# The Annotation Scheme of English-Chinese Clause Alignment Corpus

**Shili Ge**
Laboratory of Language and Artificial Intelligence, Guangdong University of Foreign Studies, Guangzhou, China 510420
geshili@gdufs.edu.cn

**Xiaoping Lin**
Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou, China 510420
lxpteresa@126.com

**Rou Song** ✉
Laboratory of Language and Artificial Intelligence, Guangdong University of Foreign Studies, Guangzhou, China 510420
College of Information Science, Beijing Language and Culture University, Beijing 100083
songrou@126.com

## Abstract

A clause complex consists of clauses, which are connected by component sharing relations and logic-semantic relations. Hence, clause-complex level structural transformations in translation are concerned with the expression adjustment of these two types of relations. In this paper, a formal scheme for tagging structural transformations in English-Chinese translation is designed. The annotation scheme include 3 steps operated on two grammatical levels: parsing an English clause complex into constructs and assembling construct translations on the clause complex level; translating constructs independently on the clause level. The assembling step involves 2 operations: performing operation functions and inserting Chinese words. The corpus annotation shows that it is feasible to divide structural transformations in English-Chinese translation into 2 levels. The corpus, which unfolds formally the operations of clause-complex level structural transformations, would help to improve the end-to-end translation of complicated sentences.

## 1 Introduction

The grammatical levels of a natural language include morpheme, word, group/phrase, clause, and clause complex. Units of a higher level are made up of units of a lower level. Therefore, the central task for machine translation is language transformations on each grammatical level between languages. So far, there have been many studies on group/phrasal- and clausal-level structures and structural transformations. However, clause-complex level (CC-level) structures and structural transformations are far less discussed.

Halliday and Matthiessen (2004) describes the structures of English clause complex based on the theory of Systemic-Functional Grammar. Wang (2012) carries out an in-depth study on the structures of Chinese complex sentence in comparison with English. Luo (1992) points out that clauses should be considered as the translation unit in English-Chinese translation. These studies are enlightening, but they are limited to theoretical illustrations and discussions. Song and Ge (2015) study clause complex for language engineering. They put forward and demonstrate the PTA (Parsing-Translating-Assembling) model for English-Chinese translation on the CC-level, which is only a tentative idea and has not been tested through corpus annotation. Ge and Song (2020) clarify the concept of Component Sharing, define clause and clause complex based on this concept, and propose the design of the annotation scheme and specification for English-Chinese Clause Alignment Corpus (ECCA Corpus). Yet, the details of the annotation scheme and specification of the ECCA Corpus still need further study and exploration, especially on the structural transformations between English and Chinese clause complexes and their annotation.

A clause complex consists of clauses, but many clauses are not connected linearly because there are shared components between them. In order to present the alignment of English and Chinese clauses, it is necessary to show how English and Chinese clauses correspond under various component sharing mechanisms. In ECCA Corpus, the correspondence relationship between English and Chinese clauses is



Proceedings of the 19th China National Conference on Computational Linguistics, pages 1035-1044,
Hainan, China, October 30 - Novermber 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

shown through the annotation process of CC-level structural transformations, including construct analyzing, construct translating, and construct and component translations assembling. The work of this paper completes the annotation scheme, including defining the operation unit of CC-level structural transformations, i.e. constructs, specifying the content of each annotation step, formalizing assembling operations, and summarizing the operation functions used and the Chinese words inserted.

It is believed that ECCA Corpus is significant for theoretical linguistics and cognitive linguistics by providing samples for comparing CC-level structures and studying structural transformations between English and Chinese. Meanwhile, the corpus is believed to be significant in application. Although machine translation has been greatly improved with data-driven approaches, it still fails to produce satisfying results when it comes across long sentences with complicated structures. This corpus explores the feasibility of and practical ways for mechanical transformations on the CC-level. It is hoped that the knowledge of CC-level structural transformations may help to improve the performance of machine translation in dealing with complicated sentences.

The remainder of this paper is organized as follows: Section 2 introduces the objective of annotation, Section 3 introduces the annotation scheme, Section 4 and 5 present operation functions and inserted Chinese words applied in annotation; Section 6 provides relevant statistical results, and Section 7 concludes the paper.

