English Language Teaching and Linguistics Studies
ISSN 2640-9836 (Print) ISSN 2640-9844 (Online)
Vol. 8, No. 3, 2026
www.scholink.org/ojs/index.php/eltls

Original Paper

A Corpus-Based Study of Dependency Relation Types in English Writing by Chinese Senior High School Students: Evidence from CLEC

Guiru Wang[1*]

[1] Dalian University of Technology, Dalian, Liaoning, China

[*] Corresponding Author

Received: April 12, 2026 Accepted: May 28, 2026 Online Published: June 8, 2026

doi:10.22158/eltls.v8n3p170 URL: http://dx.doi.org/10.22158/eltls.v8n3p170

Abstract

This study investigates the use of dependency relation types in English writing by Chinese senior high school students. Based on the ST2 sub-corpus of the Chinese Learner English Corpus, it aims to reveal the syntactic features of students' English sentence construction and provide implications for English grammar teaching. Using Python-assisted dependency parsing and frequency analysis, this study examines the frequency and distribution of dependency relation types in the English writing of Chinese senior high school students and compares them with those in a native-English corpus, GUM, within the Universal Dependencies framework. The findings show that the most frequently used dependency relation types include nsubj, Root, det, case, advmod, obj, obl, and amod. Compared with the GUM, students use fewer case relations and show less syntactic flexibility, especially in complex structures. The study suggests that grammar teaching should focus more on prepositional structures, common syntactic patterns, and complex sentence construction.

Keywords

Chinese senior high school students, dependency relation types, English writing, corpus linguistics

1. Introduction

In the context of globalization, English has become one of the major languages for international communication (Crystal, 2003). For second language learners, writing ability is also an important part of overall language proficiency (Hyland, 2003; Weigle, 2002). For Chinese senior high school students, English writing does not only reflect vocabulary knowledge and grammar use. It also shows whether students can organize words into clear and acceptable English sentences. Previous research has pointed

out that syntactic complexity and sentence structure are closely connected with second language writing ability (Lu, 2011; Ortega, 2003). Therefore, improving students' ability to construct English sentences is necessary for the development of their English proficiency.

For L2 learners, producing coherent texts is important, but it is often not easy (Crossley et al., 2016). The ability to understand sentence structures and syntactic elements also supports reading comprehension and written expression (Schoonen, 2019). In senior high school English learning, students need to understand how parts of speech, syntactic elements, and grammatical structures are related to each other. Studies have shown that syntactic complexity and linguistic knowledge are associated with L2 writing development and proficiency (Ortega, 2003; Van Gelderen et al., 2003). If students lack this kind of syntactic awareness, their understanding of English grammar may remain fragmented. This may further affect their reading comprehension, writing performance, and learning efficiency.

Sentence structure has always been a key part of English teaching. However, grammar teaching may not work well if grammar is only explained as a set of separate rules. Myhill et al. (2012) found that embedded grammar teaching can support students' writing development and metalinguistic understanding. Therefore, teachers need to connect grammar instruction with actual writing practice. In this way, students can gradually understand how English sentences are organized and how different grammatical elements work together.

In syntactic research, dependency grammar and constituency grammar are two common approaches. Constituency grammar focuses on phrase structures, while dependency grammar describes sentence structure through the relations between heads and dependents (De Marneffe & Nivre, 2019). Dependency-based methods have been widely used in natural language processing, especially in dependency parsing (Kübler et al., 2009). In recent years, such methods have also been applied in second language research to examine learners' syntactic development and writing proficiency (Ouyang et al., 2022). Dependency relation types, as an important part of dependency grammar, can show how words are syntactically connected in a sentence. The Universal Dependencies framework also provides a relatively unified annotation system for describing these syntactic relations across languages (De Marneffe et al., 2021).

However, previous studies have mainly discussed dependency distance, syntactic complexity, or cross-linguistic differences. Less attention has been given to the distribution of dependency relation types in learner writing at a specific educational stage, especially among Chinese senior high school students. Based on this gap, the present study uses the ST2 sub-corpus of the Chinese Learner English Corpus (CLEC) to analyze dependency relation types in students' English writing. The results are compared with those from GUM, a native English corpus annotated within the Universal Dependencies framework. By examining the distribution of different dependency relations, this study attempts to describe the syntactic features of Chinese senior high school students' English sentence construction and offer some suggestions for English grammar and writing instruction.

2. Literature Review and Theoretical Framework

2.1 Dependency Grammar and Dependency Relation Types

Dependency grammar is a syntactic theory based on dependency relations between words. Unlike constituency grammar, which focuses on phrase structures, dependency grammar analyzes sentence structure through directed relations between heads and dependents (Tesnière, 1959; De Marneffe & Nivre, 2019). In a dependency structure, each word is directly or indirectly connected to a governing word, usually called the head. According to dependency grammar, the finite verb is usually the core of a clause, and other sentence elements depend on it in different ways (Tesnière, 1959). Compared with phrase structure, dependency structure focuses more directly on the relations between words, and its structure is usually less complicated (De Marneffe & Nivre, 2019). Therefore, dependency analysis can be used to describe the distribution of dependency relations in treebank data (Liu, 2009).

This study uses dependency grammar to analyze English sentence structures in the writing of Chinese senior high school students. Within the Universal Dependencies framework, relation types such as nsubj, obj, obl, amod, and advmod are used to show how words are syntactically connected with each other (De Marneffe et al., 2021). By focusing on these word-to-word relations, the analysis can show how learners organize English sentences and what syntactic features appear in their writing.

2.3 Dependency Grammar in Linguistic and Second Language Research

Dependency grammar has been used in many areas of linguistic research, including syntactic description, natural language processing, cross-linguistic studies, translation studies, and language education. Its basic idea is that words in a sentence are connected through asymmetric head-dependent relations. These relations can help describe how syntactic and semantic information is organized in a sentence (Liu, 2009; de Marneffe & Nivre, 2019). In computational linguistics, dependency grammar is also closely related to dependency parsing, because it represents syntactic relations between words in a relatively direct way (Kübler et al., 2009). The Universal Dependencies framework has made this approach more widely applicable across languages. It offers a unified system for annotating parts of speech, morphological features, and syntactic dependencies, which makes cross-linguistic comparison more convenient (de Marneffe et al., 2021).

Dependency-based methods have also begun to attract attention in second language writing research. For example, Ouyang et al. (2022) used dependency distance measures to assess L2 writing proficiency and found that these measures could reflect learners' syntactic development. Gao (2023) compared dependency distance in L1 and L2 English academic writing and showed that dependency-based indicators could reveal syntactic differences between native and non-native writers. These studies indicate that dependency grammar is not only useful for linguistic description, but can also provide quantitative evidence for the analysis of learner language.

Nevertheless, existing research has mainly discussed dependency distance, syntactic complexity, or cross-linguistic differences. The distribution of specific dependency relation types in learner English has received relatively limited attention, especially in the English writing of Chinese senior high school

students. Compared with general measures of complexity, dependency relation types can show more directly what syntactic functions and structural patterns learners tend to use. Therefore, examining these relation types may help describe the sentence construction features of senior high school students more clearly and offer empirical evidence for English grammar and writing instruction.

