supervised***...

(3) (b) The ***administrator who*** the nurse from the clinic ***supervised***...

(4) (a) I ***proposed*** nearly two years ago that we ***create*** *a department.*

(4) (b) Nearly two years ago, I ***proposed*** that we ***create*** *a department.*

Last, *neg* refers to the relation between a negation word and the word it modifies. The decrease in dependency distance of this type may be explained with two reasons. One possible reason is that it plays functional roles in sentences and its position in sentence is seemingly more flexible than that of other relations. Language users may prefer to rearrange the word order and place the negation modifier in the position that is closer to the word it modifies in order to minimize the dependency distance of this type. For example, in (5a), the negation modifier *not* is very far from the word *taken*, which is predicted to induce a greater cost to working memory when listeners or readers integrate them. However, if *not* is placed closer to *taken* as in (5b), lower cognitive load would be required. Another possible reason is that some lengthy adverbial phrases that intervene between a negation word and the word that they modify as in (5a) (i.e., *too soon or too seriously*) may have been avoided. Such a structure not only complicates the syntax but also requires more working memory resources. Simple phrases could improve readability and reduce memory burden, and hence result in shorter dependency distance of the relation *neg*.

(5) (a) *It is a subject which can****not*** too soon or too seriously be ***taken*** *into consideration.*

(5) (b) *It is a subject which can****not*** be ***taken*** *into consideration too soon or too seriously.*

Another important point worth discussing is that although the MDDs showed decreasing tendencies in longer sentences with five or more words, six dependency types were found consistently increasing. A common feature of these six types (i.e., *compound: prt*, *compound*, *amod*, *det*, *nmod:poss*, and *advmod*) is that they are noun-phrase-related dependency types. The increasing MDDs of these types may be explained by the increased length and complexity of noun phrases in informational registers found by previous studies (e.g., Biber and Clark, 2002; Biber and Gray, 2011). For example, Biber and Gray (2011) found that three-noun sequences (e.g., *trade boycott campaign*) and even four-noun sequences (e.g., *mean plasma glucose value*) had become increasingly commonly used in academic writing. Such a trend may be largely attributed to the changing communicative demands under the influence of information explosion (Biber and Gray, 2011). In other words, with more information to be communicated, language users may pack more words into the noun phrases, since such a construction is tightly integrated and meets the need for the economy of expression. As a result, noun phrases have become increasingly complex, which results in longer dependency distance.

Implications. The findings in this study make at least two contributions to the existing literature. First, the decreasing trends found in the MDDs of sentences and dependency types (particularly *root*) across the 200 years further support the hypothesis that the dependency distance of human languages tends to minimize diachronically (Lei and Wen, 2020). This trend may be attributed to the limited working memory capacity as well as the principle of least effort (Zipf, 1949). A longer dependency distance means that more words would be stored in memory. When the number of words stored exceeds the memory capacity, it leads to processing difficulty (Jiang and Liu, 2015). In contrast, a shorter dependency distance is easier to produce and comprehend (Hawkins, 1994; Gibson, 1998). DDM, therefore, manifests a general cognitive tendency, i.e., reducing the complexity of syntax to decrease 'the average rate of work-expenditure over time' (Zipf, 1949) and to lessen processing costs (Lu et al., 2016), so that the most effective communication could be achieved.

In addition, it may also be of interest to consider our findings from a more macroscopic perspective of language evolution. The diachronic dependency distance minimization found in our study suggests that language has become less complex in terms of syntax, since dependency distance is taken as a measure of syntactic complexity. Also, the increasing dependency distances of noun-phrase-related dependency types may serve as another approach to the facilitation of dependency processing. These points seemingly provide further evidence to the claim that human language may be evolving toward a simplified direction (Lei and Wen, 2020). Such a tendency has also been detected at other levels of human language. For example, the loss of inflected forms has been captured in the evolution of languages such as English (Lieberman et al., 2007; Bentz et al., 2014; Zhu and Lei, 2018), German (Carroll et al., 2012) and Dutch (Knooihuizen and Strik, 2014). In addition, (Baker, 2011) found that Modern English showed a decreasing trend of verbosity. For example, those verbose words such as *men*, *women* and *children* were being replaced by the single word *people*. To summarize, as a complex adaptive system (Liu et al., 2017; Liu, 2018), language may be evolving towards simplicity in order to adapt to the limited capacity of human memory (Lei and Wen, 2020).

Conclusion

Dependency length and dependency types are two factors that may affect the MDD of a language and their effects on dependency distance minimization over a long period of time have not been extensively examined. The present study hence explored the diachronic change of dependency distance in terms of these two variables. Results showed that sentence length plays an important role in diachronic DDM. Particularly, the phenomenon of anti-DDM in short sentences was confirmed diachronically. In addition, nine types of dependency relations were found consistently decreasing in sentences of five or more words in our study. The findings in our study provide more evidence to the hypothesis that dependency distance of human languages tends to minimize, and more importantly, human languages may experience a minimized or simplified evolution.

