Research Article

Hengbin Yan, Yinghui Li*

Beyond Length: Investigating Dependency Distance Across L2 Modalities and Proficiency Levels

https://doi.org/10.1515/opli-2019-0033
Received June 11, 2019; accepted September 17, 2019

Abstract: As an important index of working memory burden and syntactic difficulty, Dependency Distance (DD) has been fruitfully applied in the context of Second Language Acquisition (SLA) to both native and non-native language production. Recent research has focused on DD as a predictor of writing performance across different proficiencies, while the other modality of second language (L2) performance – speaking – has been largely neglected. Moreover, while previous results reported significant positive correlations between DD and writing proficiency, a lack of control for important variables such as sentence/text length may have potentially inflated the outcomes of the analyses. In this study, we examine the effects of DD across the different modalities and proficiency levels, controlling for the effects of text and sentence lengths. Results from statistical analysis show that the mean DD of L2 writing is significantly shorter than L2 speech production, indicating that L2 writing may be a cognitively more demanding task than L2 speaking. Additionally, while significant statistical differences in DD were found across proficiency levels in both L2 writing and speech, the significance disappeared after controlling for sentence and text length. The implications of the present study for future research on DD in SLA are discussed.

Keywords: Dependency Distance; Cognitive Load; Second Language Acquisition; L2 Proficiency

1 Introduction

According to Dependency Grammar (Tesnière 1959; Hudson 2010), underlying the apparent linearity of sentences in natural language is a hierarchical structure in the form of dependency relations between two words that are syntactically related. Research in quantitative linguistics has proposed Dependency Distance (DD), the linear distance (number of intervening words) between two words with a dependency relation, as a universal index of working memory (WM) burden and syntactic difficulty across different languages and genres (Liu 2008; Liu et al. 2017).

With its significant role as a psychological and linguistic metric, DD is highly relevant in the study of Second Language Acquisition (SLA), where learners at different stages of language development have access to varying amounts of cognitive resources for language production. Recent studies have investigated the DD of written outputs by second language (L2) learners across different proficiency levels (Ouyang and Jiang 2018; Jiang and Ouyang 2017). However, existing research has focused exclusively on writing proficiency, while the oral proficiency of L2 learners has remained largely unexplored. Moreover, the

***Corresponding author: Yinghui Li,** School of English and Education/Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou, Guangdong, China, E-mail: liyinghui@gdufs.edu.cn
Hengbin Yan, Faculty of English Language and Culture/Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou, Guangdong, China

reliability of existing studies suffers from a lack of control for important variables such as text and sentence length, which have been found to be positively correlated with DD as well as L2 proficiency (Ferris 1994; Chodorow and Burstein 2004; Wang and Liu 2017).

To address these shortcomings, we investigate in this study the patterns of DD of L2 production across modalities (i.e. writing and speech) and proficiency levels, controlling for effects of text and sentence length. Aided by modern tools for automatic dependency parsing, DD can serve as a convenient and inexpensive metric applicable to potentially unlimited textual resources. By adopting a corpus-based approach to quantifying the relationships between modalities, L2 proficiency levels and cognitive demands as revealed through DD, we can probe into the otherwise hidden psychological/linguistic properties in both L1 and L2 data, and elucidate the intricate interplay between language and the human mind.

2 Literature Review

2.1 Dependency Distance and Cognitive Load

Cognitive load, the mental effort required for the successful completion of a task, has been known to significantly influence learners' performance in L2 processing and learning (Kirkland and Saunders 1991; Al-Shehri and Gitsaki 2010). While a proper amount of cognitive load is conducive to L2 proficiency development, cognitive overload can impede transfer of information and hinder L2 learning (Sweller, 2011). To achieve the best learning teaching and learning outcomes, it is important to first understand the cognitive load that the different teaching/learning tasks impose on L2 learners in instructional design (Paas et al. 2003).

Various psychological experiments (Bartek et al. 2011; Levy et al. 2012; Fedorenko et al. 2013) have pointed to a positive relationship between DD and cognitive load. It is found that human languages, governed by the principle of least effort (Zipf 1949), seem to have a preference for Dependency Distance Minimization (DDM), a universal tendency for natural languages to syntactically structure sentences in a way that minimizes the overall dependency distance, presumably due to constraints of working memory (Liu et al. 2017). Thus, the DD of linguistic outputs can be seen as a measure of working-memory-related cognitive load imposed on the human brain by the language production task at hand.

2.2 Dependency Distance and Modality

Existing applications of DD to SLA research has focused exclusively on writing (e.g. Fang & Liu, 2018; Jiang & Ouyang, 2017; Ouyang & Jiang, 2018), while L2 speaking has remained largely unexplored. This neglect is presumably due to the basic assumption that syntactic representations are shared between the written and the spoken modalities (Cleland and Pickering 2006), and therefore the cognitive load involved when processing them (i.e., DD) are modality-neutral. However, this assumption has been challenged by Wang & Liu (2017), who found that the mean DD of written-to-be-spoken text (consisting of scripted television materials and play scripts from British National Corpus) was significantly shorter than the DD of written text (either imaginative or informative texts). It should be noted that they investigated native (L1) written-to-be-spoken speech, which has important differences from spontaneous speech in a language learning context (Wagner 2014). As the syntactic representation is not the only factor that could introduce cognitive load either in a writing task or a speaking one, and L2 users could be no less constrained by their cognitive resources in processing L2 than L1 users would be in L1 processing, it is reasonable to empirically examine the cognitive load involved in L2 speaking and in L2 writing, hypothesizing that DD differs between these two modalities.

2.3 Dependency Distance and Proficiency

From the perspective of cognitive load constraint, it may be hypothesized that lower-proficiency learners produce written and spoken outputs with lower DD than higher-proficiency counterparts because the former have to devote more cognitive resources to additional cognitively demanding tasks such as recalling unfamiliar vocabulary and performing grammatical checks. As learners develop linguistically, such tasks may become easier, freeing up valuable cognitive resources for production of more complex sentences.

