# Generation of Anaphors in Chinese

**Ching-Long Yeh**


Ph.D.
University of Edinburgh
1995

# Abstract

The goal of this thesis is to investigate the computer generation of various kinds of anaphors in Chinese, including zero, pronominal and nominal anaphors, from the semantic representation of multisentential text. The work is divided into two steps: the first is to investigate linguistic behaviour of Chinese anaphora, and the other is to implement the result of the first part in a Chinese natural language generation system to see how it works.

The first step is in general to construct a set of rules governing the use of all kinds of anaphors. To achieve this, we performed a sequence of experiments in a stepwise refined manner. In the experiments, we examined the occurrence of anaphors in human-generated text and those generated by algorithms employing the rules, assuming the same semantic and discourse structures as the text. We started by distinguishing between the use of zero and other anaphors, termed non-zeroes. Then we performed experiments to distinguish between pronouns and nominal anaphors within the non-zeroes. Finally, we refined the previous result to consider different kinds of descriptions for nominal anaphors. In this research we confine ourselves to descriptive texts. Three sets of test data consisting of scientific questions and answers and an introduction to Chinese grammar were selected. The rules we obtained from the experiments make use of the following conditions: locality between anaphor and antecedent, syntactic constraints on zero anaphors, discourse segment structures, salience of objects and animacy of objects. The results show that the anaphors generated by using the rules we obtained are very close to those in the real texts.

To carry out the second step, we built up a Chinese natural language generation system which is able to generate descriptive texts. The system is divided into a strategic and a tactical component. The strategic component arranges message contents in response to the input goal into a well-organised hierarchical discourse structure by using a text planner. The tactical component takes the hierarchical discourse structure as input and produces surface sentences with punctuation marks inserted appropriately. Within the tactical component, the first task consists of linearising in depth-first order the message units in the discourse structure and mapping them into syntactic-oriented representations. Referring expressions, the main concern in this thesis, are generated within the mapping process. A linguistic realisation program is then invoked to convert the syntactic representation into surface strings in Chinese.

After the implementation, we sent some generated texts to a number of native speakers of Chinese and compared human-created results and computer-generated text to investigate the quality of the generated anaphors. The results of the comparison show that the rules we obtained are effective in dealing with the generation of anaphors in Chinese.

# Acknowledgements

My first thanks go to my principal supervisor, Dr. Chris Mellish, for sparking my interest in natural language generation. His thoughtful comments and insights provided resolution for difficulties at key points during this project. His prompt feedback on every draft of this thesis contributed greatly the early completion of this work. Thanks also go to my second supervisor, Dr. Matt Crocker, for helpful discussion on semantic issues. I am indebted to the following NL people Saad Al-Jabri, Alistair Knott, Nicolas Nicolov, and Osama Zaki for useful discussions, in particular, Nicolas, for patiently introducing to me the concept of natural language generation from conceptual graphs.

I would also like to thank those Taiwanese students in Edinburgh who kindly finished the questionnaires I gave to them. Without their help, the zero anaphora experiment and evaluation work could not have been finished. I thank David Ellis for proof reading the final draft of this thesis.

Thank you Show-Fen and Geng-Lun for our happy time in Edinburgh. Finally, I would like to dedicate this thesis to my parents, Yu-Yin Yeh and Mian Chang-Yeh who have always encouraged me in my studies.

## Declaration

I hereby declare that I composed this thesis entirely myself and that it describes my own research.

Ching-Long Yeh
Edinburgh
December 27, 1995

# Contents

# List of Figures

# Conventions Used in Examples

In this thesis, all the Chinese examples are presented in the following way.

- Each Chinese example is given a numerical index, starting from 1 in each chapter. The sentential units within an example are indexed by lower-case English letters, if it consists of multiple units; otherwise, the single unit is not indexed.
- Each sentential unit in an example is first given in Chinese romanisation, which is based on the *Concise English-Chinese Chinese-English Dictionary*, published by the Commercial Press and Oxford University Press, 1986. Then a word-by-word translation is given; the abbreviations used are described below. Finally, the English translation of the whole unit is given.
- In the word-by-word translation, some markers are abbreviated as below. We follow the abbreviations used in [Li & Thompson 81].

| **Abbreviation** | **Term** |
|---|---|
| ASSOC | associative (de) |
| ASPECT | aspect marker |
| BA | ba |
| BEI | bei |
| CL | classifier |
| GEN | genitive (de) |
| NOM | nominaliser (de) |

# Chapter 1

# Introduction

## 1.1 Problem and Aim

The field of natural language generation has made a great deal of progress in the generation of multisentential text in recent years [McKeown 85, Maybury 90, Dale 92, Hovy 93]. Most of the well-known systems consist of a strategic, or *what-to-say*, and a tactical, or *how-to-say* component [Thompson 77, Reiter 94]. The strategic component is concerned with selecting and organising the message contents to be generated and the tactical component maps the organised results into a sequence of surface sentences. The conceptual integration of the selected contents is the primary concern of the first component. Basically, the integration is achieved through a set of semantic relations which hold between sentences [McKeown 85, Maybury 90, Hovy 93].

When mapping into the surface form, appropriate linguistic devices must be used in order to make the generated text a cohesive unit. There are many devices which aid cohesion [Grosz & Sidner 86, Maybury 90]. Among them, connectives, such as *for example* and *however*, and coreference, in particular, anaphora, have received more attention than others in existing work. Previous work has classified connectives into categories to indicate the underlying structure [Knott & Dale 92]. Work on coreference, on the other hand, focuses on the selection of appropriate forms for anaphors [McDonald 80, Dale 92].

In this thesis, we aim at the computer generation of anaphors in Chinese. This requires an effective anaphor generation component in a Chinese natural language generation

system. In the first part of this thesis we focus on the establishment of rules for the generation of Chinese anaphors. In the following part we concentrate our attention on the implementation of a Chinese natural language system whose anaphor generation component is based on the results obtained in the first part. On completing the implementation, we then carry out an evaluation of the anaphors in the texts generated by the system.

## 1.2 Research Methodology

This research starts with establishing rules for the generation of anaphors in Chinese. Previous work suggests obtaining these rules from consulting the results of linguistic study. Most of the linguistic results employed in previous work are general principles, like the Gricean maxims [Grice 75] used in [Dale & Haddock 91, Reiter & Dale 92, Dale 92], focus theory in [Dale 92], etc. A shortcoming of this approach is that it is unclear the extent to which the resulting rules are effective in dealing with the generation of anaphors. To overcome this, we adopt an empirical approach to obtaining rules based on observations on real texts.

