# Component Sharing in English and Chinese Clause Complex

Shili Ge[1], Xiaoping Lin[2], and Rou Song[1, 3, (✉)]

[1]Laboratory of Language and Artificial Intelligence, Guangdong University of Foreign Studies Guangzhou, China 510420
[2]Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies Guangzhou, China 510420
[3]College of Information Science, Beijing Language and Culture University
Beijing, China 100083

geshili@gdufs.edu.cn, lxpteresa@126.com, songrou@126.com

## Abstract

NT Clause Complex Framework defines a clause complex as a combination of NT clauses through component sharing and logic-semantic relationship. This paper clarifies the existence of component sharing mechanism in both English and Chinese clause complexes, illustrates the differences in component sharing between the two languages, and introduces a formal annotation scheme to represent clause-complex level structural transformations. Under the guidance of the annotation scheme, the English-Chinese Clause Alignment Corpus is built. It is believed that this corpus will aid comparative linguistic studies, translation studies and machine translation studies by providing abundant formal and computable samples for English-Chinese structural transformations on the clause complex level.

## 1 Introduction

Natural language contains five grammatical levels: morpheme, word, phrase/group, clause, clause complex. Elements of higher levels are constructed out of elements of the level next below (Lyons, 1968; Crystal, 1980; Halliday and Matthiessen, 2004). In terms of the structure of clause complex or sentence, the current predominant method for analysis is phrase-based, while the clause-based structural analysis is less discussed. The phrase-based analysis has yielded fruitful results. Nevertheless, clause-based structural analysis is no less important, especially when discourse-level natural language processing is concerned. Song (2013) puts forward the NT Clause Complex Framework, which provides a framework for clause-based analysis of clause complex structures. This paper is to illustrate the concept of component sharing in this framework, and to introduce a new perspective for understanding English-Chinese translation based on component sharing.

## 2 English and Chinese Clause Complex

### 2.1 NT Clause Complex Framework

Clause complex has been discussed by many linguists and researchers, such as Halliday and Matthiessen (2004), Huang and Xiao (1996) and Hu (1990). Yet, NT Clause Complex Framework is quite different from previous discussions on clause complex. It is a theoretical framework about structures of clause complexes. Based on the framework, a clause complex is composed of NT clauses. An NT clause is a combination of a naming and a telling. Naming is the start of an utterance. Telling is defined as the component that predicates or explains the naming.

Component sharing, the mechanism through which NT clauses are combined, is at the core of this framework. Ge and Song (2020) define the concept of component sharing, and puts forward three features to identify it. It is assumed that component sharing exists in both Chinese and English clause complexes. An example and relevant analysis will be given in the next subsection.

### 2.2 Component Sharing in English and Chinese Clause Complex

Example 1 includes two English clause complexes. Both clause complexes are composed of a main clause and an attributive clause. From the perspective of traditional grammar, the two English clause complexes are different in two aspects. Firstly, the attributive clause in Example 1a is restrictive, while the one in Example 1b is non-restrictive. Secondly, the attributive clause in Example 1a is syntactically part of the noun



*Proceedings of the 2nd International Workshop on Discourse Processing*, pages 49–53

phrase, while the one in Example 1b is syntactically a separate clause.

**Example 1:**

a. Yesterday I met a foreigner who could speak very good Chinese.
b. Yesterday I met Mr. Spoon, who was shopping.

For these two clause complexes, there are 3 ways to translate them. The three translated versions are shown as follows.

Version 1 renders both main clauses and attributive clauses as punctuation clauses (a segment of Chinese text separated by commas, semicolons, periods, exclamation marks and question marks).

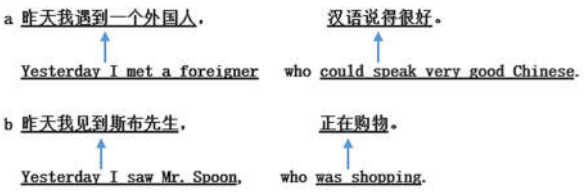


Figure 1. Translated Version 1 of Example 1

Version 2 renders main clauses as punctuation clauses, and supplements attributive clause translations with the translations of their antecedents.


