Annotating Discourse Connectives in the Chinese Treebank *

Nianwen Xue
Department of Computer and Information Science
University of Pennsylvania
xueniwen@linc.cis.upenn.edu

Abstract

In this paper we examine the issues that arise from the annotation of the discourse connectives for the Chinese Discourse Treebank Project. This project is based on the same principles as the PDTB, a project that annotates the English discourse connectives in the Penn Treebank. The paper begins by outlining range of discourse connectives under consideration in this project and examines the distribution of the explicit discourse connectives. We then examine the types of syntactic units that can be arguments to the discourse connectives. We show that one of the most challenging issues in this type of discourse annotation is determining the textual spans of the arguments and this is partly due to the hierarchical nature of discourse relations. Finally, we discuss sense discrimination of the discourse connectives, which involves separating discourse connective from non-discourse connective senses and teasing apart the different discourse connective senses, and discourse connective variation, the use of different connectives to represent the same discourse relation.

* I thank Aravind Johi and Martha Palmer for their comments. All errors are my own, of course.

1 Introduction

The goal of the Chinese Discourse Treebank (CDTB) Project is to add a layer of discourse annotation to the Penn Chinese Treebank (Xue et al., To appear), the bulk of which has also been annotated with predicate-argument structures. This project is focused on discourse connectives, which include *explicit connectives* such as subordinate and coordinate conjunctions, discourse adverbials, as well as *implicit discourse connectives* that are inferable from neighboring sentences. Like the Penn English Discourse Treebank (Miltsakaki et al., 2004a; Miltsakaki et al., 2004b), the CDTB project adopts the general idea presented in (Webber and Joshi, 1998; Webber et al., 1999; Webber et al., 2003) where discourse connectives are considered to be predicates that take abstract objects such as propositions, events and situations as their arguments. This approach departs from the previous approaches to discourse analysis such as the Rhetorical Structure Theory (Mann and Thompson, 1988; Carlson et al., 2003) in that it does not start from a predefined inventory of abstract discourse relations. Instead, all discourse relations are lexically grounded and anchored by a discourse connective. The discourse relations so defined can be *structural* or *anaphoric*. Structural discourse relations, generally anchored by subordinate and coordinate conjunctions, hold locally between two adjacent units of discourse (such as clauses). In contrast, anaphoric discourse relations are generally anchored by discourse adverbials and only one argument can be identified structurally in the local context while the other can only be de-

Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, pages 84–91,
Ann Arbor, June 2005.

rived anaphorically in the previous discourse. An advantage of this approach to discourse analysis is that discourse relations can be built up incrementally in a bottom-up manner and this advantage is magnified in large-scale annotation projects where inter-annotator agreement is crucial and has been verified in the construction of the Penn English Discourse Treebank (Miltsakaki et al., 2004a). This approach closely parallels the annotation of the the verbs in the English and Chinese Propbanks (Palmer et al., 2005; Xue and Palmer, 2003), where verbs are the anchors of predicate-argument structures. The difference is that the extents of the arguments to discourse connectives are far less certain, while the arity of the predcates is fixed for the discourse connectives.

This paper outlines the issues that arise from the annotation of Chinese discourse connectives, with an initial focus on explicit discourse connectives. Section 2 gives an overview of the different kinds of discourse connectives that we plan to annotate for the CDTB Project. Section 3 surveys the distribution of the discourse connectives and Section 4 describes the kinds of discourse units that can be arguments to the discourse connectives. Section 5 specifies the scope of the arguments of discourse relations and describes what should be included in or excluded from the text span of the arguments. Sections 6 and 7 describes the need for a mechanism to address sense disambiguation and discourse connective variation, drawing evidence from examples of explicit discourse connectives. Finally, Section 8 concludes this paper.

