International Journal of Linguistics Studies
ISSN: 2754-2599
DOI: 10.32996/ijls
Journal Homepage: www.al-kindipublisher.com/index.php/ijlss

IJLS

AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

A Quantitative Study on Academic Texts from an Interdisciplinary Perspective

Mengwei Du

Zhejiang University of Finance & Economics, China

Corresponding Author: Mengwei Du, **E-mail**: 2725269923@qq.com

Measurement studies on the syntactic complexity of academic texts have been widely conducted. Using a quantitative linguistic approach, this study explores the syntactic features of academic texts from an interdisciplinary perspective by analyzing the dependency distance (DD), mean dependency distance (MDD) and dependency types of academic texts from different disciplines in a self-constructed corpus. The corpus includes 400 abstracts of parallel comparable academic texts from eight disciplines authored by native and L2 writers. Research findings indicate that: 1) Both native (L1) and second language (L2) authors' academic texts exhibit a highly consistent distribution of dependency distances, following a long-tail distribution, , which suggests constraints imposed by human working memory. 2) Across all disciplines, abstracts written by L2 authors have significantly longer Mean Dependency Distance (MDD) compared to those written by L1 authors, indicating the significant impact of native language background on the syntactic complexity of academic texts. 3) Within the same language background, there are significant differences in dependency types across different academic disciplines, reflecting specific emphases in how authors express their research findings across disciplines. These findings further reveal the correlation between MDD and disciplinary contexts.

| KEYWORDS

Dependency Distance (DD); Mean Dependency Distance (MDD); Syntactic Complexity (SC); Dependency Types; Disciplines

| ARTICLE INFORMATION

ACCEPTED: 01 February 2025 **PUBLISHED:** 23 February 2025 **DOI:** 10.32996/ijls.2025.5.1.4

1.0 Introduction

Recent years have seen much interest in the study of syntactic complexity. The measurement methods of syntactic complexity (SC) in L2 writing studies are under constant improvement, with various linguistic units being integrated into the measurement to unveil the whole picture of SC development (Jiang et al., 2019: 2).

Dependency distance (DD) emerges as a pivotal measure in understanding syntactic complexity across languages. Initially, Hudson (1995: 16) proposes a definition of dependency distance based on memory decline and short-term memory, i.e., "the distance between words and their parents, measured in terms of intervening words." He (1995: 21) argues that "the distance between a word and its parent is relevant to the difficulty of processing a sentence." Syntactic complexity is then linked to the linear distance between dominant and subordinate words. Liu (2008:167) claims "distance is an important property of a dependency relation because of its implications for the cognitive cost of processing the dependency; likewise, the average dependency distance of a text is an important comparative measure and throws light on the cognitive demands of the language concerned relative to other languages." He (2008) examines a

hypothesis about the use of mean dependency distance (MDD) as a measure of processing complexity and finds that MDD can effectively and sensitively reflect syntactic complexity. Ouyang et al. (2022: 3) also point out some advantages of using dependency distance as a measure of L2 syntactic complexity. First, dependency distance is a cross-linguistic metric, revealing the ways in which cross-linguistic factors influence syntactic complexity. Second, it is efficient and convenient to extract dependency relations (types) and calculate dependency distance. Finally, dependency distance is useful in assessing the beginners' compositions with run-on sentences.

Studies have shown that MDD is a good predictor of syntactic complexity (Hao et al., 2022; Jiang et al., 2019; Jiang & Ouyang, 2018; Li & Yan, 2021; Ouyang et al.) In-depth studies on MDD have found that it can reflect the degree of difficulty in second language processing and effectively differentiate learners at different levels (Ouyang & Jiang, 2018). However, MDD in academic texts written by authors of the same or similar level but with different native language backgrounds has not been sufficiently investigated, and research on the disciplinary effects of MDD in academic texts has been restricted to a limited number of subject areas, such as applied linguistics (Casal, 2020); linguistics, physics, and chemistry (Gao & He, 2023); and anthropology, chemical engineering, electrical engineering, and sociology (Lu et al., 2021). Systematic and comprehensive analyses of syntactic complexity in other disciplinary areas have not yet been conducted. Furthermore, the genres favored by scholars studying academic writing are mostly published journal articles (Biber & Gray, 2016) rather than professional writing.

Based on this, the present study attempts to examine the SC of academic texts by investigating the MDD of academic texts written by authors from different native language backgrounds from an interdisciplinary perspective, exploring whether there are significant differences in the MDD of academic texts written by authors from different native language backgrounds and from different disciplines. Additionally, the study aims to explore the probability distribution of MDD and dependency relations in academic texts.

2.0 Literature review

Multiple studies on dependency distance from multi-dimensions have concluded that dependency distance and dependency direction are effective indices for language categorization (Liu, 2009, 2010; Liu & Xu, 2012; Chen & Gerdes, 2017). Meanwhile, mean dependency distance proves useful and sensitive in reflecting the degree of SC. This chapter aims to figure out the development of MDD theory and the findings of studies on it.

2.1 Dependency Distance

Dependency Grammar describes unequal syntactic relations between two words in a sentence. One of the two words acts as the governor and the other as the dependent (Heringer, 1993; Gibson, 1998; Tesnière 1959; Liu, 2009; Hudson, 2010). It is a descriptive method of natural language structures based on dependency structure (Liu, 2009), broadly endorsed within the domains of computational linguistics (Nivre, 2006) and theoretical linguistics (Hudson, 2007).

Hudson (1995: 1) introduces "dependency theory, and in particular the version of dependency theory called Word Grammar." He (1995: 2) defines dependency distance as "the number of words that separate a word from the word on which it depends" and applies it to some well-known processing difficulties. In his study, three sources of parsing difficulty have been considered: two quantitative (dependency distance and dependency density) and one qualitative (bad dependency decisions which lead to centre-embedded failures). The direction of dependency (i.e., whether the head is initial or final) is also a significant factor in parsing. Based on the relation, the following core properties of syntactic dependencies have been identified:

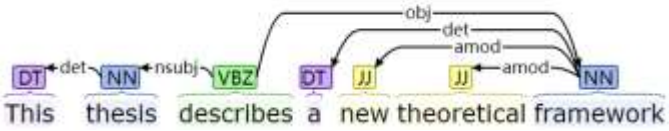

Figure 2.1 Dependency structure of sample sentence This thesis describes a new theoretical framework.

In Figure 2.1, all the words in a sentence are connected by grammatical relations. For example, the subject and the object depend on the main verb; prepositions (not exemplified in Figure 2.1) depend on the nouns or verbs that they modify; and so on. In each pair of connected words, one is called the dependent and the other is called the governor. The labeled line with an arrow is directed from the governor to the dependent.