The basic framework of the empirical work is to conduct experiments that compare anaphors occurring in human and potentially computer-generated texts, assuming the same semantic contents in both texts. To carry out the empirical study, we first of all confine ourselves to a certain type of text, here, descriptive text, and select a set of texts of this type as the test data. The test data provides human texts with which these may be compared. As for the other side, we assume that there could exist a computer system that takes the same semantic content as the human texts and can generate Chinese anaphors according to some possible anaphor generation rules. Therefore, once a rule is decided on, we can see the extent of effectiveness of the rule by comparing the anaphors occurring on both sides. At the beginning of the empirical work, a rule with simple constraints is used and an experiment is performed to see its effect. One additional constraint is then appended to the preceding rule and the same experiment is repeated. The empirical work continues in this way until the result is promising.

In the course of establishing the rules, we first focus on collecting constraints that suggest the use of zero anaphora. Having done this, the rule is made to distinguish between the two types of non-zero anaphors, namely, pronouns and nominal anaphors. In the succeeding experiments, more constraints are considered to refine pronouns and nominal anaphors within non-zeroes and then choose descriptions for nominal anaphors.

In the implementation stage of this thesis, we take ideas from well-known natural language generation systems as the backbone of our system. Decisions about anaphoric forms occur right after the message contents are selected and organised, namely, at the end of text planning. The output structure of the text planner greatly affects this decision process. Therefore, when designing the text planner, one important consideration is whether it provides sufficient information in its output structure for the decision about anaphors. As for the linguistic realisation part of our Chinese generation system, it must be able to deal with each kind of anaphors.

## 1.3 Overview of Anaphor Generation in Chinese

The decision over which anaphor to use occurs in a component within a natural language generation system. To illustrate the idea of anaphor generation, we first of all give an overview of our Chinese natural generation system, as shown in Fig. 1.1. As shown in the figure, the system accepts whatever a user wants as the input goal. The text planner then consults the planning operator library to get appropriate operators to organise a hierarchical discourse structure that satisfies the goal. At the terminal nodes of the hierarchical structure are the semantic representations of message units which altogether form the message content in response to the user's question. The hierarchical structure looks like a tree where the root is the user's goal, as shown in Fig. 1.2. The internal nodes in the trees are decomposable planning operators; the terminal nodes are the corresponding message units extracted from the domain knowledge base through a set of rhetorical predicates [McKeown 85, Maybury 90]. The message units attached to the terminal nodes in the plan tree are semantic representations where entities are represented as indices linking to entries in the domain knowledge base. Each message unit is realised as a sentence. The user model specifies the user's

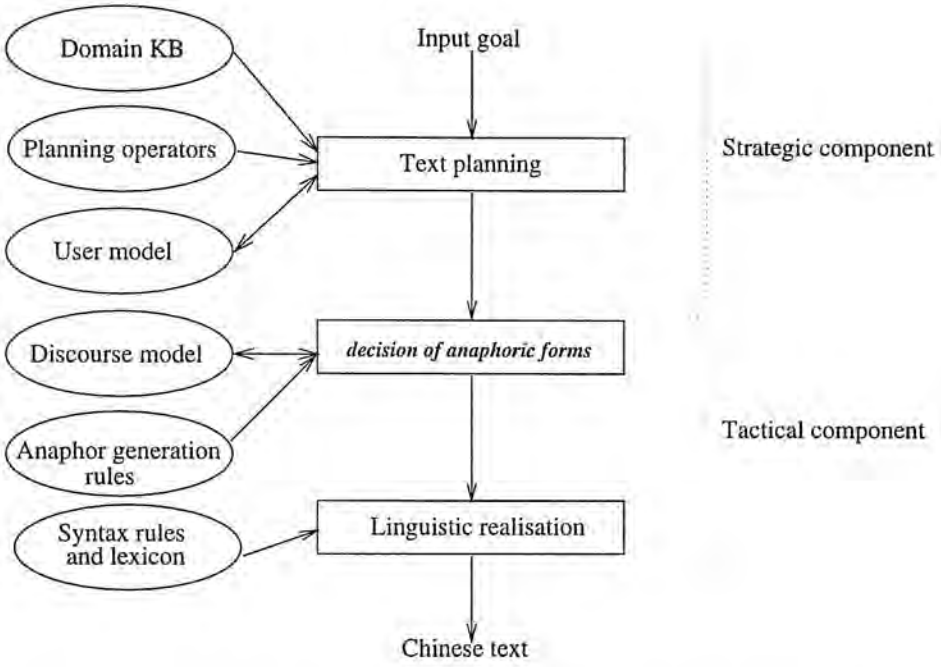



Figure 1.1: System overview of our Chinese natural language generation system.

specific requirements and is used to tailor the tree structure.

The message units in the tree in a linear sequence form the message content to be produced. Therefore, the next step is to linearise the tree using a depth-first traversal. Within the traversal, the system determines whether an entity in a message unit occurred previously in the tree. If it did occur previously, then it needs to be referred to by an anaphor and the anaphor generation component is invoked to get an anaphoric form for it. The system maintains a discourse model to record the history of the present discourse. This includes syntactic/semantic details of the preceding sentence, local and global focus space stacks [Grosz & Sidner 86], and a list of entities occurring previously. The system consults the anaphor generation rule base along with the information stored in the discourse model to get a form for an anaphor.

Within the depth-first traversal, the semantic representation of each message unit is converted into a syntactic-oriented representation. In this representation, each anaphor is featured with the anaphoric form, either zero, pronoun, or nominal, and the syntactic information about the anaphor, including the head and the modification part of the

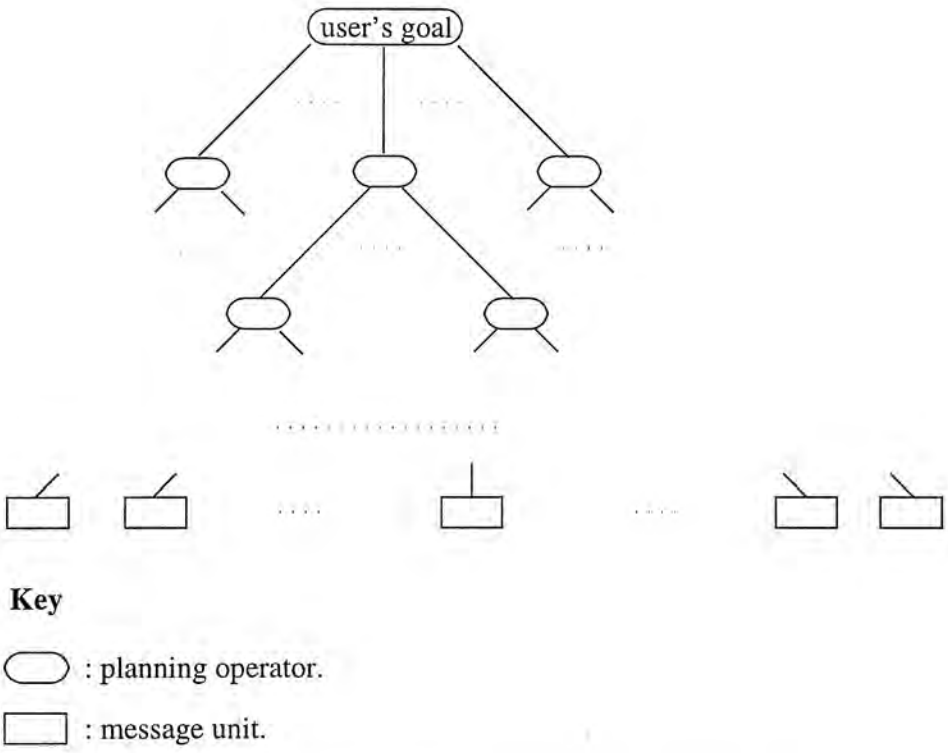



Figure 1.2: Tree representation of hierarchical discourse structure.