2.3 Corpus-based Studies on Learner English Writing

Corpus linguistics is a useful method for studying learner language, because it is based on large amounts of authentic language data. Through learner corpora, researchers can find common patterns in students' writing and describe their language use in a more objective way. In previous studies on second language writing, learner corpora have been used to analyze errors, syntactic complexity, vocabulary and phrase patterns, first language influence, and the development of grammatical competence (Granger et al., 2002; Paquot & Granger, 2012).

In the field of Chinese EFL writing research, corpus-based methods have also been widely used to compare learner English with corpora of native English writing. For example, Ai and Lu (2013) compared syntactic complexity in non-native and native university students' writing by analyzing 600 essays from the Written English Corpus of Chinese Learners and the Louvain Corpus of Native English Essays. Their study showed that corpus-based syntactic measures can reveal important differences between Chinese EFL learners and native English writers. Lu (2011) also demonstrated that corpus-based syntactic complexity measures can serve as useful indicators of ESL writing development. These studies suggest that corpus-based analysis can provide quantitative evidence for understanding learners' syntactic development in English writing.

Recent corpus-based studies have begun to analyze Chinese learners' English writing from more specific syntactic angles. For example, Shen (2023) compared syntactic complexity in English academic writing by L1 English and L1 Chinese students. The study found that the two groups differed in their use of clausal and phrasal complexity. Such research shows that corpus data can help identify syntactic features in learner writing and make comparisons with native English writing more concrete.

However, existing corpus-based studies on Chinese learners' English writing have mostly discussed writing errors, vocabulary, cohesion, syntactic complexity, or general sentence patterns. The use of dependency relation types in senior high school students' English writing has not received enough attention, especially from the perspective of dependency grammar. It is still worth asking which dependency relation types Chinese senior high school students use most often and how their use differs from native English writing. Therefore, this study focuses on dependency relation types in order to describe students' sentence construction in a more concrete way.

2.4 Research Gap

In general, previous studies have shown that dependency grammar and corpus linguistics can be used to analyze learner language. Dependency-based measures are helpful for describing learners' syntactic development, while corpus-based methods make it possible to study learner writing with real language data. However, there is still not much research on the use of specific dependency relation types in

English writing by Chinese senior high school students. Therefore, this study uses the ST2 sub-corpus of the Chinese Learner English Corpus (CLEC) to analyze the distribution of dependency relation types in students' English writing. The results are also compared with those from GUM, a native English corpus annotated within the Universal Dependencies framework. By making this comparison, the study tries to show the syntactic features of Chinese senior high school students' English sentence construction and provide some suggestions for English grammar and writing teaching.

3. Methodology

3.1 Research Questions

This study aims to investigate the use of dependency relation types in English writing by Chinese senior high school students and to compare their usage patterns with those in GUM, a corpus of native English writing. Specifically, the study addresses the following three research questions:

(1) What are the major characteristics and distributional patterns of dependency relation types in the English writing of Chinese senior high school students in the ST2 sub-corpus of CLEC?

(2) What core dependency relations characterize the basic sentence construction patterns in Chinese senior high school students' English writing?

(3) How does the use of dependency relation types by Chinese senior high school students differ from that in GUM, a corpus of native English writing?

3.2 Corpora and Tools

The learner data used in this study were drawn from the ST2 sub-corpus of the Chinese Learner English Corpus (CLEC). CLEC is a learner corpus consisting mainly of written English produced by Chinese English learners at different learning stages. It contains approximately one million words and is divided into five sub-corpora, from ST2 to ST6. The ST2 sub-corpus represents the secondary school stage and mainly consists of free compositions written by senior high school students. Therefore, it is suitable for examining the syntactic features of Chinese senior high school students' English writing.

According to the official description of CLEC, the ST2 sub-corpus contains 208,088 words after punctuation is excluded. In the present study, after tokenization and the manual removal of titles and some irrelevant text endings, the total number of tokens was 244,986. After punctuation was excluded, the number of tokens was 207,766, which is close to the official word count.

For comparison, this study used GUM as a corpus of native English writing. GUM is an open-source, multilayer corpus within the Universal Dependencies framework. It contains richly annotated English texts from various genres. The version used in this study contains 203,879 words, which is comparable to the size of the CLEC ST2 sub-corpus. Since both corpora contain English texts with diverse topics and are processed or annotated under the Universal Dependencies framework, they are suitable for comparative analysis.

The main analytical tool used in this study was Python. The Python library spacy_udpipe was used to conduct dependency parsing and extract dependency relation types from the CLEC ST2 texts.

Dependency relation types in the GUM corpus were extracted directly from its existing Universal Dependencies annotations. Microsoft Excel was used for data organization, frequency calculation, and comparative analysis.

3.3 Data Collection and Analysis

The text data used in this study were taken from the ST2 sub-corpus of the Chinese Learner English Corpus (CLEC). Before analysis, the original texts were cleaned and segmented. Tags, extra spaces, irregular titles, and irrelevant endings were removed to reduce noise in the data. The ST2 corpus was divided into 305 text units for easier processing. After the automatic cleaning procedure, the texts were checked manually, and the remaining irregular titles and unrelated endings were further deleted. This step helped make the data more suitable for later analysis.

The cleaned CLEC ST2 texts were then processed with the Python library spacy_udpipe. This tool follows the Universal Dependencies framework and was used to parse the texts and extract dependency relation types. The extracted results were saved in an Excel file for subsequent statistical analysis.

After the dependency relations had been extracted, frequency counts were carried out for all text samples. A frequency table was produced to show the overall distribution of dependency relation types in the CLEC ST2 sub-corpus. For comparison, the study also used the GUM corpus, which already provides dependency annotations under the Universal Dependencies framework. For this reason, the GUM data did not need to be parsed again. The dependency relation types were extracted directly from its existing annotations with Python.

Finally, the frequency and proportion of each dependency relation type were calculated and compared between the CLEC ST2 sub-corpus and the GUM corpus. Data visualization was used to present the distributional patterns of dependency relation types in the two corpora. Through comparison, this study identified the major characteristics of dependency relation use in Chinese senior high school students' English writing and examined how these characteristics differ from those in native English writing.

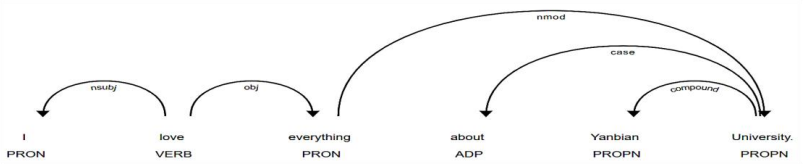

Figure 1. Example of Dependency Relations

As shown in Figure 1, the example sentence is divided into six nodes, with each node representing a single word. The part-of-speech tag of each word is indicated beneath it. The arrows in the diagram point from heads to dependents, and the labels above the arrows indicate the dependency relations between words. For instance, the pronoun "I" and the verb "love" form an nsubj, or nominal subject, dependency relation. Apart from the root node "love," Figure 1 includes five dependency relation types: nsubj, obj, nmod, case, and compound.