Recently, power-law distribution, which was proposed by Zipf (1936) to describe word frequencies and their ranks, has been applied to model individual DDs in many languages and has been found to follow certain rules. Numerous previous studies have corroborated that the distribution of dependency distances of natural languages can well fit exponential distribution or power law distribution (Ferrer-i-Cancho, 2004; Liu, 2007; Jiang and Liu, 2015; Ouyang & Jiang, 2017; Lu & Jiang, 2023). For example, Ouyang and Jiang (2017) found that the Right truncated modified Zipf-Alekseev distribution well captures the probability distribution of dependency distance of second language learners' natural languages at each grade and of native speakers' natural languages. Hao et al. (2023) further found most of the corresponding goodness of fits were excellent, demonstrating that the power law distribution is universal, not only in natural language, but also in interlanguage and the general application of the laws of quantitative linguistics.

2.2 MDD

Liu et al. (2009: 162), furthermore, propose that "distance is an important property of a dependency because of its implications for the cognitive costs of processing the dependency." They describe a method for calculating the dependency distance between the words in a text, i.e., the number of words that separate each word from the word on which it depends syntactically, using the following formula:

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

Here n is the number of words in the sentence and DD_i is the dependency distance of the i-th syntactic link of the sentence. In order to illustrate the method of analysis, the dependency distances for the different dependencies of the example sentence presented in **Error! Reference source not found.** are shown in **Error! Reference source not found.** below. The MDD of the example sentence is (1+1+0+3+2+1+4)/6=2.

Table 2.1 Dependency distance of sample sentence *This thesis describes a new theoretical framework*

Head	Dependent	Dependency type	ependency distance	ependency distance (Absolut
thesis	This	det	1	1
describes	thesis	nsubj	1	1
describes	describes	ROOT	0	0
framework	a	det	3	3
framework	new	amod	2	2
framework	theoretical	amod	1	1

describes	framework	dobj	-4	4

Mean dependency distance proves to be an essential index for predicting syntactic difficulty and writing proficiency (Jiang & Ouyang, 2018; Liu et al., 2017; Ouyang et al., 2022), as "MDD can be used to account for processing difficulty in the case of long-distance dependency and center-embedding sentences" (Liu, 2008: 171). What's more, dependency distance, in comparison with dependency length, is probably a better term that can reflect language processing as a dynamic cognitive and psychological process (Liu & Liang, 2017: 172). In addition, Ouyang et al. (2022) argue that MDD is a more promising measure of L2 syntactic complexity than traditional sentence length-based measures. They point out that the overall MDD of learner compositions (i.e., total DDs/total number of dependencies in the composition) is statistically better able to distinguish L2 learners at all neighboring language levels from Pre-A1 to C1, according to the Common European Framework of Reference for Languages (CEFR).

2.3 Previous studies on MDD

Several areas of research have been covered in the domain of syntactic complexity (SC) in second language writing research. A series of studies focus on the relationship between syntactic complexity and writing quality examining learner development in syntactic complexity over time (Lu, 2010; Lu & Ai, 2015; Jiang et al., 2019; Ouyang et al., 2022), among which several measures were used, such as mean dependency distance (MDD), mean length of clause (MLC); mean length of T-unit (MLT); mean length of sentence (MLS); dependent clauses per clause (DC/C); coordinate phrases per clause (CP/ C); T-units per sentence (T/S); complex nominal per clause (CN/C) and etc. Compared with traditional SC measures, mean dependency distance (MDD) is widely considered to be a more robust metric for analyzing syntactic complexity in academic writing (Ouyang et al., 2022). It offers insights into writing proficiency, processing speed, disciplinary effect, etc., contributing to both theoretical advancements and practical applications in L2 writing education.

Firstly, in the field of second language acquisition, mean dependency distance is considered to be an important indicator of how well learners understand the syntactic structure of the target language. Ouyang et al. (2022) combine traditional SC measures with newly-proposed MDD measures to better assess second language (L2) SC development. Based on a syntactically-annotated corpus of 400 compositions, they find the overall mean dependency distance can significantly discriminate all pairs of adjacent proficiency levels, serving as the best metric explored in their study. Likewise, Gao and He (2023) conduct a corpus-based study analyzing 400 PhD dissertation abstracts written by native English (L1) and English as a foreign language (L2) academic writers. They find that L2 writers tend to produce a longer mean dependency distance (MDD) than L1 writers, attributed to the frequent use of prepositional phrases by L2 writers. This suggests that while L2 writers have achieved native-like proficiency in extending nominal structures, they still exhibit distinct syntactic patterns.

Moreover, dependency distance can further explain the findings of the traditional SC measures from the perspective of language processing (Ouyang et al., 2022). Studies have found that DD correlates with reading time in L1 psycholinguistic research, suggesting that it is an indicator of language processing difficulty and memory burden (Gibson, 1998; Liu & Liang, 2017; Jiang et al., 2019; Chen et al., 2022; Gao & Sun, 2024). Gao and Sun (2024) explore the relationship between DD, L2 processing speed, and L2 proficiency. Using a maze task to capture processing speed, they discover that DD significantly predicts L2 processing speed at both word and sentence levels, regardless of participants' L2 proficiency levels. This study provides empirical evidence that DD is a useful predictor of L2 processing difficulty and can differentiate L2 proficiency levels. Furthermore, Liu and Liang (2017: 189) conclude that "human languages seem to present a preference for short dependency distance, which may be explained in terms of the general cognitive constraint of limited working memory," aligning with Hudson's (1995: 19) discovery that one way in which languages help their users is by adjusting their word order rules to allow dependency distances to be short. The preference for short dependency distance widely presented in natural languages is defined as dependency distance minimization (DDM) (Liu et al., 2017: 173), that is, "a propensity to syntactically structure

sentences in such a way so as to minimize its overall dependency distance."

In addition, disciplinary differences in academic writing have been noted, with certain disciplines exhibiting unique syntactic features (Durrant & Mathews-Aydınlı, 2011; Omidian, 2018; Casal, 2020; Casal et al., 2021). Durrant and Mathews-Aydınlı (2011) demonstrate a 'function-first' approach through a comparative analysis of introductions to student essays and research articles and find that both the choice to use the function and the choice of linguistic forms that realize the function vary across subject areas in research articles, but not in student essays and that research articles tend to be more formulaic in expressing the function than student essays. Their findings, using a function-first approach, have demonstrated significant differences in syntactic features across various disciplines. Moreover, Omidian (2018) examines how hard and soft science disciplines reflect their epistemological orientations through the use of recurrent word combinations in research article abstracts. A corpus-driven and mixed-methods approach revealed discipline-specific priorities in representing research across different rhetorical moves in academic abstracts.

