The Analysis of Chinese Sentence Semantic Chunk Share Based on HNC Theory

Quan Zhang[1], Chen Wu[1,2], Xiangfeng Wei[1]

[1] The Institute of Acoustics, CAS, Beijing 100080 P.R.C.
[2] The Graduate School, CAS, Beijing 100039 P.R.C.
E-mail:zhq{at}mail.ioa.ac.cn

Abstract. This paper analyses anaphora from the view of concept association vein. It probes into the problem of sentence semantic chunk share based on the sentence category expression of HNC theory. Firstly, the chunks share is divided into two kinds, one is sharing in a sentence where there is chunk extension into sentence in the sentence, and the other is sharing between two sentences. Secondly, the anticipation knowledge of sharing in one sentence is acquired by deduction from the definition of sentence category which chunk extends into sentence. Then, the classification of sharing between two sentences and corpus investigation are presented, the distribution data of the share is counted, and preliminary analysis for the data is put forward. The sentence analysis processing can benefit from the result.

Keywords: anaphora, HNC theory, sentence, sentence category, semantic chunk share.

1 Introduction

In Chinese expression, the sentence often omits its components in the unambiguous precondition. As to the specific sentence, it is ellipsis. But the omitted content usually appears in the context. The phenomena can be classified into anaphora, and be called Zero type anaphora. It constitutes the main content of the anaphora study with the pronoun anaphora and noun anaphora. Many linguistic investigations focus on this field [1]. However from the view of the Hierarchical Network of Concepts (HNC in short) theory, the phenomena can be realized as the share of semantic chunk.

In order to establish the natural language interactive engine which simulates the intelligence of human being's brain, HNC frames a space for language concepts which stresses the concept association vein, which constructs a digitized symbol system for natural language [2]. Furthermore, HNC defines some key notions, such as sentence, and formalizes them for convenience for computer processing. In general, a sentence is one linguistic unit with relatively complete content. HNC deepens this definition, and puts forward the notion of sentence category, which can express the concrete concept association knowledge for a sentence. The main chunks configuration in one sentence category provides the certain restriction for the semantic integrity by the sentence category knowledge.

The expression of HNC semantic chunks in a sentence category [2] is shown as follows:

Expression 1. The expression of HNC semantic chunks in a sentence category.

SC=GBK1+EK+GBKm(m=2-3)
SCR=SC+fKm

***Note**: where, SC is the mathematical expression for the sentence category space, and the SCR is the SC with auxiliary semantic chunks. In the HNC sentence category space, the type and quantity of the main chunk in a sentence depend on the concept association vein of the sentence, and the number of the main chunk is no more than 4; the auxiliary semantic chunks occur in the sentence according to necessity of the sentence meaning, they depend on the concept association vein of the sentence feebly.*

In the view of HNC theory, a natural language sentence express meaning with a definite sentence category, the sentence category is the concept association vein of the sentence. The sentence component unit is the semantic chunk in the sentence category system. The semantic chunk is linguistic unit which constitutes a sentence directly. It is structural concept mass in sentence, and also is the unit which bears the concept association vein of the sentence. In form, the semantic chunk can be a word, a phrase or a sentence. However, some sentence categories need a sentence to serve as their definite main semantic chunks, and it is called chunk extension into sentence (CES in short). Moreover, the CES is just used to indicate those sentence categories which can activate the association of another sentence as their main chunk, and the sentences which instance from those sentence categories. CES asks a sentence to be its main chunks, and it emphasizes the anticipation knowledge that chunk extension into sentence occurs in some sentence categories.

According to the Expression 1, we can deduce the classification of chunk share between two sentences, and the case more than two sentences can be treated as many chunk shares between two sentences. When analyzing the chunk share between sentences, we should get rid of the CES sentence category. In spit of the same in form, the chunk shares in CES sentence and between sentences are provided with different property. The chunk share between sentences is beyond one sentence processing, it belongs to sentence group processing, and the chunk share can't be predicted. While, the chunk share in a CES sentence arises in one sentence, and it belongs to one sentence processing, some chunk share can be predicted by the sentence category knowledge.