2 Overview of Chinese Discourse Connectives

With our theoretical disposition, a discourse connective is viewed as a predicate taking two abstract objects such as propositions, events, or situations as its arguments. A discourse connective can be either explicit or implicit. An explicit discourse connective is realized in the form of one lexical item or several lexical items while an implicit discourse connective must be inferred between adjacent discourse units. Typical explicit discourse connectives are subordinate and coordinate conjunctions as well as discourse adverbials. While the arguments for subordinate and coordinate conjunctions are generally local, the first argument for a discourse adverbial may need to be identified long-distance in the previous discourse.

2.1 Subordinate conjunctions

There are two types of subordinate conjunctions in Chinese, single and paired. With single subordinate conjunctions, the subordinate conjunction introduces the subordinate clause, as in (1). By convention, the subordinate clause is labeled *ARG1* and the main clause is labeled *ARG2*. The subordinate conjunction is NOT included as part of the argument. The subordinate clause generally precedes the main clause in Chinese, but occasionally it can also follow the main clause. The assignment of the argument labels to the discourse units is independent of their syntactic distributions. The subordinate clause is always labeled *ARG1* whether it precedes or follows the main clause.

Simple subordinate conjunctions: Simple subordinate conjunctions are very much like English where the subordinate clause is introduced by a subordinate conjunction:

(1) 报告 认为 ，[*conn* 如果] [*arg1* 经济 和
report believe , **if** economic and
金融 政策 得力]，[*arg2* 亚洲地区 经济
financial policy effective , Asia region economy
可望 在１９９９年开始 回升]。
expect in 1999 begin recover .

”The report believes that if the economic and financial policies are effective, Asian economy is expected to recover in 1999.”

Paired subordinate conjunctions: Chinese also abounds in paired subordinate conjunctions, where the subordinate conjunction introduces the subordinate clause and another discourse connective introduces the main clause, as in (2). In this case, the discourse connectives are considered to be paired and jointly anchor ONE discourse relation.

(2) [*conn* 如果] [*arg1* 改革 措施 不 得力 ，
if reform measure not effective ,
信心 危机 依然存在]，[*conn* 那么] [*arg2*
confidence crisis still exist , **then**
投资者 就 有 可能 把 注意力 转向其他
investor will have possibility BA attention turn other
新兴 市场]。
emerging market .

”If the reform measures are not effective, confidence crisis still exists, then investors is likely to turn their attention to other emerging markets.”

Modified discourse connectives: Like English, some subordinate conjunctions can be modified by an adverb, as illustrated in (3). Note that the subordinate conjunction is in clause-medial position. When this happens, the first argument, ARG1 in this case, becomes discontinuous. Both portions of the argument, the one that comes before the subordinate conjunction and the one after, are considered to be part of the same argument.

(3) [*arg1* 去年 初 浦东 新区 诞生 的
last year beginning Pudong new district open DE
中国 第一家 医疗 机构 药品采购 服务
China first CL medical institution drug purchase service
中心]，[*conn* 正 因为] [*arg1* 一 开始 就
center , **just because** once begin
比较 规范]，[arg2 运转 至今 ，成交
relatively standardized , operate till now , trade
药品 一亿多 元 ，没有发现一 例
medicine over 100 million yuan , not find one case
回扣]。
killback .

"It is because its operations are standardized that the first purchase service center for medical institutions in China opened in the new district of Pudong in the beginning of last year has not found a single case of kickback after it has traded 100 million yuan worth of medicine in its operation till now."

Conjoined discourse connectives: The subordinate conjunctions can be conjoined in Chinese so that there are two subordinate clauses each having one instance of the same subordinate conjunction. In this case, there is still one discourse relation, but ARG1 is the conjunction of the two subordinate clauses. This is in contrast with English, where only one subordinate conjunction is possible and ARG1 is linked with a coordinate conjunction, as illustrated in the English translation.