The following limitations may be addressed in future research. First, the present study is limited in the data of only one genre (i.e., political texts) and one language (i.e., English). Previous research has found that genre has an effect on dependency distance (Liu et al., 2009b; Oya, 2013; Wang and Liu, 2017). Therefore, we should be cautious about the generalizability of the results in the study. Future research may validate our findings with other genres of texts. Also, it may be of interest to extend the present study to texts written in other languages. Second, our results showed that nine types of dependency relations are possibly responsible for diachronic DDM. Future research may consider examining the role of different dependency types in more areas such as language typology or interpreting research. It would be particularly interesting to explore whether results are consistent.

Data availability

All data analyzed in this study are cited in this article and available in the public domain.

Received: 26 July 2022; Accepted: 14 November 2022;
Published online: 26 November 2022

References

Baker P (2011) Times may change, but we will always have money: diachronic variation in recent British English. J Engl Linguist 39:65–88

Bartek B, Lewis RL, Vasishth S, Smith MR (2011) In search of on-line locality effects in sentence comprehension. J Exp Psychol Learn Mem Cogn 37:1178–1198

Bentz C, Kiela D, Hill F, Buttery P (2014) Zipf's law and the grammar of languages: a quantitative study of old and modern English parallel texts. Corpus Linguist Linguist Theory 10:175–211

Biber D, Clark V (2002) Historical shifts in modification patterns with complex noun phrase structures: how long can you go without a verb? In: Fanego T, López-Couso MJ, Pérez-Guerra J (eds) English historical syntax and morphology. John Benjamins, Amsterdam and Philadelphia, pp. 43–66

Biber D, Gray B (2011) Grammatical change in the noun phrase: the influence of written language use. Engl Lang Linguist 15:223–250

Carroll R, Svare R, Salmons JC (2012) Quantifying the evolutionary dynamics of German verbs. J Hist Linguist 2:153–172

Carter R, McCarthy M (2006) Cambridge grammar of English. Cambridge University Press, Cambridge

Chen X, Alexopoulou T, Tsimpli I (2021) Automatic extraction of subordinate clauses and its application in second language acquisition research. Behav Res Methods 53:803–817

Cowan N (2001) The magical number 4 in short-term memory: a reconsideration of mental storage capacity. Behav Brain Sci 24:87–114

Cristofaro S (2003) Subordination. Oxford University Press, Oxford

Ferrer-i-Cancho R (2013) Hubiness, length, crossings and their relationships in dependency trees. Glottometrics 25:1–21

Ferrer-i-Cancho R (2014) Why might SOV be initially preferred and then lost or recovered? A theoretical framework. In: Cartmill EA, Roberts S, Lyn H, Cornish H (eds) The evolution of language—Proceedings of the 10th international conference (EVOLANG10). Wiley, Vienna, pp. 66–73

Ferrer-i-Cancho R (2017) The placement of the head that maximizes predictability. An information theoretic approach. Glottometrics 39:38–71

Ferrer-i-Cancho R, Gómez-Rodríguez C (2021) Anti dependency distance minimization in short sequences. A graph theoretic approach. J Quant Linguist 28:50–76

Futrell R, Levy RP, Gibson E (2020) Dependency locality as an explanatory principle for word order. Language 96:371–412

Futrell R, Mahowald K, Gibson E (2015) Large-scale evidence of dependency length minimization in 37 languages. Proc Natl Acad Sci USA 112:10336–10341

Gibson E (1998) Linguistic complexity: locality of syntactic dependencies. Cognition 68:1–76

Gildea D, Temperley D (2010) Do grammars minimize dependency length? Cogn Sci 34:286–310

Grodner D, Gibson E (2005) Consequences of the serial nature of linguistic input for sentenial complexity. Cogn Sci 29:261–290

Hawkins JA (1994) A performance theory of order and constituency. Cambridge University Press, Cambridge

Healy AF, Miller GA (1970) The verb as the main determinant of sentence meaning. Psychon Sci 20:372

Heringer H, Strecker B, Wimmer R (1980) Syntax: Fragen-Lösungen-Alternativen. Wilhelm Fink Verlag, München

Hudson RA (2010) An introduction to word grammar. Cambridge University Press, Cambridge

Hussain M, Mahmud I (2019) pyMannKendall: a python package for non-parametric Mann Kendall family of trend tests. J Open Source Softw 4:1556

Ivanova I, Ferreira VS (2019) The role of working memory for syntactic formulation in language production. J Exp Psychol Learn Mem Cogn 45:1791–1814

Jiang JY, Liu HT (2015) The effects of sentence length on dependency distance, dependency direction and the implications–Based on a parallel English–Chinese dependency treebank. Lang Sci 50:93–104

Jiang JY, Ouyang JH (2017) Dependency distance: a new perspective on the syntactic development in second language acquisition: Comment on "Dependency distance: a new perspective on syntactic patterns in natural language" by Haitao Liu et al. Phys Life Rev 21:209–210