The above hypothesis seems to be confirmed by existing corpus-based investigations of the relationship between DD and language proficiency, which again, have focused exclusively on writing. Jiang & Ouyang (2017)'s short commentary on corpus-based analysis of Chinese learners of English across eight grades showed that Mean Dependency Distance (MDD) increased from 1.845 at grade one of junior high schools to 2.466 at grade two in universities. However, there were no significant increases later in the university years, indicating that the MDD had plateaued as the values neared that of English native speakers (2.543). Using the same dataset, Ouyang & Jiang (2018) compared the distribution of DD of native and non-native speakers and found that although their proficiencies differed in theory, the distributional patterns were highly similar. They also tested whether the parameters in one of the distribution models could reflect language proficiency and found significant correlations between two of the parameters in the model and L2 proficiency. However, Ouyang & Jiang did not indicate how the parameters in the probability distribution relate to direct measurements of DD, making the results more difficult to operationalize for applications such as second language pedagogy.

2.4 Effects of Text/Sentence Length on L2 Proficiency and Dependency Distance

Text length and sentence length have been known as important predictors of language proficiency and cognitive load of language production tasks (Ferris 1994; Chodorow and Burstein 2004; Wang and Liu 2017). Essay scoring models using text length alone can account for about 30% (Ferris 1994: 418) to more than half (Chodorow and Burstein 2004: 12) of the variance in the holistic scores by humans. Ferrer-i-Cancho & Liu (2014, p. 151) argued for the necessity of treating DD as a function of sentence length. Jiang & Liu (2015), who studied the effects of sentence length on dependency distance, compared the sentence lengths of a small parallel Chinese-English treebank that were divided into sets of varying lengths and found that the DD of each text increased with the length of the sentences. However, compared with random ordering sentences, they found that the increase was relatively slow, presumably due to the constraints in working memory.

If language learning is, as Ellis (1996) pointed out, mostly sequence learning, then measures of length of the sequence, whether linear or hierarchical, should be taken into consideration. In the case of DD, only after controlling for the linear length can the differences be attributed to the hierarchical/syntactic aspect of the metric. It is therefore necessary to account for both of these variables in any rigorous investigation of DD.

3 The Present Study

In the current study, we mainly investigate the patterns of DD of L2 production across modalities (i.e. writing and speaking), with the potential confounding effects of text length and sentence length analyzed and controlled. As the DD of written outputs by L2 learners was found different across different proficiency levels (Ouyang and Jiang 2018; Jiang and Ouyang 2017), it is reasonable to hypothesize that the DD of oral production also differs among L2 learners of different proficiency levels. We thus aim to address the following research questions: (1) Do L2 learners, constrained by cognitive processing load, produce different DDs across modalities and proficiency levels? (2) To what extent do sentence length and text length influence the effects of modality and proficiency on DD in L2 productions?

4 Materials and Method

4.1 Materials

The written essays used in this study came from ETS Corpus of Non-Native Written English (Blanchard et al. 2013). The corpus is a collection of essays written by test-takers of 11 L1 backgrounds as responses to the TOEFL writing test. 1100 essays were collected from L1 backgrounds, totaling 12100 essays. Each essay has been rated by at least two experts from ETS, and then classified into three groups into proficiency levels: low, medium and high. To ensure comparability between the three levels, we randomly sampled 400 essays from each level, matching the L1 backgrounds and written prompts in each level. The statistics of the sample essays are presented in Table 1.

Table 1: Text Samples in Use from ETS Corpus of Non-Native Written English

Proficiency	# of essays	# of tokens	# of tokens per essay
low	400	91050	227.63
medium	400	136057	340.14
high	400	161914	404.79
total	1200	389021	324.18

The spoken data in our study were taken from the Spoken and Written English Corpus of Chinese Learners 2.0 (SWECCL 2.0). The corpus is a collection of written and spoken output randomly sampled from the Test for English Majors, Band 4 (TEM 4) and Band 8 (TEM 8), two national standardized tests taken by most English majors in China. While the two tests are designed by the same group of researchers to evaluate the proficiency of advanced learners, they have different types of prompted tasks and scoring criteria. To control for such variables, we include only the TEM8 data for our study. The spoken component of the TEM8 subcorpus is a collection of manually transcribed transcripts of recordings produced by test-takers (the composition of the data is shown in Table 2). Each transcript has been rated by official scorers and is given a rank of 1 (high-proficiency) to 3 (low-proficiency) based on the position of the transcript relative to other transcripts in the same scoring group.

Table 2: Composition of the selected data from the SWECCL spoken corpus

Proficiency	# of Transcripts	# of tokens	Average # of tokens/transcript
low	276	98811	334.95
medium	314	119386	358.52
high	271	110139	382.43
total	916	328336	358.45

4.2 Preprocessing

The written and spoken corpora in our study are highly comparable since they share a number of key characteristics: 1) both are collections of spontaneous outputs by English learners under standardized examination conditions; 2) both tests require students to make argumentative comments in response to a set of prompts 3) the two corpora are similar in corpus size and mean text length 4) both have collapsed score levels corresponding to the relative proficiency rank of test takers 5) the essays and words in the two corpora across the score level are roughly evenly distributed.