This study adopts the Universal Dependencies (UD) framework to classify dependency relation types.

UD provides a cross-linguistically consistent annotation system for parts of speech, morphological features, and syntactic dependencies (De Marneffe, Manning, Nivre, & Zeman, 2021). It supports multilingual parser development, cross-linguistic comparison, and syntactic analysis from a typological perspective. The main dependency relation types used in the UD framework are listed in Table 1.

Table 1. Universal Dependencies

Dependency Types			
nsubj	advcl	nmod:npmod	goeswith
nsubj:pass	advcl:relcl	nmod:tmod	conj
nsubj:outer	advmod	nmod:poss	cc
obj	vocative	acl	cc:preconj
iobj	discourse	acl:relcl	case
csubj	expl	amod	list
csubj:pass	aux	det	dislocated
csubj:outer	aux:pass	det:predet	parataxis
ccomp	cop	compound	orphan
xcomp	mark	compound:prt	reparandum
obl	nummod	fixed	ROOT
obl:npmod	appos	flat	punct
obl:tmod	nmod	flat:foreign	dep

4. Results and Discussion

4.1 Distribution of Dependency Relation Types in CLEC

Following the procedures described above, this study used spacy_udpipe to extract dependency relation types from the ST2 sub-corpus of CLEC. As shown in Figure 2, a total of 44 dependency relation types were identified in the English writing of Chinese senior high school students. The total number of dependency relations was 246,110, which is close to the number of tokens in the processed CLEC ST2 corpus, namely 244,986 tokens including punctuation. The consistency of the results shows that the extraction process was basically complete and reliable enough for the following statistical analysis.

As can be seen from Figure 2, the dependency relation types in CLEC present a clear long-tail distribution (Zipf, 1949). In other words, a small number of relations occur very often, whereas many other relation types appear only occasionally. This may reflect that Chinese senior high school students mainly rely on several basic dependency relations in English sentence construction.

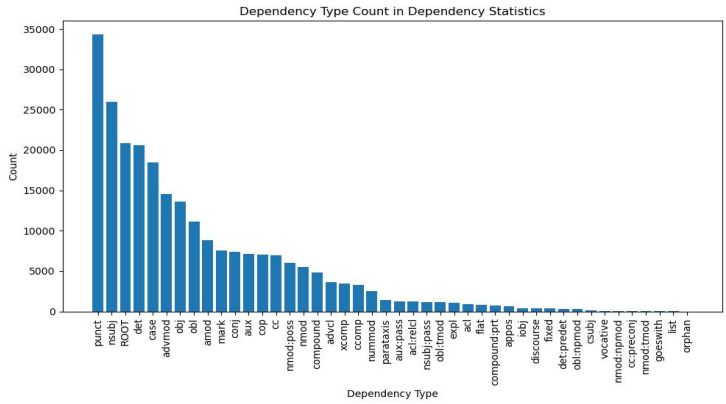

Figure 2. Dependency Type Count (CLEC)

4.1.1 High-frequency Dependency Relations

After excluding punct, eight dependency relation types occurred more than 8,000 times in CLEC, accounting for approximately 63% of all dependency relations. These high-frequency relations are nsubj, ROOT, det, case, advmod, obj, obl, and amod, as shown in Table 2. These relations mainly involve core sentence elements, including subjects, predicates, objects, and modifiers.

Table 2. Dependency Types of CLEC (High Frequency)

Dependency Types	Count
punct	34284
nsubj	25943
ROOT	20832
det	20599
case	18482
advmod	14514
obj	13606
obl	11095
amod	8806

The high frequency of nsubj and ROOT suggests that subject-verb structures constitute a basic pattern in Chinese senior high school students' English sentence construction. As the root node, ROOT is indispensable in each sentence. However, the frequency of nsubj is higher than that of ROOT, which indicates that some sentence units contain more than one subject relation. A closer examination shows that this phenomenon is partly related to students' punctuation use. For example, in sentences such as "Oh, we win, it's very good" and "Are you sure, it's not here," comma splices may lead to multiple subject relations within a single sentence. This point also reminds us that punctuation is not only a

surface problem, but is closely connected with how students judge sentence boundaries in English writing.

Multiple nsubj relations may also appear in compound sentences, complex sentences, inverted structures, and parenthetical structures. For example, in “I answered and then she asked me another question,” both “I” and “she” are labeled as nsubj. Compared with ROOT and nsubj, obj occurs less frequently, since not all sentence patterns require an object. Intransitive verbs and copular structures, for instance, do not usually take objects.

Among the modifier-related relations, det appears quite often. Determiners are used to modify or specify nouns, showing definiteness, indefiniteness, or quantity within noun phrases. The frequent occurrence of det is therefore closely related to the large number of nouns in students’ writing. By contrast, det:predet appears much less often. This may mean that students usually place a single determiner directly before a noun, while structures with more than one determiner modifying the same noun are less common in their writing.

The frequent use of case and obl is also worth discussing. In the Universal Dependencies framework, case is usually related to adpositions, and obl refers to oblique nominals, which often work as adverbial elements in a sentence. These two relations often appear together in prepositional phrases. For example, as shown in Figure 3, in the sentence “It is rich in fibre,” “in” is labeled as case, and “fibre” is labeled as obl.

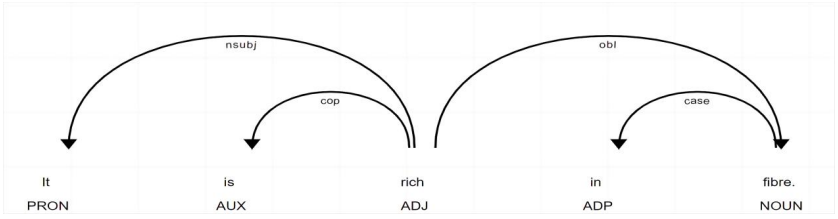

Figure 3. Example of Dependency Relations

Similarly, as shown in Figure 4, in the sentence “A healthy diet is very important for people,” “for” is labeled as case, and “people” is labeled as obl.

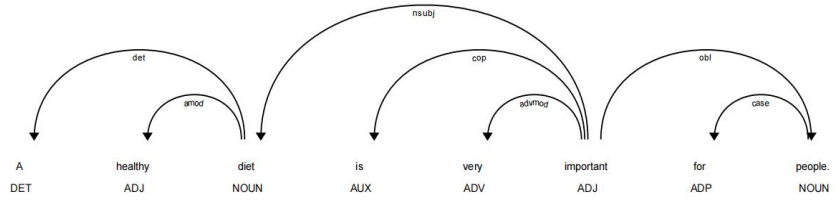

Figure 4. Example of Dependency Relations

The frequent occurrence of advmod and amod reflects students’ common use of adverbs and adjectives in English writing. Adverbs are mainly used to modify verbs, adjectives, or other adverbs, while

adjectives mainly modify nouns, pronouns, or numerals. This suggests that these modifying structures are frequently used in students' English writing.