(4) [conn 虽然] [arg1 黄春明 已经
although Huang Chunming already
十几 年 没有出版 小说集 了]，[conn
over 10 year not publish novel series AS ,
虽然] [*arg2* 从 〈城仔 落 车 〉到〈
although from " city boys miss bus " to "
售票口 〉，中间 隔 了 三十七 年]，
ticket box ", middle span AS thirty seven year ,
[conn 但] [arg2 黄春明 的 文学 内在 ，
but Huang Chunming DE literary theme ,
有些 东西 竟然 从来都 没有改变]。
some thing surprisingly ever have not change .

"Although Huang Chunming has not published a novel series for over ten years, and it spans over thirty seven years from 'City Boys Missed Bus' to 'Ticket Box', surprisingly some things in Huang Chunming's literary themes have never changed."

2.2 Coordinate conjunctions

The second type of explicit discourse connectives we annotate are coordinate discourse conjunctions. The arguments of coordinate conjunctions are annotated in the order in which they appear. The argument that appears first is labeled ARG1 and the argument that appears next is marked ARG2. The coordinate conjunctions themselves, like subordinate conjunctions, are excluded from the arguments.

(5) 近年 来，美国 每 年 糖尿病
recent years in , the U.S. every year diabetes
医疗费 约 一百亿 美元 ，印度 去年
medical expense about 10 billion dollar , India last year
糖尿病 医疗费 为
diabetes medical expenses be
六点一亿 美元 ，[*arg1* 中国 尚
six hundred and 10 million dollar , China yet
无 具体 统计]，[*conn* 但] [*arg2* 中国
not have concrete statistics , **but** China
糖尿病 人数 正 以 每 年 七十五万
diabetes population currently with every year 750,000
新 患者 的 速度 递增]。
new patient DE speed increase .

"In recent years, the medical expenses for diabetes patients in the U.S. is about 10 billion dollars. Last year the medical expenses for diabetes patients in India is six hundred and ten million dollars. China does not have concrete statistics yet, but its diabetes population is increasing at a pace of 750,000 new patients per year.

Paired coordinate conjunctions: Like subordinate conjunctions, coordinate conjunctions can also be paired, as in (6):

(6) 现代 父母 难 为的 地方 在于 [*conn* 既
modern parent difficult be DE place lie in **CONN**
] [*arg1* 无法 排除 血液 中流传的 观念]，
no way eliminate blood in flow DE tradition ,
[*conn* 又] [*arg2* 要 面对新 的 价值]。
CONN need face new DE value .

"The difficulty of being modern parents lies in the fact they can not get rid of the traditional values flowing in their blood, and they also need to face new values."

2.3 Adverbial connectives

The third type of explicit discourse connectives we annotate are discourse adverbials. A discourse adverbial differs from other adverbs in that they require an antecedent that is a proposition or a set of related propositions. Generally, the second argument is adjacent to the discourse adverbial while the first argument may be long-distance. By convention, the second argument that is adjacent to the discourse connective is labeled ARG2 and the other argument is

marked as ARG1. Note that in (7b) that first argument is not adjacent to the discourse adverbial.

(7) a. 美国 商会 广东
The U.S. Chamber of Commerce Guangdong
分会 会长 康永华 律师 说 ,
Chapter Chairman Kang Yonghua lawyer say ,
[*arg1* 克林顿 政府 已经 表示 要
Clinton Administration already indicate will
延长 中国 的 贸易 最惠国待遇] , [*conn*
renew China DE trade MFN status ,
因此] , [*arg2* 这 次 游说 的 重点 是
therefore , this time lobby DE focus be
那些 较 保守 的 议员] 。
those relatively conservative DE congressman .

