Automatic Construction of the English Sentence Pattern Structure Treebank for Chinese ESL learners

Lin Zhu[1,2*] Meng Xu[1,2*] Wenya Guo[1,2] Jingsi Yu[1,2]
Liner Yang[1,2†] Zehuang Cao[3] Yuan Huang[2] Erhong Yang[1,2]

[1]National Language Resources Monitoring and Research Center for Print Media, Beijing Language and Culture University, China
[2]School of Information Science, Beijing Language and Culture University, China
[3]Faculty of Foreign Studies, Beijing Language and Culture University, China
lineryang@gmail.com

Abstract

Analyzing long and complicated sentences has always been a priority and challenge in English learning. In order to conduct the parse of these sentences for Chinese English as Second Language (ESL) learners, we design the English Sentence Pattern Structure (ESPS) based on the Sentence Diagramming theory. Then, we automatically construct the English Sentence Pattern Structure Treebank (ESPST) through the method of rule conversion based on constituency structure and evaluate the conversion results. In addition, we set up two comparative experiments, using trained parser and large language models (LLMs). The results prove that the rule-based conversion approach is effective.

1 Introduction

Reading comprehension is a fundamental skill in English learning, pivotal for linguistic acquisition, critical thinking, and effective communication across various contexts. For Chinese ESL (English as a Second Language) learners, the ability to analyze complicated sentences represents both a central priority and a significant challenge in reading comprehension. To overcome this reading barrier, it is essential for learners to have a certain level of grammatical knowledge. Bernhardt (1993) believes that grammar is very crucial for second language learners' reading ability. Alderson(1993) considers grammatical ability an important foundation for second language learners' reading, emphasizing a vital to divide sentences into correct patterns.

Existing analysis tutorials for complicated English sentence typically contain only hundreds of example sentences, making it difficult for students to receive immediate and targeted feedback during practice, such as a book published by New Oriental Education, short for NOE300(Chen et al., 2019). Automatic syntactic analysis can compensate for this by transcending the boundaries of time and space and provide unlimited sentence analysis support. Most of the current automatic English grammar parses are designed for processing simple sentences, e.g. Enpuz[0] analyses sentences with an upper limit of 20 words in length. Based on these considerations, this paper conduct automatic grammar analysis of long and complicated sentences without length constraints. We adopts the widely recognized Sentence Diagramming theory, referring its more standardized approaches such as Grammar Revolution[1] and Sentence Analytics[2]. These methods have covered most English grammatical cases, providing vivid and detailed analysis, but their perspectives of grammar explanation are not completely suited to the learning habits of Chinese ESL learners. Therefore, we have made improvements such as grouping various clauses to-

*Equal contribution.
†Corresponding author: Liner Yang (mailto: lineryang@gmail.com)

This work was supported by the MOE (Ministry of Education in China) Project of Humanities and Social Sciences (No. 23YJCZH264), the funds of Research Project of the National Language Commission (No. ZDA145-17) and Education and Teaching Reform Project of Beijing Language and Culture University in 2024 (No. JXGG202403).

[0]https://enpuz.com
[1]https://letsdiagram.com/
[2]https://www.sentenceanalytics.com/

gether, focusing on adverbial accompaniment, etc, and, designed the English Sentence Pattern Structure (ESPS).

Treebank is the processed corpus that records the syntactic annotation of every sentence, providing word segmentation, part-of-speech tagging, syntactic structure and other information. Based on the proposed grammatical system, we aim to build the English Sentence Pattern Structure Treebank (ESPST) and further develop automatic parses. The standard methods of constructing large-scale syntactic resources are manual annotation and automatic conversion. Manual annotation can ensure the data quality but is time-consuming and labor-intensive. A practical alternative method is to utilize existing treebank resources and converting them into the target treebank by finding the mapping relationships between two forms.

Current research on automatic treebank conversion mainly focuses on the conversion between constituency treebanks and dependency treebanks. Lin (1998) proposed an early method using a headword node table to convert constituency trees into dependency trees. Xia (2001) described two algorithms for converting constituency trees into dependency trees, employing a headword filtering table method, and proposed a new algorithm for converting the generated dependency trees back into constituency trees, with the results closely resembling the original Penn Treebank (PTB). Zabokrtsky (2003), Niu (2009), and Kong (2015) also conducted research on the conversion between constituency structure and dependency structure. In Chinese, some scholars have researched the conversion between constituency treebanks, dependency treebanks, and Chinese Sentence Pattern Structure Treebanks (SPST). Among them, Zhang (2018) converted the Tsinghua Chinese Treebank (TCT) into SPST, with an overall accuracy rate of 92.9%. Xie (2022) used rule-based methods to convert the Chinese Treebank (CTB) into SPST, with an overall accuracy rate of 89.72%. These studies have proved the feasibility of interconversion between different syntactic structures.

Using conversion rules and the advanced parser of the source treebank, we can automatically generate targeted trees from raw sentences. Yet, creating these conversion rules is a very challenging task that needs careful observation and steady practice, posing a significant challenge to our research. Considering the wide use of the PTB in natural language processing, we choose to convert the constituency structure treebank into ESPST. To verify the effectiveness of the conversion rules, we conducted experiments on a manually annotated test set to compare the conversion results with the effects of trained parser and large language models (LLMs). The results indicating that the rule-based conversion method proposed in this paper is the most effective.

2 Background

In this paper's conversion, the source treebank is mainly the English PTB corpus, and the target treebank is the ESPST we designed based on Sentence Diagramming theory, as introduced in Section 3. The formulation of conversion rules necessitates a comparative analysis of the grammatical forms between these two. This section will separately introduce PTB and Sentence Diagramming.

