On Generalized-Topic-Based Chinese Discourse Structure *

Song Rou[1] Jiang Yuru[2,4] Wang Jingyi[3]
Beijing Language and Culture University[1]
Beijing University of Polytechnic Technology[2]
Beijing Forest University[3]
Beijing University of Information Science and technology[4]

Abstract: Due to the lack of external formal marks, components in Chinese discourse can hardly be categorized into the traditional syntactic system. In fact, Chinese is a typical topic-prominent language, so it should rather be analyzed from the point of topic. This paper, targeting at computer processing, raises the concepts of punctuation clause, generalized topic, discourse structure and topic clause, and reveals the properties of Chinese discourse structure based on generalized topic. The applicability of this theory has been validated in an initial experiment.

Keywords: Punctuation Clause, Generalized Topic, Discourse Structure, Topic Clause.

1. Punctuation Clause, Generalized Topic and Discourse Structure

The traditional study on syntax is based on individual sentences and the formal marks of syntactic components. But due to the lack of external formal marks, the concept of sentence in Chinese is not clear and the boundary of sentences is difficult to be defined. What's more, there are no formal means to discriminate variant types of syntactic structures. Therefore, the traditional parsing often meets difficulty when it comes to Chinese. This paper does not intend to provide a comprehensive analysis of the achievements and deficiency of the work done by the scholars in this field before. The study is rather based on the factual language phenomena in Chinese and oriented to computer processing of the language. In this paper, some concepts, including punctuation clause, generalized topic, discourse structure and topic clause, are defined, and some properties of Chinese discourse structure are raised, and initial verification done in practical application.

The basic unit of a Chinese discourse is punctuation clause (PClause). A PClause is a string of words separated by comma, semicolon, period, exclamation mark, question mark or quotation marks. Since PClauses can be identified with formal marks, and their internal structure and their relations with each other are restrained, therefore the basic conditions of processing them with computers are satisfied.

E.g. 1.1. (Adopted from newspaper news)

①突然，②他听到洗手间有流水声，③警官与特警踢开门，④将洗手间内的人猛地摔倒在地并铐住，⑤经辨认，⑥正是叶成坚。

（①Suddenly，②he heard the sound of water in the washroom.③the police officers and the special policemen kicked the door open，④wrestled the man in the washroom on the floor and handcuffed him，⑤after identifying，⑥was nobody but Ye Chengjian.[1] ）

This is a discourse fragment composed of 6 punctuation clauses.

E.g. 1.2. (Adopted from newspaper news)

①叶成坚对在珠海杀人、诱赌勒索台商游某，②以及在澳门实施的四宗持枪抢劫案的犯罪事实供认不讳，③并将私藏枪支的地点一一指认。

(①Ye Chengjian confessed murders in Zhuhai, seducing and blackmailing a Taiwan businessman named You, ② and the four armed robbery in Macao，③ and identified the places where he illegally hid

* This study is supported by National Natural Science Foundation of China, subject No. 60872121

[1] With a purpose to show the structure of PClauses in Chinese, the translation of Chinese works may not appear very standard in English. The same applies hereinafter.

the guns.)

This is a discourse fragment composed of 3 punctuation clauses.

The discourse structure in Chinese is a kind of syntactic structure of a PClause sequence, which is composed of a generalized topic and a number of comments. Generalized topic refers to a syntactic component of a PClause. The subsequent parts of the punctuation clause after it and the neighbor PClauses may be comments about it. Usually a generalized topic is nominal, functioning as the subject, object or attributive in the clause in traditional grammar. In this case, the comments answer "what" and "how" about the topic. The generalized topic can also be verbal, playing the part of the central component of a verb phrase. In some cases, the generalized topic can even be adverbial or an individual preposition. That's why the word "generalized" is adopted. For sake of simplicity, generalized topic will simply be referred to as topic in later sections.

E.g. 1.3. (Adopted from *A Tale of Old Man Xing and His Dog*, by Zhang Xianliang)

她收起了手中的针线，进到屋里，把炕扫了扫，上炕跪坐在炕头，低着脑袋，两手垂在两膝之间，像一个犯人在审讯室里一样静等着。

(She collected her needlework, went into the house, swept the kang[2], got on it and sat down, lowered her head, let her hands dangle between her knees and waited quietly like a prisoner in the hearing room.)