"Lawyer Kang Yonghua, chairman of the Guangdong Chapter of the U.S. Chamber of Commerce, says that since the Clinton Administration has already indicated that it will renew China's MFN status, the focus of the lobby this time is on those relatively conservative congressmen."

b. [*arg1* 中国 批准 的 外企 中 ,
China approve DE foreign enterprise in ,
工业 项目 占 七成
industry project account for seventy percent,
, 其中 加工 工业偏 多
among them processing industry excessive
], 这 与 中国 劳动力 素质 、成本
, this with China labor force training , cost
较 低 的 国情 相吻合 ,
relatively low DE state of affairs consistent ,
[*conn* 从而] [*arg2* 吸纳 了 大量
therefore absorb ASP big volume
劳动力] 。
labor force .

"In the foreign enterprises that China approved of, industry projects accounts for seventy percent of them. Among them processing projects are excessively high. This is consistent with the current state of affairs in China where the training and cost of the labor force is low. Therefore they absorbed a large portion of the labor force."

2.4 Implicit discourse connectives

In addition to the explicit discourse connectives, there are also implicit discourse connectives that must be inferred from adjacent propositions. The arguments for implicit discourse connectives are marked in the order in which they occur, with the argument that occurs first marked as ARG1 and the other argument marked as ARG2. By convention a punctuation mark is reserved as the place-holder for the discourse connective. Where possible, the annotator is asked to provide an explicit discourse connective to characterize the type of discourse relation. In (8), for example, a coordinate conjunction 而"while" can be used in the place of the implicit discourse connective.

(8) [*arg1* 其中 出口 为 一百七十八点三亿美元
among them export be 17.83 billion dollar
, 比 去年 同 期 下降
, compared with last year same period decrease
百分之一点三] [**conn=**而 ;] [*arg2* 进口
1.3 percent ; import
一百八十二点七亿美元 , 增长
18.27 billion dollar , increase
百分之三十四点一] 。
34.1 percent .

"Among them, export is 17.83 billion, an 1.3 percent increase over the same period last year. Meanwhile, import is 18.27 billion, which is a 34.1 percent increase."

3 Where are the discourse connectives?

In Chinese, discourse connectives are generally clause-initial or clause-medial, although localizers are clause-final and can be used as discourse connective by themselves or together with a preposition. Subordinate conjunctions, coordinate conjunctions and discourse adverbial can all occur in clause-initial as well as clause-medial positions. The distribution of the discourse connectives is not uniform, and varies from discourse connective to discourse connective. Some discourse connectives alternate between clause-initial and clause-medial positions. The examples in (9) show that 尽管"even though", which forms a paired connective with 但是"but", occurs in both clause-initial (9a) and clause-medial (9b) positions.

(9) a. [*conn* 尽管] [*arg1* 亚洲 一些 国家 的
even though Asia some country DE
金融 动荡 会 使 这些 国家 的
financial turmoil will make these country DE
经济 增长 受到 严重 影响] ,
economy growth experience serious impact ,
[*conn* 但] [*arg2* 就整 个 世界 经济 而言
but to whole CL world economy
, 其他 国家 的 强劲 增长 势头 会
, other country DE strong growth momentum will
弥补 这 一 损失] 。
compensate this one loss .

"Even though the financial turmoil in some Asian countries will affect the economic growth of these countries, as far as the economy of the whole world is concerned, the strong economic growth of other countries will make up for this loss."

b. [*arg1* 展望 虎年 , 中国 的
look ahead Year of Tiger , China DE
经济 列车] [*conn* 尽管] [*arg1* 会
economy train **even though** will

有 颠簸 起伏]，[*conn* 但] [*arg2* 只要
have ups and downs , **but** as long as
调控 措施 适时 、得当 ，相信 会 沿着
adjust measure timely , proper , believe will along
预设 的 轨道 稳健 前行]。
expect DE track steady advance .

"Looking ahead at the Year of Tiger, even though China's economic train will have its ups and downs, as long as the adjusting measures are timely and proper, we believe that it will advance steadily along the expected track."