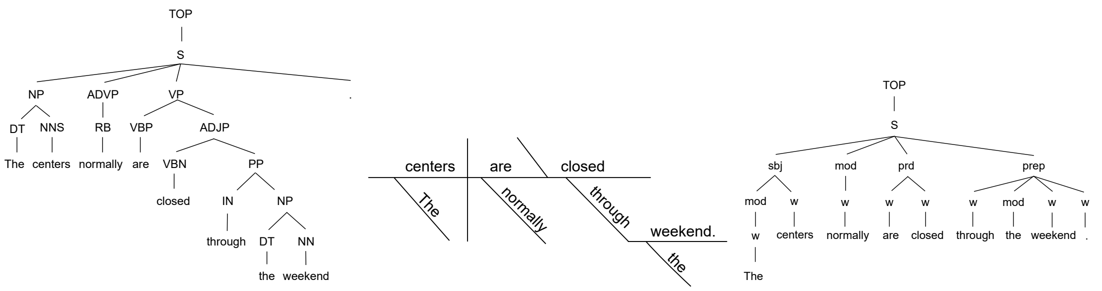
2.1 Penn Treebank

The English PTB corpus, particularly the section of the corpus corresponding to the articles of Wall Street Journal (WSJ), is one of the most known and used constituency structure corpus for the evaluation of models for sequence labelling. It is a corpus consisting of over 4.5 million words of American English. The material annotated by PTB includes such wide-ranging genres as IBM computer manuals, nursing notes, Wall Street Journal articles, and transcribed telephone conversations. The large amount of data produced by the project continues to provide an available resource for computational linguists, natural language programmers, corpus linguists and others interested in empirical language studies.

According to Marcus(1993), PTB corpus is annotated for part-of-speech (POS) information and skeletal syntactic structure. Considering the notable differences between the syntactic structure annotations provided in the PTB and the grammar accustomed by Chinese ESL learners, this paper exclusively focuses on the POS annotations of the PTB. The majority of the output of the PTB consists of tagged and bracketed versions. As shown in Figure 1a, compared to the ESPST as proposed in Section 3, the PTB

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1223-1238, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

we focus on only annotates part-of-speech information and hierarchical structure information, lacking in the depiction of syntactic relationships between sentence components, which can be extracted based on rules.



(a) Constituency Structure (b) Sentence Diagramming (c) English Sentence Pattern Structure

Figure 1: Example of three formats.

2.2 Sentence Diagramming

Sentence Diagramming is a pictorical representation of a sentence's grammatical structure, which is used to teaching difficult written language. The model shows the relations between words and the nature of sentence structure and can be used to help recognize which potential sentences are grammatically correct.

Most Sentence Diagramming methods in pedagogy are based on the work of Alonza Reed and Brainerd Kellog(1886). Sentences in the Reed-Kellogg system are diagrammed according to the following forms: the diagram begins with a horizontal line called the base; the subject is written on the left, the predicate on the right, separated by a vertical bar; the verb and its object are separated by a line that ends at the baseline; modifiers, as well as prepositional phrases, are placed on slanted lines below the word they modify. These basic diagramming conventions are augmented for other types of sentence structures, e.g. for coordination and subordinate clauses.

A specific example of sentence diagramming is illustrated in Figure 1b. Based on the direct modifying function of adverbials on the predicate meaning, *normally* is attached below the predicate *are*. Above the horizontal line is the simplified main component of the sentence, "centers are closed".

To further deepen our understanding of the components and rules of Sentence Diagramming, we referenced an exceptional work. The Grammar Revolution project, developed by Elizabeth O'Brien, aims to redefine traditional methods of grammar learning by offering an innovative perspective through interest. In Grammar Revolution, 11 lessons unfold sequentially: Basic Sentence Diagramming, Modifiers, Prepositional Phrases, Coordinating Conjunctions, etc. By transforming abstract grammatical concepts into concrete, visual patterns, this project not only makes grammar learning more engaging but also opens new avenues for learners who feel intimidated by or disinterested before. This has bolstered our confidence in applying Sentence Diagramming theory to the research for Chinese ESL learners.

3 The English Sentence Pattern Structure

Based on the theories of Sentence Diagramming discussed in Section 2.2, and in conjunction with Chinese ESL learners' cognitive habits, we have defined 14 grammatical labels involving sentence components and logical relationships. As shown in the Table 1, the 14 labels are categorized as main components, supplementary components, and relational components. These 14 components can all be represented by diagrams, which will not be elaborated here.

3.1 Main Components

For every sentence or clause, we analyze its main components, including the subject, predicate, direct object, indirect object, and predicative. These five main components constitute three basic sentence patterns: subject-verb-object, subject-verb-indirect object-direct object, and subject-linking verb-predicative.

Categories	Labels	Explanation	Categories	Labels	Explanation
Main Components	sbj	Subject	Supplementary Components	mod	Modification
	prd	Predicate		advcla	Adverbial Accompaniment
	obj	Object		cla	Clause
	pred	Predicative		wh	Relative Connectives
	iobj	Indirect Object	Relational Components	sencoo	Sentential Coordination
Supplementary Components	todo	Todo Infinitive		phrcoo	Phrasal Coordination
	prep	Prepositional Phrase		cc	Coordinating Conjunction

Table 1: The 14 grammatical labels.

The predicate refers to the main verb and its modifying elements. For example, in the sentence *He has not yet seen the bird*, the predicate would be *has not yet seen*, which includes the verb *has* and *seen* carrying tense information and modifiers *not yet*.

The linking verbs that introduce predicative indicates the subject's state, quality, characteristics, or nature, including verbs such as *be, remain, feel*, and their variant forms. Moreover, the predicative may be a noun, adjective, certain adverbs, non-finite verbs, prepositional phrases, or clauses.

3.2 Supplementary Components

To facilitate learners' grasp of the main skeleton of sentences, the *mod* label encompasses various components with a modifying function without further detailed subdivision, which includes adverbs, adjectives, numerals, quantity phrases, possessive pronouns, post-modifiers led by gerunds or past participles, appositives, etc.

Infinitives are typically used to express purpose or intention or as a complement to another verb. Prepositional phrases act as adverbials of time, place, manner, etc. For example, in *I came here to see the exhibition*, the *to* leads an infinitive, indicating the purpose of coming here is to see the exhibition; in *I look forward to seeing you soon*, *to* is a preposition as part of the phrase *look forward to*, followed by the gerund form *seeing*, rather than an infinitive.