Further preprocessing steps were taken to ensure that the corpus data are suitable for the subsequent computation of syntactic features as well as for statistical analysis. Both corpora in the current study left many lexical features unedited to preserve the texts as they were written/spoken, which could pose problems to NLP tools. Using a Python script, we removed from the SWECCL corpus disfluency markers such as ‘um’ and ‘uh’, which may inflate sentence length and text length. However, repetitions in general (repeating the previous word/phrase in an attempt to self-repair utterances) were retained because these represent inherent features of spoken language whose occurrences may depend on the cognitive load of the speaking task (Oomen and Postma 2001). Removal of such repetitions might take away important variables in the relationship between speech production and cognitive demand. Spelling errors abound in the unedited ETS corpus essays, which might affect the accuracy of automatic dependency parsers. We thus filtered out 491 essays with a spelling error rate of more than 7.2% (two standard deviations above the mean rate). The misspelled words were identified by looking up the lemmatized form of each word in a large dictionary. The average spelling error rate (number of misspelled words divided by number of all words) of all texts was 2.68%.

The Stanford Parser (Chen and Manning 2014; Klein et al. 2003) was chosen for dependency parsing over other dependency parsers due to its widespread popularity in natural language processing tasks and its accuracy in parsing non-standard learner texts, with a reported accuracy of up to 92.1% (Geertzen et al. 2013). Following Crossley (2013, p. 177) and Biber et al. (2016, p. 652), we filtered out texts shorter than 100 words in both corpora due to the unreliability of textual indices when applied to very short texts. The punctuation marks in the SWECCL corpus were provided and proofread by corpus transcribers. The ETS corpus, on the other hand, came with an original, unedited version and a version with manually corrected tokenization and punctuation marks. As the punctuation marks in many of the original essay texts produced by the learners were incorrect, which proved problematic for sentence demarcation and thus dependency parsing, we opted to use the tokenized version. In addition, we filtered out essays with no punctuation marks (treated as one long sentence even in the tokenized version), and discarded essays that combined a string of long sentences into a single one, which may inflate dependency distance as well as sentence length metrics.

4.3 Computation of DD

Following the formula proposed in Liu (2008), we calculated the Mean Dependency Distance (MDD) of a sentence. As is shown in formula (1), to compute MDD, we obtain the sum of the absolute value of the DD of each dependency relationship, divided by the number of words minus 1 (since the root in the sentence does not have a governor and is given a DD of zero). The DD between a pair of syntactically related words is the difference in the linear positional indices between the governor and the dependent in the syntactic relationship, with adjacent words having a DD of 1 (rather than 0 as in Gibson, 2000). In the formula, n is the number of words in the sentence and DD_i is the ith dependency link in the sentence. The formula sums up the absolute value of the DD of all dependency links in the sentence divided by the number of links.

$$\mathrm{MDD(sentence)} = \frac{1}{n-1}\sum_{i=1}^{n} |DD_i| \tag{1}$$

For a collection of sentences in a text, the sum of all DDs in the text is divided by n - s, where s is the number of sentence (whose roots have a DD of zero by definition).

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

We illustrate the computation of DD with an example sentence (whose dependency structure is shown in Figure 1). A dependency relation is a labeled link pointing from a governor to a dependent. In Figure 1, the dependency link between *John* and *wanted* is represented by a pointed arc labelled with the relation name *nsubj*, with the arrow pointing from the governor *wanted* to the dependent *John*.

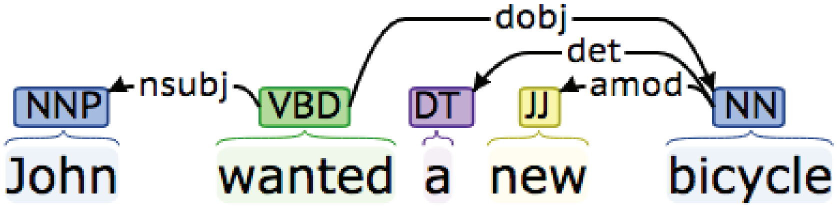

Figure 1: Dependency structure of an example sentence: John wanted a new bicycle.

The dependency structure as parsed by the Stanford Parser is given as follows:
root (ROOT-0 , wanted-2)
nsubj (wanted-2 , John-1)
det (bicycle-5 , a-3)
amod (bicycle-5 , new-4)
dobj (wanted-2 , bicycle-5)

The parsed output is a collection of triplets of typed dependency relations in the form of *relation (dependent, governor)*. For example, the triplet *nsubj (wanted-2 , John-1)* means that the word *John*, whose position index in the sentence is 1, is the dependent of its governor *wanted*, whose index is 2, with a dependency relation of nominal subject (*nsubj*) between them. The dependency distance between the two words can be obtained simply by subtracting their position indices: 2-1=1. The DD of the sentence can then be obtained by summing up the each of the triplets in this way, with the exception of *root (ROOT-0 , wanted-2)*, which is excluded from the calculation because it is only a dummy node signifying that *wanted* is the root of the sentence. Thus, according to formula 1, the DD of the sentence can then be computed as follows: (1 + 2 + 1 + 3) / 4 = 1.75.

4.4 Statistical Methods

To control for the potential confounding effects of sentence/text length, one approach used in previous studies (Crossley and McNamara 2012: 120) is to limit text length to a predefined range, and ensure there is no significant correlation between text/sentence length and the target variable (e.g. proficiency). However, such manual filtering leaves out a large proportion of the samples and may introduce artificial bias to the sample that can make estimating the effect size of the variables more difficult. In this study, we adopted another statistically grounded approach that uses Multiple Hierarchical Regression, which have been widely applied in SLA studies (e.g. Hoang & Boers, 2018; Qin & Uccelli, 2016; Wolfe et al., 2015). Hierarchical regression shows whether independent variables of interests can explain a statistically significant amount of variance of the dependent variable after controlling for other variables. After entering text/sentence length as a control variable in the initial models, the model can be compared with later models with the predictor variable, thus revealing the true effect size of the predictor variable.

5 Results

5.1 Dependency Distance and Modalities

We compared the MDDs between the writing and speech modalities to see if there are any differences. We entered text and sentence length as control variables to the regression model and used the stepwise method to locate the significant predictor. Results showed that sentence length, which accounted for 46.0% of the variance of DD, was the only significant predictor in the first model (Table 3).