Localizers are a class of words that occur after clauses or noun phrases to denote temporal or spatial discourse relations. They can introduce a subordinate clause by themselves or together with a preposition. While the preposition is optional, the localizer is not. When both the preposition and the localizer occur, they form a paired discourse connective anchoring a discourse relation. Example (10) shows the preposition 当and the localizer 时form a paired discourse connective equivalent to the English subordinate conjunction "when".

(10) 日前 ，[*conn* 当] [*arg1* 记者 在这里
a few days ago , **when** reporter at here
专访 欧盟欧洲 委员会 驻华
interview exclusively EU Europe Commission to China
代表团 团长魏根深 大使 ，请 他
delegation head Wei Genshen ambassador , ask he
评价 这 一 年 来 双方 的 合作
comment this one year since two sides DE cooperation
成果] [**conn** 时]，[*arg2* 他 毫 不
accomplishment **when** , he little no
迟疑 地 说 ： "欧盟同 中国 的 政治
hesitate DE say : ' EU with China DE political
关系 、贸易 关系 以及在投资 等 方面
relation , trade relation and at investment etc. aspect
的 合作 在一九九七年都 取得 了
DE cooperation in 1997 all achieve ASP
显著 的 发展 。"]
significant DE progress . "

"A few days ago, when this reporter exclusively interviewed Wei Genshen, head of the EU Europe Commission delegation to China, and asked him to comment on the accomplishment of the cooperation between the two sides in the past year, without any hesitation he said: 'There was significant progress in the political relations, trade relations, and the cooperation in trade, etc. between EU and China.' "

4 What counts as an argument?

This section examines the syntactic composition of arguments to discourse connectives in Chinese. Arguments of discourse relations are propositional situations such as events, states, or properties. As such an argument of a discourse relation can be realized as a clause or multiple clauses, a sentence or multiple sentences. Typically, a subordinate conjunction introduces clauses that are arguments in a discourse relation. Discourse adverbials and coordinate conjunctions, however, can take one or more sentences to be their arguments. The examples in (11) shows that arguments to discourse connectives can be a single clause (11a), multiple clauses (11b), a single sentence (11c) and multiple sentences (11d) respectively.

(11) a. [*conn* 尽管] [*arg1* 今年 一 至
even though this year January to
十一月 中国 批准 利用 外资
November China approve utilize foreign investment
项目 数 和 合同 外资
project number and contract foreign investment
金额 都 比 去年 同 期
amount both compared with last year same period
有所下降]，[*conn* 但] [*arg2* 实际 利用
have decrease , **but** actually use
外资 金额 仍 比
foreign investment amount still compared with
去年 同 期 增长 了
last year same period increase ASP
百分之二十七点零一] 。
27.01 percent .

"Even though the number of projects that use foreign investment that China approved of and contractual foreign investment both decreased compared with the same period last year, the foreign investment that has actually been used increased 27.01 percent."

b. [*conn* 由于] [*arg1* 茅台酒 制作 工艺
because Maotai Liquor brew process
复杂 ，生产 周期 长]，[*conn*
complicated , production cycle long ,
因而] [*arg2* 其 产量 十分有限
therefore its production volume very limited
] 。
.

"Because the brewing process of Maotai liquor is complicated and its production cycle is long, its production volume is very limited."

c. [*arg1* 中国 乒乓球 运动员没有参加
Chinese table tennis athlete not participate
第二十九 和 三十 届
twenty-ninth and thirtieth CL
世乒赛] 。[*conn* 因此]
word table tennis tournament . **therefore**
，[*arg2* 复制 的 金牌 中包括 将 要
, replicate DE gold medal in include will will
举行的 第四十五 届
hold DE forty-fifth CL
世乒赛 金牌] 。
world table tennis tournament gold medal .