Figure 2. Translated Version 2 of Example 1

Version 3 renders main clauses as punctuation clauses. Meanwhile, the translations of attributive clauses are inserted to the left of the translations of the antecedents. The Chinese auxiliary word "的" is inserted to link them.

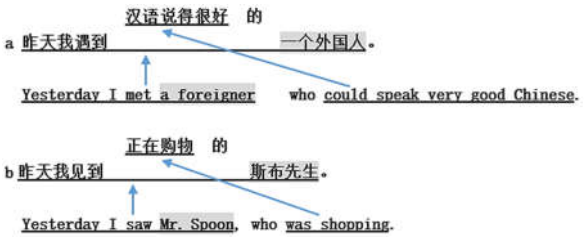


Figure 3. Translated Version 3 of Example 1

In all the three translated versions, the relative words "who" are omitted without translation.

The three translated versions suggest that component sharing exist in both Chinese and English clause complexes. The following is an illustration with the first translated version.

In the first translated version, the first Chinese clauses of both clause complexes are semantically complete, while the second clauses are semantically incomplete with the lack of agents for verbs. However, a Chinese native speaker will intuitionally identify the missing agents as "一个外国人" (a foreigner) and "斯布先生" (Mr. Spoon). Hence, two semantically complete clauses will be formed in mind, namely "这个外国人汉语说得很好" (the foreigner could speak very good Chinese) and "斯布先生正在购物" (Mr. Spoon was shopping). It should be noted that "一个外国人" (a foreigner), an indefinite form, has been changed into its definite form "这个外国人" (the foreigner).

The interclausal relationship in two Chinese translations can be better illustrated with the following newline-indent schemas.

a 昨天我遇到一个外国人，
　　| 汉语说得很好。
b 昨天我见到斯布先生，
　　| 正在购物。

Figure 4. Newline-Indent Schemas of Translated Version 1

Newline-indent schema is used in NT Clause Complex Framework to present the component sharing relationship. In Figure 4, as the punctuation clause "汉语说得很好" (could speak very good Chinese) share a component in the first clause, it is indented to the right of the component. The vertical bar "|" is used to mark the left boundary of the component. The same is true of the schema of Example 1b. The newline-indent schema clearly shows that the two clauses in Example 1a share "一个外国人" (a foreigner), while the two clauses in Example 1b share "斯布先生" (Mr. Spoon).

Such component sharing also exists in the English originals. The following is newline-indent schemas of the English originals.

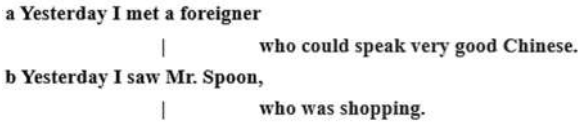


Figure 5. Newline-Indent Schemas of English Originals

Each of the two clause complexes above contains two NT clauses:

a. Yesterday + I met a foreigner
   a foreign + who could speak very good Chinese
b. Yesterday + I met Mr. Spoon
   Mr. Spoon + who was shopping

Direct concatenation of antecedents and attributive clauses does not make syntactically well-formed English clauses. In other words, in each group of NT clauses above, the first NT clause is well-formed while the second isn't. A few mechanical adjustments need to be made to turn them into well-formed clauses. In this example, the relative words "who" need to be omitted and the indefinite noun form should be changed into the definite form. Adjusted versions of the second NT clauses are as follows:

a. the foreigner could speak very good Chinese.
b. Mr. Spoon was shopping.

## 3 Component Sharing Patterns and Clause-Complex Level Structural Transformations

Corpus annotation has revealed that both English and Chinese have 4 types of component sharing patterns, which include branch pattern, graft pattern, postposition pattern and influx pattern. However, the distribution of these patterns is quite different between the two languages. Such differences lead to clause-complex level structural transformations in English-Chinese translation. In this section, the graft pattern, as well as structural transformations on the clause complex level under this pattern, will be introduced.

### 3.1 Graft Pattern

A graft pattern occurs when a non-naming component in a clause is stated by a component after it. In this case, the former is defined as the graft naming, while the latter is the graft telling. The newline-indent schema of the graft pattern is shown in the following.