Casal (2020) reveals that Applied Linguistics trended towards high levels of complexity and Economics towards lower levels across the assessed measures in RA Introductions in an analysis which included two Engineering disciplines. This suggests that these disciplinary differences detected here exist beyond social sciences. What's more, Casal et al. (2021) examine syntactic complexity across different academic research article part-genres from a cross-disciplinary perspective and highlight a significant large effect of both discipline and part-genre on all eight syntactic complexity indices, as well as a significant but small effect size for the interaction of move and discipline on the complexity measures. Their study underscores the importance of considering disciplinary norms and expectations when analyzing academic texts.

To summarize, MDD has been verified to be good at measuring syntactic complexity (Jiang & Ouyang, 2018; Jiang et al., 2019; Li & Yan, 2021; Hao et al., 2022; Ouyang et al., 2022), discriminating between learners with different proficiency levels (Ouyang & Jiang, 2018), reflecting L2 processing difficulty (Hudson, 1995; Liu & Liang, 2017; Jiang et al., 2019; Chen et al., 2022; Gao & Sun, 2024), reacting to different disciplines (Durrant & Mathews-Aydınlı, 2011; Omidian, 2018; Casal, 2020; Casal et al., 2021) and etc.

These studies are important because they provide evidence for the strong connection that exists between MDD and SC, and explore the MDD of academic texts from multi-dimensions. What is largely missing in the described studies, however, is a systematic analysis of the link between MDD and the disciplines across different author's native language backgrounds. Specifically, more empirical evidence is needed to develop a better understanding of academic writing across different disciplines through the lens of the strong connection that exists between MDD and SC.

2.4 Current research

The current study intends to make a contribution to the field of syntactic complexity and specifically to this unclear issue, addressing the difference in syntactic complexity between native English (L1) and English as a foreign language (L2) academic writers whose level of English is considered to be similar. The present study focuses on MDD of academic writing from different disciplines. This makes a novel contribution due to the scarcity of systematic and comprehensive studies on the relationship between MDD and disciplines in the literature. As an additional contribution to the literature, the study includes dependency relations in the analysis and further explores the question of whether there is some difference in dependency relations across disciplines. The research questions are as follows:

1) What is the distribution of DD in academic texts authored by individuals from different language backgrounds?

2) Are there significant differences in the Mean Dependency Distance (MDD) of academic texts across different language backgrounds and disciplines?

3) Do the dependency types of academic texts vary significantly across different language backgrounds and disciplines?

3.0 Methodology

3.1 Research corpora

In light of the previous discussions, this study aims to investigate whether MDD can be used to discriminate different disciplines and native language backgrounds. The abstract is chosen as the part-genre of analysis. Given that authors typically view the abstract section as a concise overview of their research and its findings, abstracts serve as an optimal platform for examining how authors across different disciplines commonly highlight various aspects of their research and identify which elements are frequently regarded as pivotal promotional tools for a paper. Moreover, choosing the abstract section as the unit of analysis allows for the inclusion of more texts from different disciplines in the corpus, let alone the labor-intensive cleaning-up of the raw texts. Consequently, we believe that an investigation into disciplinary disparities within this specific section can yield valuable insights into how authors in diverse academic fields align with readers' expectations through the crucial role played by abstracts in publicizing and summarizing attached articles.

A self-built corpus is adopted in this study. The corpus consisted of 400 PhD dissertation abstracts from eight disciplines (literature, economics, law, education, science, engineering, agriculture, and medicine). These eight disciplines encapsulate most of the humanities and natural sciences. Within each field, half of the dissertations were selected from the Chinese National Knowledge Infrastructure (CNKI) and the other half from the ProQuest Dissertations & Theses (PQDT). Specifically, there are 200 abstracts for each of the first and second language learners' papers. Each amount of each discipline is shown in Table 3.1.

Table 3.1 The distribution of PhD dissertation abstracts across eight disciplines

Discipline		L1	L2	Total
Humanities	Literature	25	25	200
	Economics	25	25	
	Law	25	25	
	Education	25	25	
Natural Sciences	Science	25	25	200
	Engineering	25	25	
	Agriculture	25	25	
	Medicine	25	25	
Total		200	200	400

3.2 Methods

In the phase of corpus processing, this study initially utilizes the Wordless to extract comprehensive information on DD and dependency types from the L1 and L2 corpora. The Altmann-Fitter is then employed to fit the DD data, thereby obtaining the overall distribution of DD for both corpora. Subsequently, Wordless is employed once more to process texts from a range of disciplines on an individual basis, thus enabling the calculation of the MDD for each discipline. Ultimately, the data is analyzed using the statistical software package SPSS, with particular attention paid to the significance testing. Independent sample tests are conducted to ascertain whether significant differences in MDD exist between texts authored by individuals with disparate native language backgrounds and between texts

from distinct disciplinary domains.

Obviously, understanding the syntactic differences between native and second language speakers is crucial for developing effective language teaching methodologies. Dependency relations, which describe the grammatical relationships between words in a sentence, serve as a valuable indicator of syntactic complexity and variation. This study compares the frequencies of various dependency relations in L1 and L2 corpora to uncover patterns and potential areas of difficulty for L2 learners. The collected corpus has been imported into Wordless, and the information on the L1 and L2 sub-corpora is shown in Table 3.2.

Table 3.2 Descriptive statistics of the corpus

Sub-corpora	Tokens	Type-Token Ratio (TTR)	Texts
L1	52789	0.17	200
L2	172876	0.07	200

It is evident that the word count of abstracts written by 200 second language (L2) authors is nearly three times that of abstracts written by native language (L1) authors. However, despite the significantly lower word count in L1 abstracts compared to L2, the Type-Token Ratio (TTR) in L1 is higher. TTR is a useful quantitative measure of text diversity and lexical richness; a higher TTR indicates greater text diversity and richer vocabulary. This suggests that even with comparable language proficiency levels, abstracts written by L1 authors exhibit higher lexical richness than those by L2 authors. Now, how does the syntactic complexity compare between these groups?

4.0 Results

4.1 Distribution of DD

Constrained by human working memory capacity, natural languages tend to minimize DD, which causes the distribution of DDs to follow certain Zipf-like laws. To answer the first research question, the individual dependency distances of the academic texts written by authors from different language backgrounds are extracted.

The Waring model and the Zipf-Alekseev model are of great importance in the field of linguistics, where they are used to model word frequency distributions. Köhler and Altmann (2000) demonstrated the applicability of the Waring model in fitting distributions of syntactic units. In contexts where values are constrained within specific ranges, truncated models are frequently the preferred option. Liu (2007) posited that the Right truncated modified Zipf-Alekseev model is an effective tool for fitting probability distributions of dependency distances. Our findings corroborate this conclusion. Undoubtedly, the coefficient of determination, R^2, is a standard measure of goodness of fit (Liu at al., 2009; Lu & Jiang, 2023). Table 4.1 presents the fitting outcomes of L1 and L2 corpora using two truncated models.