”Chinese athletes did not attend the twenty-ninth and the thirtieth world table tennis tournaments. Therefore, The replicated gold medals also include the gold medals in the yet-to-be-held forty-fifth world tournament.”

d. [*arg1* 回归 后 对 澳门 的 未来 发展 是
return after for Macao DE future prospect be
利 还是弊 ？有 五成三 的 人
plus or minus ? have 53 percent DE people
回答 不 知道]。[*conn* 但] [*arg2* 对于能 不
answer not know . **but** to can not
能 接受 和 港 澳 一样，以 「
can accept like Hong Kong Macao same , with ’
一 国 两 制 」解决 台湾 问题，则
one country two system ’ resolve Taiwan issue ,
有 二成七 的 民众 表示 「不 知道 」
have 27 percent DE people indicate ’ not know ’
，五成九 的 民众 表示 「不 能 接受 」
, 59 percent DE people indicate ’ not can accept ’
] 。
.

”Is the return of sovereignty (to China) a plus or minus for Macao’s future? 53 percent of people say they don’t know. But to the question of whether they accept the resolution of the Taiwan issue with ’one country, two systems’ like Hong Kong and Macao, 59 percent of the people say ’they cannot accept’ . ”

5 Argument Scope

Determining the scope of an argument to a discourse connective has proved to be the most challenging part of the discourse annotation. A lot of the effort goes into deciding when certain text units should be included in or excluded from the argument of a discourse connective. Under our annotation scheme, the prepositional phrases, which generally precede the subject in a Chinese clause, are included in the argument of a discourse connective, as illustrated in (12a). The material in the main clause that embeds a discourse relation, however, are excluded, as in (12b).

(12) a. 另外 ，[*arg1* 在休闲 文化 生活缺乏
in addition , in recreation culture life lack
的 东莞]，[*conn* 除非] [*arg1* 很 有
DE Dongguan , **unless** very have
教育 热诚]，[*conn* 否则] [*arg2*
education enthusiasm , **otherwise**
很 难 留住教师]。
very difficult keep teacher .

” In addition, in Dongguan where recreational activities are lacking, unless they are very enthusiastic about education, it is very hard to keep teachers.”

b. 任志刚 还 表示 ，[*conn* 由于] [*arg1*
Ren Zhigang also indicate , **because**
香港 和 美国 息 差 达
Hong Kong and the U.S. interest discrepancy reach
一百二十五点]，[*arg2* 如果市场 对
125 point , if market in
香港 经济 前景 充满 信心 ，
Hong Kong economic prospect full of confidence ,
仍 有 减 息 空间]。
still have reduce interest space .

”Ren Zhigang also indicated that because the interest discrepancy between Hong Kong and the U.S. reaches 125 point, if the market is fully confident in the economic prospect of Hong Kong, there is still room for reducing interest rates.”

A lot of the challenge in determining the scope of an argument stems from the fact that discourse structures are recursive. As such identifying the scope of an argument is effectively determining how the discourse relations are hierarchically organized. This is illustrated in (13), where the discourse relation anchored by the coordinate conjunction 但”but” is embedded within the discourse relation anchored by the subordinate conjunction 如果”if”. The ambiguity is whether the conditional clause introduced by ”如果” has scope over one or two of the clauses coordinated by 但”but”.

(13) 报告 认为 ，[**conn** 如果] [**arg1** 经济 和
report believe , **if** economy and
金融 政策 得力]，[**arg2** [*arg1* 亚洲地区
finance policy effective , Asia region
经济 可望 在１９９９年开始 回升]，[*conn*
economy expect in 1999 begin recover ,
但] [*arg2* 不 会 象 墨西哥和 阿根廷 在
but not will like Mexico and Argentina in
１９９４－１９９５年金融 危机后 那样
1994 to 1995 finance crisis after like that
出现 高速 Ｖ形 大 回升]]。
occur high-speed V-shaped big recovery .

”The report believes that if the economic and financial policies are effective, the economy of Asia is expected to recover, but there will not be a V-shaped high-speed recovery like the one after the financial crisis of Mexico and Argentina in 1994 and 1995.”