Table 3: Hierarchical multiple regression analyses of the effects of modality on DD

Model	Independent variable	R^2	ΔR^2	F	p
1	sentence length	0.460	0.460	1801.560	0.000
	text length	0.460	0.000	0.354	0.552
2	modality	0.601	0.140	741.980	0.000

In the second model, modality was introduced as the predictor variable. Now with sentence length and text length controlled, modality explained an additional 14.0% of the variation of DD, resulting in a total of 60.0% of the variance. An ANOVA comparison of the two models revealed that the main effect of modality was significant, indicating significant differences in DD between L2 writing and speech, even after accounting for sentence length. A two-sample t-test further showed that the mean sentence length of L2 essays was significantly shorter than that of speech, t(1482.5)=-11.352, $p<0.0001$. In addition, a much higher percentage of long sentences (comprised of 30 words or more) were found in the spoken corpus (20.3%) than in the written corpus (8.1%). To determine whether the differences in sentence length resulted from the way clauses were combined (i.e. coordinating vs. subordinating conjunctions, see Greenbaum & Nelson, 1995, p. 16), we counted the number of coordinating conjunction (cconj) dependency relations between verbal clauses by the Stanford Parser. Results showed that L2 speech had a significantly larger number of coordinating conjunctions than L2 writing, t(1543.3)=4.3803, $p < 0.0001$.

5.2 Dependency Distance and L2 Proficiency

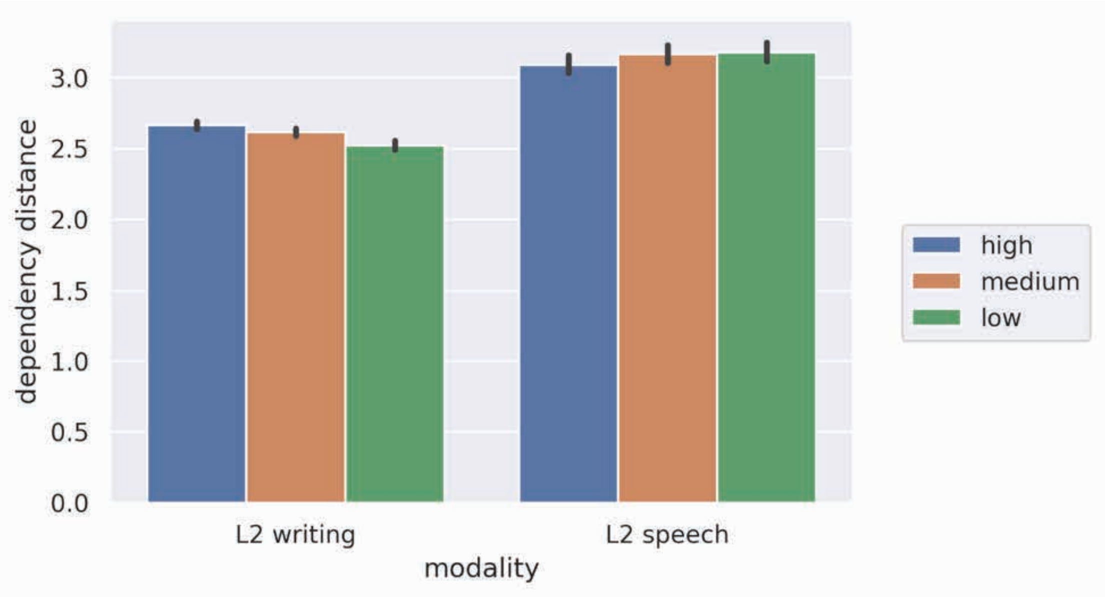

Figure 2: Dependency distances across proficiency levels in L2 writing and speech