## 2 Clause-Complex Level Structural Transformation

The ECCA Corpus is designed to annotate CC-level structural transformations between English and Chinese. In most linguistic theories, a clause complex is generally regarded as a group of clauses combined together based on logic-semantic relations. This being the case, CC-level structural transformations during translation should involve only reordering of clauses, which are usually organized in different logical ways between languages. However, there is another important transformation that should be noticed, i.e. the transformation of naming-telling structural relations.

> **Example 1**: There are fewer than 100 potential customers for supercomputers priced between $15 million and $30 million – presumably the Cray-3 price range.
>
> **Chinese Translation**: 价格 在1500万美元至3000万美元之间 的 超级计算机 的潜在客户不到100家，这个区间是克雷 3 号机大概的价格范围。
>
> **Machine Translation**: 价格 在1500万美元到3000万美元之间 的 超级计算机 的潜在客户不到100家——大概是Cray-3的价格区间。

In Example 1, the English clause complex contains a "modified component & modifying component" structure and a "described component & describing component" structure. As stated in Fang et al. (2016), the modifying and describing components are tellings, while the modified and described ones are namings. The two namings are highlighted in grey. The modifying telling, which closely follows its modified naming, "supercomputers", is marked with a single underline. The describing telling, which closely follows its described naming, "between $15 million and $30 million", is marked with a wave underline. It can be seen that the described naming is embedded inside the previous modifying telling. In the Chinese translation, the translation of the modifying telling "priced between $15 million and $30 million" is reordered and placed before the translation of its modified naming, "supercomputers". Thus, the translation of the describing telling, "– presumably the Cray-3 price range", could not share its described naming as it does in the English text. To deal with the problem, the described naming is reproduced in the Chinese translation as a generalized form "这个区间" and combined with the translation of its describing telling into a new clause. However, the machine translation does not reproduce the described naming and thus fails to translate the "described component & describing component" structure correctly. This example shows that the adjustment of naming-telling relationship is no less important than logic-semantic relationship adjustment in CC-level structural transformations.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1035-1044,
Hainan, China, October 30 - Novermber 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Previous corpus studies prove that naming-telling structures are prevalent in both Chinese and English clause complexes. Although the two languages share the same types of naming-telling structures, they have different distributions of the structure types (Ge and Song, 2016). As a result, naming-telling structure adjustment is often necessary in English-Chinese translation. Meanwhile, the two languages arrange clauses in different logical ways, which leads to the other kind of structural transformations.

To sum up, the annotations of CC-level structural transformations are to demonstrate the adjustment of naming-telling structures and logical expressions in English-Chinese translation.

## 3 Design of the Annotation Scheme

The CC-level structural transformations of Example 1 are illustrated in Figure 1.

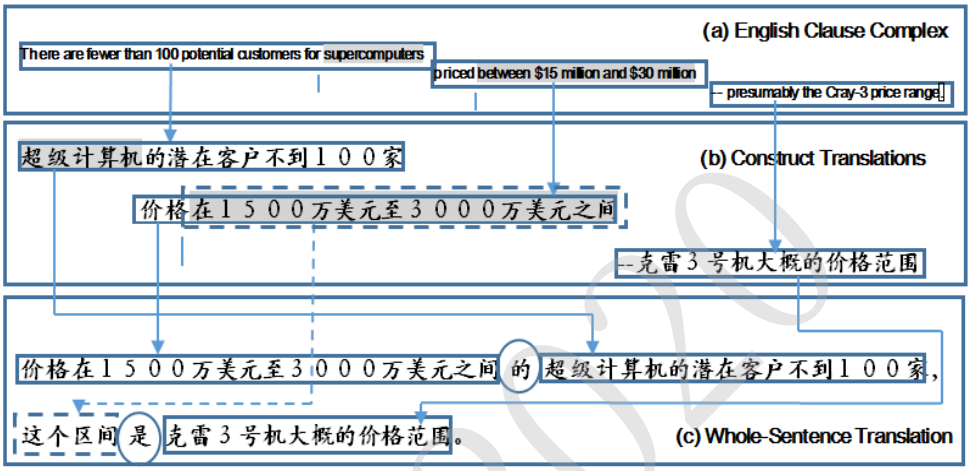


Figure 1. CC-Level Structural Transformation of Example 1

In Figure 1-(a), the English clause complex is firstly segmented into three constructs based on naming-telling structural analysis. The grey parts are namings, whose left-boundaries are marked by the symbol "|" below the line. Tellings modifying or describing these namings take up new lines and are indented to the right after their namings. This way of demonstrating the naming-telling relationship is called newline-indent schema.