In this paper, we will focus on the two type chunk share. This paper includes six parts. This is the first part. In the second part, the CES sentence category is explained, and the third part gives out the result of chunk share in CES sentence. The forth and fifth part discuss the chunk share between sentences, the type of the chunk share between sentence and the distribution data in real corpus about the type are presented separately. At the end, the sixth part sums up the full paper.

2 The CES sentence category

There are more than one verb in a modern Chinese sentence. In order to deal with the verbs, linguistics has done many investigations on it. They put forward the some new notions of sentence[3], such as Liandong(连动句), Jianyu(兼语句) and subject-predicate phrase acting as object. Here are some examples of the special sentences.

Example 1 Liandong sentence
大家扛着锄头跑了。
Dajia kang zhe chutou pao le.
People run away with lifting hoe with hands.

Example 2 Jianyu sentence
这种可能性鞭策我们更加努力。
Zhezhong keneng xing biance women gengjia nuli.
This feasibility encourages us to work hard.

Example 3 sentence with subject-predicate phrase acting as object
我希望大家来
Wo xiwang dajia lai.
I wish everybody come here.

In another hand, from the view of HNC theory[4], the words "鞭策(encourage)"and "希望(wish)" can provide the association for another sentence to act as main chunk in the Example 2 and 3, the examples are CES sentences. But the following sentence, Example 4, although it is a Jianyu sentence in the term of linguistics, it isn't a CES sentence in the terms of HNC as the "有(have)" can't give the prediction knowledge that the chunk need extend into sentence. In this case, HNC thinks that it is composed of two

sentence, and the chunk “朋友(friend)” is regarded as being shared by the two sentences. The Example 1 is same as this one, and the shared chunk is “大家(People)”. Huang[5] explained the Liandong and Jianyu in terms of HNC.

Example 4 another Jianyu sentence

他有个朋友住在杭州。
Ta you ge pengyou zhuzai HnagZhou.
He has a friend who lives in HangZhou.

The investigation has been done on CES sentence in HNC. Xue[6] probed into the problem that the verbal phrase serve as sentence component, and brought forward the range of CES sentence preliminarily. Miao[7] discussed the sentence category knowledge systemically. He made the CES definition more clear based on Xue’s work, and divided the CES into two categories: the one is unconditional, and the other is conditional. The unconditional group includes 8 sentence categories and the conditional group includes 2 sentence categories. The CES can be merged into the sentence knowledge as a kind of transcendent knowledge. The knowledge is provided to computer to process the natural language sentence automatically. These CES sentence categories are Action CES, Information transfer, Decision CES, Consequence Response, Extended Principal and Subordinate Relation, Extended Equivalence Relation, Extended Substitution, Extended Equivalence Substitution, and Common Response, Common Effect. The last two are conditional CES sentence categories.

Example 5 Some example sentences of CES sentence categories

Action CES (the same as Example 2 and Example 6).

Information transfer:
成都十六家企业 倡议 为平抑物价 干 实事。
Chengdu shiliu jia qiye changyi wei pingni wujia gan shishi.
The sixteen corporations in Chengdu propose that they do something to restrain the price.

Decision CES:
西方人 认为，这种安排 是 合理的分工，是 理所当然的现象。
Xifang ren renwei, zhezhong anpai shi heli de fengong, shi lisuodangran de xianxiang.
The western consider that this arrangement is rational and is a kind of phenomena which go without saying.

Consequence Response:
黄山丁感动 哭了。
Huang Dingshan gandong ku le.
Dingshan HUANG cries by sensation.

Extended Principal and Subordinate Relation:
他 带领 公司的科技人员着手研制。
Ta dialing gongsi de keji renyuan zhoushou yanzhi.
He leads the researchers in the corporation to start to develop it.

Extended Equivalence Relation：
中兴与河北省邮管局 将合作 建设保定双频网络。
Zhongxing yu Hebei sheng you guan ju jiang hezuo jianshe Baoding shuangpin wangluo.
Zhongxing Inc. and the post agency of Hebei Province will cooperate to build the bifrequency network in Boding.

Extended Substitution：

陈先生用标准的普通话 代表 我们向主人 道了谢。

Chen xiansheng yong biaozhun de Putonghua daibiao women xiang zhuren dao le xie.