Overall, the high-frequency dependency relations show that Chinese senior high school students' English sentence construction relies strongly on basic dependency relations, especially subject-verb, verb-object, and modifier-head relations. This result is consistent with previous research showing that dependency-based measures can provide useful evidence for analyzing syntactic development in L2 writing (Ouyang et al., 2022).

4.1.2 Medium-frequency Dependency Relations

In the medium-frequency range, defined as dependency relations occurring more than 2,000 times but fewer than 8,000 times, 12 relation types were identified: mark, conj, aux, cop, cc, nmod:poss, nmod, compound, advcl, xcomp, ccomp, and nummod, as shown in Table 3. These 12 relation types account for approximately 31% of all dependency relations. They can be grouped into four major categories: modifying relations, auxiliary and copular relations, coordinating relations, and subordinate and clausal relations.

Table 3. Dependency Types of CLEC (Medium-frequency)

Dependency Types	Count
mark	7568
conj	7348
aux	7133
cop	7054
cc	6951
nmod:poss	5984
nmod	5470
compound	4848
advcl	3599
xcomp	3460
ccomp	3271
nummod	2505

4.1.2.1 Modifying Relations

The modifying relations in this range include nmod, nmod: poss, compound, and nummod. Their relatively frequent use suggests that Chinese senior high school students often expand noun phrases through nominal modifiers, possessive structures, compounds, and numeral modifiers.

The relation nmod refers to a nominal modifier, usually a noun phrase or a prepositional phrase that modifies another nominal element. For example, in the phrase "a book on the shelf," "on the shelf"

functions as a nominal modifier of the noun “book,” as shown in Figure 5.

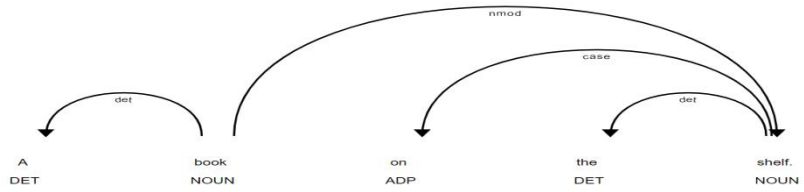

Figure 5. Example of Dependency Relations

The subtype nmod:poss indicates a possessive relation between a nominal modifier and the noun it modifies. For instance, in “John’s car,” “John’s” modifies “car” and expresses possession, as shown in Figure 6.

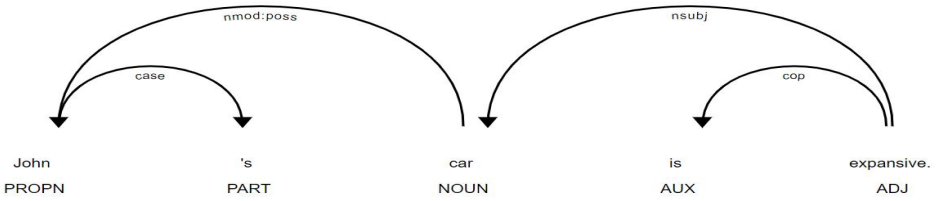

Figure 6. Example of Dependency Relations

The relation compound represents a compound structure formed by two or more words. For example, in “apple tree,” “apple” modifies “tree” and is labeled as compound, as shown in Figure 7.

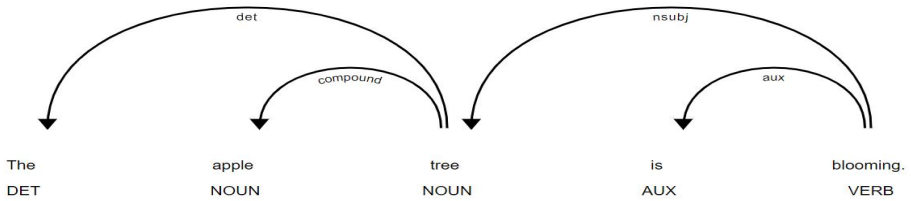

Figure 7. Example of Dependency Relations

The relation nummod indicates numeral modification. For example, in the sentence “There are two different kinds of diets: Western diet and Chinese diet,” “two” is labeled as nummod, as shown in Figure 8.

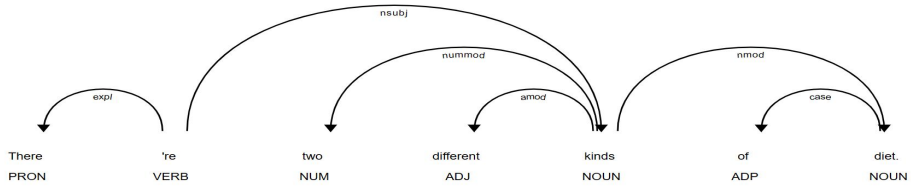

Figure 8. Example of Dependency Relations

These results indicate that Chinese senior high school students use nominal modification in English writing. However, compared with high-frequency relations such as nsubj, ROOT, and det, these modifying relations occur less frequently, suggesting that noun phrase expansion is less prominent than basic sentence-level relations in students' writing.

4.1.2.2 Auxiliary Relations

In this frequency range, aux and cop are the main relation types. Their use shows that students are not only using main verbs, but are also beginning to use auxiliary verbs and copular structures in their sentences. Auxiliary verbs are related to tense, modality, and voice, while copular verbs connect the subject with the following predicative part. In this sense, students' sentence patterns are not restricted to simple subject-verb-object structures.

The relation cop can be seen in copular sentences. For example, in "I am very happy to be Chinese," "am" links the subject "I" with the predicative adjective "happy." Therefore, it is labeled as cop in Figure 9.

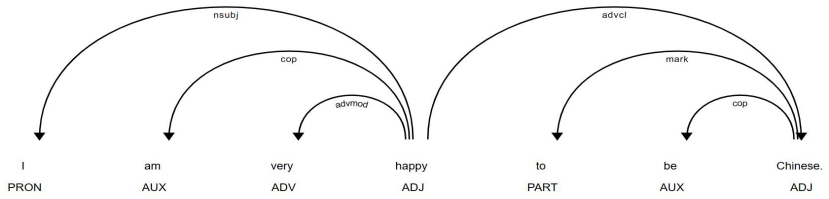

Figure 9. Example of Dependency Relations

Aux is the relation used for auxiliary verbs, including modal auxiliaries. In the sentence "Many people can not answer this question exactly," for instance, "can" functions as a modal auxiliary and is labeled as aux in Figure 10.

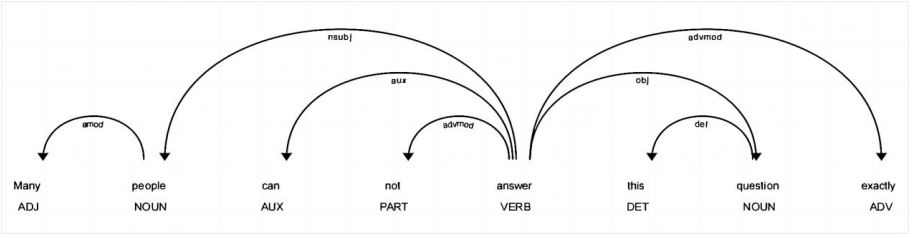

Figure 10. Example of Dependency Relations

The use of aux and cop shows that Chinese senior high school students have used some basic auxiliary and copular structures in their English writing. However, compared with core relations such as nsubj, ROOT, and obj, these two relations occur less often. This means that students still rely more on core sentence relations, while auxiliary and copular structures are not as prominent in their writing.