Each line in Figure 1-(a) is considered as one construct for making up the English clause complex, and they are translated independently in Figure 1-(b). Each line of translations in Figure 1-(b) is called a construct translation. Contruct translations are also displayed in the newline-indent schema, with the translations of tellings indented to the right side after the translations of their namings. The arrows between Figure 1-(a) and 1-(b) start from each English construct and point to their Chinese counterparts. Figure 1-(c) shows the whole-sentence translation. The solid line arrows between Figure 1-(b) and 1-(c) start from each construct translation, and point to their new positions in the whole-sentence translation. The dash line arrow starts from the translation of a naming and points to its generalized form. The circles in Figure 1-(c) mark the insertion of the particle "的" and the linking verb "是".

The graphic demonstration in Figure 1 clearly displays how the English clause complex is transformed step by step into a Chinese one. However, the demonstration is quite complicated, hard to be annotated and not convenient for statistical analysis. Hence, a more formal annotation scheme for annotating structural transformations is designed.

The formal annotation scheme follows the 3 steps in the graphic demonstration: (1) segment English clause complexes into constructs and display them in newline-indent schema; (2) translate independently each construct into Chinese; (3) rearrange construct translations for a whole-sentence translation.

The structural transformations are to be annotated at the end of each line of the whole-sentence translation. The parts that make up the whole-sentence translation are encoded as numbers, and the operations implemented on these parts are tagged as operation functions. In this way, structural

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1035-1044,
Hainan, China, October 30 - Novermber 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

transformations could be annotated formally. The following is a detailed illustration of the designs.

**Whole-Sentence Translation of Example 1**:
价格在1500万美元至3000万美元之间的超级计算机的潜在客户不到100家，//2+的+1
这个区间是克雷3号机大概的价格范围。//sum(2.2)+是+delt(3)

As shown above, structural transformations are tagged after the symbol "//" at the end of each line. The numbers represent the parts making up the whole-sentence translation. For example, the number 2 of "2+的+1" represent the second line of construct translations, namely "价格在1500万美元至3000万美元之间". The number 2.2 of "sum(2.2)+是+delt(3)" represent the second section of the second line of construct translations, namely "在1500万美元至3000万美元之间". In the annotation scheme, the translations of namings are usually processed as a single unit. When the translation of a naming is positioned within a construct translation, the construct translation is segmented by the translation of this naming into several parts, which includes the naming translation, the parts before and/or after the naming translation. These segments are named as component translations. The component translations on the $n^{th}$ line are encoded from left to right as n.1, n.2, and n.3 etc. In this example, the second line of construct translation contains the translation of a naming at its end, and thus it is divided into two components. The component before the naming translation is encoded as 2.1, while the naming translation is encoded as 2.2. From this example, it can be seen that the parts making up a whole-sentence translation include construct translations and component translations. These two types of constituents in translations are the basic units to be dealt with by operation functions, and thus they are called operation units in this paper.

As for operation functions, they are used to mark the operations implemented on operation units. The symbol "+" means linking two operation units. The function "sum(2.2)" means turning the encoded component 2.2, namely "在1500万美元至3000万美元之间", into a more generalized expression "这个区间". The function "delt(3)" means deleting the dash in the translation of the encoded construct 3, namely "–克雷3号机大概的价格范围". The designing of operation functions will be discussed in detail in section 4.

Additionally, it is noted that the translation of every construct in the second step is independent of its context. Certainly, the disambiguation of a certain word still need reference to its context, but it is not allowed to add extra words, delete words or change the structures based on the context.

## 4 Operation Functions

There are two types of operations for CC-level structural transformations: (1) processing and assembling the operation units, and (2) inserting Chinese words. The first type of operation is annotated as operation functions, which will be discussed in this section. The second type of operation will be discussed in Section 5.

Operation functions are written in the format of FunctionName(x) or FunctionName(x,y), in which FunctionName specifies the operation to be implemented, while x and y specify the objects to be processed, which are all called operation units.