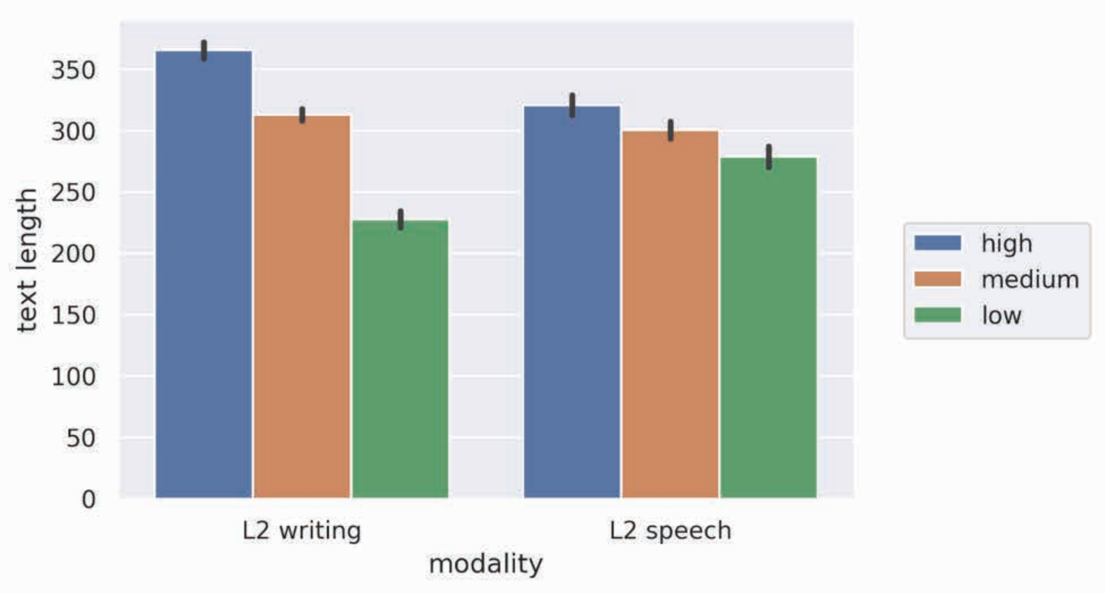

Figure 3: Text lengths across proficiency levels in L2 writing and speech

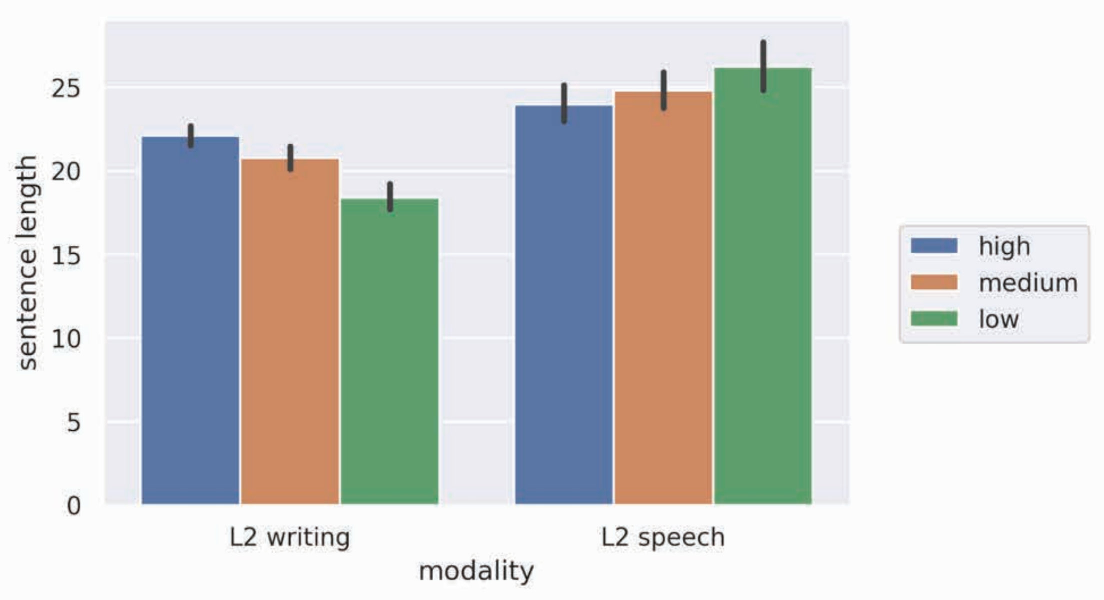

Figure 4: Sentence lengths across proficiency levels in L2 writing and speech

As can be seen in Figure 2, the MDD of L2 writing appears to increase as the proficiency level advances from low to high, while an opposite trend is seen for speech. To test whether the differences were statistically significant, two ANOVA tests were performed on each of the modalities, with the DD as the dependent variable and the proficiency level as the independent variable. The omnibus test was significant for writing (F=27.06, p<0.0001), but not for speech (F=2.058, p>0.05). A post-hoc analysis was conducted to compare all paired means of proficiency levels in writing using Tukey Honest Significant Differences (TukeyHSD) with Bonferroni correction. Significant differences were found between all the pairs ($p < 0.05$). The results on writing confirm the findings of Jiang & Ouyang (2017), who reported that DD tends to increase with proficiency levels in L2 writing. However, these differences across the proficiency levels may again have been due to the differences in text length, which appeared to increase as proficiency level advances, as shown in Figure 3, and sentence length, which exhibited a trend similar to that of DD across the proficiency levels (shown in Figure 4).

Table 4: Hierarchical multiple regression analyses of the effects of proficiency on DD in writing

Model	Independent variable	R^2	ΔR^2	F	p
1	sentence length	0.510	0.510	1255.1855	0.000
	text length	0.514	0.004	9.3474	0.002
2	proficiency	0.514	0.000	0.288	0.750

Table 5: Hierarchical multiple regression analyses of the effects of proficiency on DD in speech

Model	Independent variable	R^2	ΔR^2	F	p
1	sentence length	0.422	0.422	665.645	0.000
	text length	0.423	0.001	2.134	0.144
2	proficiency	0.425	0.001	1.171	0.310

We performed a set of hierarchical multiple regression analyses to determine whether proficiency level contributed a unique amount of variance to DD independent of the effects of sentence and text length. The results for analyses of writing is shown in Table 4. We first entered sentence length and text length as the control variables, both of which contributed a statistically significant variance (51.4% in total) to DD. However, most of the contribution came from sentence length (51.0%), while that from text length was extremely small (0.37%). Adding proficiency level to the statistical model did not contribute significant variance to the model, F(2, 1195)= 0.288, p>0.05, with an increased R^2=0.00023, or 0.023%. The same hierarchical regression procedure was applied to the speech data (Table 5). It was found that sentence length accounted for 42.2% of the variance of DD, while no significant amount of variance was found attributable to text length, F(1, 909)=2.134, p > 0.05. Similar to L2 writing, proficiency level did not turn out to be a significant predictor for DD in L2 speech, F(2, 907)= 1.171, p > 0.05. The statistical analysis shows that proficiency level was not a good predictor of DD in either L2 writing or speech. Sentence length was found to be a strong predictor of DD in both modalities, while text length also contributed a small amount of variance in writing. Proficiency level, on the other hand, did not contribute a significant amount of variance to the regression model after controlling for sentence and text length.

6 Discussion

The present study investigated the DD produced by L2 learners across two modalities and three L2 proficiency levels, with the potential confounding effects of text length and sentence length controlled.

Our findings first show that sentence length (but not text length) was a factor closely related to DD produced by L2 learners in either written or oral production. With text/sentence length accounted for statistically, the remaining variance of DD contributed by modality differentiation may be due to the different cognitive processing load imposed on the L2 learners by the two different tasks. We thus postulate that, under examination conditions, writing may be a cognitively more demanding task for learners than speaking. Cowan (1999) proposed that working memory is temporally part of long-term memory. The number of chunks that can be held in WM is a function of the cognitive load derived from comprehension, reformulation and production processes. In L2 spoken tasks, the speaker relieves the chunks as soon as it is uttered in real time. On the contrary, L2 writing under time pressure requires the writer to hold accumulating chunks of information in the focus of their attention during the writing task, before finally reformulating them into a coherent text. As more items take up the writer’s WM, less space is available for production. As WM and DD are correlated, the reduction in production WM leads to reduced DD.