entity. After the traversal, the linguistic realisation component takes the syntactic-oriented representations as input and maps them into the surface string of text in Chinese.

## 1.4 Scope of Thesis

This thesis concentrates on the use of zero, pronoun and nominal anaphors in Chinese generated text. We are not concerned with lexical anaphora [Tutin & Kittredge 92] where the anaphor and its antecedent share meaning components, while the anaphor belongs to an open lexical class. For example, *flower* can be used as a lexical anaphor for *rose* [Tutin & Kittredge 92]. Nor does our work consider the factor of the types of user when choosing nominal descriptions [Reiter 90]. We assume a single level for the user.

This work does not aim at a general account for anaphoric phenomena occurring in various kinds of text. Instead we focus on investigating anaphors occurring in descriptive texts and the corpus selected for this work is therefore confined to this type of written text.

This work, though it includes the implementation of a Chinese natural language generation system, does not intend to invent any new ideas on natural language generation system apart from its treatment of anaphor generation. The Chinese generation system is mainly employed as the framework to test the anaphor generation rules we establish. Thus we adopt concepts from existing natural language generation systems to build up our system so that it can generate descriptive texts. This system involves the same essential components as in the conventional natural language generation system, namely, text planner and linguistic realisation component, and the accompanying knowledge bases. It does not focus on other issues of natural language generation, such as user modelling, tailoring responses to various level of users, recovery mechanisms, etc.

## 1.5 Contributions

The main contribution of this thesis is a set of computational rules for the generation of anaphors in Chinese. In contrast to other anaphor generation work that established rules from linguistic principles [Dale 92, Horacek 95], our rules were established by integrating linguistic results with observations on real data. The experiments we carried out for the establishment of anaphor generation rules not only show the effectiveness of the rules, but also can be used as the framework to study the generation of anaphors for other types of text or other languages.

This work focuses on investigating the generation of zero, pronominal and nominal anaphors in Chinese, which contrasts with previous work on other languages, like English [Dale 92], French [Tutin & Kittredge 92], or German [Block & Horacek 90]. This work can provide a starting basis towards the study of anaphor generation in a multilingual environment.

In practical terms, this work provides a successful implementation of anaphor generation rules in a Chinese natural language generation system. This work presents generated texts to show the kinds of anaphors can be generated by the system. In addition, it provides an evaluation of the generated anaphors to show the quality of the output.

## 1.6 Thesis Organisation

The remainder of this thesis is organised as follows. First, surveys on linguistic and computational linguistic aspects are given in Chaps. 2 and 3. In Chaps. 4 to 6, the empirical work on zero, pronominal anaphora and nominal descriptions is presented. After the anaphor generation rules are established, we show the implementation of the Chinese natural language generation system in Chap. 7. This enables us to investigate the behaviour of the anaphor generation rules in a real system. In Chap. 8, we present the evaluation of anaphors in the texts generated by our Chinese generation system. Finally, Chap. 9 summarises the results and suggests areas for future research.

# Chapter 2

# Relevant Linguistic Background

## 2.1 Introduction

In this chapter, we briefly introduce the linguistic background relevant to the research in this thesis. We start by describing the importance of topic in Chinese grammar and topic as a discourse element. We then introduce various kinds of anaphors in Chinese, including zero, pronominal and nominal anaphors. After introducing the above concepts, we present a survey of previous linguistic studies on Chinese anaphora. Then we describe the scope of anaphora investigated in this thesis.

## 2.2 Topic Prominence in Chinese

Chinese is termed a *topic-prominent* language in that in addition to the grammatical relation of "subject" and "direct object", the description of Chinese must also include the element "topic" [Chao 68, Li & Thompson 81]. The topic of a sentence is what the sentence is about and always comes first in the sentence; [1] the rest of the sentence is comment upon the topic. The subject of a sentence is the noun phrase that has a "doing" or "being" relationship with the verb in the sentence. For example, in (1), zheke shu (this tree) and yezi (leaf) are the topic and subject of the sentence,

[1] Note that sentence here refers to a single unit in a text which ends up with a comma, or a full stop mark, such as a sentential mark and question mark. The structure of a sentence can be a simple declarative construction, a presentative construction, a question, a comparative construction, a serial verb construction, and a complex stative construction. See [Li & Thompson 81] for detailed descriptions. Later in Sec. 2.4, we use quoted sentence to represent a unit of complete meaning in Chinese discourse.

respectively.

(1) zheke shu yezi hen da.
this tree leaf very big
This tree, (its) leaves are very big.

By distinguishing topics and subjects in sentences, we have the following types of sentences: sentences with both subject and topic, sentences in which the subject and the topic are identical, sentences with no subject, and sentences with no topic, which are exemplified in (2) to (5), respectively [Li & Thompson 81].

(2) naben shu wo yijing du guo le.
that book I already read ASPECT
That book I have already read.
(3) wo xihuan chi pingguo.
I like eat apple
I like to eat apple.
(4) naben shu $\phi$ yijing chuban le.
that book (someone) already publish ASPECT
That book, (someone) has published it.
(5) jin-lai le yige ren.
enter-come ASPECT one person
A person came in.

Sentence (5) is an example of a "presentative sentence" which "presents" an indefinite noun phrase in discourse. Topics in Chinese sentences must be either definite or generic [Li & Thompson 81]. Consequently, though the only noun phrase in (5), yige ren (one person) is clearly the subject of the verb jin-lai (come in), it is not the topic because it is neither definite nor generic. It introduces a previously unknown entity, i.e., new information, into the discourse.

Topic, as a discourse element, can simply relate to some part in the preceding sentence, introduce a subtopic which is related to what has been discussed, or reintroduce a topic that has been mentioned earlier. All of the above cases involve a noun phrase that refers

to an object mentioned earlier in the sentence or in a previous sentence. This noun phrase is called an anaphor. In addition to topic, anaphors in general can occur in other positions in a sentence. In Chinese, anaphors can be in one of zero, pronominal and nominal forms. In the next section, we give an overview of various kinds of anaphors.