Twenty operation functions are designed, which involve 6 types of operations: link, reorder, add, delete, rewrite, and substitute. The 20 operation functions are listed in Table 1.

| Operation Types | Operation Functions |
|---|---|
| Link | concatenate(x,y) (i.e. x+y) |
| Reorder | demonstrated with the codes of operation units |
| Add | corcj(x), corcj2(x) , prd(x) |
| Delete | ignore(x) (i.e. *x), delcj(x) , delcj2(x) , delpn(x) , delt(x) |
| Rewrite | det(x) , ndet(x) , sum(x) , pron(x) , rel(x) , paren(x) , n2v(x) |
| Substitute | rpw(x,y) , r2n(x,y) , n2r(x,y) |

Table 1. Operation Functions

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1035-1044,
Hainan, China, October 30 - Novermber 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Of all these functions, link and reorder are common operations in almost all processed whole-sentence translations. The usage of these two functions is shown above in Example 1. Other functions are divided into two types based on their adjustments to clause complex structures. Some of the two classes are discussed with examples in the following subsections. Due to limited space, the functions not discussed in this paper can be referred to in Song et al. (2020).

### 4.1 Operation Functions for Transforming Naming-Telling Structures

Due to different distributions of naming-telling structural types, it is often necessary to transform naming-telling structures during English-Chinese translation. Generally, there are 3 ways to rearrange English tellings in Chinese translations: (1) inserting the telling translation as a modifier on the left of its naming translation, (2) keeping the telling translation as a statement or a description on the right of its naming translation, (3) reproducing the naming and rendering it another way before linking it with the telling translation. Of these 3 ways, the previous two requires only the link and reorder operations. When it comes to the third way, extra processing is needed, namely to reproduce the naming and render it in certain forms. This is because in a clause complex, a naming, if referred to more than once, should take different forms for its respective occurrence. To be more specific, a naming usually appear at first in its full name or its indefinite form, and then appear in its definite form, as a pronoun, or as a more generalized form. The operation functions det(x), ndet(x), pron(x) and sum(x) are specially designed for rewriting a naming. Table 2 presents the definitions of operation functions used to transform naming-telling structures.

| Operation Types | Operation Functions | Definition |
|---|---|---|
| Rewrite | det(x) | change x into its definite form |
| Rewrite | ndet(x) | change x into its indefinite form |
| Rewrite | pron(x) | change x into a corresponding pronoun |
| Rewrite | sum(x) | change x into a more generalized term |
| Rewrite | rel(x) | concretize x based on the current context |
| Delete | ignore(x) | delete the relative pronoun/adverb in x |
| Substitute | rpw(x,y) | replace the relative pronoun/adverb in x with y |

Table 2. Operation Functions for Transforming Naming-telling Structures

The usage of sum(x) has been illustrated in Example 1. The usage of ignore(x) and rpw(x,y) will be discussed in the following.

Since attributive clauses do not have clear semantic meanings by themselves, they need special treatment in annotation. In an attributive clause, the relative pronoun is only a formal substitute for the antecedent, and it is meaningless by itself. As a result, attributive clauses cannot be translated independent of context theoretically. To handle the problem, it is specified that relative pronouns in capitalized forms should be used to occupy the positions where the translations of antecedents should have been in construct translations.

In most cases, capitalized relative pronouns occupy the positions of a subject at the beginning of construct translations. Hence, the ignore(x) function is used to delete the capitalized relative pronouns before construct translations are linked with the translations of their namings.

Sometimes, capitalized relative pronouns occupy positions in the middle of construct translations. In this case, the function rpw(x,y) should be used to replace relative pronouns with the translations of their antecedents. Such substitutions are operable since capitalized relative pronouns are identifiable with their special forms. Example 2 shows the usage of this function.

**Example 2**: The Company has proposed an internal reorganization plan in Chapter 11 bankruptcy proceedings, under which it would remain an independent company.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1035-1044,
Hainan, China, October 30 - Novermber 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

**(1) Newline-Indent Schema of English Clause Complex**:
The Company has proposed an internal reorganization plan in Chapter 11 bankruptcy proceedings,
| under which it would . . . company.