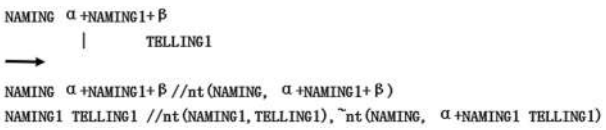


Figure 6. Newline-Indent Schema of Graft Pattern

In the first line of the schema, NAMING represents a naming while α+NAMING1+β represents its telling. The telling α+NAMING1+β can be divided into 3 parts, including α, NAMING1 and β. Under the graft pattern, the NAMING1 component is a must, while the α and β components may be null. In the second line of the schema, TELLING1 is indented to right after NAMING1. The vertical bar "|" indicates the boundary of the graft naming. Instead of stating the NAMING in the first line, TELLING1 chooses a new start of utterance, that is NAMING1. Hence NAMING1 and TELLING1 constitute a graft pattern, with NAMING1 being the graft naming and TELLING1 being the graft telling. The component sharing patterns in Example 1a and 1b are the graft pattern.

### 3.2 Clause-Complex Level Structural Transformations and Annotation Scheme

Ge and Song (2016) point out that Chinese is rich in branch patterns while English is rich in graft patterns. It means that many graft patterns in English cannot be directly converted into Chinese ones. Hence, it is often necessary to carry out

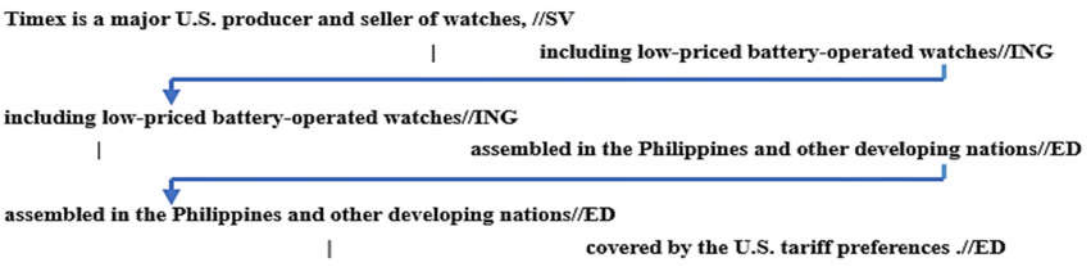


Figure 7. Newline-Indent Schema of Example 2

clause-complex level structural transformations when translating English clause complexes of graft pattern. The following shows an English clause complex of nested graft patterns and illustrates how such a structure is transformed into a Chinese one step by step.

**Example 2:**
Timex is a major U.S. producer and seller of watches, including low-priced battery-operated watches assembled in the Philippines and other developing nations covered by the U.S. tariff preferences.

Due to limitation of the layout, the newline-indent scheme in Figure 7 is an adjusted one. Each line except for the first line is duplicated for the convenience of presenting naming-telling relationship. The clause complex is broken down into four constructs, each taking up a line in the schema. The first construct is a subject-predicate clause and is tagged with SV. The second one is a telling in the form of present participle, with "watches" as its naming. The third and fourth constructs are both tellings in the form of past participle, with "low-priced battery-operated watches" and "other developing nations" as their namings respectively.

The newline-indent schema suggests that the clause complex is made up of the following 4 NT clauses.

(1) Timex + is a major U.S. producer and seller of watches,
(2) watches + including low-priced battery-operated watches
(3) low-priced battery-operated watches + assembled in the Philippines and other developing nations
(4) other developing nations + covered by the U.S. tariff preferences.

Each of the NT clauses above corresponds to a well-formed clause as follows. Some adjustments are made to produce well-formed NT clauses. The curly bracket suggests that the form of the word is revised. The square brackets suggest that words are added.

(1) Timex is a major U.S. producer and seller of watches,
(2) [The] watches {including -> include} low-priced battery-operated watches
(3) [The] low-priced battery-operated watches [are] assembled in the Philippines and other developing nations
(4) [The] other developing nations [are] covered by the U.S. tariff preferences.

To translate this clause complex into Chinese, linear concatenation of the translation of each construct cannot produce a sound whole-sentence translation. It also doesn't work to just reorder each telling before its naming. Instead, clause-complex level structural transformations must be made, including reproducing shared namings in proper forms and then combining them with their tellings.

A formal annotation scheme has been designed to represent the process of structural transformations, and will be illustrated with this example in the following.