## 2.3 Anaphora in Chinese

In Chinese, anaphors can be classified as zero, pronominal and nominal forms, as exemplified in (6) by $\phi_1^i$, ta$^i$ (he) and nage ren$^i$ (that person), respectively [Chen 87]. [2] Zero anaphors are generally noun phrases that are understood from the context and do not need to be specified. In contrast, in this thesis, we use the term non-zero anaphors to denote those that are specified in discourse, namely, pronominal and nominal anaphors.

(6)a. Zhangsan$^i$ jinghuang de wang wai pao,
Zhangsan frightened NOM towards outside run
Zhangsan was frightened and ran outside.

b. $\phi_1^i$ zhuangdao yige ren$^j$,
(he) bump-to a person
(He) bumped into a person.

c. ta$^i$ kanqing le na ren$^j$ de zhangxiang,
he see-clear ASPECT that person GEN appearance
He saw clearly that person's appearance.

d. $\phi_2^i$ renchu na ren$^j$ shi shui.
(he) recognise that person is who
(He) recognised who that person is.

According to [Li & Thompson 81], zero anaphors can be classified as intrasentential or intersentential. Intrasentential zero anaphora occur mainly in topic-prominent constructions, namely, sentences having a topic but not a subject, as the $\phi$ in (7). In this

[2] We use a $\phi_a^b$ to denote a zero anaphor, where the subscript $a$ is the index of the zero anaphor itself and the superscript $b$ is the index of the referent. A single $\phi$ without any script represents an intrasentential zero anaphor. Also note that a superscript attached to an NP is used to represent the index of the referent.

sentence, the noun phrase, fangzi (house), is the topic while the subject is not present. In sentences of this sort, subjects, in general, refer to general classes or unspecified noun phrases. In English, *you*, *they* (or more formally *one*) is used in this function. This kind of zero anaphor occurs specifically in topic-prominent constructions; they have nothing to do with entities in previous sentences in discourse.

(7) fangzi $\phi$ zhaohao le.
house (someone) build-finish ASPECT
The house, (someone) has finished building it.

In the intersentential case, antecedent and anaphor are located in different sentences. Depending upon the distance between the sentences containing antecedent and anaphor, it can further be divided into two types: immediate and long distance. The former is where the sentences containing the antecedent is immediately followed by the one containing the anaphor, such as $\phi_1^i$ and $\phi_2^i$in (6b) and d. For the long distance type, the sentences containing the antecedent and anaphors, on the other hand, are not in immediately succeeding order, such as $\phi_1^k$ in (8e), whose antecedent occurs four sentences away in (8a).

(8)a.pangxie$^i$ you sidui buzu$^j$,
crab have four-pair walking-foot
A crab has four pairs of feet.
b.$\phi_1^j$ sucheng "tuier",
(they) common-called "tuier"
(They) are commonly called "tuier."
c.youyu meitiao "tuier"$^j$ de guanjie$^k$ zhi neng xiang xia wanqu,
since every "tuier" ASSOC joint only can towards down bend
Since every "tuier"'s joint can only bend downwards,
d.$\phi_1^k$ bu neng xiang qianhou wanqu,
(it) not can towards forward-backward bend
(it) can't bend backward or forwards.
e.$\phi_1^i$ paxing shi,
(it) crawl ASPECT

(When) (it) crawls,

f. $\phi_2^i$ bixu xian yong yibian buzu de zhijian zhua di,

(it) must first use one-side walking-foot ASSOC fingertip grasp-on ground

(it) must use the tips of feet on one side to grasp the ground.

g. $\phi_3^i$ zai yong ling yibian de buzu zhishen qilai,

(it) then use another one-side ASSOC walking-foot straight-rise upwards

(It) then uses the feet on the other side to move upwards.

h. $\phi_4^i$ ba $\phi_5^i$ shenti tui guoqu.

(it) BA (it) body push get-through

(It) pushes (its) body towards one side.

Since Chinese has no inflection, conjugation, or case markers, the pronominal system is relatively simple, as shown in Table 2.1 [Li & Thompson 81].

Table 2.1: Pronominal system in Chinese.

| Number | Person | Pronoun |
|---|---|---|
| singular | first | wo (I/me) |
| singular | second | ni (you) |
| singular | third | ta (he/she/it/him/her) |
| plural | first | women (we/us) |
| plural | second | nimen (you) |
| plural | third | tamen (they/them) |

Note that all the third-person pronouns in the table seem to have same word ta. In fact, they simply share the same pronunciation. In Chinese, there are three tas written differently to distinguish between male, female and neutral. A third-person pronoun can be used to replace an intersentential zero pronoun, except for first- and second-person pronouns, without changing the meaning of the sentence. For example, in (8) all of the zero anaphors can be replaced by third person pronouns. Though the resulting meaning of each sentence is unchanged, the whole discourse becomes less coherent.

Nominal anaphors do not have unique forms like their zero and pronominal counterparts. The descriptions of nominal anaphors can be the same as the initial reference, parts of the information in the initial reference can be removed, new information can be added to the initial reference, or even a different lexical item can be used for a nominal anaphor. For example, an initial reference, tuoyuanxing da tong (oval big barrels),

occurs in (9a), while, a reduced description, da tong (big barrels), is used in (9b) to refer to the same entity, and later in (9i) the full description is reused.

(9)a.shashuicheshang$^i$ tuoyuanxing da tong$^j$ shi zhuang shui yong de,
sprinkle-water-car-above GEN oval-shape big barrel is fill water use NOM
Big oval barrels on sprinkler trucks are used to fill with water.

b.women xiwang da tong$^j$ neng duo zhuang dian shui.
we hope big barrel can more fill some water
We hope that the big barrel can be filled with more water.

c.yuanxing da tong$^k$ zhuang shui zui duo,
spherical big barrel fill water most much
Spherical big barrels can be filled with much more water.

d.danshi yuanxing tong$^k$ rongyi zuoyou gundong,
but spherical barrel easily left-right roll
However, spherical barrels easily roll.

e.$\phi^j$ buyi wengude anzhuang zai shashuicheshang$^i$,
(they) not-easily stably install on sprinkler-truck
(They) are not easily installed on sprinkler trucks.

f. ruguo ba tong$^l$ zuocheng tuoyuanxing de,
if $\phi$ BA barrel make-into oval-shape NOM
If (we) have a barrel made into oval shape.

g.$\phi^l$ zhuang shui bijiao duo,
(it) fill water more much
(It) can be fill with much more water.

h.$\phi^l$ hai bi yuanxing tong$^k$ wending.
(it) moreover compare spherical barrel stable
And (it) is much more stable than spherical barrels.

i. tuoyuanxing da tong$^j$ de ling yige youdian shi zhizuo rongyi.
oval-big-barrel GEN another a advantage is make easily
Another advantage of oval big barrels is they are easier to make.