Table 4.1 Fitting models for the dependency distances in L1 and L2 corpora

Corpus	ncated modified Zipf-Alekseev (x-max, α fixed)	Right truncated Waring (b,n)
L1	0.9981	0.9996
L2	0.9962	0.9990

The Right truncated modified Zipf-Alekseev model (a, b; n = x-max,α fixed) and the Right truncated Waring (b,

n) model both demonstrate robust performance across both corpora, with R^2 values exceeding 0.99. It is noteworthy that the Right truncated Waring (b, n) model demonstrates an exceptional level of fit (R^2 > 0.999). In light of the aforementioned findings, our study employs the Right truncated Waring model to represent the distribution of dependency distances, as illustrated in Table 4.2.

Table 4.2 Fitting the Right truncated Waring to the dependency distances of L1 and L2 corpora

					5				

Figure 4.1 Right truncated Waring Model Fit in L1 corpus

Figure 4.2 Right truncated Waring Model Fit in L2 corpus

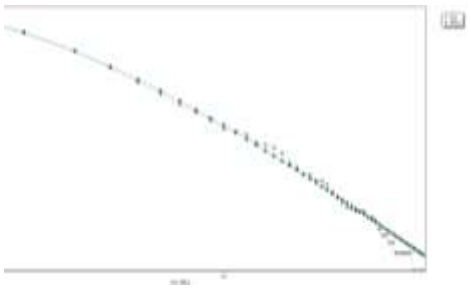

Figure 4.3 Log-log plot of Right truncated Waring Model fit in L1 corpus

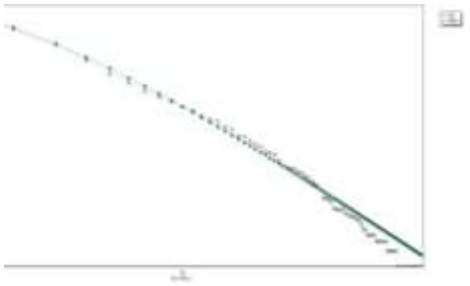

Figure 4.4 Log-log plot of Right truncated Waring Model fit in L2 corpus

These results demonstrate that academic texts authored by both L1 and L2 writers exhibit analogous power-law distributions in dependency distance characteristics, as depicted in Figures 4.1 to 4.4. This phenomenon could be governed by human cognitive load capacities and the principle of minimal effort (Hudson, 1996; Köhler and Altmann, 2000; Ferrer-i-Cancho, 2004; Liu, 2007; Lu & Jiang, 2023). In academic writing, authors tend to prefer shorter dependency distances to reduce cognitive load for both themselves and their readers. It is noteworthy that despite differing language backgrounds, similar regularities are observed compared to native speakers, indicating universality in language usage patterns.

Given the consistent distribution of dependency distances, further exploration is needed to understand how the average dependency distance varies across academic texts in different disciplines authored by speakers with diverse language backgrounds.

4.2 Distribution of MDD across language backgrounds and disciplines

Using the Wordless tool, the two sub-corpora L1 and L2 underwent processing to extract dependency

relationships and dependency distances for each sentence within the corpus. Subsequent statistical analysis yielded the average dependency distance for each discipline, as presented in Table 4.3.

Table 4.3 Descriptive statistics of MDD in L1 and L2 sub-corpora

Discipline		L1	L2	Mean
Humanities	Economics	2.36	2.55	2.46
	Education	2.32	2.71	
	Law	2.36	2.62	
	Literature	2.29	2.49	
Natural Sciences	Agriculture	2.38	2.70	2.54
	Engineering	2.40	2.82	
	Medicine	2.43	2.68	
	Science	2.34	2.53	
Mean		2.36	2.64	2.50

Based on the data presented in the table, several key insights emerge regarding the Mean Dependency Distance (MDD) in academic writing across different author backgrounds and disciplines. Firstly, it is evident that second language (L2) authors tend to exhibit longer MDD values compared to native language (L1) authors, with the former being 2.36 and the latter 2.64. The independent sample t-test further confirms the significance of MDD variation ($t(14)=-6.588$, $p<0.01$) across different language backgrounds as shown in Table 4.4.

Table 4.4 Comparison of MDD between L1 and L2

	L1 ($n=8$)		L2 ($n=8$)		*MD*	*t*(14)
	M	*SD*	*M*	*SD*		
MDD	2.36	0.44	2.64	0.11	-0.28	-6.588

*$p<0.01$

Inspections of the two group means indicate that the average MDD of students' English composition given by native writers (2.36) is significantly lower than that given by second language writers (2.64). The difference between the means is -0.28. This suggests that, despite potential linguistic challenges, L2 writers tend to use more complex syntactic structures in their academic abstracts, indicating, on the one hand, that they strive for accuracy and clarity, and, on the other hand that L2 writers have reached or even surpassed the level of native writers. The greater degree of complexity in L2 abstracts also aligns with other studies (e.g., Guo et al., 2013; Crossley & McNamara, 2014; Gao & He, 2023), despite variations in the syntactic complexity metrics employed.

Secondly, in the interdisciplinary comparison, although the natural sciences had slightly higher MDD values than the humanities, the overall difference was not statistically significant. This observation indicates that the two general fields exhibit comparable levels of syntactic complexity in academic discourse, highlighting the importance of clarity and technical precision in different disciplinary contexts. See Table 4.3.

Within disciplines, however, native language authors demonstrate consistent MDD values across different fields, with medical texts registering the highest MDD at 2.43 and literary texts the lowest at 2.29. This consistency suggests that the syntactic demands within disciplines are relatively stable among L1 authors. Conversely, among L2 authors, engineering texts exhibit the highest MDD (2.82), surpassing even the highest MDD observed among L1 authors,

while literary texts by L2 authors (2.49) mirror the lowest MDD seen in native language literature. This variability within disciplines for L2 authors underscores the nuanced challenges and adaptations made in academic writing across different fields. See Figure 4.5.