In this example, each of the seven PClauses has the topic "她(she)" as appears in the first PClause, and make comment about "她(she)", answering the questions about her behavior and what she is like. They compose a discourse structure. The first PClause is composed of one topic and one comment, while the rest have comment only but no topic. This discourse structure can be expressed below.

{她[收起了手中的针线，	{She [collected her needlework,
进到屋里，	went into the house,
把炕扫了扫，	swept the kang,
上炕跪坐在炕头，	got on it and sat down,
低着脑袋，	lowered her head,
两手垂在两膝之间，	let her hand dangle between her knees,
像一个犯人在审讯室里一样静等着。]}	waited quietly like a prisoner in the hearing room.]}

For sake of visual cognition, the PClauses are put in different lines and are indented after the topic that they comment. This way of expression is called indented new-line representation. What is quoted by the "[]"marks is some comments, the left of which is the topic. And what is quoted by the "{}"marks is the discourse structure.

E.g. 1.4. (Adopted from *Fortress Besieged* by Chien Chung-Shu)

{她[{只穿[绯霞色抹胸，	{She [{was wearing only [a scarlet top,
海蓝色贴肉短裤,]}]}	and navy blue, skin-tight shorts,]}]}

These two PClauses both comment on what "她只穿(she was wearing only)". "只穿(was wearing only)" is one topic, and "绯霞色抹胸(a scarlet top)", and "海蓝色贴肉短裤(navy blue, skin-tight shorts)" are two comments, answering the question of what was being worn only. The topic and its two comments, when combined together, constitute a discourse structure, which is in turn the comment of "她(she)", answering the question of what she was like. In other words, this discourse structure and "she" constitute an external discourse structure.

E.g. 1.5. (Adopted from *Fortress Besieged* by Chien Chung-Shu)

{鸿渐{吓得[头颅几乎下缩齐肩，
眉毛上升入发，]}}

[2] a kind of bed in some parts of China

{Hung-chien[{was so horrified that[his forehead nearly shrank into his eyebrows,
(as) his eyebrows rose up to his hairline,]}]}

These two PClauses both comment on the extent of his being horrified. The verbal structure of verb + auxiliary "吓得[3]（was so horrified）" is the topic. The topic and its two comments constitute a discourse structure, which is in turn the comment of "鸿渐(Hung-chien)".

E.g. 1.6. (Adopted from *A Tale of Old Man Xing and His Dog*, by Zhang Xianliang)

{全队三百多口子[{都[张着嘴要吃，伸起手要穿。]}]}
{More than 300 people of the team [{all [need feeding, need clothing.]}]}

The two PClauses comment on what "全队三百多口子(more than 300 people all)". The generalized topic "都(all)" has two comments "张着嘴要吃(need feeding)" and "伸起手要穿(need clothing)". They both answer "all" what. "都(all)" and the two comments constitutes a discourse structure, commenting on what "全队三百多口子(more than 300 people all)" were like. The two form external discourse structure.

E.g. 1.7. (Adopted from Preamble of CONSTITUTION OF THE PEOPLE'S REPUBLIC OF CHINA)

{本宪法[{以法律的形式[确认了中国各族人民奋斗的成果，
规定了国家的根本制度和根本任务，]}
是国家的根本法，
具有最高的法律效力。]}

{This Constitution, [{in legal form, [affirms the achievements of the struggles of the Chinese people of all nationalities,
(and) defines the basic system and basic tasks of the state;]}
(it) is the fundamental law of the state,
(and) has supreme legal authority.]}

The adverbial "以法律的形式(in legal form)" in the first PClause is the generalized topic. The section after it "确认了中国各族人民奋斗的成果(affirms the achievements of the struggles of the Chinese people of all nationalities)" and the second PClause "规定了国家的根本制度和根本任务(defines the basic system and basic tasks of the state)" are its two comments, answering what is done "in legal form". These three constitute a discourse structure. This structure, together with the third and the fourth PClauses, are all comments on the subject of the first PClause "本宪法(this Constitution)" , answering what "本宪法(this Constitution)" is about. These three comments, together with "本宪法(this Constitution)" form the external discourse structure.