In (10), danche (bike) is used as the initial reference in a, while another word jiaotache is used subsequently in b to refer to the same object [Chen 86]. In (11), an initial

reference, jiaozi (dumpling), occurs in a; later on, in b, it appears in a description, panli de jiaozi (the dumplings in the plate), with new information added.

(10)a.yige xiaohaizi$^i$ qi danche$^j$ chuxian le.

one kid ride bike appear ASPECT

A boy appeared, riding a bike.

$\cdots$ (sentences) $\cdots$

b.ta$^i$ xian ba jiaotache$^j$ fang xialai,

he first BA bike put down

First he put the bike down.

$\cdots$ (sentences) $\cdots$

(11)a.suoyi $\phi$ chi jiaozi$^i$ de shihou

therefore (we) eat dumpling ASSOC time

Therefore, when eating dumplings,

b.$\phi$ yao ba panli de jiaozi$^i$ shishi yong kuaizi fandong,

(we) must BA plat-in ASSOC dumpling constantly use chopstick turn-over

(we) must use chopsticks to constantly turn over the dumplings in the plate.

c.$\phi$ buyao rang tamen$^i$ nian zai yiqi.

(we) no-must let they stick at together

(We) must not let them stick together.

The surface structure of the description for a nominal anaphor is a noun phrase of one of the following schemas [Li & Thompson 81].

> associate phrase + classifier phrase/measure phrase + relative clause + adjective + noun, or
> associate phrase + relative clause + classifier phrase/measure phrase + adjective + noun.

Both schemas have the same elements, only with classifier phrase/measure phrase and relative clause in the reverse order. All of the above elements except the head noun are optional, as illustrated in (12) [Li & Thompson 81].

(12) wo de neige zhuzai Meiguo de hao pengyou

I ASSOC that-classifier live-at America NOM good friend

that good friend of mine who lives in the United States

In (12), the noun phrase wo (I) together with the associative particle de is an associate phrase; neige (that) is a classifier phrase; zhuzai Meiguo de (living in the United States) is a relative clause; hao (good) is an adjective; and pengyou is a head noun. The associative phrase, schematised as a noun phrase followed by the particle *de*, indicates that this noun phrase and the head noun in the above schema are "associated" or "connected" in some way [Li & Thompson 81], as exemplified in (12) and (13).

(13)a.na-ge fandian de cai
that-CL restaurant ASSOC food
the food of that restaurant
b.xuexiao de jiaoyuan
school ASSOC teaching-staff
the school teaching staff
c.chenshan de kouzi
shirt ASSOC button
the buttons on (the) shirt

Classifier phrases and measure phrases are used to indicate the referentiality or quantity of an object. They are composed of determinative/number and classifier/measure. The choice of classifier is determined by the head noun. For example, jia and zhan in (14a) and b are classifiers accompanied with feiji (aeroplane) and deng (lamp).

(14)a.wu-jia-feiji
five-CL-aero plane
five aeroplanes
b.zhei-zhan-deng
this-CL-lamp
this lamp

A noun phrase can be embedded in a locative phrase of the following structure [Chao 68, Li & Thompson 81]:

zai noun-phrase locative-particle

For example, in (15), zai fangzi-houmian is a locative phrase, where houmian is a locative particle. The coverb zai in a locative phrase is obligatory except in the case of presentative constructions where a locative phrase is placed at the beginning of a sentence. [3] In presentative sentences, zai is optional, as illustrated in (16).

(15)tamen zai fangzi-houmian xiuli dianshiji.
they at house-behind repair television
They repair television behind the house.

(16)(zai) wuzi-li you sange ren.
(at) house-in exist three person
In the house, there are three people.

## 2.4 Punctuation Marks in Written Chinese

Punctuation marks are considered an important element towards successful writing in Chinese [Liu 84]. There are two groups of punctuation marks in Chinese; one is used to separate sentences, with which we are concerned in this paper, and the other is used to denote quotations, emphasis, ellipsis, etc. [Liu 77]. There are seven marks in the former group, among which the sentential mark and comma are essential for creating text, while others are auxiliaries of the two marks [Liu 77]. A sentential mark is used to indicate the full stop of a "sentence"; a comma within a "sentence" indicates a temporary stop. [4] The sentential mark has two auxiliaries, the question (?) and exclamation (!) marks, which are used to express full stops of "sentences" with certain tones. The comma has three auxiliaries, dun-hao, semicolon (;) and colon (:), which are used to express very brief pause, two closely related sentences and summary for the either preceding or succeeding sentences, respectively. Basically, both sentential marks and commas are sufficient for expressing full and temporary stops in "sentences" [Liu 77]. The existence of auxiliaries is to create delicate distinctions among the use of punctuation marks and, on the other hand, allow readers to easily grasp writers' ideas, emotions, etc. In this paper, we focus on investigating the generation of sentential marks and commas.

[3] A coverb in a sense is like a preposition in English [Chao 68, Li & Thompson 81].

[4] Here we use a quoted sentence to distinguish from the usual sense of sentence in English.

A sentential mark is used to represent the full stop of a "sentence," while a comma is used to indicate that a "sentence" is not yet finished. Obviously, to make proper use of them, we first have to understand the above conditions. In other words, we have to exploit what a "sentence" is. Although there is no clear definition for "sentence," a commonly employed idea is that a "sentence" is a meaning-complete unit [Liu 84, Liu 77, Yu 55]. For example, both (17) and (18) are "sentences."

(17) a. women dajia yici lai wan.
we all-of-us together come play
We all come to play together.

(18) a. dao tian ming shi,
till sky light ASPECT
Till sun rise,
b. ta$^i$ cai ba wo de yifu chuanhao,
she then BA I ASSOC clothes wear-good
she then put my clothes on,
c. $\phi_1^i$ cui wo qu shangxue.
(she) hurry I go school
(She) hurried me off to to school.

The number of sentences in a "sentence" may range from one, several, as in (17) and (18), for example, to even longer size [Yu 55]. A sentence can be a "sentence", or can combine with others to form a larger "sentence." For example, in (19), all individual sentences are meaning-complete and each can be a "sentence"; however, if they combine together, as in (19), then they result in altogether another "sentence" with a different meaning in which the first and second sentences are the cause of the final one.

(19)a.wo ai kan dianying,
I like see movie
I like to see movies.
b.ta ai ting jingxi,
he like listen Peking-opera

He likes to listen to Peking opera.

c. zanmen liangren zong coubudao-yikuaier.

we two-people always can-not-put-together

We both always can not be together.