**(2) Construct Translations**:
该公司已在第 1 1 章破产程序中提出了一个内部重组计划，
| 根据WHICH 它将仍为一个独立公司。

**(3) Whole-Sentence Translation**:
该公司已在第 1 1 章破产程序提出了一个内部重组计划，//1
根据该计划它将仍为一个独立公司。//rpw(2,sum(1.2))

The second line in Example 2-(1) is an attributive clause, with "an internal reorganization plan" as its antecedent. In this example, the antecedent is a naming while the attributive clause is its telling. In Example 2-(3), the result of "sum(1.2)" is a generalized term for "一个内部重组计划", namely "该计划"(this plan). The function "rpw(2,sum(1.2))" means replacing "WHICH" in the second line of construct translations with "该计划".

### 4.2 Operation Functions for Transforming Logical Expressions

English and Chinese clause complexes differ in logical expressions in the following 3 aspects: (1) clausal order, (2) the use of logical conjunctions, and (3) naming sharing of logically-related clauses. These differences may give rise to different translation problems, and thus different functions are designed to deal with them.

| Operation Types | Operation Function | Definition |
|---|---|---|
| Substitute | r2n(x,y) | replace the pronoun in x with the corresponding noun in y |
| Substitute | n2r(x,y) | replace the noun in x with the corresponding pronoun in y |
| Add | corcj(x) | add the matched conjunction for the first one in x |
| Add | corcj2(x) | add the matched conjunction for the second one in x |
| Delete | delcj(x) | delete the first conjunction in x |
| Delete | delcj2(x) | delete the second conjunction in x |
| Delete | delpn(x) | delete the relevant pronoun in x |

Table 3. Operation Functions for Transforming Logical Expressions

Firstly, English and Chinese clause complexes have different clausal orders. The differences lie in two aspects: (1) In English, main clauses are usually placed before subordinate clauses, while it is the opposite in Chinese. (2) In English, quotation verbs are placed after or between quotations, while in Chinese, quotation verbs are usually placed before quotations. In the annotation scheme, the operation of reorder is demonstrated by the line numbers referring to clause translations. Sometimes, the reorder of clauses is accompanied with the necessity of changing referential order. The two functions r2n(x,y) and n2r(x,y) are specially designed for dealing with this situation.

**Example 3**: Yields may blip up again before they blip down because of recent rises in short-term interest rates.

**(1) Newline-Indent Schema of English Clause Complex**:
Yields may blip up again
before they blip down
because of recent rises in short-term interest.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1035-1044,
Hainan, China, October 30 - Novermber 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

**(2) Construct Translations**:
收益率可能会再次上升
它们在下降之前
因为最近短期利率上升。

**(3) Whole-Sentence Translation**:
因为最近短期利率上升，//3
收益率在下降之前，//r2n(2,1)
它们可能会再次上升。//n2r(1,2)

In Example 3, the English clausal orders should be adjusted in the Chinese translation. The rearrangement of clausal orders is displayed in Figure 2.

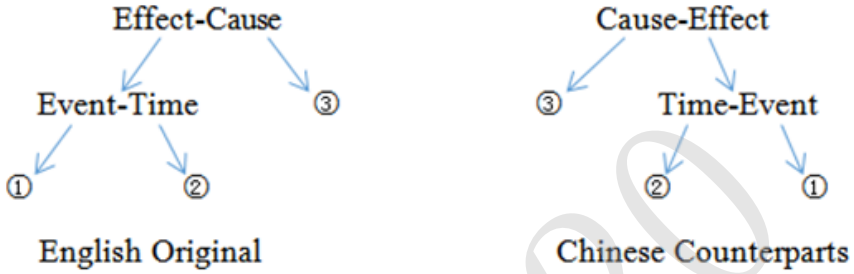


Figure 2. Logical Orders of Clauses in Example 3 and Its Chinese Counterparts

The exchange of clausal orders is demonstrated with the exchange of orders of line numbers. As the first line shown in Example 3-(3), the number "3" at its end means that this line comes from the third line of construct translations. Meanwhile, the interclausal order between the first and second line of construct translations has also been changed in Example 3-(3). The first line of construct translations with the noun "收益率" is placed after the second line with the pronoun "它们". However, in general terms, the line with a pronoun is supposed to appear after the line with the noun it refers to. Hence, it is necessary to exchange the noun and pronoun concerned in the two lines. The function r2n(2,1) means replacing the pronoun "它们" in second line of construct translations with the corresponding noun "收益率" in the first line. The function n2r(1,2) means replacing the noun "收益率" in first line of construct translations with the corresponding noun "它们" in the second line.