Adverbial accompaniment, describing subsidiary actions or states that occur concurrently with the main action, is an integral part in sentence. It can be expressed through various grammatical forms, such as present participle phrases and past participle phrases. Formally, when adverbial clauses are positioned at the beginning or middle of a sentence, they are often separated from the main clause by commas.

Various types of clauses, such as adverb clauses, adjective clauses, noun clauses, etc., are uniformly classified as *cla*. The relative connectives of clauses can be a single word, such as *because, if, when, although*, or phrases such as *even though, in order that*.

3.3 Relational Components

In the ESPS, we further define the logical relationships of *coordination*. Coordination refers to the structural equivalence of two or more sentence components, which jointly function as a more significant unit and semantically represent various meanings such as *alliance*, *contrast*, and *progression*. Within a sentence, the coordination of sub-sentences or clauses is defined as *sencoo*, and the coordination of phrases is defined as *phrcoo*, with the coordinating conjunctions guiding these two types of relationships defined as *cc* (such as *and* or *but*). To enhance parsing efficiency, for phrase coordination, we currently focus only on the coordination of subjects, predicates, predicatives, and objects, which are directly related to the basic structure and meaning.

4 Constructing the English Sentence Pattern Structure Treebank

In English, the techniques of constituency structure are relatively mature and have yielded promising results. Therefore, we choose the constituency structure as the source treebank and construct the ES-PST through rule-based conversion. Specifically, this involves formulating conversion rules for the 14

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1223-1238, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

Labels	Explanation	constituency Structure	Conversion Rules	Examples
sbj	subject	S → NP	<sbj>NP</sbj>	It wasn't Black Monday.
prd	predicate	S/SINV → VP → (VP → VERB)	<prd>VERB</prd>	The equity market was illiquid.
obj	object	S/SINV → VP → (VP → VERB + NP)	<obj>NP</obj>	They received approvals for development.
pred	predicative	VP → (VERB + NP/PP/ ADVP/ADJP/SBAR) and VERB is linking verb	<pred>NP/PP/ADVP/ ADJP/SBAR</pred>	It wasn't Black Monday.
iobj	indirect object	S/SINV → VP → (VP → VERB + NP1 + NP2)	<iobj>NP1</iobj> <obj>NP2</obj>	She gave me a book.

Table 2: Sample conversion rules of main components. VERB includes the labels of *VBP, MD, VBD, VBZ, VBN, VB,* and *VBG*.

grammatical labels described in Section 3 and handling certain exceptional cases. However, compared to ESPST, the PTB lacks the depiction of syntactic relationships between sentence components. Thus, formulating the rules can also be considered as the precise correspondence between part-of-speech information and syntactic component information in English. The following elaborates on the conversion rules we have developed.

4.1 Conversion Rules of Grammatical Components

In the constituency structure, information on sentence components is scattered among part-of-speech labels, which cannot be directly correlated on a one-to-one basis. Based on this, we have compiled detailed rules for converting grammatical components. For each label, we only present a representative conversion rule here, with other specific rules available in the appendix A.

4.1.1 Conversion of Main Components

The selected rules for converting the five main components from constituency structure is shown in Table 2.

In the constituency structure, the noun phrase *NP* under the sentence *S* and the inverted sentence *SINV* is typically the subject of the sentence. If there is no *NP* at this position, then a sentence *S* or a clause *SBAR* at the same level is matched as the subject.

Three rules for predicate conversion correspond to scenarios where the same predicate part has one, two, or three verbs. These scenarios involve different hierarchical relationships in the constituency structure tree.

The conversion rules for direct objects, indirect objects, and predicatives are closely related to the rules for predicate: within the same level after a predicate in the constituency structure, if there is one *NP*, it is matched as a single direct object; if there are two *NP*, the first is matched as an indirect object and the latter as a direct object; *NP, PP, ADVP, ADJP, SBAR* at the same level as the linking verb are matched as predicatives.

4.1.2 Conversion of Supplementary Components

The selected rules for converting the six supplementary components from constituency structure is shown in Table 3.

For the preposition *to*, if its parent node is a verb phrase (*VP*), then this verb phrase matches as *todo*; if its parent node is a prepositional phrase (*PP*), then this prepositional phrase matches as *prep*. Moreover, combinations of other prepositions with noun phrases, sentences, or adjective phrases also match as *prep*.

As previously mentioned, the *mod* label encompasses various components with a modifying function. The conversion rules correspond to these nine types of modifiers: adverbs and phrases serving as adverbials, adjectives and adjective phrases, numerals, and quantity phrases, possessive pronouns, nouns

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1223-1238, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Labels	Explanation	constituency Structure	Conversion Rules	Examples
todo	Todo Infinitive	S/SBAR → VP → TO + VP	<todo>VP</todo>	And the link with stocks began to fray again.
prep	Prepositional Phrase	PP → IN/TO + NP/S/ADJP	<prep>PP</prep>	At the end of the day, 251.2 million shares were traded.
mod	Modification	NP/NML → JJ/JJS/ ADJP/RBR/PDT + NP/NN/NNS/NNP/NNPS	<mod>JJ/JJS/ADJP/ RBR/PDT</mod>	I wouldn't expect an immediate resolution to anything.
advcla	Adverbial of Accompaniment	S1/VP1 → S2/VP2 → PP/VP3 → VBG/VBN + XP	<advcla>S2/VP2</advcla>	Noting others' estimates, he said October.
cla	Clause	SBAR	<cla>SBAR</cla>	When the dollar is in a free-fall, even central banks can't stop it.
wh	Relative Connectives	SBAR → WHNP	<wh>WHNP</wh>	Speculators are calling for a degree of liquidity that is not there in the market.

Table 3: Sample conversion rules of supplementary components. *XP* stands for any component.

Labels	Explanation	constituency Structure	Conversion Rules	Examples
sencoo	Sentential Coordination	S1 → S2 + CC + S3	<sencoo>S1</sencoo>	But the build-up of S&P futures sell orders weighed on the market, and the link withstocks began to fray again.
phrcoo	Phrasal Coordination	S → NP1 → NP2 + CC + NP3	<phrcoo>NP1</phrcoo>	Many money managers and some traders had already left their offices.
cc	Coordinating Conjunction	CC in sencoo/phrcoo	<cc>CC</cc>	Many money managers and some traders had already left their offices.