## 2.5 Previous Linguistic Studies on Chinese Anaphora

*Zero anaphora.* An early set of studies focused on investigating the behaviour of zero anaphora in adverbial clause constructions and correlative structures, as exemplified by (20) and (21), respectively [Li & Thompson 78, Liu 81].

(20)a. Zhangsan$^i$ zou le *yihou,*

Zhangsan leave ASPECT after

After Zhangsan left,

b. $\phi^i$ jiu mei huilai guo.

(he) then not come-back ever

(he) has never returned.

(21)a. *yinwei* Lao Li$^i$ hen mang,

because Lao Li very busy

Because Lao Li is very busy,

b. *suoyi* ta$^i$ bu neng lai kan ni.

therefore not can come see you

he can't come to see you.

These two constructions can be distinguished by the occurrence of correlative markers: there is a pair of correlative markers in a correlative structure, such as *yinwei* and *suoyi* in (21), while there are no such markers in the other construction. Li and Thompson [Li & Thompson 78] propose that a zero anaphor is obligatory in adverbial clause constructions, while optional in the other. Later in [Liu 81], Liu extended the above condition by taking into consideration the positions of the subject and the correlative marker in the constructions. He defined that a construction has parallel structure if the markers either both precede the subjects or both follow the subjects; for

example, (21) has parallel structure. [5] The extended condition is that a zero anaphor is obligatory in parallel structures which are subject-initial; otherwise it is optional. He further noticed that in parallel structures which are not subject-initial, the pronoun is preferred, such as (21).

Later work provided a more general account of zero anaphora by taking discourse into account [Tai 78, Li & Thompson 79, Chen 84, Chen 86, Chen 87]. According to Li and Thompson's study [Li & Thompson 78, Li & Thompson 79], a zero anaphor can occur in almost any position in the sentence, with their antecedents occurring in any grammatical slot. To achieve a full account of zero anaphora, all linguistic information, including lexical, syntactic, semantic and pragmatic, and even world knowledge are required [Li & Thompson 79, Chen 84, Chen 86]. Tai in his study [Tai 78] observed some examples and proposed that an anaphor can be zeroed across segment boundaries if both the anaphor and the antecedent are subjects, and the segments containing the anaphor and the antecedent are adjacent and have the same type of description. His study provides an initial idea for the use of zero anaphora, but the problems are that the definition of segment is not clear at all and the observational data was very limited. Li and Thompson found that, from their experiments [Li & Thompson 79], zero anaphors commonly occur in the situation of a "topic chain," where a referent is referred to in the first sentence, and then several more sentences follow talking about the same referent, namely, the topic, but with it omitted; see (8e) to h in Sec. 2.3 for example.

In [Li & Thompson 79], Li and Thompson also formulated a negative rule stating that zero anaphors are not allowed in certain syntactic positions, regardless of discourse factors. [6] First, for example, the NP right after a coverb cannot be zeroed. The coverb in Chinese is to some extent similar to the preposition in English, as **gen** (with) in the following sentence. The NP following the coverb, **ta** (he) in (22), for example, cannot be a zero anaphor. Another syntactic constraint is that the pivotal NP in a serial verb construction cannot be zeroed. For example, in the following serial verb construction, where **quan ta** (urge him) and **bie he jiu** (not to drink) are the consecutive verb phrases in the construction, **ta** (he) is the pivotal NP which can not be zeroed.

[5] Here, the subject is the same as the topic in the sentence.

[6] See [Li & Thompson 81], for similar conclusion.

(22) wo gen ta xue Yingwen.

I with he learn English

I learn English with him.

(23) wo quan ta bie he jiu.

I urge he not drink wine

I urged him not to drink.

Chen, in his thesis [Chen 84], proposed that a zero anaphor is triggered by the satisfaction to a high extent of two conditions: the predictability condition (PC) and the negligibility condition (NC). The PC is determined by the following parameters. First is the availability of competing nouns in the discourse in terms of the syntactic, semantic and pragmatic aspects of the sentence where the anaphor is embedded, and the world knowledge of the participants of the discourse. Second is whether the sentences embedding the anaphor is closely related with the preceding one as a grammatical unit rather than as two independent units. [7] The third parameter concerns the position of the anaphor in the sentence. According to his observation, he found that most zero anaphors occur in the topic/subject position, 75% (43/57), then the direct object position, 19% (11/57), and the indirect object, 5% (3/57). The NC is concerned with the noteworthiness of any specific mention of the anaphor, which is measured in terms of the following three parameters in the discourse. First, when an anaphor is a non-specific or generic reference in Chinese, it is considered to have high negligibility. [8] Second is the positions of anaphors in main vs. subordinate sentences. He claimed that the subject of the subordinate sentence has higher negligibility than that of the main sentence when they are of identical reference. Third is the animacy of referents: an inanimate referent has higher negligibility that an animate referent. This study does not provide clear rules for the purpose of generation. However, it can be employed to aid in the establishment of anaphor generation rules.

Later, Chen proposed a notion of "continuity" of referents in discourse to give a more specific account for zero anaphora [Chen 86, Chen 87]. Continuity of referent has lo-

[7] Here Chen borrowed the idea of degree of connection between two successive sentences in [Li & Thompson 79], which is described later on in this section.

[8] Detailed descriptions about non-specific and generic reference can be found in [Li & Thompson 81, Chen 86].

cal and global perspectives. The local perspective has to do with the position of the antecedent and anaphor in their respective sentences. The other perspective considers the linear and hierarchical relationship between sentences containing antecedent and anaphor in the discourse structure. The linear relationship along with the local continuity is able to account for zero anaphors occurring in successive sentences. In other words, he considered the positions of anaphors in their respective sentences and the positions of their antecedents in the preceding sentences in order to give an account for zero anaphora. According to his observation for successive sentences, the most frequent use of zero anaphora is to encode anaphors in the topic position, with their antecedents most possibly in the preceding topic or object positions within the same "sentence" boundary. As for other cases of zero anaphora, the possibility of occurrence is very low. He further employed an RST (Rhetorical Structure Theory) [Mann & Thompson 87] analysis as the basis of the hierarchical relation of global continuity. The long-distance zero anaphora are thus analyzed according to the discourse structure of RST trees. From his analysis, he found that if the level of nonterminal in the RST tree connecting the sentences containing the antecedent and the anaphor is lower, then it is more possible that zero anaphora are used.

Chen found quite a few zero anaphors in his data could not be explained in terms of high continuity of referents [Chen 86]. The referents of one kind of zero anaphor in these cases can be inferred through the information available in the context. Cataphoric zero anaphors present another problem. The factors affecting the use of these anaphors can be captured to some extent by the psychological weight of noteworthiness of the referents in the discourse. For example, inanimates are less noteworthy than animates; adjunct sentences are usually less noteworthy than nuclear sentences for the communicative goals of the interlocutors. As a principle, the less noteworthy a referent is, the more inclined to zero form is its anaphor.