4.1.2.3 Connecting Relations

The main connecting relations in the data are cc and conj. In the Universal Dependencies framework, cc is related to coordinating conjunctions, and conj is used for the elements connected through

coordination. The occurrence of these two relations shows that students can join parallel words, phrases, or clauses in English writing.

For example, in “It contains a lot of fruit and vegetables,” “and” is labeled as cc, while “vegetables” is labeled as conj, as shown in Figure 11. With this structure, students can express addition, parallel relations, or contrast more clearly.

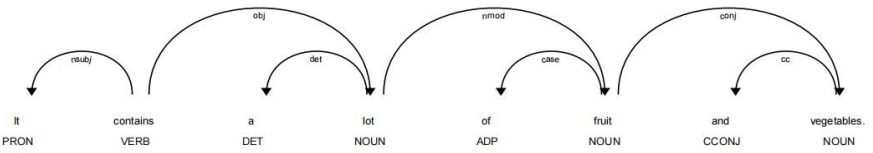

Figure 11. Example of Dependency Relations

4.1.2.4 Subordinate Relations

Among the medium-frequency relations, mark, advcl, xcomp, and ccomp are closely related to subordination and clausal structures. Their occurrence shows that Chinese senior high school students do not depend only on simple sentence patterns. They have also begun to use more complex structures, such as subordinate clauses, infinitive clauses, and clausal complements.

The relation mark usually appears with words that introduce subordinate structures, such as “that,” “if,” “whether,” “because,” and “while.” It can also appear with “to” before an infinitive clause. For example, in “The Chinese diet is considered to be the healthiest in the world,” “to” connects “considered” with the following infinitive phrase and is labeled as mark in Figure 12.

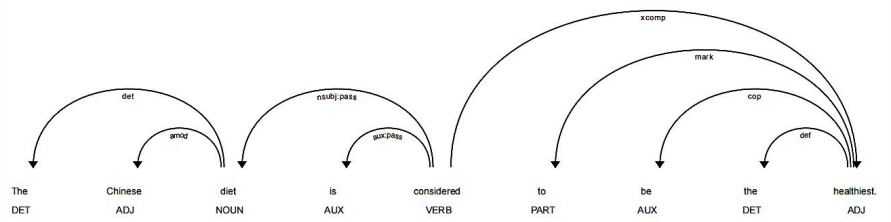

Figure 12. Example of Dependency Relations

Advcl is related to adverbial clause modifiers, which often express time, condition, reason, purpose, and other meanings. In the sentence “If they want to be healthy, they should have a healthy diet at first,” “want” is labeled as advcl because it is the main verb of the conditional clause. In the same example, “healthy” is labeled as xcomp, since it functions as an open clausal complement, as shown in Figure 13.

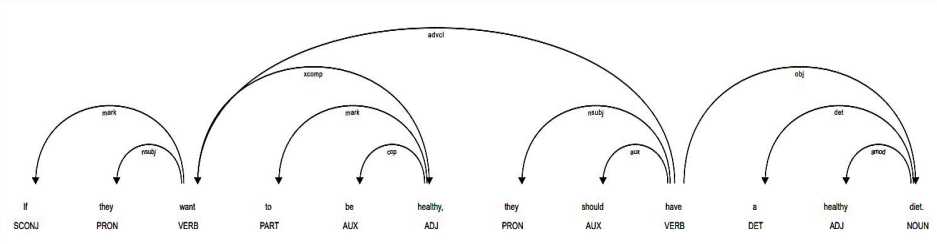

Figure 13. Example of Dependency Relations

The relation ccomp is used when a clause serves as the complement of a verb. For example, in "I believe that she is honest," the clause "that she is honest" functions as the complement of "believe," so the relation can be labeled as ccomp, as shown in Figure 14.

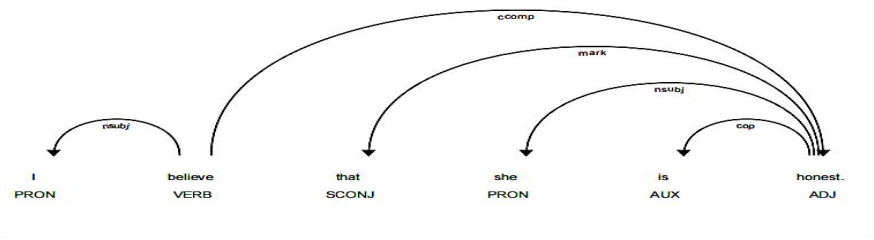

Figure14. Example of Dependency Relations

The oThe use of these subordinate and clausal relations shows that Chinese senior high school students can use some complex sentence structures in English writing. However, compared with basic sentence relations, these structures appear less often. In other words, complex clauses are present in students' writing, but they are not a major feature.

The data also show that students often use coordination to connect sentence elements. Still, coordination alone cannot fully show the complexity of sentence construction. Therefore, it is necessary to look at subordinate and clausal relations, such as mark, advcl, xcomp, and ccomp.

4.1.3 Low-frequency Dependency Relations

In the low-frequency range, defined as relation types occurring fewer than 2,000 times, 23 dependency relation types were identified in CLEC. These low-frequency relation types include parataxis, aux:pass, acl:relcl, nsubj:pass, obl:tmod, expl, acl, flat, compound:prt, appos, iobj, discourse, fixed, det:predet, obl:npmod, csubj, vocative, nmod:npmod, cc:preconj, nmod:tmod, goeswith, list, and orphan. Together, they account for approximately 6% of all dependency relations. In addition, eight relation types did not occur in the corpus, namely nsubj:outer, csubj:pass, csubj:outer, advcl:relcl, flat:foreign, dislocated, reparandum, and dep, as shown in Table 4.

Table 4. Dependency Types of CLEC (Low Frequency)

Dependency Types	Count	Dependency Types	Count
parataxis	1403	fixed	384
aux:pass	1264	det:predet	345
acl:relcl	1228	obl:npmod	308
nsubj:pass	1200	csubj	167
obl:tmod	1179	vocative	57
expl	1045	nmod:npmod	49
acl	931	cc:preconj	42
flat	855	nmod:tmod	41
compound:prt	743	goeswith	30
appos	628	list	21
iobj	428	orphan	1
discourse	409		

The low frequency of these relations is closely connected with the complexity and specificity of the structures they represent. Passive-related relations, such as aux:pass, nsubj:pass, and csubj:pass, appear relatively rarely in the data. These relations are used in passive constructions, where the receiver or object of an action becomes the grammatical subject. Since passive structures are different from common active sentence patterns, students may find them more difficult to use in writing. This may be one reason for their lower frequency.