E.g. 1.8. (Adopted from *Fortress Besieged* by Chien Chung-Shu)

{学生[{把[分数看得太贱，功课看得太容易]}]}
{The students[{took[grades as too cheap, courses as too easy]}]}

The preposition "把[4]" in the first PClause is the generalized topic. "分数看得太贱(took grades as too cheap)" and "功课看得太容易(took courses as too easy)" comment on what and its result. These three then constitute a discourse structure, making comments on "学生(the students)". They form the external discourse structure.

E.g. 1.9. (Adopted from *Royal Tramp (Lu Ding Ji)* by Louis Cha)

顾炎武在城中买了{一份邸报，
[上面详列明史一案中获罪诸人的姓名。]}

Gu Yanwu bought at the town{a piece of court bulletin,

[3] the word "得" in Chinese is an auxiliary, indicating result.
[4] the word "把" in Chinese is a preposition. It is used in transitive structure, introducing the object.

[(it) listed in detail the names of the criminals accused in the case of Ming Dynasty history.]}

The discourse structures in other examples of this section are embedding, while this example is of overlapping type. The first PClause “顾炎武在城中买了一份邸报(Gu Yanwu bought at the town a piece of court bulletin)” is a discourse structure. The object “一份邸报(a piece of court bulletin)” is not the topic in this PClause, but it is the topic of the second PClause “上面详列明史一案中获罪诸人的姓名(it listed in detail the names of the criminals accused in the case of Ming Dynasty history)” and the two form another discourse structure. The two structures are overlapping, they share one component “一份邸报(a court bulletin)”.

2. The static property of Chinese discourse structure

From the examples in the previous section, we can notice the characteristics of Chinese discourse structure:

(1) A generalized topic and a comment group constitute a discourse structure. A comment group is composed of a number of comments.

(2) A comment can be the part of a PClause that follows the topic, or a whole PClause, or another discourse structure. Therefore, the discourse structure is embedded in a recursive way to the right.

Using Context-Free Grammar, the rules are

① DiscourseStructure→GeneralizedTopic CommentGroup
② CommentGroup→Comment
③ CommentGroup→Comment CommentGroup
④ Comment→PClauseTail
⑤ Comment→PClause
⑥ Comment→DiscourseStructure
⑦ GeneralizedTopic→
⑧ PClauseTail→
⑨ PClause→

Here PClauseTail is the tail of the PClause where the generalized topic appears. In these rules, ①-⑥ are generating rules for discourse structure, comment group and comment respectively. ⑦⑧⑨ are the generating rules for generalized topic, PClause tail and PClauses. The right part of these rules is related to terminal symbols and is not listed here.

Statistics on the corpora show that in genuine Chinese texts, there are a large number of PClauses whose subject is missing. This phenomenon is regarded as zero anaphora or elision in traditional language study. But as a matter of fact, the nature of this phenomenon is that there is more than one comment that corresponds to a topic. Since it is a topic, it is natural that there are a lot of comments. There are pauses between the comments and the result is that several PClauses are formed. Neither is this phenomenon zero anaphora nor ellipses, but topic sharing.

Take 1.8 as an example. The following is its generating process (the numbers following the arrow are rule ID).

DiscourseStructure
→①GeneralizedTopic CommentGroup
→⑦本宪法 CommentGroup
→③本宪法 Comment CommentGroup
→⑥本宪法 DiscourseStructure CommentGroup
→①本宪法 GeneralizedTopic CommentGroup CommentGroup

→⑦本宪法 以法律的形式 CommentGroup CommentGroup

→③本宪法 以法律的形式 Comment CommentGroup CommentGroup

→④本宪法 以法律的形式 PClauseTail CommentGroup CommentGroup

→⑧本宪法 以法律的形式 确认了中国各族人民奋斗的成果，CommentGroup CommentGroup

→②本宪法 以法律的形式 确认了中国各族人民奋斗的成果，Comment CommenrGroup

→⑤本宪法 以法律的形式 确认了中国各族人民奋斗的成果，PClause CommenrGroup

→⑨本宪法 以法律的形式 确认了中国各族人民奋斗的成果，规定了国家的根本制度和根本任务，CommentGroup

→③本宪法 以法律的形式 确认了中国各族人民奋斗的成果，规定了国家的根本制度和根本任务，Comment CommentGroup

→⑤本宪法 以法律的形式 确认了中国各族人民奋斗的成果，规定了国家的根本制度和根本任务，PClause CommentGroup

→⑨本宪法 以法律的形式 确认了中国各族人民奋斗的成果，规定了国家的根本制度和根本任务，是国家的根本法，CommentGroup

→②本宪法 以法律的形式 确认了中国各族人民奋斗的成果，规定了国家的根本制度和根本任务，是国家的根本法，Comment

→⑤本宪法 以法律的形式 确认了中国各族人民奋斗的成果，规定了国家的根本制度和根本任务，是国家的根本法，PClause

→⑨本宪法 以法律的形式 确认了中国各族人民奋斗的成果，规定了国家的根本制度和根本任务，是国家的根本法，具有最高的法律效力。

This nature describes the internal relations of a discourse structure. Therefore it is termed static nature.