Jiang XL, Jiang Y (2020) Effect of dependency distance of source text on disfluencies in interpreting. Lingua 243:102873

Knooihuizen R, Strik O (2014) Relative productivity potentials of Dutch verbal inflection patterns. Folia Linguistica 35:173–200

Lei L, Jockers ML (2020) Normalized dependency distance: proposing a new measure. J Quant Linguist 27:62–79

Lei L, Wen J (2020) Is dependency distance experiencing a process of minimization? A diachronic study based on the State of the Union addresses. Lingua 239:102762

Levy R (2008) Expectation-based syntactic comprehension. Cognition 106:1126–1177

Liang JY, Fang YY, Lv QX, Liu HT (2017) Dependency distance differences across interpreting types: implications for cognitive demand. Front Psychol 8:2132

Lieberman E, Michel JB, Jackson J, Tang T, Nowak MA (2007) Quantifying the evolutionary dynamics of language. Nature 449:713–716

Liu HT (2007) Probability distribution of dependency distance. Glottometrics 15:1–12

Liu HT (2008) Dependency distance as a metric of Language comprehension difficulty. J Cogn Sci 9:159–191

Liu HT (2010) Dependency direction as a means of word-order typology: a method based on dependency treebanks. Lingua 120:1567–1578

Liu HT (2018) Language as a human-driven complex adaptive system. Phys Life Rev 26-27:149–151

Liu HT, Hudson R, Feng ZW (2009a) Using a Chinese treebank to measure dependency distance. Corpus Linguist Linguist Theory 5:161–174

Liu HT, Xu CS (2011) Can syntactic networks indicate morphological complexity of a language? Europhys Lett 93:28005

Liu HT, Xu CS, Liang JY (2016) Dependency length minimization: puzzles and promises. Glottometrics 33:35–38

Liu HT, Xu CS, Liang JY (2017) Dependency distance: a new perspective on syntactic patterns in natural languages. Phys Life Rev 21:171–193

Liu HT, Zhao YY, Li WW (2009b) Chinese syntactic and typological properties based on dependency syntactic treebanks. Poznan Stud Contemp Linguist 45:509–523

Lu Q, Xu CS, Liu HT (2016) Can chunking reduce syntactic complexity of natural languages? Complexity 21:33–41

Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D (2014) The Stanford CoreNLP natural language processing toolkit. In: Proceedings of the 52nd annual meeting of the association for computational linguistics: system demonstrations, Baltimore, pp. 55–60

Miyake A, Shah P (eds) (1999) Models of working memory: mechanisms of active maintenance and executive control. Cambridge University Press, Cambridge

Ouyang JH, Jiang JY (2018) Can the probability distribution of dependency distance measure language proficiency of second language learners? J Quant Linguist 25:295–313

Oya M (2013) Degree centralities, closeness centralities, and dependency distances of different genres of texts. In: Selected papers of the 17th conference of Pan-Pacific association of applied linguistics, pp. 42–53

Poiret R, Liu HT (2020) Some quantitative aspects of written and spoken French based on syntactically annotated corpora. J French Lang Stud 30:355–380

Raeburn VP (1979) The role of the verb in sentence memory. Mem Cogn 7:133–140

Savoy J (2015) Text clustering: an application with the State of the Union addresses. J Assn Inf Sci Technol 66:1645–1654

Temperley D (2007) Minimization of dependency length in written English. Cognition 105:300–333

Temperley D (2008) Dependency-length minimization in natural and artificial languages. J Quant Linguist 15:256–282

Wang L, Liu HT (2013) Syntactic variations in Chinese–English code-switching. Lingua 123:58–73

Wang L, Liu HT (2016) Syntactic differences of adverbials and attributives in Chinese-English code-switching. Lang Sci 55:16–35

Wang YQ, Liu HT (2017) The effects of genre on dependency distance and dependency direction. Lang Sci 59:135–147

Yngve VH (1960) A model and an hypothesis for language structure. Proc Am Philos Soc 104:444–466

Zhu HR, Lei L (2018) Is modern English becoming less inflectionally diversified? Evidence from entropy-based algorithm. Lingua 216:10–27

Zhu HR, Lei L (2022) A dependency-based machine learning approach to the identification of research topics: a case in COVID-19 studies. Libr Hi Tech 40:495–515

Zipf GK (1949) Human behavior and the principle of least effort. Addison-Wesley Press, Boston

Acknowledgements

This study was supported by an MOE (Ministry of Education of China) Foundation Project of Humanities and Social Sciences: Linguistic Complexity-based Research on Text Classification (Grant No. 21YJC740085).

Author contributions

All authors contributed to the design and implementation of the work. All authors were involved in the analysis and interpretation of data for the work. XYL and LL drafted and revised the work. XYL and HRZ contributed to the tables and figures. LL and HRZ proofread the manuscript. All authors approved the final version.

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Correspondence and requests for materials should be addressed to Lei Lei.

Reprints and permission information is available at http://www.nature.com/reprints