Subordinate clause relations, including acl, acl:relcl, csubj, nsubj:outer, csubj:outer, and advcl:relcl, are also not common in the CLEC ST2 corpus. Such structures can make sentences more flexible and layered, but they usually involve modification, embedding, or clause-level organization. Their limited occurrence shows that Chinese senior high school students still use complex subordinate structures relatively cautiously in English writing.

Some modifying structures also appear less often, such as temporal modifiers like obl:tmod and nmod:tmod, nominal modifiers like obl:npmod and nmod:npmod, and pre-determiners such as det:predet. These relations are more specific in function, so they are naturally less common in students' compositions.

Loose joining relations, including parataxis, list, dislocated, orphan, and reparandum, also have low frequencies. This may be related to the nature of the corpus, since these structures are more likely to appear in spoken language or informal discourse, while the CLEC ST2 texts are written compositions. Other relations, such as flat, compound:prt, fixed, goeswith, expl, discourse, and vocative, are also infrequent. These relations are often connected with compound expressions, unanalyzed structures, discourse markers, or special dependents.

Taken together, the low-frequency relations show that Chinese senior high school students do use some complex or less common structures, including passive constructions, subordinate clauses, and discourse-related structures. However, these structures are not prominent in their writing. Their sentence construction still depends more on common and basic syntactic relations.

4.2 Comparison between CLEC and GUM

To better understand how Chinese senior high school students use dependency relation types, the CLEC results were compared with those from the GUM corpus. GUM is annotated under the Universal Dependencies framework, which is the same framework used in the analysis of CLEC. This makes the two corpora comparable in terms of dependency relations. Universal Dependencies provides a cross-linguistically consistent way to describe syntactic relations (McDonald et al., 2013).

In the original statistical output of GUM, 53 labels were found, including an additional label "-". After checking the relevant sentences, this label was considered a non-standard label caused by annotation or data processing. For example, in the sentence "This system can also be used longitudinally to study how the workforce's composition changes over time, which is particularly valuable for evaluating the effectiveness of policies meant to broaden participation or improve retention of faculty," the word "workforce's" was labeled as "-". Since this label does not belong to the standard Universal Dependencies relation types, it was excluded from the analysis.

After this label was removed, 52 dependency relation types remained for comparison. Figure 15 shows that GUM also has a long-tail distribution. That is, a few dependency relations appear very frequently, while many others occur only occasionally. This pattern is similar to CLEC, but GUM still contains a wider range of low-frequency relations.

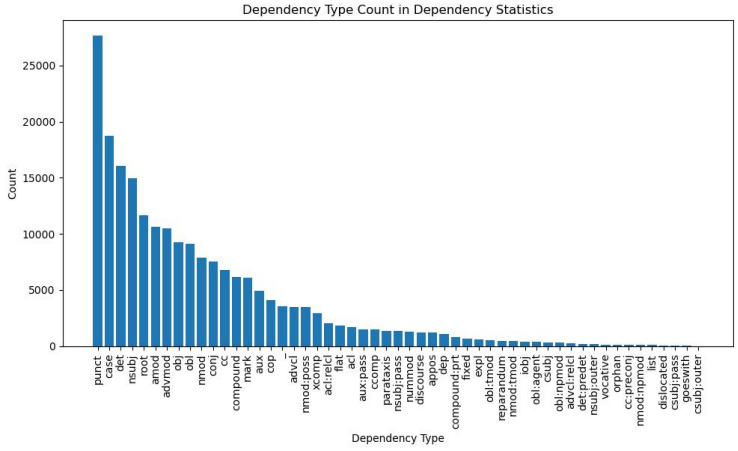

Figure 15: Dependency Type Count(GUM)

4.2.1 Overall Distribution

At the overall level, the GUM corpus contains all 52 dependency relation types considered in this study, whereas the CLEC ST2 corpus contains 44 types. This difference suggests that GUM may display a wider range of syntactic structures and greater diversity in sentence construction. In other words,

compared with the learner corpus, GUM, as a corpus of native English writing, shows greater syntactic diversity.

However, the overall distribution patterns of the two corpora are similar in that both show a small number of highly frequent dependency relations and a large number of low-frequency relations. This similarity reflects the general distributional tendency of syntactic dependency relations in English texts. The main difference lies in the proportion and range of less frequent relations, with GUM showing greater diversity.

4.2.2 Comparison of High-frequency Relations

After excluding punct, eight dependency relation types occurred more than 8,000 times in GUM, accounting for approximately 57% of all dependency relations. These high-frequency relations are case, det, nsubj, ROOT, amod, advmod, obj, and obl, as shown in Table 5.

Table 5. Dependency Types of GUM (High Frequency)

Dependency Types	Count
punct	27635
case	18716
det	16021
nsubj	14973
ROOT	11656
amod	10652
advmod	10490
obj	9229
obl	9128

The high-frequency dependency relations in GUM are largely consistent with those in CLEC. Both corpora contain case, det, nsubj, ROOT, amod, advmod, obj, and obl among the most frequent relations, although their rankings differ. This suggests that the English writing of Chinese senior high school students and native English writing share similar core dependency relations in sentence construction.

One notable difference is the ranking and frequency of case. In GUM, case is the most frequent dependency relation after punct, whereas in CLEC it ranks lower. Since case is closely related to the use of prepositions, this difference suggests that Chinese senior high school students may use prepositional structures less frequently than those found in native English writing. This finding is consistent with previous corpus-based studies showing that Chinese EFL learners differ from native speakers in their use of English prepositions and that L1 influence may partly affect preposition use (Yuan, 2014; Li, 2023).

Another difference is that amod ranks higher in GUM than in CLEC. This may indicate that adjectival

modification is more prominent in GUM. However, the presence of amod in the high-frequency range of both corpora also shows that adjective-noun modification is a common structure in both learner English and native English writing.

4.2.3 Comparison of Medium-frequency Relations

In the medium-frequency range of GUM, defined as relation types occurring more than 2,000 times but fewer than 8,000 times, 11 dependency relation types were identified: nmod, conj, cc, compound, mark, aux, cop, advcl, nmod:poss, xcomp, and acl:relcl. These 11 relation types account for approximately 32% of all dependency relations, as shown in Table 6.

Table 6. Dependency Types of GUM (Medium-frequency)

Dependency Types	Count
nmod	7877
conj	7555
cc	6760
compound	6128
mark	6109
aux	4887
cop	4115
advcl	3510
nmod:poss	3505
xcomp	2918
acl:relcl	2007

The medium-frequency relations in GUM are generally similar to those in CLEC. Both corpora contain relations related to nominal modification, coordination, auxiliary and copular structures, and subordinate or clausal structures. This suggests that Chinese senior high school students use many basic and intermediate syntactic patterns that are also common in native English writing.

Nevertheless, some differences can still be observed. CLEC shows relatively higher frequencies of ccomp and nummod, while GUM contains more instances of acl:relcl. The higher occurrence of acl:relcl in GUM may suggest that relative clause structures are more frequently used in native English writing. This indicates that although Chinese senior high school students use some clausal structures, relative clause modification is less frequent in CLEC than in GUM.