The above finding is in line with the results of a recent study on the effects of modality on L2 performance, which reported that L2 learners produce higher syntactic complexity in spontaneous production in speaking than in writing (Cho 2018). However, the finding differs somewhat from a related study (Wang and Liu 2017: 142), which reported longer MDD for more formal, written English text than less formal, or written-to-be-spoken text. Wang and Liu argued that written-to-be-spoken texts have smaller DD because they are intended to be syntactically easier to process for listeners who have limited working memory capacity. The differences in the results from the two studies might be explained by the different nativeness status (L2 vs. L1) and genre (spontaneous vs. scripted speech) investigated.

Two reasons may account for the difference in MDD between L2 spontaneous speech and written text. First, the greater mean sentence length in L2 speech may have resulted from disfluencies such as repetitions of words and phrases, as well as loosely structured run-on sentences, which are common in spontaneous speech. This is in contrast with the written-to-be-spoken/scripted genre in Wang & Liu (2017), where the author has much more planning and control (Wagner 2014). The following two example sentences with disfluencies and run-on structure are taken from the spoken corpus:

In in most people’s point of view Beijing is an ancient city with a three hun … three thousand years’ history but now Beijing is … should make people know the aspect of modernization and Beijing should and people in Beijing … should learn some foreign cultures and when they welcome the foreign visitors.

In English, I think we can the government can can open English changing class no matter the old people or the young people can can come to this kind of English changing class into learn the English.

The analysis above also points to the potential bias of DD as a metric when L2 spoken data was involved. As DD is a concept closely linked with sentence length, the boundaries of sentences should be carefully defined, as they can have a significant impact on the computed DD. In L2 spoken production, issues such as disfluencies and unintended repetitions may potentially inflate sentence length, calling for careful corpus preprocessing and statistical control.

Another reason for the difference in MDD between the two modalities may be the tendency for L2 speakers to combine independent sentences into longer sentences using paratactic conjunctions, potentially resulting in greater dependency distances. Monologues make more extensive use of parataxis (coordination of clauses using conjunction words such as *and*, *or*, *but*) than academic writing (Greenbaum and Nelson 1995: 16). Our preliminary comparisons between the two modalities appeared to be in alignment with existing findings, with L2 speech relying more heavily on parataxis than L2 writing.

As for our second finding, it appears that, at least in L2 language production, DD is a function of sentence length rather than language proficiency. DD alone may be insufficient for differentiating the proficiency levels of L2 language while sentence length can predict a large amount of variance in DD. The finding that learners of varying proficiency produce similar written and spoken output of similar DD calls into question whether DD can be an effective index of L2 proficiency. Empirically, previous studies (Jiang and Ouyang 2017; Ouyang and Jiang 2018) have pointed to a potential relationship between DD and L2 proficiency, but did not establish a direct link between the two. Ouyang and Jiang (2018), for example, concluded that L2 writing proficiency is correlated with two parameters in the probability distribution of DD, but did not investigate the direct relationship between DD and L2 proficiency. More importantly, the strong effects of sentence length on DD have not been taken into consideration in existing studies. To investigate effects attributable to DD and cognitive load constraints, future investigation must strictly control for sentence length when examining the interaction between DD and other variables.

Theoretically, the most direct link between DD and L2 proficiency is DD's postulated use as an index of syntactic complexity. Given the long-established association between syntactic complexity and L2 language development (see a review in Lu, 2011), one may justifiably assume that DD is closely related to proficiency development. However, since syntactic complexity may encompass different subcomponents that develop at different stages of language learning (Bulté and Housen 2014), it appears still unclear which aspect of the complexity DD is measuring, and whether we can use a single measure for the two modalities, when the kinds of complexity in writing are different from speaking (Biber et al. 2011). In addition, even if DD is indeed an adequate metric of an important aspect of syntactic complexity, its measurement of the complexity may not always be directly reflected in human judgement of L2 production quality. For example, Crossley and McNamara (2014) found that, while the syntactic complexity in learner writing, as measured by the multifaceted computational indices computed using Coh-Metrix (Graesser et al. 2011), developed as a function of time spent on study, such development was not significantly correlated with human rating of the writing. Of the 11 syntactic indices investigated, only one is found to be predictive of human judgement of writing quality. Incidentally, this feature, the number of all clauses produced in the writing, is intrinsically related to sentence length. Our findings point to the potentially intricate relationships between DD and L2 language proficiency, which, we believe, will require much further investigation before conclusive results can be attained.

7 Conclusions

In this study, we investigated the effects of proficiency levels on L2 modalities and proficiency levels. Significant differences in dependency distance were found between L2 writing and speech. Contrary to earlier results in related L1 genres, the MDD of L2 essays is found to be shorter than that of L2 speech. Based on our results, we postulate that L2 writing tasks under time pressure taxes working memory more than

L2 speaking tasks. This is because L2 speaking releases information chunks from WM as the words were uttered, while in timed writing, chunks of information were held in the writer's WM before reformulating them into a coherent text.

Proficiency level has no significant effect on DD in L2 writing or L2 speech after controlling for sentence length. This result also fails to confirm previous findings that reported significant correlational relationships between DD and L2 writing (Jiang and Ouyang 2017). These differences are probably due to a lack of control for sentence length in the early study. Our analysis thus again highlights the importance of accounting for sentence length, a factor known to be strongly correlated with L2 proficiency levels as well as DD but neglected by most earlier investigations of L2 studies.