An important element of the annotation scheme is the coding of component translations. Generally, each line of translations will be coded with the numbers of the lines they take up. However, if a construct translation contains a naming translation, it will be segmented into

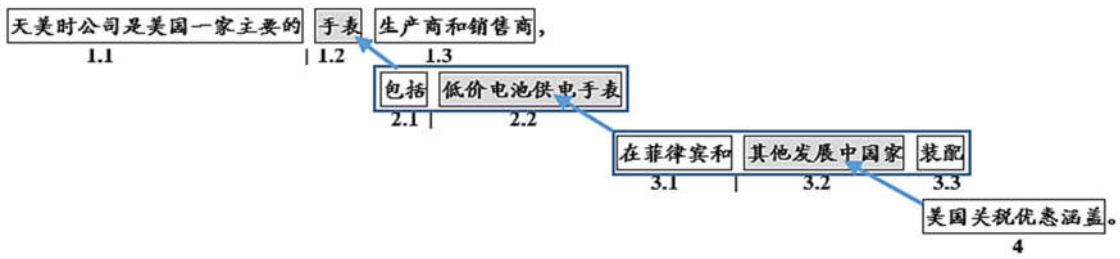


Figure 8. Construct Translations of Example 2

天美时公司是美国一家主要的手表生产商和销售商，//1
所涉手表包括在菲律宾和美国关税优惠涵盖的其他发展中国家装配的低价的电池供电的手表。//det(1.2)+2.1+3.1+4+的+3.2+3.3+的+2.2

Figure 9. Newline-Indent Schema of Whole-Sentence Translation of Example 2

several component translations.

For instance, in Example 2, the second English construct is a telling, and it shares "watch" in the first construct as its naming. As is shown in Figure 8, the translation of "watch", namely "手表", thus segments the translation of the first construct into 3 components, including "天美时公司是美国一家主要的" (Timex is a major U.S), "手表" (watch) and "生产商和销售商" (producer and seller). The three component translations are coded as 1.1, 1.2 and 1.3 respectively. The same is true of coding of component translations in the second and third constructs.

The annotation of structural transformations follows two procedures and is shown in Figure 9. Firstly, assemble construct translations into a whole-sentence translation and display it in newline-indent schema to present the naming-telling relationship. Secondly, formal tags should be added to represent the structural transformations. The whole-sentence translation of Example 2 has two constructs. The first one is the translation of the first English construct. Hence it is tagged as "1". The second is a combination of multiple component translations. In the schema, the numbers such as 1.2 and 2.1 refer to component translations of constructs. The symbol "det(x)" suggests an operation function, which is used to change a noun phrase into its definite form. For example, "det(1.2)" means changing the indefinite form of the component translation 1.2, namely "手表" (watch), into its definite form "所涉手表" (the watch).

## 4 English-Chinese Clause Alignment Corpus

Following the formal annotation scheme, the English-Chinese Clause Alignment Corpus is built. The annotation objects of this corpus include the syntactic categories of English constructs, naming-telling relationship between constructs, translations of constructs, and the processes to combine translation units into whole-sentence translation. Operation functions for transforming features of translation units, as well as inserted Chinese words, are designed. Wall Street Journal newspapers in Penn Treebank are chosen for annotated. So far, about 5000 English clause complexes have been annotated, which comprise about 12000 English NT clauses.

The annotations present component sharing between namings and tellings in English clause complexes, and the clause-complex level structural transformations between English and Chinese. The corpus aims to provide formal and computable material for comparative linguistic studies, translation teaching and machine translation. It is still under construction and will be expanded in the future.

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

## References

Crystal, D. (1980). *A first dictionary of linguistics and phonetics*. Cambridge: Cambridge University Press.

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.

Hu, Z. (1990). Clause and Compound Sentence. In Hu, Z. (ed.), *Language System and Function: Proceedings of 1989 Beijing Systemic-Functional Workshop* (pp. 130-141). Beijing: Beijing University Press.

Huang, G. & Xiao J. (1996). *English complex sentence*. Xiamen: Xiamen University Press.

Lyons, J. (1968). *Introduction to theoretical linguistics*. Cambridge: Cambridge University Press.

Song, R. (2013). Stream model of Generalized Topic Structure in Chinese text. *Studies of the Chinese Language*, (6), 483-494.