Given our bottom-up approach in which discourse connectives anchor binary discourse relations, we do not explicitly annotate hierarchical structures between the arguments. However, such discourse relations can be deduced when some discourse relations are recursively embedded within another as arguments to another discourse connective.

6 Sense Disambiguation

Although discourse connectives are often considered to be a closed set, some lexical items in Chinese can be used as both a discourse connective and a non-discourse connective. In this case it is important to tease them part. There are also discourse connectives that have different senses, and it is potentially beneficial for certain NLP applications to disambiguate these senses. Machine Translation, for example, would need to translate the different senses into different discourse connectives in the target language. The examples in (14) shows the different senses of 而, which can be translated into "while" (14a), "but" (14c), "and" (14d) and "instead" (14e). Note that in (14e) it is important for the first argument to be negated by 不"not". In (14b), however, it is not a discourse connective. It does not seem to contribute any meaning to the sentence and is probably just there to satisfy some prosodic constraint.

(14) a. 1997年发达 国家 经济 形势
1997 developed country economic situation
的 特点 是 [*arg1* 美国 增长 强劲]
DE characteristic be U.S. grow strongly
[*conn* 而] [*arg2* 日本 经济 疲软] , 美国
while Japan economy weak , U.S.
经济 增长率 估计 为 百分之三点七 ,
economic growth estimate be 3.7 percent ,
日本 仅 为 百分之零点八 。
Japan only be 0.8 percent .

"The economic situation in developed countries in 1997 is that the U.S. (economy) grows strongly while the Japanese economy is weak. The U.S. economic growth rate was estimated to be 3.7 percent while the Japanese economy grows at 0.8 percent."

b. 水东 开发区 位 于
Shuidong Development Zone located at
粤西 地区 的 茂名市
western Guangdong region DE Maoming city
境内 , 面积 八十多 平方公里 ,
territory , coverage over eighty square kilometer ,
是 适应 乙烯 工程 的 需要 [? 而] 建立
be suit ethylene project DE need **?** establish
的 一 个 后继 加工 基地 。
DE one CL downstream process base .

"Shuidong Development Zone, located in Maoming City of western Guangdong occupies an area of over eighty square kilometers. It is a downstream processing base established to meet the need of the ethylene project."

c. 能 生产 [*arg1* 中国 不 能 生产] [*conn*
can produce China not can produce
而] [*arg2* 又 很 需要] 的 药品 的
but again badly need DE drug DE
企业
enterprise

"Enterprises that can produce drugs that China badly needs but cannot produce"

d. 吉林省 珲春市 市长 金硕仁 说
Jilin Province Huichun City mayor Jin Shuoren say
: "国际 社会 的 支持 和
: " international community DE support and
参与 , 对于 珲春 的 开发
participation , to Huichun DE development
开放 起 了 [*arg1* 积极]
opening to the outside play DE positive
[*conn* 而] [*arg2* 关键] 的作用 。"
and key DE role . "

"Jing Shuoren, mayor of Huichun City of Jilin Province said: "The support and participation of the international community played a positive and key role in Huichun's development and opening up to the outside."

e. [*arg1* 这 当然 不 是 历史 的 巧合]
this certainly not be history DE coincidence
, [*conn* 而] [*arg2* 是 历史 的
, **instead** be history DE
积累 和 转接] 。
accumulation and transition .

"This certainly is not historical coincidence. Instead it is historical accumulation and transition."

7 Discourse Connective Variation

The flip side of sense disambiguation is that one discourse relation is often realized with different discourse connectives due to the long evolution of the Chinese language and morphological processes like *suoxie*, which is one form of abbreviation. The examples in (15) shows the different variations of the discourse relation of concession. The different forms of the discourse connective are so similar that they can hardly be considered to be different discourse connectives. In principle, any combination of part 1 and part 2 from Table 7 can form a paired discourse connective, subject to some non-discourse related constraints. In (15a), for example, the abbreviated 虽 can only occur in clause-medial positions. (15b) shows the second part of the paired discourse connective can be dropped without changing the semantics of the discourse relation. (15c) shows that the second part of the paired discourse connective can be combined with another discourse connective.