However, the whole-sentence translation above is not optimal. A better whole-sentence translation is shown as the following.

因为最近短期利率上升，//3
所以收益率在下降之前可能会再次上升。//corcj(3)+r2n(2,1)+delpn(n2r(1,2))

In this new whole-sentence translation, line 2 and line 3 in the original whole-sentence translation are combined into one by deleting the pronoun "它们". The deletion of the pronoun is tagged as delpn(n2r(1,2)), which means deleting the pronoun in the result of n2r(1,2). With the operation of this function, the result of n2r(1,2), namely "它们可能会再次上升", is turned into "可能会再次上升". Meanwhile, the conjunction "所以" is added, matching that of the third line of construct translations. This addition of a conjunction is tagged as corcj(3).

Example 3 shows relevant functions for dealing with English-Chinese differences on clausal orders and on the use of logical conjunctions.

## 5 Inserted Chinese Words

The inserted Chinese words are function words such as linking verbs, particles, conjunctions and prepositions. The Chinese words can be classified into 2 types based on their functions: (1) words indicative

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1035-1044,
Hainan, China, October 30 - Novermber 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

of argumentative relations between operation units, and (2) words indicative of logical relations between operation units.

| Words Indicative of Argumentative Relations Between Operation Units | |
|---|---|
| 的 | connect a modifier with its modified noun |
| 是 | connect a naming with its descriptive noun phrase |
| 即 | connect a naming with its parallel noun phrase |
| 有 | introduce a naming, functioning like “there be” in English |
| 在 | connect a naming with its telling as a location |
| 将 | inserted before a naming whose telling is about the processing or recognition of the naming |
| 在…中 | a frame in which a naming used as a location is embedded |
| **Words Indicative of Logical Relations Between Operation Units** | |
| 使 | logical component inserted before a telling which is the logical result of a behavior or an event |
| 而 | logical component inserted between two verb phrases which are in the cause-effect relations |
| 就 | logical component inserted before a telling which is the logical result of a behavior or an event |
| 并 | logical component suggesting a further movement |

Table 4. Inserted Chinese Words

Of words indicative of argumentative relations between operation units, the three most frequently used Chinese words are “的”, “是” and “即”. “的” is inserted between an attribute and its modified noun phrase. “是” and “即” are often inserted between a naming and its telling which is usually an appositive in English clause complexes. “是” is applied when the telling is describing the property of its naming, while “即” is applied when the telling has the same reference as its naming. The usage of “的” and “是” has been shown in Example 1. Example 4 is to show the usage of the logical conjunction “并”.

**Example 4**: Mrs. Yeargin concedes that she went over the questions in the earlier class, adding: “I wanted to help all” students.

**(1) Newline-Indent Schema of English Clause Complex**:
Mrs. Yeargin concedes
that she went over the questions in the earlier class,
adding:
“I wanted to help all” students.

**(2) Construct Translations**:
叶尔金太太承认
她在上一届的班里复习了那些问题，
补充说：
“我想帮助所有”学生。

**(3) Whole-Sentence Translation**:
叶尔金太太承认她在上一届的班里复习了那些问题，//1+2
并补充说：//并+3
“我想帮助所有学生。”//4



Proceedings of the 19th China National Conference on Computational Linguistics, pages 1035-1044,
Hainan, China, October 30 - Novermber 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

There are two clauses in this example. One clause is constituted by lines 1 and 2 in Example 4-(1), and the other is constituted by the naming in line 1, i.e. “Mrs. Yeargin”, and the telling, lines 3 and 4. Semantically speaking, the second clause is the continuation of the first one, involving the action to be taken after that of the first clause. In the English clause complex, the logical relation is presented by using an infinite verb for the action in the second clause, namely “adding”, to lower the grammatical hierarchical level of the clause. However, in Chinese, there is no such grammatical device as changing verb forms. Therefore, the logical conjunction “并” is added for connecting the two clauses logically.