4.2.4 Comparison of Low-frequency Relations

In the low-frequency range of GUM, defined as relation types occurring fewer than 2,000 times, 32 dependency relation types were identified. These relations include flat, acl, aux:pass, ccomp, parataxis, nsubj:pass, nummod, discourse, appos, dep, compound:prt, fixed, expl, obl:tmod, reparandum,

nmod:tmod, iobj, obl:agent, csubj, obl:npmod, advcl:relcl, det:predet, nsubj:outer, vocative, orphan, cc:preconj, nmod:npmod, list, dislocated, csubj:pass, goeswith, and csubj:outer, as shown in Table 7.

Table 7. Dependency Types of GUM (Low Frequency)

Dependency Types	Count	Dependency Types	Count
flat	1794	iobj	383
acl	1676	obl:agent	352
aux:pass	1491	csubj	323
ccomp	1461	obl:npmod	301
parataxis	1364	advcl:relcl	241
nsubj:pass	1338	det:predet	183
nummod	1293	nsubj:outer	180
discourse	1240	vocative	122
appos	1220	orphan	92
dep	1041	cc:preconj	91
compound:prt	826	nmod:npmod	91
fixed	666	list	76
expl	595	dislocated	52
obl:tmod	510	csubj:pass	18
reparandum	442	goeswith	16
nmod:tmod	426	csubj:outer	6

As can be seen from Table 7, low-frequency dependency relations account for about 11% of all dependency relations in GUM, which is almost twice as high as that in CLEC. This shows that native English writing contains more varied syntactic structures. Some of these structures are related to passive voice, relative clauses, clausal subjects, discourse expressions, and nominal modification.

The main difference between CLEC and GUM does not lie in the basic dependency relations, but in the use of less frequent and more complex ones. In the CLEC ST2 corpus, Chinese senior high school students use basic relations such as nsubj, ROOT, det, obj, advmod, and amod quite often. However, they do not use complex structures as frequently as native English writers. In comparison, GUM contains more types of dependency relations and a larger proportion of low-frequency structures.

This comparison shows that students have already mastered some basic patterns of English sentence construction. At the same time, their writing still lacks enough variety in complex dependency relations. Therefore, in grammar and writing teaching, teachers may need to help students move from basic sentence patterns to more varied structures, especially prepositional phrases, passive voice, subordinate clauses, and other complex sentence forms.

5. Pedagogical Implications

The findings of this study have several implications for English grammar and writing instruction at the senior high school level. In the CLEC ST2 sub-corpus, Chinese senior high school students mainly rely on a small group of high-frequency dependency relations, including nsubj, ROOT, det, case, advmod, obj, obl, and amod. By contrast, some complex and low-frequency relations appear less often in CLEC than in GUM, the native English reference corpus. This difference indicates that students have generally mastered some basic syntactic patterns, but they still need more practice with varied and complex sentence structures.

5.1 Emphasizing Basic Dependency Relations and Punctuation

The high-frequency relations found in CLEC are closely connected with basic sentence relations, especially subject-verb, verb-object, and modifier-head relations. In grammar teaching, teachers can make these core relations clearer to students by showing how subjects are linked to verbs, how verbs take objects, and how modifiers depend on the words they modify. This is also related to previous research, which suggests that dependency-based measures can be used to examine syntactic development and writing proficiency in L2 writing (Ouyang et al., 2022).

Another point worth noticing is that nsubj occurs more frequently than ROOT. This may be partly caused by students' inappropriate use of punctuation. For example, when students join several clauses with commas, several subject relations may appear within what is treated as one sentence. For this reason, punctuation should not be taught only as a matter of writing convention. It is also closely related to sentence boundaries and sentence organization. In classroom practice, teachers may use examples from learner corpora to help students recognize run-on sentences, comma splices, and unclear sentence boundaries. As Myhill et al. (2012) [9] argue, grammar teaching becomes more effective when it is connected with writing practice and students' metalinguistic awareness. Corpus-based examples may therefore help students improve both grammatical accuracy and sentence organization.

5.2 Strengthening the Teaching of Prepositional Structures

The comparison between CLEC and GUM shows a clear difference in the use of case, which appears more frequently in GUM. In the Universal Dependencies framework, case is usually related to prepositions. This difference may suggest that Chinese senior high school students use prepositional structures less often, or not as flexibly as native English writers. This is an issue worth noting, since prepositions are highly frequent in English. Kennedy (1998) points out that about one preposition appears in every eight words, and Mindt and Weber (1989) report that prepositions account for 12.21% of the Brown Corpus and 12.34% of the LOB Corpus.

Because prepositions are both frequent and functionally important, grammar teaching should give more attention to prepositional structures. Teachers may focus on common prepositions and their typical collocational patterns, such as "be interested in," "be good at," "depend on," and "be different from." Instead of teaching prepositions as isolated words, teachers can place them in sentence contexts and guide students to observe how they combine with nouns, pronouns, adjectives, and verbs. This kind of practice may help students become more familiar with prepositional phrases and use them more

naturally in their own writing.

5.3 Enhancing Instruction on Complex Sentence Structures

Some low-frequency dependency relations appear relatively rarely in students' writing. These relations are usually related to passive constructions, relative clauses, clausal subjects, and other subordinate structures, including aux:pass, nsubj:pass, acl:relcl, csubj, and advcl:relcl. Their limited occurrence shows that students still rely mainly on basic sentence patterns, and more complex structures are not yet used very often.

In grammar teaching, complex sentence structures can be introduced step by step. Teachers may begin with common structures, such as the passive voice, attributive clauses, adverbial clauses, and object clauses. After students become familiar with the basic forms, they can be guided to use these structures in writing. For example, students may first identify relative clauses in model sentences, then practice combining simple sentences, and finally use relative clauses in paragraph writing. Through this process, complex structures are no longer treated only as separate grammar points, but as practical choices for improving sentence variety and expression.

Teachers should also remind students that complex sentences are not always better. Accuracy and appropriateness still matter. Students need to know when a complex structure is needed and how it contributes to the meaning of the sentence. This may help them develop syntactic flexibility while maintaining grammatical accuracy.

5.4 Using corpus-based and Cross-linguistic Approaches

The findings also point to the value of corpus-based teaching in senior high school English writing instruction. Corpora contain real examples of language use, so students can observe how words and sentence structures appear in actual contexts. With the help of corpus examples, students may notice repeated grammatical patterns and compare their own writing with native or reference corpus data. This approach is in line with data-driven learning, which encourages learners to discover language patterns through authentic examples (Johns, 1991; Boulton, 2009).

In classroom practice, teachers can choose examples from both learner corpora and reference corpora to illustrate common problems in sentence construction. For example, by comparing CLEC examples with GUM examples, students may find differences in preposition use, modifier placement, passive constructions, and subordinate clauses. Such comparison can make grammar learning more concrete and connect it more closely with the problems students meet in their own writing.