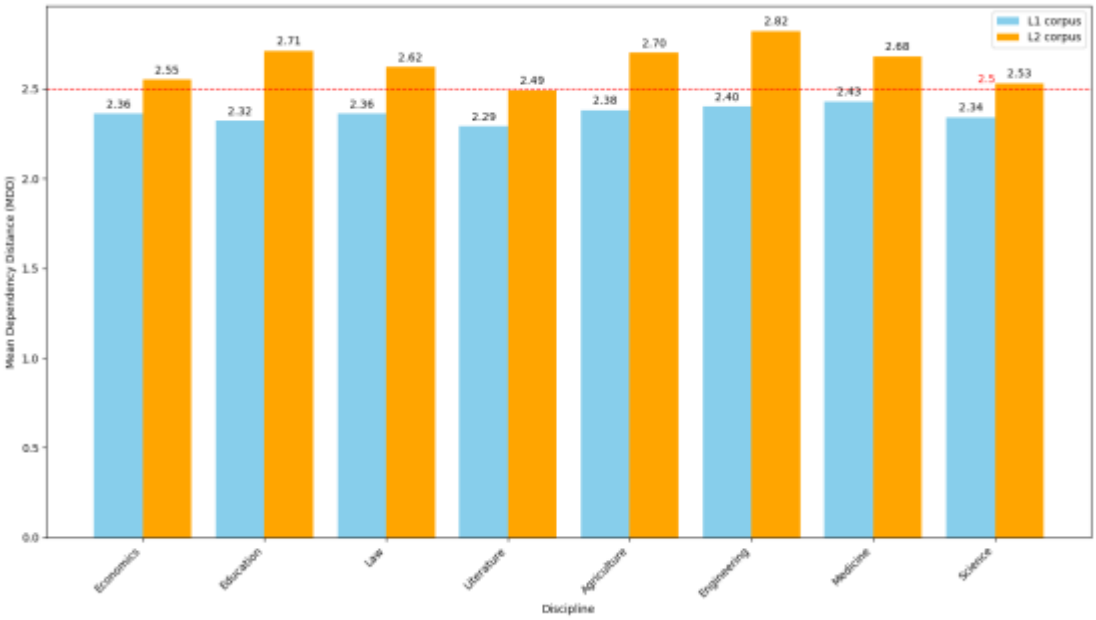

Figure 4.5 Figure Description of MDD across disciplines in L1 and L2 sub-corpora

In addition, Figure 4.5 illustrates that MDD in each discipline of L2 is higher than that in L1. MDD values across disciplines in L1 are all below 2.5, whereas in L2, nearly all disciplines have MDD values above 2.5. The lowest mean difference in L2 is observed in academic literature, with a value of 2.49, which remains higher than the highest mean difference in L1 found in academic medicine, with a value of 2.43. Briefly, these findings underscore the influence of linguistic background on syntactic complexity in academic writing, with L2 authors often striving to achieve comparable levels of precision and complexity as their L1 counterparts, albeit with observable variations across disciplines.

To sum up, the significant differences between the MDDs indicate different syntactic complexities between academic texts written by second language writers and native English writers, as well as between the eight disciplines. However, it remains unknown how native English writers utilize specific syntactic structures to differ from non-native English writers in terms of dependency distance. Therefore, a more detailed investigation into dependency relations closely associated with MDD differences is needed.

4.3 Distribution of Dependency types across language backgrounds and disciplines

The dependency relations were parsed and counted, yielding frequencies for each relation type. To ensure a fair comparison, the frequencies in the L2 corpus were adjusted for corpus size differences. The most frequent dependency relations were identified and compared across the two corpora.

Table 4.5 Comparison of Dependency Type Frequencies Between L1 and L2 Corpora

L1		L2	
Dependency type	Frequency	Dependency type	Frequency(adjusted)
prep	6960	prep	6928
pobj	6614	pobj	6637
amod	5509	amod	5855
det	5194	det	5711
compound	4858	compound	5133
nsubj	2699	conj	2847
dobj	2663	cc	2767
conj	2076	dobj	2413
cc	2070	nsubj	2289
advmod	1586	advmod	1515
aux	1317	aux	1178
auxpass	835	nmod	972
nmod	826	auxpass	646
nsubjpass	765	nsubjpass	586
nummod	691	advcl	506
advcl	605	nummod	493
acl	597	ccomp	461
relcl	555	attr	454
mark	541	mark	439
poss	477	acl	414

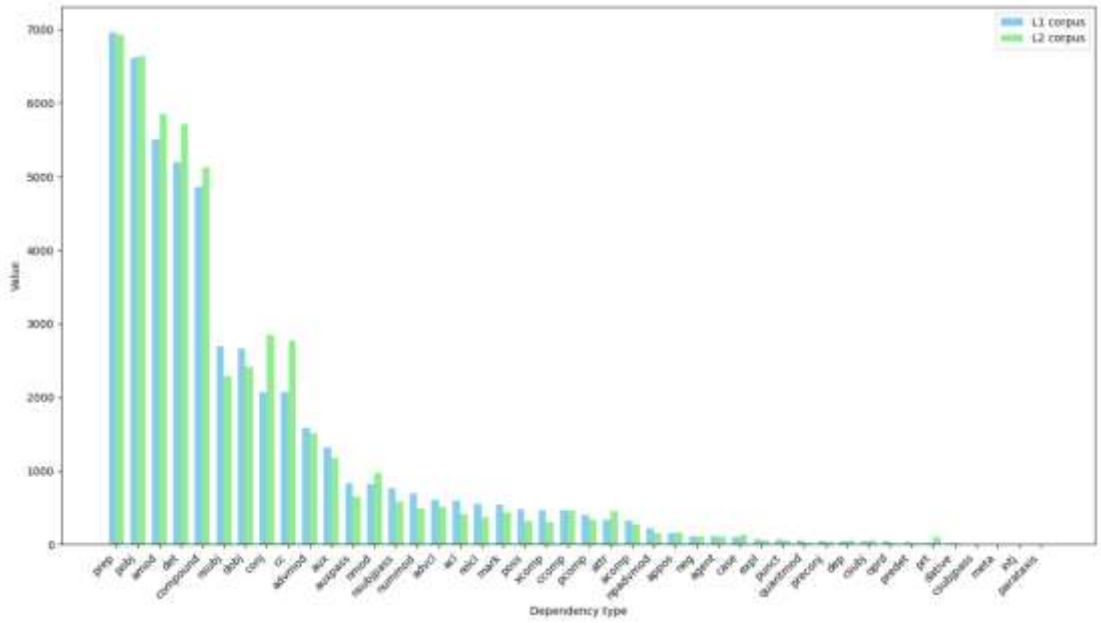

Figure 4.6 Comparison of Dependency Type Frequencies Between L1 and L2 Corpora

The results demonstrate that the frequency distributions of L1 and L2 remain largely consistent. A comparison of the frequency of dependency relations between L1 and L2 authors reveals a high degree of consistency in the use of basic syntactic structures, including *prep* (prepositional modifier), *pobj* (object of a preposition), *amod* (adjectival modifier), and *det*

(determiner). This indicates that both native and non-native authors utilize these fundamental components consistently in their academic writing.