## 6 Statistical Data

So far, we have annotated 2108 clause complexes on 136 documents from English Penn Treebank. Of the annotated clause complexes, 336 contain only one clause. Of the clause complexes containing more than one clause, 532 do not involved CC-level structural transformations. Therefore, only a total of 1240 clause complexes are annotated with relevant functions and Chinese words, accounting for 58.82% of the 2108 clause complexes.

| Function | *x | pron(x) | sum(x) | det(x) | delt(x) | delpn(x) |
|---|---|---|---|---|---|---|
| Freq. | 361 | 136 | 103 | 56 | 32 | 23 |
| Function | rpw(x,y) | corcj(x) | r2n(x,y) | paren(x) | n2r(x,y) | delcj(x) |
| Freq. | 20 | 20 | 16 | 14 | 11 | 9 |
| Function | n2v(x) | prd(x) | rel(x) | ndet(x) | corcj2(x) | delcj2(x) |
| Freq. | 8 | 6 | 6 | 4 | 1 | 1 |

Table 5. Frequency of Operation Functions in ECCA Corpus

The frequency of each operation function in ECCA Corpus is shown in Table 5. The number of each inserted Chinese word is also counted. The most frequently used words, “的”, “是” and “即”, appear for 486, 112 and 22 times, respectively. Other inserted Chinese words are used for less than 5 times.

## 7 Conclusions and Discussions

Component sharing relations and logic-semantic relations are organized differently in English and Chinese clause complexes. As a result, during English-Chinese translation, it is necessary to adjust the expressions of these two relations with some structural transformations on the clause complex level. This paper divides English-Chinese clause complex translation into two grammatical levels. On the clause complex level, an English clause complex is parsed into constructs, and the translations of these constructs are assembled into a whole-sentence translation. On the clause level, each construct is translated independently. The two-level translation mechanism, including operation functions and inserted Chinese words used in the assembling step, has been designed formally and proved feasible with corpus manual annotation.

By designing the two-level translation mechanism, this paper follows a common strategy for AI problem solving, namely to decompose a complicated task into sequential simple tasks. It is believed that this mechanism could reduce the demanded data scale and calculation complexity for machine-learning-based machine translation, since the task of translating a sentence is decomposed into simple tasks of translating and assembling shorter constructs. Meanwhile, although the mechanism cannot produce perfect results in some cases, it is an explainable translation process and thus is worth further exploring.

The work present in this paper is only initial. In the future, efforts will be made to enlarge the corpus size, improve the quality of annotated translations, provide multiple translation alternatives, design algorithms for realizing operation functions and discover linguistic knowledge based on the ECCA Corpus.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1035-1044,
Hainan, China, October 30 - Novermber 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

## Acknowledgements

This research is supported by National Natural Science Foundation of China (61672175).

## References

Fang, F., Ge, S., & Song, R. (2016). Error analysis of English-Chinese machine translation. In Sun, M., Huang, W., Lin, H., Liu, Z., & Liu, Y. (eds.), *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data* (pp. 35-49). Gewerbestrasse: Spring.

Ge, S., & Song, R. (2016). The naming sharing structure and its cognitive meaning in Chinese and English. In Xiong, D., Duh, K., Agirre, E., Aranberri, N., & Wang, H. (eds.), *Proceedings of the 2nd Workshop on Semantics-Driven Machine Translation (SedMT 2016)* (pp. 13-21). Stroudsburg: Association for Computational Linguistics (ACL).

Ge, S., & Song, R. (2020). English-Chinese clause alignment corpus tagging system based on corpus annotation. *Journal of Chinese Information Processing, 34*(6), 27-35.

Halliday, M. A., & Matthiessen, C. M. I. M. (2004). *An introduction to functional grammar third edition*. London: Edward Arnold.

Luo, X. (1992). Unit of transfer in translation. *Foreign Language Teaching and Research, 4*, 32-37.

Song, R., & Ge, S. (2015). English-Chinese translation unit and translation model for discourse-based machine translation. *Journal of Chinese Information Processing, 29*(5), 125-136.

Song, R., Ge, S., Chen, X., & Lin, X. (2020). English-Chinese clause alignment corpus annotation guidelines. *Technical Report of Collaborative Innovation Center for Language Research & Service of Guangdong University of Foreign Studies*. Guangzhou.

Wang, L. (2012). *The complete works of Wang Li volum 8: Chinese grammar theory*. Beijing: Zhonghua Book Company.