This nature can cover most examples in the preceding section except example 1.9. This is because the overlapping type of the discourse structure in the example 1.9 can not be represented by Context-Free Grammar.

3. Dynamic Property of Chinese Topic Clause

3.1. Topic Structure and Topic Clause

In this paper, the structure formed by a comment and its topic is called a topic structure. A topic structure as comment can be combined with an external topic and form an external topic structure. If the topic of a comment is the outmost layer of a discourse structure, it is then called the topic clause. In most cases, every PClause corresponds to a topic clause.

E.g.3.1. (Adopted from the Biology Section of China Encyclopedia)

c_1 澳洲肺鱼产卵期很长，(the spawning season of neoceratodus forsteri is quiet long)

c_2 一般以 9～10 月为旺期。(usually September and October are most productive period)

c_3 卵大，(eggs are big)

c_4 卵径 6～7 毫米，(eggs are 6-7 mm in diameter)

c_5 具胶质膜，(have gelatinous membrane)

c_6 无粘性。(are not sticky)

c_7 卵产于植物中间，(the eggs are laid among plants)

c_8 一部分沉入水底。(some sink deep in the water)

Here, the outmost topic is “澳洲肺鱼(neoceratodus forsteri)”.

The topic clause of c_1 is c_1 itself. The comment is “产卵期很长(the spawning season of is quiet

long)".

c_2"一般以 9～10 月为旺期(usually September and October are most productive period)"is a comment, and its topic is"产卵期(the spawning season)". The topic structure composed of the two is the comment on "澳洲肺鱼(neoceratodus forsteri)"，therefore "澳洲肺鱼产卵期一般以 9～10 月为旺期" is the topic clause of c_2。

c_3"卵大(eggs are big)"is the comment on the topic "澳洲肺鱼(neoceratodus forsteri)". The topic clause of c_3 is "澳洲肺鱼卵大".

c_4"卵径 6～7 毫米(eggs are 6-7 mm in diameter)" is the comment on the topic "澳洲肺鱼(neoceratodus forsteri)". The topic clause of c_4 is "澳洲肺鱼卵径 6～7 毫米"。

c_5"具胶质膜(have gelatinous membrane)"is the comment on the topic "卵(eggs)". The topic structure "卵具胶质膜(eggs have gelatinous membrane)" composed of the two is the comment on "澳洲肺鱼(neoceratodus forsteri)". And the topic clause of c_5 is "澳洲肺鱼卵具胶质膜"。

c_6"无粘性(are not sticky)" is the comment on the topic "卵(eggs)". The topic structure "卵无粘性(eggs are not sticky)" composed of the two is the comment on "澳洲肺鱼(neoceratodus forsteri)". And the topic clause of c_6 is "澳洲肺鱼卵无粘性"。

c_7"卵产于植物中间(the eggs are laid among plants)" is the comment on the topic "澳洲肺鱼(neoceratodus forsteri)". The topic clause of c_7 is "澳洲肺鱼卵产于植物中间"。

c_8"一部分沉入水底(some sink deep in the water)"is the comment on the topic "卵(eggs)"， The topic structure "卵一部分沉入水底(eggs some sink deep in the water)" composed of the two is the comment on the topic "澳洲肺鱼(neoceratodus forsteri)". And the topic clause of c_8 is "澳洲肺鱼卵一部分沉入水底"。

The purpose of analyzing a PClause sequence is to find out its discourse structure. If the topic clause of every PClause is constructed, the topic of each comment at every layer is then found out, and consequently the entire discourse structure will be clear. The next section provides an approach to finding out the topic clause of PClauses.