However, certain relations exhibit significant discrepancies. Initially, a notable distinction emerges with L2 authors employing *amod*, *det*, and *compound* structures more frequently than their L1 counterparts. Each of these categories shows an approximate increase of 300 occurrences in L2 compared to L1. This trend suggests that L2 authors may place a greater emphasis on descriptive precision and syntactic complexity in their academic writing, potentially influenced by educational strategies emphasizing these linguistic elements. Secondly, L1 authors demonstrate a comparable prevalence of *nsubj* (nominal subject) and *dobj* (direct object) usage, whereas L2 authors exhibit a heightened frequency of *dobj* usage. This may be indicative of a greater reliance on the object in the construction of complex sentences among L2 authors. Additionally, there are notable differences in the use of connectives with parallel structures between L1 and L2 authors. With the same volume, L2 authors employed both *conj* (conjunct) and *cc* (coordination) approximately 800 times more frequently than L1 authors, which may be related to their preference for expressing juxtapositions. Furthermore, the utilization of auxiliary verbs, which are employed to form tenses, inflections and moods, differs considerably from that of the passive voice. The use of the term *aux* (auxiliary) and *auxpass* (passive auxiliary) is more prevalent among native writers, which may be attributed to their greater flexibility in expressing tense and inflection changes. Conversely, L2 learners may be influenced by their native language to favor the use of the active voice and may exhibit less comfort than L1 authors in utilizing complex tenses and inflections.

Briefly, the analysis of dependency relation frequencies reveals syntactic similarities and differences between native and second language authors. L2 learners demonstrate a tendency to overuse certain dependency relations, such as *det*, *conj* and *cc*, while underusing others, like *dobj*. These findings underscore the nuanced differences in syntactic strategies between native and non-native writers, reflecting varying approaches to language use and textual coherence in academic contexts. Understanding these patterns can inform language instruction and curriculum development to better support L2 learners in achieving syntactic proficiency comparable to native speakers in academic contexts.

In addition, it is worth investigating whether the discrepancies between the L1 and L2 corpora with regard to dependency types are attributable to factors other than the authors' native language background. Furthermore, it remains to be explored whether disciplinary factors not only impact MDD but also influence the preference for the utilization of dependency types. This study will therefore further examine the usage of dependency types across different disciplines. To facilitate comparison, the frequency of dependency types in each discipline will be adjusted to align with the tokens within each discipline to ensure a balanced analysis.

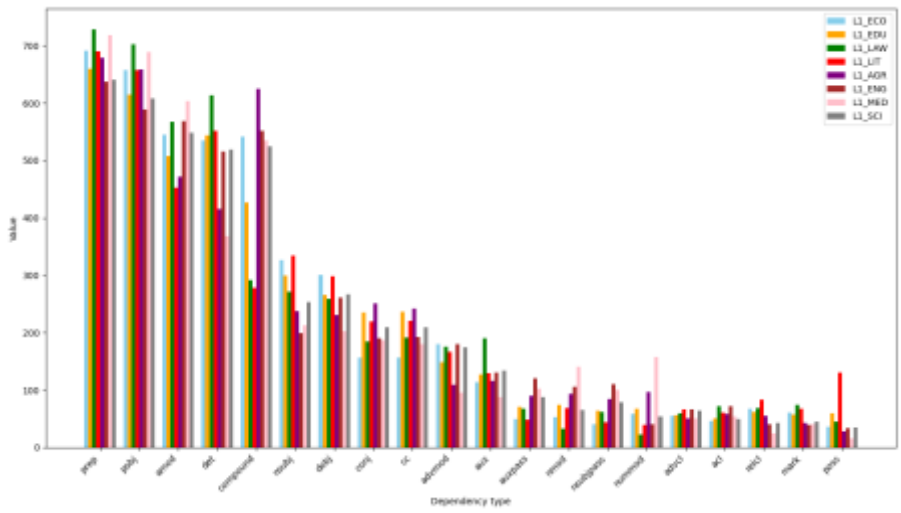

Figure 4.7 Comparison of Dependency Types Across Disciplines in L1 Corpus

The bar chart illustrates the distribution of dependency relations in academic texts written by L1 authors across various disciplines, revealing significant trends and variations. High-frequency dependencies, such as *prep* and *pobj*, are consistently prevalent across all disciplines, underscoring the essential role of prepositional phrases in constructing complex and precise sentences. Similarly, *amod* and *det* are frequently used, reflecting the importance of descriptive detail and specificity in academic texts. Notably, these dependencies are particularly prominent in law and medicine, where detailed descriptions are crucial. In addition, the highly frequent use of *compound* structures across natural science highlights the significance of complex noun phrases in conveying detailed information efficiently. It is noteworthy that agricultural texts, which are not significant in terms of frequency among the top four dependency types, exhibit a high prevalence of *compound*. This significantly contributes to the high MDD (2.38) observed in this research area. Conversely, texts in law and literature, which frequently employ the first four dependency types, exhibit the lowest frequency of *compound* dependencies and lower MDDs of 2.36 and 2.29. This finding gives us insights to the relation between dependency types like *compound* and MDD.

Moderate-frequency dependencies, such as *nsubj* and *dobj*, also show variations across disciplines, with humanities texts exhibiting higher usage frequencies. Specifically, in *nsubj* and *dobj* dependencies, the top three highest frequencies are observed in texts belong to economics and literature, whereas agricultural and medical texts demonstrate the lowest frequencies. In contrast, the highest usage frequencies for the *conj* and *cc* dependencies are observed in agricultural texts, with economic text showing the lowest frequency. High frequencies of them suggest a tendency towards more complex and coordinated sentence structures. Regarding the use of *advmod* (adverbial modifier), there is a noticeable distinction between humanities and scientific texts. Humanities texts, such as those in literature and law, display relatively consistent usage frequencies. However, scientific and engineering texts show much higher frequencies, nearly double that of agricultural and medical texts. This indicates a stronger reliance on adverbial modifiers to add nuance and detail in scientific writing.

Furthermore, low-frequency dependencies offer insights into disciplinary differences. For example, *aux* is most prevalent in the field of law, likely due to the necessity for particular descriptive constructs within this domain. Likewise, *auxpass*, *nmod* (nominal modifier) and *nsubjpass* (passive nominal subject) show higher usage in natural science like engineering and medicine, where nuanced tense and voice constructions are more prevalent compared to humanities disciplines. What's more, the utilization of *nummod* (numeric modifier) also exhibits considerable variation across disciplinary domains, with medical texts representing the most prevalent instance, followed by agricultural texts. This is primarily attributable to the fact that research in these two disciplines frequently entails the administration of reagents, the calculation of ratios of chemical elements, the determination of temperatures, and other such variables where data utilization is commonplace. Other dependencies, such as, *advcl* (adverbial clause modifier), *relcl* (relative clause modifier) and *poss* (possession modifier), are less frequently used but slightly higher in literature, suggesting the complex syntactic structures in literary analysis and critique.

In conclusion, the distribution of dependency relations in L1 authors' academic texts varies across disciplines, reflecting different syntactic strategies and priorities. High-frequency dependencies highlight the universal importance of prepositional phrases and descriptive details, while moderate-frequency dependencies reveal different emphases on subject-object relationships and sentence complexity. Low-frequency dependencies further underscore the unique syntactic needs of different disciplines. These findings underscore the distinct syntactic strategies employed in different academic disciplines, highlighting the varying emphasis on specific dependency relations based on the disciplinary context.