(15) a. [*arg1* 王翔] [*conn* 虽] [*arg1*
Wang Xiang **although**
年过半百] , [*conn* 但] [*arg2* 其
over fifty years old , **but** his

gloss	discourse connectives
although	[1] 虽然, 虽说, 虽 [2] 但是, 但, 还是, 可是, 却, 然而, 不过
because	[1] 因为, 因, 由于 [2] 所以
if	[1] 如果, 若, 假如 [2] 就
therefore	因此, 于是

Table 1: Discourse connective variation

充沛 的 精力 和 敏捷 的 思维 ，给
abundant DE energy and quick DE thinking , give
人 以 一 个 挑战者 的 印象]。
people with one CL challenger DE impression .

"Although Wang Xiang is over fifty years old, but his abundant energy and quick thinking gives people the impression of a challenger."

b. [*arg1* 外在 的 环境] [*conn* 虽然]
external DE environment **although**
[*arg1* 改变 了]，[*arg2* 内心 那 份 渴望
change ASP , heart that CL long for
记忆 与 归属 的 需求 很
memory and sense of belonging DE need very
难 改变]。
difficult change .

"Although the external environment has changed, the need of longing for memory and sense of belonging is very difficult to change."

c. [*arg1* 大陆 政策] [*conn* 虽然] [*arg1*
mainland policy **although**
动辄得咎]，[*conn* 但 却] [*arg2* 是
vulnerable to criticism , **but but** be
所有政策 的 基础]，任何候选人 都 无法
all policy DE basis , any candidate all cannot
忽视 。
ignore .

"Although the mainland policy is vulnerable to criticism, it is the basis of all policies and no candidate afford to ignore it."

8 Conclusion

We examined the range of discourse connective we plan to annotate for the Chinese Discourse Treebank project. We have shown that while arguments to subordinate and coordinate conjunctions can be identified locally, arguments to discourse adverbials may be long-distance. We also examined the distribution of the discourse connectives in Chinese and the syntactic composition and the scope of the arguments in discourse relations. We have shown the most challenging issue in discourse annotation is determining the text span of a discourse argument and this is partly due to the hierarchical nature of discourse structures. We have discussed the need to address sense disambiguation and discourse connective variation in our annotation of Chinese discourse connectives.

References

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In *Current Directions in Discourse and Dialogue*. Kluwer Academic Publishers.

William Mann and Sandra Thompson. 1988. Rhetorical Sturcture Theory. *Text*, 8(3):243–281.

E. Miltsakaki, R. Prasad, A. Joshi, and B. Webber. 2004a. The Penn Discourse Treebank. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

E. Miltsakaki, R. Prasad, A. Joshi, and B. Webber. 2004b. The Penn Discourse Treebank. In *Proceedings of the NAACL/HLT Workshop on Frontiers in Corpus Annotation*, Boston, Massachusetts.

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1).

B. Webber and A. Joshi. 1998. Anchoring a lexicalized tree-adjoining grammar for discourse. In *In ACL/COLING Workshop on Discourse Relations and Discourse Markers*, Montreal, Canada.

Bonnie Webber, Alistair Knott, Matthew Stone, and Aravind Joshi. 1999. Discourse Relations: A Structural and Presuppositional Account using Lexicalized TAG. In *Meeting of the Association of Computational Linguistics*, College Park, MD.

Bonnie Webber, Aravind Joshi, Matthew Stone, and Alistair Knott. 2003. Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–587.

Nianwen Xue and Martha Palmer. 2003. Annotating the Propositions in the Penn Chinese Treebank. In *The Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing*, Sapporo, Japan.

Nianwen Xue, Fei Xia, Fu dong Chiou, and Martha Palmer. To appear. The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*.