3.2. Stack Model of Dynamic Generation of the Topic Clause

The topic clause of PClause c_i of Ex.3.1 is marked as c_i'. They are listed below.

c_1'．澳洲肺鱼产卵期很长，(the spawning season of neoceratodus forsteri is quiet long)

c_2'．澳洲肺鱼产卵期一般以 9～10 月为旺期。(the spawning season of neoceratodus forsteri usually September and October are most productive period)

c_3'．澳洲肺鱼卵大，(neoceratodus forsteri’ s eggs are big)

c_4'．澳洲肺鱼卵径 6～7 毫米，(neoceratodus forsteri’ s eggs are 6-7mm in diameter)

c_5'．澳洲肺鱼卵具胶质膜，(neoceratodus forsteri’ s eggs have gelatinous membrane)

c_6'．澳洲肺鱼卵无粘性。(neoceratodus forsteri’ s eggs are not sticky)

c_7'．澳洲肺鱼卵产于植物中间，(neoceratodus forsteri’ s eggs are laid among plants)

c_8'．澳洲肺鱼卵一部分沉入水底。(some eggs of neoceratodus forsteri sink deep in the water)

The generation of each ci′ is exemplified below.

$c_1'=c_1$;

The topic of c_2 is "产卵期(the spawning season)" in c_1'. Delete the part of c_1' right to the topic and replace it with c_2，and we will have c_2';

The topic of c_3 is "澳洲肺鱼(neoceratodus forsteri’)" in c_2'. Delete the part of c_2' right to the topic and replace it with c_3，and we will have c_3'.

The topic of c_4 is "澳洲肺鱼(neoceratodus forsteri’)" in c_3'. Delete the part of c_3' right to the topic and replace it with c_4，and we will have c_4'.

The topic of c_5 is "卵(eggs)" in c_4'. Delete the part of c_4' right to the topic and replace it with c_5，and

we will have c_5'.

The topic of c_6 is "卵(eggs)" in c_5'. Delete the part of c_5' right to the topic and replace it with c_6, and we will have c_6'.

The topic of c_7 is "澳洲肺鱼(neoceratodus forsteri')" in c_6'. Delete the part of c_6' right to the topic and replace it with c_7, and we will have c_7'.

The topic of c_8 is "卵(eggs)" in c_7'. Delete the part of c_7' right to the topic and replace it with c_8, and we will have c_8'.

Generally, given a PClause sequence $\{c_1,\ldots,c_n\}$, if the first PClause is a complete structure of topic-comment, then

(1) the topic clause of the first PClause is the PClause itself;
(2) if the topic of a subsequent PClause is missing, then the topic should be in the topic clause of its previous PClause;
(3) the topic clause of every subsequent PClause can be generated recursively by stack operation.

Note the topic clause of c_i as c_i', and the topic clause of c_{i+1} as c_{i+1}',

（3.1）if the topic of c_{i+1} is missing, and $c_i'=\alpha A\beta$, where A is the topic of c_{i+1}, then $c_{i+1}'=\alpha A c_{i+1}$。

（3.2）if the topic of c_{i+1} is not missing, then $c_{i+1}'=c_{i+1}$.

If we regard the beginning and the end of a topic clause as the bottom and the top of a stack respectively, then the removal and connection of the components in the generation process of topic clause are typical stack operations. Therefore the recursive law of such generation can be called the stack model

The stack model can not only applied to embedded discourse structure, but also some overlapping structures such as instance 1.4. Details are not given here. Our investigation into corpora (about 340,000 Chinese characters) of different registers shows that more than 95% PClauses meet the model.

From the stack model, it can be seen that the key to generate the topic clause of a PClause is to identifying which component of the topic clause of the previous PClause is its topic. This would require to uncover the constraints for forming the discourse structure.

4. Constraints on Discourse Structure

4.1. Acceptability and completeness of Topic Clause

A topic structure is composed of a topic and its comments. Therefore mostly it is acceptable. A topic clause is not only acceptable, but also complete with necessary syntactic and semantic components. Taking advantage of this nature, the filtering of topic-seeking for a PClause can be boiled down to the judgment of the acceptability and completeness of a single clause. For example, the topic clause of PClause 7 in example 3.1 is:

c_7'.澳洲肺鱼卵产于植物中间，(neoceratodus forsteri' s eggs are laid among plants)

and PClause c_8 is

一部分沉入水底。(some sink deep in the water)