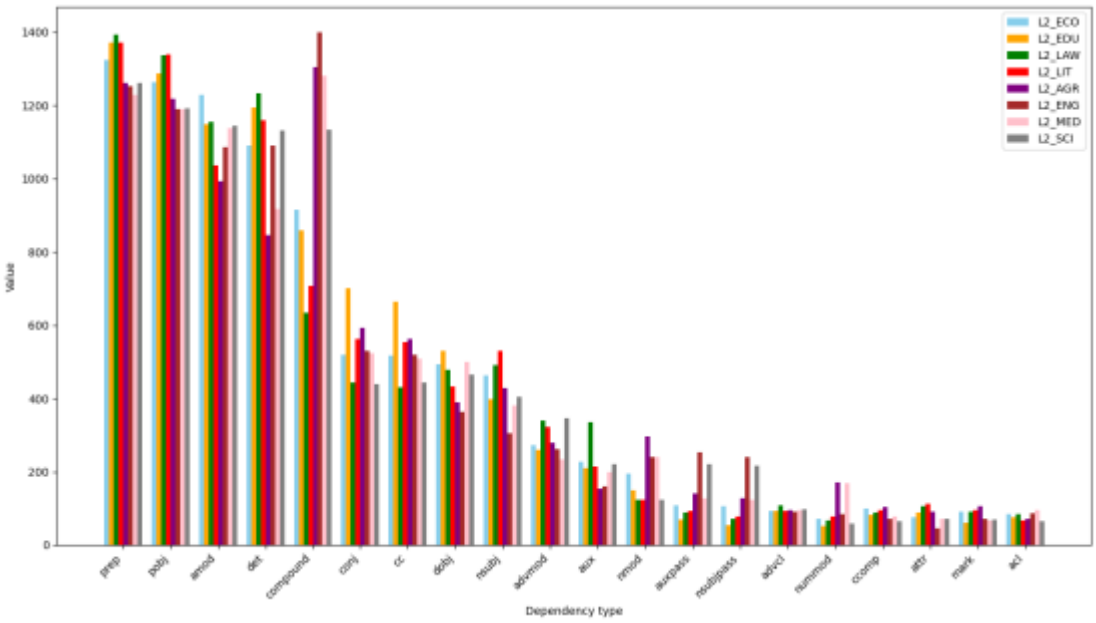

Figure 4.8 Comparison of Dependency Types Across Disciplines in L2 Corpus

Figure 4.8 offers a comprehensive examination of the prevalence of distinct dependency types in academic texts produced by L2 authors across a range of disciplines. The data demonstrate a frequency distribution of dependency types that is largely consistent with L1, particularly the four most frequently used dependency types. However, there are notable differences in the discipline-specific use of some dependency types. The first noteworthy observation is the amazing utilization of the *compound* dependency type by L2 authors, particularly in natural science texts. In fact, the frequency of *compound* use in engineering texts is the highest among all dependency types. In contrast, the lowest frequency of *compound* is observed in law texts, representing less than half of that observed in engineering texts. The markedly elevated frequency of *compound* dependency utilization in the natural sciences is approximately twice that observed in the humanities, thereby reinforcing the correlation between the dependency type and syntactic complexity previously illustrated in Figure 4.7.

With regard to moderate-frequency dependencies, notable difference is that *nsubj* and *dobj* dependencies appear before *conj* and *cc* dependencies. The highest frequencies for *nsubj* and *dobj* dependencies are observed in economic and literary texts. This indicates that L2 authors in these fields prioritize the establishment of clear subject-object relationships in order to structure their sentences in an effective manner. In contrast, the most prevalent instance of *conj* and *cc* dependencies can be observed in agricultural and educational texts, while economic texts demonstrate the lowest frequencies for these dependencies, indicating the presence of distinct syntactic preferences.

The utilization of low-frequency dependencies in the L2 corpus is essentially analogous to that observed in L1. In particular, the highest frequency of occurrence of the dependency *aux* in the L2 corpus is in the field of law, which is significantly higher than in other disciplines. Furthermore, L2 authors demonstrate comparable patterns to those observed in L1 authors with regard to the usage of *nmod*, *auxpass*, and *nsubjpass* dependencies. Notably, natural sciences texts exhibit a higher usage of these dependencies compared to humanities. This phenomenon also elucidates why natural sciences have longer MDDs and more complex syntactic structures compared to humanities.

Briefly, the chart highlights the diverse syntactic strategies employed by L2 authors across different academic disciplines. The variations in dependency type usage reflect the specific linguistic and rhetorical demands of each field, offering valuable insights for targeted academic writing instruction and support for L2 learners.

5.0 Conclusion

By employing traditional syntactic complexity measures and dependency distance measures, we have explored

the syntactic complexity in a self-built corpus of 400 abstracts of Master's and Doctoral theses written by native and second language learners across eight disciplines.

In terms of DD distribution, it is found that academic texts authored by both L1 and L2 writers exhibit analogous power-law distributions in dependency distance. In academic writing, authors tend to prefer shorter dependency distances, which could be governed by human cognitive load capacities and the principle of minimal effort.

Research findings also indicate a significant difference in the MDD of abstracts written by authors from different language backgrounds ($p < 0.01$). Specifically, academic texts authored by L2 authors exhibit longer MDD compared to those written by L1 authors (Mancilla et al., 2017; Casal & Lu, 2021; Gao & He, 2023). Moreover, across all disciplines, texts in the natural sciences display longer MDD compared to those in the humanities, regardless of whether the author is an L1 or L2 speaker. This observation implies higher syntactic complexity in natural sciences texts because of the specialized and complex nature of these disciplines, which contradicts some previous studies (eg. Gao & He, 2023).

Furthermore, this study identifies specific characteristics of the use of dependent types. On one hand, a significant difference is observed between the preference of L1 and L2 authors for using dependency types. To illustrate, L2 authors frequently employ dependency types such as *compound*, *conj*, and *cc* in their academic writing. This indicates that authors from disparate language backgrounds do not utilize dependency types in a uniform manner, even when they have attained an equivalent level of proficiency. This may be attributed to the influence of their native language background. On the other hand, differences in the use of dependency types are observed in academic texts from different disciplines written by authors from the same language background. In L1 corpus, there is a notable prevalence of the use of *prep*, *pobj*, and *det* dependencies in law texts, as well as *nummod* in medical texts. It is noteworthy that in L2 corpus, *compound* is used with greater frequency in natural science texts, while *conj*, *cc*, and *dobj* dependencies are used with the highest frequency in educational texts.