Mr. Chen represent us to thank the host in stand Chinese mandarin.

Extended Equivalence Substitution,

产卵后，雌雄鹤 轮流 孵卵。

Chanluan hou, cixiong he lunliu fu luan.

After laying eggs, the male crane and female crane incubate the eggs alternatively.

Common Response (the same as Example 3).

Common Effect (the same as Example 7).

3 The Chunk Share in CES Sentence

The CES sentence category is introduced above. In the previous investigation on CES sentence category, they deal with the sentence extension chunk just as one chunk, and don't think of its property as sentence. In fact, the extension chunk is also a sentence. From the view of computer processing sentence, it also needs to be processing as sentence. However, it is an important component in the association vein of the CES sentence. So, it is necessary to take into account the CES sentence and the chunk extending sentence together. We investigate the chunk share in the CES sentence category, which need to process the chunk extending sentence as a sentence. The chunk share in CES sentence categories is a kind of transcendent knowledge, which is benefited to the computer processing sentence. In the convenience of narrating in this paper, we us the Ep to represent the CES sentence and Er for the chunk extending sentence. They are both in on sentence in the terms of HNC. The following focuses on the chunk share between Ep and Er sentence.

We will use the Action CES sentence and Common Effect to illuminate the shared chunk and the unshared chunk separately.

- **Action CES sentence**

The concrete sentence category expression is X03J = X03A + X03 + X03BC, where X03BC is the CES. Some verbs with the meaning of impelling, forcing and so on often arouse this kind sentence. X03BC is a sentence; the subject of the sentence is the shared chunk between Ep and Er. In fact, this case belongs to Jianyu sentence. The example is shown below.

Example 6

我强迫 自己 不去看她。

Wo qiangpo ziji bu qu kan ta.

I force myself not to look at her.

In the instance, the shared chunk is underlined.

- **Common Effect sentence**

It is a conditional CES sentence category. The concrete expression is Y0J = YB + Y + YC, when the verb which serves as the kernel of the eigen chunk Y bears the meaning of the information appearance, the YC chunk extends into sentence. There isn't any shared chunk between the Ep and Er. The YC is the information which appears about YB. It seems that there may be a shared chunk between Ep and Er. But this share is connotative in the sentence. YC is new information which is caused by YB. Then there isn't any shared chunk between Ep and Er in the sentence category.

Example 7

表面的冰逐渐融化 显露出 岩石 继续落入 大气层。

Biaomian de bing zhujian ronghu xianlu chu yanshi jixu luo ru daqiceng.

The surface ice thaws gradually, the rock appears, they fall into aerosphere sequentially.

In the CES sentence categories, there are six ones with shared chunk, and 4 ones with unshared chunk.

The sentence categories with shared chunk are: Action CES, Consequence Response, Extended Principal and Subordinate Relation, Extended Equivalence Relation, Extended Substitution, Extended Equivalence Substitution. The sentence categories with unshared chunk are: Information transfer, Decision CES, and Common Response, Common Effect.

4 The Chunk share between two sentences

Here we focus on the chunk share between sentences. There are two kinds chunk in Expression 1, the main chunk (GBK and EK) and the auxiliary chunk (fK). The tow kind of chunk are both can be shared between sentences. Then, the chunk share between sentences can be classified into two types, the main chunk share and the auxiliary chunk share. The main chunk share should include two types, generalized object chunk (GBK in short) share and the eigen chunk (EK in short) share theoretically. However, the eigen chunk bears the concept association vein. If the eigen chunk is elided, the sentence will hard be understood. So, the eigen chunk is shared rarely. And we do not take account of the eigen chunk share, concentrate only on the GBK chunk share. As the chunk with its own construction, it is necessary to think over the chunk share from another view, the whole and part. Therefore, there are four types of the chunk share. More detail is shown in Table 1.

Table 1. The basic types of chunk share.