Table 4: Sample conversion rules of relational components.

or noun phrases modifying another noun, post-modifiers led by gerunds or past participles, appositives, reflexive pronouns.

Adverbial accompaniments are typically led by the present participle *VBG* or past participle *VBN*. Comsidering the characteristics of their parent nodes, we define the conversion rule as *S/VP → S1/VP1 → PP/VP2 → VBG/VBN + XP*.

The *cla* label directly corresponds to the *SBAR* label in the PTB. The *wh* label appear in various forms, which can be single words, such as *IN* in *SBAR → IN + S*; or phrases, such as *WHNP* in *SBAR → WHNP*.

4.1.3 Conversion of Relational Components

The selected rules for converting the three relational components from constituency structure is shown in Table 4.

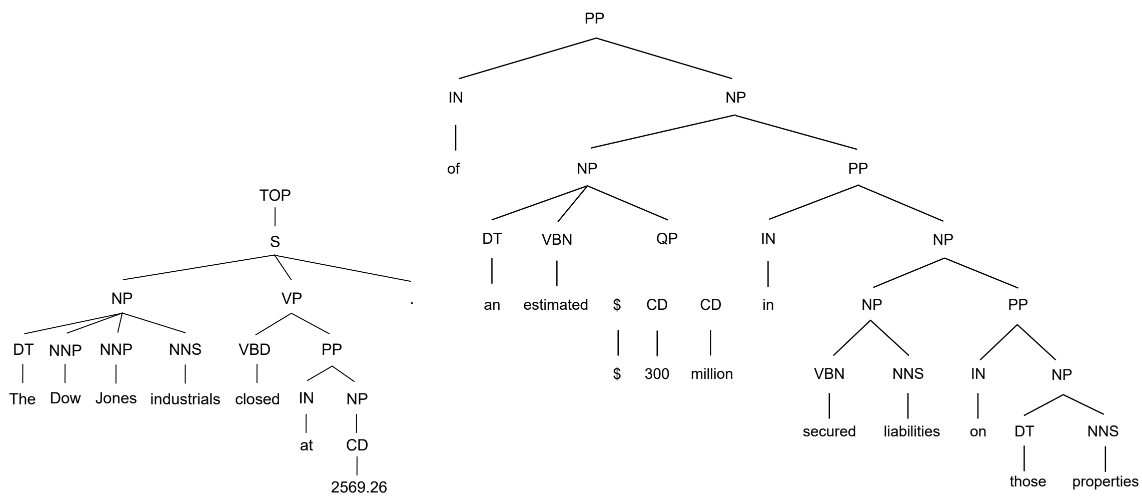
Sentence coordination encompasses coordination of sub-sentence and clauses, while phrase coordination encompasses coordination of predicates, subjects, objects, and predicatives. The conjunctions of sentence coordination and phrase coordination are matched as coordinating conjunctions.

4.2 Conversion Rules of Special Cases

In practice, we found that although the conversion rules cover the majority of standard structures, there are several special cases require appropriate handling.

- When the headword (i.e., the main noun in a noun phrase) has multiple modifiers, these modifiers should all point to the last noun individually. Noun phrases often follow a certain hierarchical structure, where the noun placed last (except in some cases of post-modifying attributes) is the

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1223-1238, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers



(a) Special Case of Proper Nouns.

(b) Special Case of Nest Restrictions.

Figure 2: Examples of special cases.

Dataset	Source	Numbers	Average length	Train	Dev	Test
PTB	Wall Street Journal articles, etc.	800	24.43	640	80	80
NOE	GRE, GMAT, LSAT, etc.	200	41.26	160	20	20
Total	-	1000	27.79	800	100	100

Table 5: Statistics of datasets.

headword, with preceding modifiers sequentially modifying and specifying it. Adding this rule aids in maintaining the grammatical correctness of the sentence.

- We stipulate that adjacent *NNP/NNPS* (proper nouns in PTB) are considered a joint unit for division or matching. In sentences containing place names, personal names, or specific terms, when two or more proper nouns are closely connected, they usually form a single semantic unit, expressing a compound concept or a concrete entity. For instance, as shown in Figure 2a, the modifier of the noun *industrials* includes *The* and *Dow Jones*, rather than *The*, *Dow*, and *Jones* as three separate modifiers.

- To maintain the clarity of sentence structure, we have imposed restrictions on the nesting of certain labels. The specific rules are as follows: *prep* and *todo* do not nest within *prep* or *todo* but can nest other labels; *wh* do not nest any other labels. For instance, in Figure 2b, *of an estimated $300 million in secured liabilities on those properties* would be converted to a *prep* label, without further conversion for *in secured liabilities on those properties* and *on those properties*. This approach is adopted to prevent the potential for ambiguity, which can arise from complex tag nesting structures in handling complicated sentences.

5 Experiments

Based on the rules, we completed the conversion of ESPST for a total of 39,406 sentences in constituency structure and manually annotated 1000 sentences as a test set to evaluate the conversion results. Among the 1000 sentences, we calculated the annotation consistency to measure the rules and data quality. To further verify the effectiveness of the treebank conversion method, we set up two additional experiments: a trained parser and a LLMs analysis.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1223-1238, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Categories	Labels	P	R	F1	Categories	Labels	P	R	F1
Main Components	sbj	99.22	97.10	98.15	Supplementary Components	todo	98.06	95.60	96.82
	prd	92.53	88.86	90.66		advcla	9.82	73.49	17.33
	pred	58.98	64.17	61.47		cla	96.77	97.82	97.29
	obj	86.19	90.71	88.39		wh	97.69	96.12	96.90
	iobj	38.64	85.00	53.12	Relational Components	sencoo	98.72	92.77	95.65
Supplementary Components	mod	86.87	84.77	85.81		phrcoo	99.30	78.33	87.58
	prep	95.16	94.36	94.76		cc	96.82	82.56	89.12
Avg. P					82.48				
Avg. R					87.26				
Avg. F1					84.81				

Table 6: Main results of rule-based conversion in 1000 sentences.