According to the stack model, the options for the topic clause of c8 are:

（1） 一部分沉入水底。(some sink deep in the water)
（suppose that the topic of c_8 is not missing）

（2） 澳洲肺鱼一部分沉入水底。(neoceratodus forsteri some sink deep in the water)
（suppose that the topic of c_8 is "澳洲肺鱼(neoceratodus forsteri)"）

（3） 澳洲肺鱼卵一部分沉入水底。(neoceratodus forsteri' eggs some sink deep in the water)

（suppose that the topic of c_8 is “卵(eggs)”）

（4） 澳洲肺鱼卵产于一部分沉入水底。(neoceratodus forsteri’ eggs some are laid sink deep in the water)

（suppose that the topic of c_8 is “产于(be laid)”）

（5） 澳洲肺鱼卵产于植物一部分沉入水底。(neoceratodus forsteri’ eggs some are laid plant sink deep in the water)

（suppose that the topic of c_8 is “植物(plant)”）

（6） 澳洲肺鱼卵产于植物中间一部分沉入水底。(neoceratodus forsteri’ eggs some are laid among plant sink deep in the water)

（suppose that the topic of c_8 is “中间(middle)”）

Chinese intuition tells us that (1) is not complete, and (4)(5)(6) are not acceptable, so the candidates are (2) and (3) only. We see that if we can formalize our intuition, we can considerably narrow down the scope of options.

The topic and the comment of a topic clause are often from different PClauses, and the components in a topic clause that have discourse functions (such as discourse conjunctions) can affect the acceptability of the topic clause. This problem needs to be addressed in separate study.

4.2. Semantic Constraints

E.g. 4.1.他买了一个钱包，是名牌产品。(He bought a wallet, (it) is a brand product.)

The topic of the second PClause could be “他(he)” or “钱包(a wallet)”. We can eliminate the first possibility by using semantic constraints, because a person can not be a product.

4.3. Syntactic Constraints

An investigation into corpora shows that the syntactic relations of the topic and the comments are of the following types:

(1) If the relation of a topic and its comment in the same PClause is subject-predicate, then the same relation is true of it with its comments in other PClauses (see example 1.3);

(2) If the relation of a topic and its comment in the same PClause is predicate-object, preposition-object or attribute-central, then the relation of it and its comment in other PClauses is of the same type or subject-predicate type (see example 1.4 and 1.8).

(3) If the relation of a topic and its comment in the same PClause is adverbial-central or predicate-complement, then its relation with its comment in other PClauses is the same (see example 1.5, 1.6 and 1.7).

(4) If a component is not the topic of the PClause where it is appears, but is the topic of other PClauses, then it must be the object or attribute in the PClause where it appears and its relation with the comments in other PClauses is subject-predicate (see example 1.9).

In addition, adjectives, numbers in partition in respect of quantity and some adverbs (such a adverbs indicating degree) cannot function as general topics.

4.4 Context Constraints

E.g. 4.2.他有个朋友，很阔气。(He has a friend, (who is)very generous with money.)

The topic of the second PClause could be “他(he)” or “个朋友(a friend)”. Whether it is “he is generous with money” or “his friend is generous with money”, it will present no problem either semantically or syntactically. However, abundant instances and analyses show that if

(1) the structure of the topic clause of the previous PClause is SVO;
(2) the core verb of the topic clause of the previous PClause has a sense of “owing” or “introducing”; and
(3) the second PClause is an adjective phrase but does not fall into the category of mental state

then the topic of the second PClause is the object rather than the subject of the topic clause.

According to this constraint, the topic of the second PClause is "个朋友(a friend)"

4.5. Cognition Constraints

Theoretically, there is no limit to the size of a discourse structure. Countless layers could be embedded or overlapped. For example, we could have the following discourse structure.

E.g. 4.3 圆周率整数部分是 3，

小数部分第一位是 1，

后面是 4，

再后面是 1，

再后面是 5，

……

(The circumference ratio's

integer part is 3,

the first number in the fraction part is 1,

followed by 4,

followed by 1,

followed by 5

…)

Here "圆周率(circumference ratio)", "1", "4", "1", "5" all are topics. They could go on with no limit.