*Pronominal anaphora.* A third-person pronoun can be used to replace a zero anaphor without loss of meaning. Depending on the context, a pronoun is more preferred over the other in certain situations. The question is: what are the appropriate conditions of using pronominal anaphora rather than zero ones? Again, it appears there is no clear rule governing the use of pronominal anaphora in discourse. Tai [Tai 78] claimed that

an anaphor that cannot be zeroed can be pronominalised regardless of the grammatical functions of the antecedent and the anaphor. However, he did not provide conditions on when to use pronouns rather than nominal anaphors. Li and Thompson, by conducting some experiments of native speaker's impressions regarding the use of the pronominal form, claimed that "the perception of the degree of connection between the clauses (sentences) in discourse affects the occurrence of pronominal anaphora" [Li & Thompson 79]. They further formulated a basic principle governing the occurrence of pronominal anaphora in Chinese discourse: "the degree of preference for the occurrence of pronominal anaphora in a clause (sentence) inversely corresponds the degree of connection with the preceding clause (sentence)." They listed some general conditions for connection:

- when the involved sentences contain a switch of information from background to foreground, or vice versa, the connection decreases;
- when the second sentence is marked with adverbial expressions, which signal the beginning of a new "sentence" rather that a connected sentence, the degree of connection decreases; and
- when the two sentences in question are spoken by different participants, the connection decreases.

After examining a number of test data, Chen claimed two major discourse features to explain the use of pronouns in "topic chains" [Chen 86]: the place of a minor discontinuity and the high noteworthiness of referent. In a discourse, a minor discontinuity happens in a place, where a switch of referent from the previous sentence occurs, a sentence and its previous one are talking about different referents, with separate schemes, or with separate goals. This feature in a sense is related to the conditions for connection listed above. The noteworthiness in the second feature is similar to the one described previously in zero anaphora, namely, animate referents and the position in nuclear sentences are more noteworthy than their counterparts. Chen found in his test data that the percentage of pronouns referring to inanimate objects is far lower than nominal anaphors. Thus he concluded that the lower in noteworthiness an anaphor stands, the less likely a pronoun is chosen than other anaphors.

*Nominal anaphora.* The previous linguistic work on Chinese anaphora focus, most of all, on zeros, and then pronouns. Nominal anaphors receive the least attention. In [Chen 86], Chen identified the low continuity of referent in discourse as the major discourse-pragmatic factor for the use of a nominal anaphor. A low continuity in a discourse results, for example, from a switch of referent from the prior topic. Another situation of low continuity in a discourse occurs when the discourse is disrupted through the discontinuity of space, time, event, or world. In his work, he did not distinguish what kind of descriptions should be used for nominal anaphors.

Huang, in [Huang 91, Huang 94], has developed a pragmatic theory for the analysis of anaphora within a neo-Gricean framework of conversational implicature [Grice 75]. In his theory, several principles based on Grice's theory of conversational implicature [Grice 75] specify the preference order of local interpretation as zero, pronominal and then nominal anaphora. The interpretation is constrained by the presumption of disjointness of arguments of a predicate, information saliency and general consistency constraints on conversational implicature. He has applied his theory to the interpretation of anaphors occurring in Chinese conversations [Huang 94]. However, he did not investigate how to apply his theory to the choice of anaphoric form.

In brief, previous linguistic studies investigated phenomena of Chinese anaphora and proposed principles or constraints for the interpretation and use of anaphors. The results, however, are not sufficient for natural language generation purposes because, first of all, they are not represented in computational forms. Furthermore, previous works did not demonstrate the effectiveness of the results.

## 2.6 Scope of Anaphora

In the following, we describe the scope of anaphora investigated in this thesis. First, of the two types of zero anaphors, intra- and inter-sentential, we are only concerned with the latter. In our test data, we will ignore the occurrence of intra cases. [9] Second, as is clarified later, no long distance zero anaphors would be produced by our generation rules. This does not mean that we ignore long distance zero anaphors

[9] This is three sets of articles selected from two scientific question-answer books and a Chinese grammar book. See Chapter 4 for detailed descriptions.

completely. Instead the fact is that in our experiment we found that they occurred far less frequently than their non-zero counterparts in the test data. The decision not to generate them is based on this experimental result. Another reason for not generating them is that, from the computational point of view, it is impractical to spend a lot of effort on a few cases. Third, for pronominal anaphora, we deal only with the generation of third-person pronouns. First- and second-person pronouns have unique forms for the initial and subsequent references if they are non-zerod. In other words, if they repeat in a discourse, they are either zeroed or appear in the same form as the initial form in the discourse. For the latter case, we treat them as a special case of nominal anaphora. Finally, as mentioned in Sec. 2.3, nominal anaphors may have various forms. To simplify our work, we assume the description of the initial reference to be a maximal description for the subsequent references throughout the rest of the discourse. This means that the descriptions of nominal anaphors must either be the same as or a reduction of the initial one.

## 2.7 Summary

In this chapter, we first illustrated the topic-comment sentential structure in Chinese as the basis of discourse units and the characteristics of the essential component, anaphor, in the discourse to provide a basic idea about what would happen in the text generated by our generation system. We introduced the concept of punctuation marks in Chinese text as the basis for the investigation of anaphora in this thesis. A survey on previous relevant linguistic study shows how complicated the factors are for the use of anaphors in Chinese. The linguistic study also reveals a fact that most of the factors affecting the use of anaphors are discourse-oriented. Since we do not intend to pursue the sophisticated behaviour used by human writers for our generation system, we therefore make some simplifications on the anaphors.

# Chapter 3

# Anaphora in Natural Language Generation

## 3.1 Introduction

After introducing what linguistic studies have been done on Chinese anaphora, in this chapter, we discuss what relevant work on anaphora has been produced in natural language generation. First of all, we look at the position of a referring expression component where anaphors are created in a natural language generation system. Since a referring expression component operates on the message structures produced by a text planner, we describe some text planners in some existing systems. As mentioned in the preceding chapter, discourse-oriented factors are most significant for the use of anaphors in Chinese. We therefore give an overview of Grosz and Sidner's theory [Grosz & Sidner 86] to investigate the influence of discourse structure on referring expression in English. Then we introduce work done on the generation of referring expressions in previous work on natural language generation. After gaining a basic understanding of how anaphors can be dealt with in a natural language generation system, we then summarise the problems and propose an approach to coping with them in our work.