5.1 Dataset

The total 39,406 sentences in constituency structure are from the test division of PTB. Our dataset for evaluating the conversion results consists of 1000 sentences including 800 from PTB23 and 200 from NOE300. NOE300 is a book compiled to help Chinese students with reading long and complex sentences in tests, selecting examples of such sentences that appear in the GRE (Graduate Record Examinations), GMAT (Graduate Management Admission Test), and LSAT (Law School Admission Test). For the subsequent two comparative experiments in Section 5.3, we constructed the train, validation, and test sets from the dataset in an 8:1:1 ratio. The specific information about the 1000 dataset is shown in Table 5.

5.2 Evaluation of Rule Conversion

We manually annotated the ESPS of the 1000 sentences and calculated the Fleiss' Kappa score (Fleiss, 1971) of annotation agreement on two annotators to be 0.88, indicating that the grammatical labels are scientifically sound. By categorizing the components, we examine the specific results of automatic conversion, with overall result presented in the Table 6 below and detailed data available in the appendix B.

Our findings include:

1. The overall conversion results are satisfactory, indicating that the conversion rules for handling the sentence's main components, supplementary components, and relational components are scientifically valid, resulting in a high-quality treebank. The F1 scores on both datasets exceed 80, demonstrating the transfer-ability of this method to texts in other domains.

2. Conversion results for PTB23 data outperform those for NOE300 data. Under the same conversion rules, the F1 score for 800 PTB23 data conversions is 86.61, while for 200 NOE300 data conversions is 80.82, a difference of 5.79. As for the reason, the latter's sentences are, on average, about twice as long as the former's and are grammatically more complex, potentially leading to cases not covered by certain rules. Moreover, the constituency structure form of NOE300 used to generate conversion results was produced by the Berkeley Constituency Parser(Kitaev and Klein, 2018), which may introduce bias.

3. Among the 14 components, subjects, predicate, prepositional phrases, infinitives, clause, sentence coordination, and relative connectives have the better conversion results, with overall F1 scores exceeding 90. This indicates that their conversion rules can cover most grammatical cases, with constituency structure corresponding accurately to the respective ESPS, resulting in a low error rate in conversion results. Additionally, conversion results for components like object, modification, phrasal coordination, and coordinating conjunction are also considerable.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1223-1238, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

4. Predicatives in the PTB treebank appear in various forms, including noun phrases, adjective phrases, adverbial phrases, and prepositional phrases, for which we have written conversion rules to match these cases. However, in many sentences' constituency structures, nodes at the same level as and following the linking verb can often match multiple predicatives, only one of which is correct. For example, in the sentence *Diamond Shamrock is the operator, with a 100% interest in the well*, conversion rules match two predicatives of noun phrases and prepositional phrases, but only the former is correct. This issue may explain the relatively worse conversion results for predicatives.

5. In constituency structure trees, many temporal adverbials appear as NP(noun phrases), affecting conversion results for components like direct and indirect objects. In particular, indirect objects are significantly impacted, with high recall but low precision due to the small base. Observations reveal many temporal adverbials such as *last week*, and*yesterday*, *tomorrow morning* being matched as indirect objects. This issue can be partially resolved by defining a list of prohibited words: among the 155 misclassifications, 59 are temporal adverbials, indicating that introducing a prohibited list of temporal adverbials could resolve about one-third of this kind of issues.

6. Concomitant adverbials are form-flexible, making the conversion task challenging. Conversion rules matching verb phrases led by present or past participles yield many wrong components, such as predicates. In 1000 sentences, the label appeared 61 times in the manually annotated results but over 200 times in the conversion results, indicating a lower accuracy rate in the conversion. This suggests the weak correspondence between such sentence structure labels and part-of-speech information requires alternative matching approaches.

5.3 Comparative Experiments

We conducted two sets of experiments to compare the effectiveness of the treebank conversion rules proposed in this paper.

5.3.1 Setup

Experiment 1: This experiment primarily investigates the performance of an automatic syntactic parser trained on ESPST generated through rule-based conversion, aiming to explore the practical value of the method we propose. Drawing on Kitaev's neural network model[3](Kitaev et al., 2018) based on self-attention mechanisms, we trained an automatic syntactic parser for ESPS. The training set is the ESPST of 39,406 sentences converted from constituency structure trees in PTB, and the testing set is the 1000 sentences introduced in Section 5.1. The model employs an encoder-decoder architecture, using the pre-trained model Bert for the encoding phase and incorporating part-of-speech and positional information as auxiliary inputs to the model. The encoder sums the word representations $[w_1, ..., w_n]$, part-of-speech representations $[m_1, ..., m_n]$, and positional representations $[p_1, ..., p_n]$ to obtain word embeddings, which are then encoded using a multi-head attention mechanism. The decoder employs the CKY algorithm (Kasami, 1966; Younger, 1967; Cocke, 1969) to generate the ESPST.

Experiment 2: To test the syntactic analysis capability of LLMs on complex English sentences, we conducted prompt-based experiments on the general-domain GPT-4. The testing set is also the 1000 sentences introduced in Section 5.1. The full prompt (shown in Appendix C) given to GPT-4 for each testing sentence consisted of the following ordered elements:

- **Syntactic Labels**, introducing the 14 grammar labels and some necessary explanations;
- **Special Rules**, describing the special rules, which significantly impact the generation results;
- **Task**, explaining the task of analyzing sentences based on the ESPS;
- **Examples**, giving three complex English sentences and their correct output results. These three sentences are not included in the test set. These examples provided the LLMs with detailed information on the output format and the handling of punctuation;

[3]https://github.com/nikitakit/self-attentive-parser

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1223-1238, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

Models	PTB			NOE			TOTAL		
	P	R	F1	P	R	F1	P	R	F1
Trained Parser	85.21	85.35	85.28	79.73	78.37	79.04	82.92	82.69	82.80
GPT-4	41.38	38.51	39.89	35.31	33.45	34.35	40.17	37.50	38.78
Rule-based Conversion	83.44	90.04	86.61	80.55	81.10	80.82	82.48	87.26	84.81

Table 7: Results of comparative experiments in 1000 test dataset.