The advantage of a corpus-based investigation of DD is that it allows for a systematic and thorough examination of large collections of learner language and helps reveal learners' psychological and linguistic properties that are otherwise implicit or only obtainable through laborious experiments.

However, the limitations inherent in the DD approach should also be noted here. DD is one of many potential predictors for proficiency in learner language. L2 proficiency is a composite metric correlated with variables of complexity, accuracy and fluency (Housen et al. 2012) and the relationships between such variables and DD merit further investigation. Additionally, the current study investigates L2 production of the argumentative genre under the examination context. While a strict control for topic and context allows for a more rigorous and comparable design, future work may further investigate whether the same results generalize to other genres of L2 production.

Acknowledgements: This study was supported by Guangdong Planning Office of Philosophy and Social Science (Grant No. GD18YWW02), Innovative School Project in Higher Education of Guangdong, China (GWTP-GC-2017-01), Department of Education of Guangdong Province (19QN16) and the Foundation for Distinguished Young Talents in Higher Education of Guangdong (2015WQNCX030).

References

Al-Shehri, Saleh & Christina Gitsaki. 2010. Online reading: a preliminary study of the impact of integrated and split-attention formats on L2 students' cognitive load. *ReCALL* 22(3). 356–375. doi:10.1017/s0958344010000212.

Bartek, Brian, Richard L. Lewis, Shravan Vasishth & Mason R. Smith. 2011. In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 37(5). 1178–1198. doi:10.1037/a0024194. http://doi.apa.org/getdoi.cfm?doi=10.1037/a0024194.

Biber, Douglas, Bethany Gray & Kornwipa Poonpon. 2011. Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly* 45(1). 5–35. doi:10.5054/tq.2011.244483.

Biber, Douglas, Bethany Gray & Shelley Staples. 2016. Predicting Patterns of Grammatical Complexity Across Language Exam Task Types and Proficiency Levels. *Applied Linguistics* 37(5). 639–668. doi:10.1093/applin/amu059.

Blanchard, Daniel, James Carlson, Brent Bridgeman, Marna Golub-smith & Ruth Greenwood. 2013. TOEFL11 : A Corpus of Non-Native English. *ETS Research Report Series* 2. i–15.

Bulté, Bram & Alex Housen. 2014. Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing* 26. 42–65. doi:10.1016/j.jslw.2014.09.005. https://linkinghub.elsevier.com/retrieve/pii/S1060374314000666.

Chen, Danqi & Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 740–750.

Cho, Minyoung. 2018. Task complexity, modality, and working memory in L2 task performance. *System*. Elsevier Ltd 72. 85–98. doi:10.1016/j.system.2017.10.010. https://doi.org/10.1016/j.system.2017.10.010.

Chodorow, Martin & Jill Burstein. 2004. *Beyond Essay Length: Evaluating E-Rater®'S Performance on Toefl® Essays*. *ETS Research Report Series*. doi:10.1002/j.2333-8504.2004.tb01931.x.

Cleland, Alexandra A. & Martin J. Pickering. 2006. Do writing and speaking employ the same syntactic representations? *Journal of Memory and Language* 54(2). 185–198. doi:10.1016/j.jml.2005.10.003. https://linkinghub.elsevier.com/retrieve/pii/S0749596X05001269.

Cowan, Nelson. 1999. An embedded-processes model ofworking memory. In Akira Miyake & Priti Shah (eds.), *Models ofWorking Memory: Mechanisms ofActive Maintenance and Executive Control*, 62–101. New York, NY: Cambridge University Press.

Crossley, Scott. 2013. Applications of Text Analysis Tools for Spoken Response Grading. *Language Learning & Technology* 17(172). 171–192. http://llt.msu.edu/issues/june2013/crossleymcnamara.pdf.

Crossley, Scott & Danielle McNamara. 2012. Predicting second language writing proficiency: the roles of cohesion and linguistic sophistication. *Journal of Research in Reading* 35(2). 115–135. doi:10.1111/j.1467-9817.2010.01449.x. http://doi.wiley.com/10.1111/j.1467-9817.2010.01449.x.

Crossley, Scott & Danielle McNamara. 2014. Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*. Elsevier Inc. 26. 66–79. doi:10.1016/j.jslw.2014.09.006. http://dx.doi.org/10.1016/j.jslw.2014.09.006.

Ellis, Nick C. 1996. Sequencing in SLA. *Studies in Second Language Acquisition* 18(1). 91–126. doi:10.1017/S0272263100014698. https://www.cambridge.org/core/product/identifier/S0272263100014698/type/journal_article.

Fang, Yu & Haitao Liu. 2018. What factors are associated with dependency distances to ensure easy comprehension? A case study of ba sentences in Mandarin Chinese. *Language Sciences*. Elsevier Ltd 67. 33–45. doi:10.1016/j.langsci.2018.04.005. https://doi.org/10.1016/j.langsci.2018.04.005.

Fedorenko, Evelina, Rebecca Woodbury & Edward Gibson. 2013. Direct Evidence of Memory Retrieval as a Source of Difficulty in Non-Local Dependencies in Language. *Cognitive Science* 37(2). 378–394. doi:10.1111/cogs.12021. http://doi.wiley.com/10.1111/cogs.12021.

Ferrer-i-Cancho, Ramon & Haitao Liu. 2014. The risks of mixing dependency lengths from sequences of different length. *Glottotheory* 5(2). 143–155. doi:10.1515/glot-2014-0014.

Ferris, Dana R. 1994. Lexical and Syntactic Features of ESL Writing by Students at Different Levels of L2 Proficiency. *TESOL Quarterly* 28(2). 414–420. doi:10.2307/3587446.

Geertzen, Jeroen, Theodora Alexopoulou & Anna Korhonen. 2013. Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCamDat). *Proceedings of the 31st Second Language Research Forum (SLRF)*, 240–254. Somerville, MA: Cascadilla Press.