Additionally, this study serves as a plot study for pedagogical practices. Some of the pioneers of corpus-based studies have asserted that the findings contribute to pedagogical practices, curriculum design, and materials development in EFL and ESL academic contexts, particularly in discipline-specific thesis and dissertation writing modules (Lu & Ai, 2015; Nasseri, 2021; Casal et al., 2021). Notwithstanding the accumulating evidence on the importance of investigating the specific linguistic features of academic writing in ESL or EAP/ESP programs, systematic inquiries remain scarce. Broad awareness that the complexity of academic research writing is not a fixed, uniform entity and that such variability may occur in academic writing across different language backgrounds, as well as in the practice of research writing across different disciplines, offers a valuable foundation for the development of pedagogical strategies, enabling learners to establish achievable objectives and receive constructive feedback on their writing choices.

At the same time, however, it is important to acknowledge some of the limitations of this study, some of which may be addressed through future research. Firstly, future studies should include L2 learners from a range of linguistic backgrounds in order to test the usefulness of dependency distance. Secondly, although our study has encompassed a more expansive range of disciplines, there are still some disciplines that have not been investigated. Additionally, the fact that our study focuses on the abstract section of the paper may potentially impact the generalizability of the results. Ultimately, further research is required to elucidate the relationship between MDD, dependency types and syntactic complexity.

References:

[1] Biber, D. & Gray, B. *Grammatical Complexity in Academic English: Linguistic Change in Writing*[M]. Cambridge: Cambridge University Press, 2016.

[2] Casal, J. E. *An integrated corpus and genre analysis approach to writing research and pedagogy: Development of graduate student genre knowledge (Unpublished doctoral dissertation)*[D]. The Pennsylvania State University: University Park, PA, 2020.

[3] Casal, J. E., Lu Xiaofei, Qiu Xixin, Wang Yuanheng & Zhang Genggeng. Syntactic Complexity across academic research article part-genres: A cross-disciplinary perspective[J]. *Journal of English for Academic Purposes*, 2021, 52: 100996.

[4] Chen, R., Deng, S., & Liu, H. Syntactic Complexity of Different Text Types: From the Perspective of Dependency Distance Both Linearly and Hierarchically[J]. *Journal of Quantitative Linguistics*, 2022, 29(4): 510-540.

[5] Crossley, S. A., & McNamara, D. S. Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners[J]. *Journal of Second Language Writing*, 2014, 26: 66-79.

[6] Durrant, P., & Mathews-Aydınlı, J. A function-first approach to identifying formulaic language in academic writing[J]. *English for Specific Purposes*, 2011, 30(1): 58-72.

[7] Ferrer i Cancho, Ramon. Euclidean distance between syntactically linked words[J]. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 2004, 70(5): 056135.

[8] Gao, J., & Sun, P. P. Dependency distance reflects L2 processing difficulty: Evidence from the relationship between dependency distance, L2 processing speed, and L2 proficiency[J]. *International Journal of Bilingualism*, 2024.

[9] Gibson, E. Linguistic complexity: Locality of syntactic dependencies[J]. *Cognition*, 1998, 68(1): 1-76.

[10] Guo, L., Crossley, S. A., & McNamara, D. S. Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study[J]. *Assessing Writing*, 2013, 18(3): 218-238.

[11] Hao, Y., Wang, X., Bin, S., & Liu, H. A probability distribution of dependencies in interlanguage[J]. *Poznan Studies in Contemporary Linguistics*, 2023, 59(1): 65-93.

[12] Hudson, R. *An introduction to word grammar*[M]. Cambridge: Cambridge University Press, 2010.

[13] Hudson, R. *Language networks: The new word grammar*[M]. Oxford: Oxford University Press, 2007.

[14] Hudson, R. Measuring syntactic difficulty[J]. *Manuscript*, 1995.

[15] Hudson, R. The difficulty of (so-called) self-embedded structures[J]. *UCL Working Papers in Linguistics*, 1996, (8): 1-33.

[16] Jiang, J., & Ouyang, J. Minimization and Probability Distribution of Dependency Distance in the Process of Second Language Acquisition[J]. *Quantitative Analysis of Dependency Structures*, 2018: 167-190.

[17] Jiang, J., Bi, P., & Liu, H. Syntactic complexity development in the writings of EFL learners: Insights from a dependency syntactically-annotated corpus[J]. *Journal of Second Language Writing*, 2019, 46(C): 1-13.

[18] Köhler R, Altmann G. Probability distributions of syntactic units and properties[J]. *Journal of Quantitative Linguistics*, 2000, 7(3): 189-200.

[19] Lewis R. A theory of grammatical but unacceptable embeddings[J]. *Journal of Psycholinguistic Research*, 1996, 25(93): 116.

[20] Liu H. Probability distribution of dependencies based on a Chinese dependency Treebank[J]. *Journal of Quantitative Linguistics*, 2009, 16(3): 256-273.

[21] Liu, H. Probability distribution of dependency distance[J]. *Glottometrics*, 2007, 15: 1-12.

[22] Liu, H., Hudson, R., Feng, Z. Using a Chinese treebank to measure dependency distance[J]. *Corpus Linguistics and Linguistic Theory*, 2009, 5(2): 161-174.

[23] Liu, H., Xu, C., & Liang, J. Dependency distance: A new perspective on syntactic patterns in natural languages[J]. *Physics of Life Reviews*, 2017, 21: 171-193.

[24] Lu, F., & Jiang, Y. Probability distribution of dependency distance and dependency type in translational language[J]. *Humanities and Social Sciences Communications*, 2023, 10(1):1-10.

[25] Lu, X. Automatic analysis of syntactic complexity in second language writing[J]. *International Journal of Corpus Linguistics*, 2010, 15: 474-496.

[26] Lu, X., & Ai, H. Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds[J]. *Journal of Second Language Writing*, 2015, 29: 16–27.

[27] Mancilla, R. L., Polat, N., & Akcay, A. O. An investigation of native and nonnative English speakers' levels of written syntactic complexity in asynchronous online discussions[J]. *Applied Linguistics*, 2017, 38(1), 112-134.

[28] Nasseri, M. Is postgraduate English academic writing more clausal or phrasal? Syntactic complexification at the crossroads of genre, proficiency, and statistical modelling[J]. *Journal of English for Academic Purposes*. 2021, 49: 100940.

[29] Nivre, J. *Inductive Dependency Parsing*[M]. Dordrecht: Springer, 2006.

[30] Omidian, T. A cross-disciplinary investigation of multi-word expressions in the moves of research article abstracts[J]. *Journal of English for Academic Purposes*, 2018, 36: 1-14.

[31] Ouyang, J., Jiang, J., & Liu, H. Dependency distance measures in assessing L2 writing proficiency[J]. *Assessing Writing*, 2022, 51: 1-14.

[32] Temperley D. Minimization of dependency length in written English[J]. *Cognition*, 2007, 105(2): 300-333.

[33] Tesnière, L. *Éléments de syntaxe structurale*[M]. Paris: Klincksieck, 1959.

[34] Zipf G. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*[M]. Cambridge, MA: Addison-Wesley Press, 1949.