- **Testing Data**, giving the 1000 sentences to be tested, one at a time.

For the 1000 responses returned by GPT-4, we first extract the parenthetical syntactic trees to clean the data. Upon observation, the pairing of parentheses in these responses are not standardized, where there are cases of missing or redundant parentheses. Therefore, we resort to manual proofreading to adjust the format before calculate the results.

5.3.2 Results and Analysis

The overall performance of methods of trained parser, LLMs analysis, and rule-based conversion are listed in Table 7.

The conclusions drawn from the table are as follows: The method proposed in this paper, rule-based conversion from constituency structure, shows the best effect, with an F1 value 2.01 higher than that of the trained parser and 46.63 higher than the LLMs analysis. This indicating that the rule-based conversion algorithm has certain advantages in automatic treebank construction, and data generated through conversion rules demonstrates significant utility in training parses.

Compared to the trained parser, the method based on conversion rules does not rely on the manually annotated 1000 sentences, allowing it to be transferred to texts in other domains, showing more robust universality. The LLMs analysis experiment performed poorly, possibly due to the general-domain LLMs' lower accuracy in complex sentence syntactic analysis tasks or the model's unfamiliarity with the grammar labeling system tailored for Chinese ESL learners. Additionally, the length of the test sentences and the bracket format of the treebank may have contributed to the reduced accuracy.

In summary, the rule-based conversion algorithm proposed in this paper has certain advantages in the automatic construction of ESPST, showing vital accuracy and universality in analysis.

6 Conclusion

We developed an ESPS rooted in Sentence Diagramming theory, which is suited to Chinese ESL learners. Through rule-based conversion from the PTB, we constructed the ESPST and evaluated its effectiveness. Comparative experiments, including parser training and LLMs analysis, showed that our treebank conversion rule-based method yielded the best results. This work provides a new perspective on efficient English grammar learning of long and complicated sentence. However, our conversion process have shortcomings, such as the rules for indirect objects, predicatives, and adverbial clauses, which need further refinement. In the future, we will continue to optimize the conversion results and build an analysis platform for ESPS, realizing the visualization of automatic syntactic analysis.

References

Elizabeth B Bernhardt. 1993. *Reading development in a second language: Theoretical, empirical, & classroom perspectives.* ERIC.

Qi Chen, Yi Ge, and Yuzhen Yan. 2019. *New Oriental Education: 300 Detailed Explanations and Practices for Long and Complex GRE/GMAT/LSAT Sentences*. Zhejiang Education Publishing House.

John Cocke. 1969. *Programming languages and their compilers: Preliminary notes*. New York University.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Tadao Kasami. 1966. An efficient recognition and syntax-analysis algorithm for context-free languages. *Coordinated Science Laboratory Report no. R-257*.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. *arXiv preprint arXiv:1805.01052*.

Nikita Kitaev, Steven Cao, and Dan Klein. 2018. Multilingual constituency parsing with self-attention and pre-training. *arXiv preprint arXiv:1812.11760*.

Lingpeng Kong, Alexander M Rush, and Noah A Smith. 2015. Transforming dependencies into phrase structures. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 788–798.

Dekang Lin. 1998. A dependency-based method for evaluating broad-coverage parsers. *Natural Language Engineering*, 4(2):97–114.

Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Zheng-Yu Niu, Haifeng Wang, and Hua Wu. 2009. Exploiting heterogeneous treebanks for parsing. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 46–54.

Alonzo Reed and Brainerd Kellogg. 1886. *Higher lessons in English*. Scholars' Facsimiles & Reprints.

Dianne Wall and J Charles Alderson. 1993. Examining washback: the sri lankan impact study. *Language testing*, 10(1):41–69.

Fei Xia and Martha Palmer. 2001. Converting dependency structures to phrase structures. In *Proceedings of the first international conference on Human language technology research*.

Chenhui Xie, Zhengsheng Hu, Liner Yang, Tianxin Liao, and Erhong Yang. 2022. Automatic construction of sentence pattern structure treebank. In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pages 464–474.

Daniel H Younger. 1967. Recognition and parsing of context-free languages in time n3. *Information and control*, 10(2):189–208.

Zdeněk Žabokrtskỳ and Otakar Smrz. 2003. Arabic syntactic trees: from constituency to dependency. In *10th Conference of the European Chapter of the Association for Computational Linguistics*.

YinBing Zhang, Jihua Song, Weiming Peng, Yawei Zhao, and Tianbao Song. 2018. Automatic conversion of phrase structure treebank to sentence structure treebank. *Journal of Chinese Information Processing*, (5):31–41.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1223-1238, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

A Full Conversion Rules

Full conversion rules in this work is shown in Table 8-10.

Labels	Explanation	constituency Structure	Conversion Rules	Examples
sbj	subject	S → NP	<sbj>NP</sbj>	It wasn't Black Monday.
		SINV → NP	<sbj>NP</sbj>	At 2:43 p.m. EDT, came the sickening news.
		S → S/SBAR + VP	<sbj>S/SBAR</sbj>	The thing that they have done is a question.
prd	predicate	S/SINV → VP → (VP → VERB)	<prd>VERB</prd>	The equity market was illiquid.
		S/SINV → VP → (VP → VERB1 + VP→ (VP → VERB2))	<prd>VERB1 + VERB2</prd>	At the end of the day, 251.2 million shares were traded.
		S/SINV → VP → (VP → VERB1 + VP → (VP → VERB2 + VP → (VP → VERB3)))	<prd>VERB1 + VERB2 + VERB3</prd>	Several traders could be seen shaking their heads.
obj	object	S/SINV → VP → (VP → VERB + NP)	<obj>NP</obj>	They received approvals for development.
		S/SINV → VP → (VP → VERB + VP → (VP → VERB + NP))	<obj>NP</obj>	He could watch updates on prices and pending stock orders.
		S/SINV → VP → (VP → VERB + VP → (VP → VERB + VP → (VP → VERB + NP)))	<obj>NP</obj>	The suppliers haven't been filling their quotas to the full extent.
pred	predicative	VP → (VERB + NP/PP/ ADVP/ADJP/SBAR) and VERB is linking verb	<pred>NP/PP/ADVP/ ADJP/SBAR</pred>	It wasn't Black Monday.
iobj	indirect object	S/SINV → VP → (VP → VERB + NP1 + NP2)	<iobj>NP1</iobj> <obj>NP2</obj>	She gave me a book.
		S/SINV → VP → (VP → VERB + VP → (VP → VERB + NP1 + NP2))	<iobj>NP1</iobj> <obj>NP2</obj>	She has told us the news.
		S/SINV → VP → (VP → VERB + VP → (VP → VERB + VP → (VP → VERB + NP1 + NP2)))	<iobj>NP1</iobj> <obj>NP2</obj>	He might have offered his colleague some help.