	Whole share	Part share
Main chunk	Type I	Type II
Auxiliary chunk	Type III	Type IV

Further analysis of Table 1 is still necessary. In HNC, the main chunk and auxiliary chunk are different two kind of chunk. The main chunk and auxiliary chunk can't share whole each other. Then, the chunk share is main chunk with main chunk and auxiliary chunk with auxiliary wholly only. The part chunk share is more complex, it occurs when the chunk is composed of complex structure. As the chunk with complex structure, the portion of chunk (or whole chunk) is shared by other sentence wholly or partly. And just part of chunk is shared, it does not involve in the property of the whole chunk. So, the main chunk and auxiliary chunk can share partly across. This is just the possibility. The language is not the realization of the possibility only, and many things of the possibility do not appear in the real language. We simplify the part chunk share according corpus. By the simplification, the Type II is used to refer to the part share just main chunk, not take account of the cross share between the main chunk and auxiliary. The Type IV refers to the share from auxiliary chunk part to a whole main chunk. In brief, we take account of the whole chunk share for the auxiliary chunk only.

In the main chunk share, except the part auxiliary chunk shares as the main chunk, there is a problem about the chunk transformation when main chunk share takes place, i.e. the GBKm in the first sentence is shared as GBKn in the second sentence in the Expression 1. Here the GBKm may be same as the GBKn, and may be different. The Table 2 presents the detail about this aspect. The number of the in generalized object chunk the sentence is not more than three.

Table 2. The basic types for the generalized object chunk.

		The sharing sentence		
		GBK1	GBK2	GBK3
The source sentence	GBK1	11	12	13
	GBK2	21	22	23
	GBK3	31	32	33

In the Table 2, the source sentence refers to the sentence which provides the main chunk, and the sharing sentence is the sentence which shares the main chunk. The Table 2 lists the code of the main chunk share. In addition, it is not necessary that the main chunk number of the source sentence equals to the number of sharing sentence.

Summarizing the uppers, we analyze the chunk share between sentences, present the classification of the chunk share, and simplify it according the real Chinese. The following investigation is based on this.

5 The distribution of chunk share between sentences

The Table 3 presents the distribution data for each chunk sharing types about real Chinese essays which is picked from the corpus. The examples are the instances for the chunk sharing types.

Table 3. The distribution data for the each chunk sharing types.

	Type I				Type II	Type III	Type IV
	11	21（31）	12	22			
C1	30	2	0	0	4	8	0
C2	231	23	1	1	11	19	6
C3	488	15	1	0	7	30	1
Total	749	40	2	1	22	57	7

***Note**: the total of sentences in the all essay is 1807. C1, C2, C3 refer to the three essays from the corpus.*

Example 8 Tpye I subtype 11

他 || 拥有了 || 自己的公司——特斯拉电气公司，$ 0 || 并争取获得 || 以交流电为基础的 新电气技术 的专利。

Ta yongyou le ziji de gongsi -- tesila dianqi gongsi, bing zhengqu huode yi jiaoliudian wei jichu de xin dianqi jishu de zhuanli.

He has one's own companies --Tesla belongs to electric company, and tries to obtain the patent taking alternating current as new electric technology of the foundation.

Note: "||" separates the chunks, "$" separates the sentences. The source chunk is underlined, "0" is used for the sharing chunk position. In this instance, "他(He)" , the GBK1 in the first sentence, is shared as GBK1 in the second sentence. The following is same.

Example 9 Tpye I subtype 21

特区的高速发展 || 带动了 || 全国的对外开放， $ 0 || 形成了 || 全面开放的新格局， $ 0 || 有力地促进了 || 改革和现代化建设事业。

Tequ de gaosu fazhan daidong le quanguo de duiwai kaifang, xingcheng le quanmian kaifang de xin geju, youli de cujin le gaige he xiandaihua jianshe shiye.

The fast growth of the country's special economic zones spurs the nationwide opening-up, which has stimulated the country's reform and modernization drive.

Example 10 Tpye II

他 || 化名 || 邓斌，$ 0 || 任 || 中共广西前敌委员会书记， $ 0 同张云逸等 ||~ 于１２月 ~|| 发动 || 百色起义.

Ta hua ming Deng Bin, ren zong gong Guangxi qian di weiyuanhui shuji, tong Zhang Yunyi deng yu shier yue fadong Baise qiyi.