Gibson, Edward. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Wayne A. O'Neil, Yasushi Miyashita & Alec Marantz (eds.), *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, 95–126.

Graesser, Arthur C., Danielle S. McNamara & Jonna M. Kulikowich. 2011. Coh-Metrix. *Educational Researcher* 40(5). 223–234. doi:10.3102/0013189X11413260. http://edr.sagepub.com/cgi/doi/10.3102/0013189X11413260.

Greenbaum, Sidney & Gerald Nelson. 1995. Clause relationships in spoken and written English. *Functions of Language* 2(1). 1–21. doi:10.1075/fol.2.1.02gre. http://www.jbe-platform.com/content/journals/10.1075/fol.2.1.02gre.

Hoang, Ha & Frank Boers. 2018. Gauging the association of EFL learners' writing proficiency and their use of metaphorical language. *System*. Elsevier Ltd 74. 1–8. doi:10.1016/j.system.2018.02.004. https://doi.org/10.1016/j.system.2018.02.004.

Housen, Alex, Folkert Kuiken & Ineke Vedder (eds.). 2012. *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*. John Benjamins Publishing.

Hudson, Richard. 2010. *An introduction to word grammar*. Cambridge University Press.

Jiang, Jingyang & Haitao Liu. 2015. The effects of sentence length on dependency distance, dependency direction and the implications-Based on a parallel English-Chinese dependency treebank. *Language Sciences*. Elsevier Ltd 50(866). 93–104. doi:10.1016/j.langsci.2015.04.002. http://dx.doi.org/10.1016/j.langsci.2015.04.002.

Jiang, Jingyang & Jinghui Ouyang. 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. *Physics of Life Reviews*. Elsevier B.V. 21. 209–210. doi:10.1016/j.plrev.2017.06.018. http://dx.doi.org/10.1016/j.plrev.2017.06.018.

Kirkland, Margaret R. & Mary Anne P. Saunders. 1991. Maximizing Student Performance in Summary Writing: Managing Cognitive Load. *TESOL Quarterly* 25(1). 105. doi:10.2307/3587030. https://www.jstor.org/stable/3587030?origin=crossref.

Klein, Dan, Christopher D. Manning. & Christopher D. Manning. 2003. Accurate unlexicalized parsing. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 423–430. Stroudsburg, PA, USA: Association for Computational Linguistics.

Levy, Roger, Evelina Fedorenko, Mara Breen & Edward Gibson. 2012. The processing of extraposed structures in English. *Cognition* 122(1). 12–36. doi:10.1016/j.cognition.2011.07.012. https://linkinghub.elsevier.com/retrieve/pii/S0010027711002010.

Liu, Haitao. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science* 9(2). 159–191. http://www.papersearch.net/view/detail.asp?detail_key=1r300030.

Liu, Haitao, Chunshan Xu & Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*. Elsevier B.V. 21. 171–193. doi:10.1016/j.plrev.2017.03.002. http://dx.doi.org/10.1016/j.plrev.2017.03.002.

Lu, Xiaofei. 2011. A Corpus-Based Evaluation of Syntactic Complexity Measures as Indices of College-Level ESL Writers' Language Development. *TESOL Quarterly* 45(1). 36–62. doi:10.5054/tq.2011.240859. http://openurl.ingenta.com/content/xref?genre=article&issn=0039-8322&volume=45&issue=1&spage=36.

Oomen, Claudy & Albert Postma. 2001. Effects of time pressure on mechanisms of speech production and self-monitoring. *Journal of Psycholinguistic Research* 30(2). 163–184.

Ouyang, Jinghui & Jingyang Jiang. 2018. Can the Probability Distribution of Dependency Distance Measure Language Proficiency of Second Language Learners? *Journal of Quantitative Linguistics*. Routledge 25(4). 295–313. doi:10.1080/09296174.2017.1373991. http://doi.org/10.1080/09296174.2017.1373991.

Paas, Fred, Juhani E. Tuovinen, Huib Tabbers & Pascal W. M. Van Gerven. 2003. Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. *Educational Psychologist* 38(1). 63–71. doi:10.1207/S15326985EP3801_8. http://www.tandfonline.com/doi/abs/10.1207/S15326985EP3801_8.

Qin, Wenjuan & Paola Uccelli. 2016. Same language, different functions: A cross-genre analysis of Chinese EFL learners' writing performance. *Journal of Second Language Writing*. Elsevier Inc. 33. 3–17. doi:10.1016/j.jslw.2016.06.001. http://dx.doi.org/10.1016/j.jslw.2016.06.001.

Sweller, John. 2011. Cognitive Load Theory. *Psychology of learning and motivation*, 37–76. Academic Press. doi:10.1016/B978-0-12-387691-1.00002-8. https://linkinghub.elsevier.com/retrieve/pii/B9780123876911000028.

Tesnière, Lucien. 1959. *Eléments de syntaxe structurale*. Paris: Klincksieck.

Wagner, Elvis. 2014. Using Unscripted Spoken Texts in the Teaching of Second Language Listening. *TESOL Journal* 5(2). 288–311. doi:10.1002/tesj.120. http://doi.wiley.com/10.1002/tesj.120.

Wang, Yaqin & Haitao Liu. 2017. The effects of genre on dependency distance and dependency direction. *Language Sciences*. Elsevier Ltd 59(866). 135–147. doi:10.1016/j.langsci.2016.09.006. http://dx.doi.org/10.1016/j.langsci.2016.09.006.

Wolfe, Edward W., Tian Song & Hong Jiao. 2015. Features of difficult-to-score essays. *Assessing Writing*. Elsevier Inc. 27. 1–10. doi:10.1016/j.asw.2015.06.002. http://dx.doi.org/10.1016/j.asw.2015.06.002.

Zipf, George Kingsley. 1949. *Human behavior and the principle of least effort: an introduction to human ecology*. New York: Hafner.