Table 8: Full conversion rules of main components. VERB includes the labels of *VBP, MD, VBD, VBZ, VBN, VB,* and *VBG*.

Labels	Explanation	constituency Structure	Conversion Rules	Examples
todo	Todo Infinitive	S/SBAR → VP → TO + VP	<todo>VP</todo>	And the link with stocks began to fray again.
prep	Prepositional Phrase	PP → IN/TO + NP/S/ADJP	<prep>PP</prep>	At the end of the day, 251.2 million shares were traded.
mod	Modification	NP/NML → JJ/JJS/ ADJP/RBR/PDT + NN/NNS/NP/NNP/NNPS	<mod>JJ/JJS/ADJP/ RBR/PDT</mod>	I wouldn't expect an immediate resolution to anything.
		(NP/ADJP/PP/S/INTJ/ VP → RB)/ADVP	<mod>RB/ADVP</mod>	These stocks eventually reopened.
		NP → DT + XP	<mod>DT</mod>	The equity market was illiquid.
		NP → NN1/NP1/NNP1/ VBP/VBG/NNS1 + NN2/NNS2/NP2/NNP2	<mod>NN1/NP1/NNP/ VBP/VBG/NNS1</mod>	The equity market was illiquid.
		NP → NP1+ ,/: + NP2 + punctuation	<mod>NP2</mod>	He is Howard Rubel, an analyst at Lawrance Inc. .
		NP → QP/CD + NNS/NN	<mod>QP/CD</mod>	At the end of the day, 251.2 million shares were traded.
		NP→ PRP$ + NN/NNP/ NNS/NP/NNPS/VBG	<mod>PRP$</mod>	Several traders could be seen shaking their heads.
		NP → NP1/PP → VBG/VBN + XP	<mod>NP1/PP</mod>	The book lying on the table is mine.
		NP → NP1 + NP2 and Reflexive pronouns in NP2	<mod>NP2</mod>	It is index of the stock market itself.
advcla	Adverbial of Accompaniment	S1/VP1 → S2/VP2 → PP/VP3 → VBG/VBN + XP	<advcla>S2/VP2</advcla>	Noting others' estimates, he said October.
cla	Clause	SBAR	<cla>SBAR</cla>	When the dollar is in a free-fall, even central banks can't stop it.
wh	Relative Connectives	SBAR → WHNP	<wh>WHNP</wh>	Speculators are calling for a degree of liquidity that is not there in the market.
		SBAR → IN + S	<wh>IN</wh>	There came news that the UAL group couldn't get financing for its bid.
		SBAR → WHADVP	<wh>WHADVP</wh>	When the dollar is in a free-fall, even central banks can't stop it.
		SBAR → WHPP	<wh>WHPP</wh>	But nobody knows at what levell the futures and stocks will open today.

Table 9: Full conversion rules of supplementary components. *XP* stands for any component.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1223-1238, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

Labels	Explanation	constituency Structure	Conversion Rules	Examples
sencoo	Sentential Coordination	S1 $\rightarrow$ S2 + CC + S3	<sencoo>S1</sencoo>	But the build-up of S&P futures sell orders weighed on the market, and the link withstocks began to fray again.
		SBAR1 $\rightarrow$ SBAR2 + CC + SBAR3	<sencoo>SBAR1</sencoo>	He said that he had not yet seen the bid but that he would review it.
phrcoo	Phrasal Coordination	S $\rightarrow$ NP1 $\rightarrow$ NP2 + CC + NP3	<phrcoo>NP1</phrcoo>	Many money managers and some traders had already left their offices.
		S $\rightarrow$ VP $\rightarrow$ (VP + CC + VP) / (VBD + CC + VBD)	<phrcoo>VP</phrcoo>	Mr. Shidler's company specializes in commercial real-estate investment and claims to have $1billion in assets.
		VP $\rightarrow$ VERB + (NP1 $\rightarrow$ NP2 + CC + NP3)	<phrcoo>NP1</phrcoo>	A portion will be used to repay its bank debt and other obligations.
		VP $\rightarrow$ VERB + (NP1 $\rightarrow$ NP2 + CC + NP3)/ (PP1 $\rightarrow$ PP2 + CC + PP3)/ (ADJP1 $\rightarrow$ ADJP2/JJ1 + CC + ADJP3/JJ2)/ (ADVP1 $\rightarrow$ ADVP2/RB1 + CC + ADVP3/RB2) and VERB is linking verb	<phrcoo>NP1/PP1/ ADJP1/ADVP1</phrcoo>	Mr.Simpson is a developer and a former senior executive of LJ. Hooker.
cc	Coordinating Conjunction	CC in sencoo/phrcoo	<cc>CC</cc>	Many money managers and some traders had already left their offices.

Table 10: Full conversion rules of relational components.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1223-1238, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

B Detailed Conversion Results

The specific performance of the conversion rules on PTB23 and NOE300 is shown in Table 11.