Using the name Deng Bin, he served as secretary of the Guangxi Front Committee of the CPC. Along with Zhang Yunyi and others he staged the Baise Uprising in December.

Note: "~" indicates the auxiliary chunk. In the third sentence, "他(He)" in the frontal sentence serves as a part of GBK1 chunk.

Example 11 Tpye IV

刘邓大军 进入 大别山地区后 ~||~ ，$ 对国民党在长江以南的广大统治区 ~|| 0 || 形成了 || 直接威胁.

Liu Deng dajun jinru Dabieshan diqu hou, dui Guomindang zai Changjiang yi nan de guangda tongzhi qu xingcheng le zhije weixie.

After entering the Dabie Mountain area, the troops of Liu and Deng constituted a direct threat to the vast Kuomintang-ruled areas south of the Yangtze River.

We can draw some conclusions from the Table 3:

1. There are a lot of main chunk share in Chinese sentences.
2. In whole main chunk share, the subtype 11 is more common. The subtype 21 (or 31) appears sometimes, and the other subtypes are met seldom.
3. The part share of main chunk appears in the essays, but not common. In the essays, all the part share of main chunk belongs to part-to-whole type, i.e. the part of a main chunk in a sentence is shared as a whole main chunk in another sentence, and vice versa. There is not any part-to-part chunk share.
4. Based on the statistical data, the main chunk share is much more common than the auxiliary chunk share. The reason which causes this result is that the amount of the auxiliary is much less than main chunk, even in some sentences, there isn't any auxiliary chunk. In the investigated essays, there are some condition auxiliary chunks about time and place which are shared among sentence group, which provide the narration background for the sentence group.
5. In part auxiliary chunk share, data which the part of auxiliary chunk is share as the whole main chunk is presented, and it is took account of only.

In this part, we examine the validity of the classification for the chunk share with the concrete Chinese sentence, and present some typical example sentences and the statistical data. By the result, the classification well represents the share of Chinese sentence chunk, and it is suitable to serve the Chinese sentence analysis.

6 Conclusion

According to the definition of sentence based on HNC, we investigate the chunk share in CES sentence and between sentences. The anticipated knowledge of chunk share in the CES sentence is presented for each CES sentence category which the HNC put forward. Then, we focus on the chunk share between sentences. We classify the chunk share between sentences by deduction, and simplify the classification in the light of the occurrence of the real Chinese sentences. We also test the validity of the classification, and count the data for each type of the chunk share between sentences. Moreover, we explain the data preliminarily. We put forward some results for zero type anaphora from the view of concept association vein.

Meanwhile, there are also some works which need be further studied yet. Firstly, we should consult more linguistic anaphora investigation to deepen the chunk share study. Secondly, we should probe into the recovering strategy for the sharing content based on this paper.

Acknowledgments. The paper is supported by the National Fundamental Research Project (973 Project) (Grant No. 2004CB318104) and The Knowledge Innovation Engineering Project of The Institute of Acoustics, CAS (Grant No. 13CX04).

References

1. Jiujiu Xu: The study of modern Chinese context anaphora. Chinese Social Science Press, Beijing (2003)
2. Zengyang Huang: The Fundamental theorem and mathematic physics expression of the language concept space. Ocean press, Beijing (2004)
3. Yushu Hu: Modern Chinese. Shanghai Education Press, Shanghai (1995)
4. Zengyang Huang: The Hierarchical Network of Concepts theory. Tsinghua University Press, Beijing (1998)
5. Zengyang Huang: Thesis 17: The Liandong Sentence and Jianyu Sentence. http://www.hncnlp.com(1998)
6. Kan Xue: The investigation of Sentence degradation into chunk and Chunk extension into sentence in modern Chinese. Master degree dissertation, China Renmin University, Beijing (1999)
7. Chuanjiang Miao: The studies on the knowledge of sentence category in HNC theory. doctor degree dissertation, The Institute of Acoustics, CAS, Beijing (2001)
8. Quan Zhang: The Analysis of Chinese Verbs with Chunk Extension into Sentence. In: The collection of the 7th Chinese Lexical Semantics Workshop, Xinzhu (2006)