But the study on factual corpora have discovered that the maximum layer of embedding or overlapping is 5, and if we shall return from the deeper layers, the maximum number of the layers that can be jumped back is 3. This has much to do with people's cognition ability. The following is an example of 5 layers of embedding and overlapping with 3 layers of maximum return. The underlined words are the generalized topics. The numbers in the brackets to the right of the PClauses indicating the depth of the embedding and overlapping. PClause "行销于外(release them)" reaches the fifth layer in depth, but the next PClause "官府追究之时(when the authorities started investigating)" returns to layer2, retreating 3 layers.

Ex. 4.4. (Adopted from *Royal Tramp* by Louis Cha)

程维藩从杭州坐船到南浔之时，（0）(Cheng Weifan, on the long boat journey from Hangzhou to Nanxun)

反覆推考，（1）(thought things over)

已思得良策，（1）(had come up with a good plan.)

心想这部《明书辑略》流传已久，(1)(thought the book had already been in circulation for some time)

隐瞒是瞒不了的，（2）(It was therefore too late for concealment)

惟有施一个釜底抽薪之计，（2）(the only expedient left was to play a trick)

一面派人前赴各地书铺，(3) (on one hand, send people to go to the bookshops all over the country)

将这部书尽数收购回来销毁，（4） (buy back and then destroy all copies of the book)

一面赶开夜工，（3）(on the other hand, work day and night)

另镌新版，（4）(make a new printing mould)

删除所有讳忌之处，（4）(remove all the offensive bits)

重印新书，（4）(reprint the book)

行销于外。（5）(release them)

官府追究之时，（2）(when the authorities started investigating)

将新版明史拿来一查，（3）(inspect the new edition of Ming History)

发觉吴之荣所告不实，（3）(find Wu's charges to be groundless)

便可消一场横祸了。（2）(can avert a hideous disaster)

5. Initial Application of Discourse Structure based on General Topic

5.1. Discourse Structure in Encyclopedia

The herein discussed Chinese discourse structure based on general topic has been initially applied and tested in the analyses of encyclopedia texts.

The entries in encyclopedia are expository, covering people, places, species, events, devices and terms etc. in various subjects. Because the different aspects of an object must be exposed, the leading role of the topic is very obvious. It frequently occurs that many PClauses are used to comment on one topic, and the comments on different aspects of an object are often presented as embedded or overlapping structures. In order to mine the information of the object described, it is necessary to analyze the governing scope of a topic. In other words, it is necessary to locate the object commented by every PClause. Therefore the discourse structure must be analyzed. Take 3.1 for example, we must be clear about for "what" September and October are the active period, the eggs of "what" are big, the eggs of "what" are 6-7 mm in diameter, "what" has gelatinous membranes and so on.

5.2. An Experiment on Discourse Structure in Encyclopedia

The experiment object of the paper is the entries about various fishes in the biology volume of China Encyclopedia. The objective is the find the topic clause of every Pclause.

There are 224 entries about fishes in this volume, each one with a title, viz. the name of the order, family, genera and species of a fish. The first PClause in the text does not mention the name, but introduces the genera information of it. The name is not necessarily mentioned in later Pclauses. For example,

澳洲肺鱼

Neoceratodus forsteri; Queensland lungfish

角齿鱼目角齿鱼科新角齿鱼属的 1 种(见图澳洲肺鱼外形),是现代肺鱼中最大的种类。体长约 125 厘米,重达 10 千克。体呈长梭形，覆盖大而薄的圆鳞。……

(A member of the family Ceratodontidae and order Ceratodontiformes (see picture of Neoceratodus forsteri). (It) is the biggest extant lungfish species in the world. (Its)Body length (is) about 125 centimeters, (it) weighs as much as 10 kilograms. (Its) Body is elongated, covered with big and thin round scales …)

In the experiment, the entry names (both in Chinese and English) and bracketed information are deleted. But the entry title is added to the left of the first PClause, connected by a "是(is) ". For example, the first PClause of the above example of neoceratodus forsteri is changed into "澳洲肺鱼是角齿鱼目角齿鱼科新角齿鱼属的 1 种，", the rest remains unchanged.

The experiment selected 3999 PClauses of 86 entries as training data, and 577 PCauses of 13 entries as open-test data. The input of the experiment is the topic clause of a PClause and its next PClause, and the output is the topic clause of the second PClause. In other words, the target of the experiment is to decide the topic of the PClause within a limited scope under the scheme of stack model.