Cross-linguistic comparison can also be used in grammar teaching. Chinese and English differ in word order, modification patterns, prepositional use, and clause structure. Because of these differences, Chinese learners may be influenced by their first language when they write English sentences. From the perspective of interlanguage development, learner language often shows systematic features shaped by both the first language and the target language. Contrastive analysis and language transfer studies also show that differences between the two languages may affect second language learning and sometimes lead to negative transfer in learner production (Lado, 1957; Odlin, 1989). Therefore, teachers can

compare Chinese and English sentence structures in class, especially in areas where students often make mistakes, such as subject-verb agreement, prepositional phrases, passive constructions, and subordinate clauses.

Taken together, the pedagogical implications of this study suggest that grammar teaching should go beyond the explanation of isolated rules. For senior high school students, syntactic analysis, corpus-based examples, and cross-linguistic comparison can be combined in writing instruction. In this way, students may develop a clearer understanding of English sentence construction and become more accurate and flexible in their writing.

6. Conclusion

This study analyzed dependency relation types in English writing by Chinese senior high school students. The ST2 sub-corpus of CLEC was used as the learner corpus, and GUM was used as the native English reference corpus. Through Python-assisted dependency parsing and statistical analysis, the study described the distribution of dependency relation types in learner English and compared it with native English writing.

The results show that the most frequent dependency relations in students' writing include nsubj, ROOT, det, case, advmod, obj, obl, and amod. These relations are mainly related to basic sentence structures, such as subject-verb, verb-object, and modifier-head patterns. This means that students' English sentence construction still depends largely on core dependency relations.

The comparison between CLEC and GUM shows that the two corpora share several high-frequency dependency relation types. This suggests that Chinese senior high school students have already used many basic English sentence patterns in their writing. However, GUM contains a wider variety of dependency relation types and more low-frequency complex structures. For example, case occurs more often in GUM than in CLEC, which may mean that students do not use prepositional structures as frequently or flexibly as native English writers. Relations connected with passive constructions, relative clauses, and other complex structures also appear less often in CLEC. These differences show that students' syntactic flexibility still needs to be further developed.

Based on the above results, English grammar and writing instruction at the senior high school level should give more attention to basic dependency relations, prepositional structures, and complex sentence patterns. Corpus-based teaching and cross-linguistic comparison may also help students develop stronger syntactic awareness and improve their English writing.

This study applies dependency relation analysis to the English writing of Chinese senior high school students, which provides another perspective for learner English research. However, the study is still limited in scope, as it only uses the ST2 sub-corpus of CLEC and compares it with GUM. Future research may include larger learner corpora, learners from different educational stages, and more detailed analyses of specific dependency relations.

References

Crystal, D. (2003). *English as a Global Language* (2nd ed.). Cambridge: Cambridge University Press.

Hyland, K. (2003). *Second Language Writing*. Cambridge: Cambridge University Press.

Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

Lu, X. (2011). A Corpus-Based Evaluation of Syntactic Complexity Measures as Indices of College-Level ESL Writers' Language Development. *TESOL Quarterly*, *45*(1), 36-62.

Ortega, L. (2003). Syntactic Complexity Measures and their Relationship to L2 Proficiency: A Research Synthesis of College-level L2 Writing. *Applied Linguistics*, *24*(4), 492-518.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, *32*, 1-16.

Schoonen, R. (2019). Are reading and writing building on the same skills? The relationship between reading and writing in L1 and EFL. *Reading and Writing*, *32*(3), 511-535.

Van Gelderen, A., Schoonen, R., De Glopper, K., Hulstijn, J., Snellings, P., Simis, A., & Stevenson, M. (2003). Roles of linguistic knowledge, metacognitive knowledge and processing speed in L3, L2 and L1 reading comprehension: A structural equation modeling approach. *International Journal of Bilingualism*, *7*(1), 7-25.

Myhill, D., Jones, S., Lines, H., & Watson, A. (2012). Re-thinking grammar: The impact of embedded grammar teaching on students' writing and students' metalinguistic understanding. *Research Papers in Education*, *27*(2), 139-166.

De Marneffe, M. C., & Nivre, J. (2019). Dependency Grammar. *Annual Review of Linguistics*, *5*, 197-218.

Kübler, S., McDonald, R., & Nivre, J. (2009). *Dependency Parsing*. Cham: Springer.

Ouyang, J., Jiang, J., & Liu, H. (2022). Dependency distance measures in assessing L2 writing proficiency. *Assessing Writing*, *51*, 100603.

De Marneffe, M. C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, *47*(2), 255-308.

Tesnière, L. (1959). *Éléments de syntaxe structurale*. Paris: Klincksieck.

Liu, H. (2009). Probability Distribution of Dependencies Based on a Chinese Dependency Treebank. *Journal of Quantitative Linguistics*, *16*(3), 256-273.

Gao, N., & He, Q. (2023). A Corpus-Based Study of the Dependency Distance Differences in English Academic Writing. *SAGE Open*, *13*(3), 1-12.

Granger, S., Hung, J., & Petch-Tyson, S (editors). (2002). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Paquot, M., & Granger, S. (2012). Formulaic Language in Learner Corpora. *Annual Review of Applied Linguistics*, *32*, 130-149.

Ai, H., & Lu, X. (2013). A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. In: Díaz-Negrillo A, Ballier N, Thompson P (editors), *Automatic Treatment and Analysis of Learner Corpus Data* (pp. 249-264). Amsterdam: John Benjamins Publishing Company.

Shen, C., Guo, J., Shi, P., Qu, S., & Tian, J. (2023). A corpus-based comparison of syntactic complexity in academic writing of L1 and L2 English students across years and disciplines. *PLOS ONE*, *18*(10), e0292688.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley Press.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., & Lee, J. (2013). Universal Dependency Annotation for Multilingual Parsing. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Volume 2) Short Papers. Sofia: Association for Computational Linguistics, 2013:92-97.

Yuan, H. C. (2014). A Corpus-based Study on the Influence of L1 on EFL Learners' Use of Prepositions. *Theory and Practice in Language Studies*, *4*(12), 2513-2521.

Li, Q. (2023). A Corpus-Based Study on the Use of High-Frequency Prepositions between Chinese English Learners and Native Speakers. *Curriculum and Teaching Methodology*, *6*(6), 104-113.

Kennedy, G. D. (1998). *An Introduction to Corpus Linguistics*. London/New York: Longman.

Mindt, D., & Weber, C. (1989). Prepositions in American and British English. *World Englishes*, *8*(2), 229-238.

Selinker, L., & Lamendella, J. T. (1978). Two perspectives on fossilization in interlanguage learning. *Interlanguage Studies Bulletin*, *3*(2), 143-191.

Johns, T. F. (1991). Should you be persuaded: Two samples of data-driven learning materials. *English Language Research Journal*, *4*, 1-16.

Boulton, A. (2009). Data-driven learning: reasonable fears and rational reassurance. *Indian Journal of Applied Linguistics*, *35*(1), 81-106.

Lado, R. (1957). *Linguistics across Cultures: Applied Linguistics for Language Teachers*. Ann Arbor: University of Michigan Press.

Odlin, T. (1989). *Language Transfer: Cross-Linguistic Influence in Language Learning*. Cambridge: Cambridge University Press.