<table>
<tr><td rowspan="11">PTB23</td><td>Categories</td><td>Labels</td><td>P</td><td>R</td><td>F1</td><td>Categories</td><td>Labels</td><td>P</td><td>R</td><td>F1</td></tr>
<tr><td rowspan="5">Main Components</td><td>sbj</td><td>99.24</td><td>98.02</td><td>98.62</td><td rowspan="4">Supplementary Components</td><td>todo</td><td>100.00</td><td>100.00</td><td>100.00</td></tr>
<tr><td>prd</td><td>92.81</td><td>89.90</td><td>91.33</td><td>advcla</td><td>10.94</td><td>73.53</td><td>19.05</td></tr>
<tr><td>pred</td><td>60.58</td><td>65.97</td><td>63.16</td><td>cla</td><td>99.84</td><td>99.69</td><td>99.77</td></tr>
<tr><td>obj</td><td>84.46</td><td>91.20</td><td>87.70</td><td>wh</td><td>99.75</td><td>98.27</td><td>99.00</td></tr>
<tr><td>iobj</td><td>38.89</td><td>100.00</td><td>56.00</td><td rowspan="3">Relational Components</td><td>sencoo</td><td>100.00</td><td>94.34</td><td>97.09</td></tr>
<tr><td rowspan="2">Supplementary Components</td><td>mod</td><td>86.44</td><td>85.31</td><td>85.87</td><td>phrcoo</td><td>99.03</td><td>82.26</td><td>89.87</td></tr>
<tr><td>prep</td><td>99.39</td><td>96.56</td><td>97.96</td><td>cc</td><td>96.73</td><td>85.55</td><td>90.80</td></tr>
<tr><td colspan="5">Avg. P</td><td colspan="5">83.44</td></tr>
<tr><td colspan="5">Avg. R</td><td colspan="5">90.04</td></tr>
<tr><td colspan="5">Avg. F1</td><td colspan="5">86.61</td></tr>
<tr><td rowspan="11">NOE300</td><td>Categories</td><td>Labels</td><td>P</td><td>R</td><td>F1</td><td>Categories</td><td>Labels</td><td>P</td><td>R</td><td>F1</td></tr>
<tr><td rowspan="5">Main Components</td><td>sbj</td><td>99.17</td><td>94.46</td><td>96.75</td><td rowspan="4">Supplementary Components</td><td>todo</td><td>92.31</td><td>83.72</td><td>87.80</td></tr>
<tr><td>prd</td><td>91.81</td><td>86.25</td><td>88.95</td><td>advcla</td><td>6.17</td><td>73.33</td><td>12.29</td></tr>
<tr><td>pred</td><td>56.35</td><td>61.21</td><td>58.68</td><td>cla</td><td>91.56</td><td>94.55</td><td>93.03</td></tr>
<tr><td>obj</td><td>90.43</td><td>89.60</td><td>90.02</td><td>wh</td><td>95.27</td><td>93.60</td><td>94.43</td></tr>
<tr><td>iobj</td><td>37.50</td><td>50.00</td><td>42.86</td><td rowspan="3">Relational Components</td><td>sencoo</td><td>96.43</td><td>90.00</td><td>93.10</td></tr>
<tr><td rowspan="2">Supplementary Components</td><td>mod</td><td>87.82</td><td>83.62</td><td>85.67</td><td>phrcoo</td><td>100.00</td><td>69.64</td><td>82.11</td></tr>
<tr><td>prep</td><td>85.28</td><td>88.87</td><td>87.04</td><td>cc</td><td>97.01</td><td>76.47</td><td>85.53</td></tr>
<tr><td colspan="5">Avg. P</td><td colspan="5">80.55</td></tr>
<tr><td colspan="5">Avg. R</td><td colspan="5">81.10</td></tr>
<tr><td colspan="5">Avg. F1</td><td colspan="5">80.82</td></tr>
</table>

Table 11: Detailed conversion results.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1223-1238, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

C Prompt in LLMs Experiment

The full prompt in the LLMs experiment is shown in Figure 3.

Prompt

Syntactic Labels: Subject <sbj>, Predicate <prd>, Object <obj>, Indirect Object <iobj>, Predicative <pred>, Infinitive <todo>, Prepositional Phrase <prep>, Clause <cla>, Modifier <mod> (adverbs and phrases as adverbials, adjectives, adjective phrases, numerals and quantifier phrases, possessive pronouns, nouns or noun phrases modifying another noun, post-modifiers led by gerunds or past participles, appositives, reflexive pronouns, etc.), Subordinating Conjunction <wh>, Coordinating Conjunction <cc>, Sentence Coordination <sencoo>, Phrase Coordination <phrcoo> (coordination of subjects, predicates, predicatives, and objects), Adverbial Clause <advcla>.

Special Rules: The prep tag (prepositional phrase) and todo tag (to-infinitive) do not nest within prep tags and todo tags but can nest other tags; wh tags (clause introducers) do not nest any other tags.

As a linguist expert, you are adept at analyzing sentences based on the aforementioned syntactic structure tagging system. Please refer to the provided examples and follow the established format to analyze the sentence I provide to you.

Sentence: Under these deals , the RTC sells just the deposits and the healthy assets .
Output: (sen (prep (w Under)(mod (w these))(w details)(w ,))(sbj (mod (w the))(w RTC))(prd (w sells))(obj (phrcoo (mod (w just))(mod (w the))(w deposits)(cc (w and))(mod (w the))(mod (w healthy))(w assets)(w .))))
Sentence: A Candian bank bought another thrift , in the first RTC transaction with a foreign bank .
Output: (sen (sbj (mod (w A))(mod (w Candian))(w bank))(prd (w bought))(obj (mod (w another))(w thrift)(w ,))(prep (w in)(mod (w the))(mod (w first))(mod (w RTC))(w transaction)(w with)(mod (w a))(mod (w foreign))(w bank)(w .)))
Sentence: Two of the four big thrifts were sold to NANB Crop. , Charlotte , N.C. , which has aggressively expanded its markets , particularly in Texas and Florida .
Output: (sen (sbj (w Two)(prep (w of)(mod (w the))(mod (w four))(mod (w big))(w thrifts)))(prd (w were)(w sold))(prep (w to)(w NCNB)(w Corp.)(w ,)(mod (w Charlotte)(w ,)(w N.C.)(w ,))(cla (wh (w which))(prd (w has)(mod (w aggressively))(w expanded))(obj (mod (w its))(w obj)(w ,))(mod (w particularly))(w in)(w Texas)(w and)(w Florida)(w .))))

Please analyze the given sentence: ...

Figure 3: Prompt in LLMs experiment.