For the training data, each PClause is replenished manually into a topic clause, and then the words are segmented. In this way, the training topic clause set G is obtained. The principle of testing is described below. For each tested PClause c and the topic clause d of its proceeding PClause, word segmentation is done separately. String d is cut at different places, the tails are replaced with c every time. Thus a number of candidate topic clauses of c are obtained. Then the similarity reckoning is made about the candidate topic clauses and the topic clauses in G. The one with the maximum similarity is chosen as the result for output.

In order to solve the problem of data sparse in the calculation, semantic generalization is made about related words. The semantic categories employed are：subjects of fishes (e.g. neoceratodus forsteri,

alopias), part (e.g. head, scale, fin), position (e.g. back, abdomen), location (e.g. front, upper), shape (e.g. fusiformis, cylindrical), size (e.g. big, short), color (e.g. red, light blue). environment (e.g. pond, near sea), geographical region(e.g. the Pacific, Huanghai), season(e.g. early spring, autumn), number (e.g. 3, 1-3) etc. Verbs are rarely generalized.

The result of the initial experiment showed the accuracy rate for open test was 78%. If add the title of a text to the beginning of every PClause in the text, 66% accuracy rate can be got as a baseline. The result of the experiment is not high indeed and there is room for improvement. Since the experiment principle was the similarity of the topic clauses, in essence only the stack model and the acceptability of topic clause are used. Semantic constraints, syntactic constraints, context constraints and cognitive constraints are not employed. In addition, word segmentation is not entirely correct, and the semantic generalization is quite rough. 78% accuracy rate of under such rough conditions has initially proved the applicability of the theoretical system.

6. Discussion

This paper employs discourse structure of topic-comment in analyzing Chinese, takes PClauses as the basic discourse unit, and extends the concept of topic to generalized topic. As a result, the properties of Chinese discourse structure are proposed. Investigations into large amount of language data have proved that this theoretical system is natural and applicable to Chinese, which is also backed up by initial experiment. Of course, the theory need to be improved, and the various types of constraints under the theory framework need to be further uncovered. More and detailed study needs to be done along this path.

References

[1] CHEN Ping（1987），Discouse Analysis of Zero anaphora in Chinese，*Zhongguo Yuwen*， No.5,1987.

[2] CHU Chauncey C.（1998），A Discourse Grammar of Mandarin Chinese，Peter Lang Publication Inc. New York.

[3] HUANG He yan, CHEN Zhao xiong（2002），The Hybrid Strategy Processing Approach of Complex Long Sentence，*Journal of Chinese Information Processing*，Vol.16, No.3.

[4] HOU min, SUN Jian-jun（2005），Zero Anaphora in Chinese and How to Process it in Chinese-English MT，*Journal of Chinese Information Processing*，Vol.19, No.3.

[5] HUANG Jian-cuan，SONG Rou（2008）,A Research on the Annotation of Punctuated Clauses，*Frontiers of Content Computing*，Edited by SUN Mao-song and CHEN Qun-xiu， Tsinghua University Press, Beijing.

[6] LI Xing; ZONG Cheng-qing（2006），A Hierarchical Parsing Approach with Punctuation Processing for Long Chinese Sentences，Journal of Chinese Information Processing，Vol.20, No.4.

[7] MAO Qi, LIAN Le-xin, ZHOU Wen-cui, YUAN Chun-fang（2007），Chinese Syntactic Parsing Algorithm Based on Segmentation of Punctuation，*Journal of Chinese Information Processing*，Vol.21, No.2.

[8] SONG Rou（1992），The Delesion of the Fronts of Clauses in Chinese Narratives，*Journal of Chinese Information Processing*，Vol.6, No.3.

[9] SONG Rou（2008），Research on Properties of Syntactic RelationBetween P-Clauses in Modern Chinese，*Chinese Teaching in the World*, No.2, 2008.

[10] SONG Rou, WANG Jingyi（2008），Syntactic Relation Between P-Clauses in Modern Chinese and Annotated Corpus, CCID & Lancaster University Joint Workshop on Corpus Linguistics & Machine Translation Applications, 2008. Beijing.

[11] XING Fu-yi（1997），Chinese Gramma，Northeast Normal University Press，Changchun.

[12] XU Yu-long（2004），Towards a Functional-Pragmatic Model of Discourse Anaphora Resolution，Shaihai Foreign Language Education Press，Shaihai.