## 3.2 Referring Expression Components in Natural Language generation Systems

Thompson, in an early study [Thompson 77], proposed a system architecture for natural language generation that divides a system into strategic, or *what-to-say*, and tactical, or *how-to-say* components, as shown diagrammatically in Fig. 3.1. In a recent study [Reiter 94], Reiter proposed a pipelined architecture, as shown in Fig. 3.1, as a consensus account for a number of recently-developed systems. Though these two architectures are different in granularity, they are functionally equivalent to each other. In general, Reiter's architecture can be referred to as a specialisation of Thompson's by seeing that content determination along with sentence planning correspond to the strategic component and the remainder to the tactical component, as shown in the figure.

The job of the strategic component is to obtain and arrange the message contents to be produced into a well-organised discourse structure. This involves two functions that (1) determine what information should be communicated to the hearer and (2) organise this information in a rhetorically coherent manner. Most existing systems use an integrated module, a text planner, to perform the above two functions. In general, the text planner employs the notion of planning from artificial intelligence [Tate 85] to obtain a plan satisfying a user's goal. On accepting a user's goal as input, the text planner is invoked to get an appropriate operator by consulting the planning operator library to achieve the goal. If the operator is not a primitive one, the planner further decomposes it into lower-level ones by consulting the operator library. Otherwise, the planner extracts the relevant knowledge from the domain knowledge base and attaches to the operator as the associated message content. The planner thus produces a hierarchical discourse structure with message content attached to the leaf nodes.

The tactical component takes the hierarchical discourse structure as input and converts the message contents in it into a sequence of surface sentences. Within this component, the first task is to linearise the message contents in the discourse structure and then map the conceptual structures of the message contents into semantic structures. This includes, as summarised in Reiter [Reiter 94]:

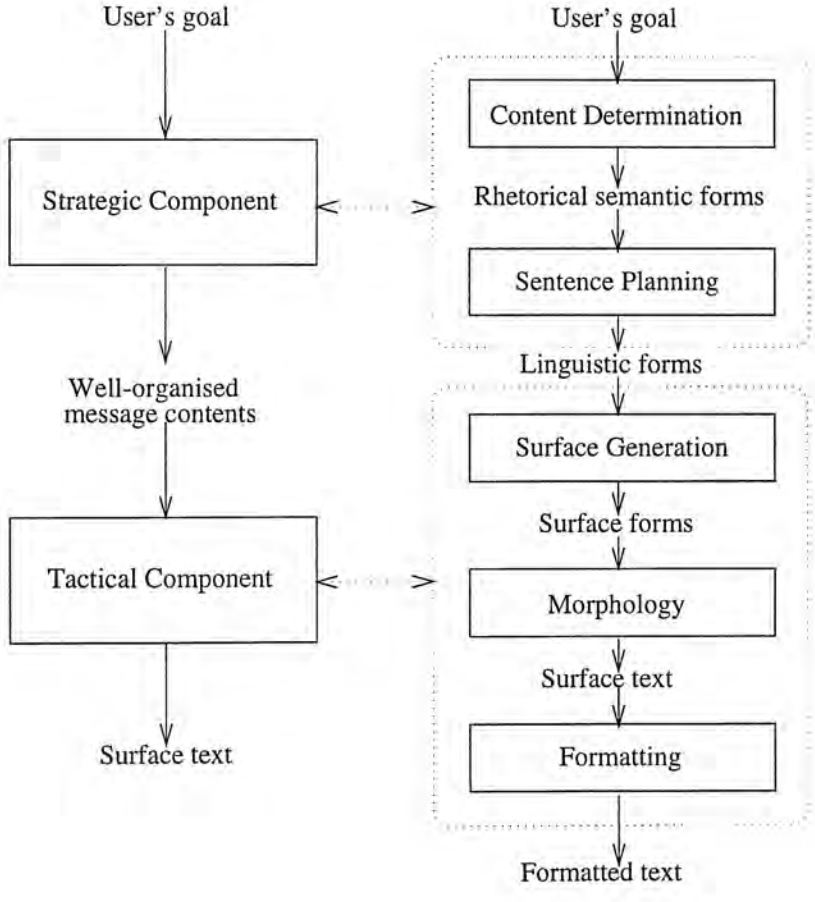



Figure 3.1: Comparison of two general system architectures for natural language generation

1. Mapping domain concepts and relations into content words and grammatical relations;

2. Generating referring expressions for individual domain entities.

3. Grouping propositions into sentences.

Among them, the second task is the main goal of this thesis. The remaining task of the tactical component is to convert the semantic structures into syntactic forms and then surface text, with optionally some sort of formatting. In brief, the position of the referring expression component is right after the text planner in a natural language generation system.

## 3.3 Text Planning

As described in Sec. 3.2, the referring expression component takes the message structures produced by a text planner as its input. In this section, we introduce some popular text planners to investigate the characteristics of the message structures. We start with the planner in McKeown's TEXT system [McKeown 85]. Then we introduce a class of planners based on the idea of RST. Finally, we introduce the text planner in Maybury's TEXPLAN system [Maybury 90] whose approach will be utilised by our Chinese natural language system.

### 3.3.1 Planning in the TEXT system

The text planner in the TEXT system makes use of script-like structure, schemata, to organise multisentential text. A schema is a standard pattern of discourse structure, which is obtained by the analysis of sample texts. There are four schemata identified in [McKeown 85], each of which is composed of a sequence of rhetorical predicates, for example, the `Identification` schema as shown in Fig. 3.2. [1] The control of the planner here is different to the one described previously. On accepting a user's question, a single schema is selected according to the information available in the

[1] Note that in the schema, "{ }" means optional, "/" indicates alternatives, "+" indicates that the item may appear 1 to n times, and "*" indicates that the item is optional and may appear 0 to n times.

```
Identification (class & attribute / function)
{Analogy / Constituency / Attributive / Renaming / Amplification}*
Particular-illustration / Evidence+
{Amplification / Analogy /Attributive}
{Particular-illustration / Evidence}
```

Figure 3.2: `Identification` Schemata in TEXT.

**Schema selected: identification**

```
identification
analogy
particular-illustration
amplification
evidence
```

**The text content**

An aircraft carrier is a surface ship with a DISPLACEMENT between 78000 and 80800 and a LENGTH between 1039 and 1063. Aircraft carriers have a greater LENGTH than all other ships and a greater DISPLACEMENT that most other ships. Mine warfare ships, for example, have a DISPLACEMENT of 320 and a LENGTH of 144. All aircraft carriers in the ONR database have REMARKS of 0, FUEL TYPE of BNKR, FLAG of BLBL, BEAM of 252, ENDURANCE RANGE of 4000, ECONOMIC SPEED of 12, ENDURANCE SPEED of 30 and PROPULSION of STMTURGRD. A ship is classified as an aircraft carrier if the characters 1 through 2 of its HULL NO are CV.

Figure 3.3: A selected and filled schema and the corresponding text.

domain knowledge base to satisfy the question. Then the selected schema is filled by matching the predicates in the schema against the knowledge base relevant to the question. For example, given a question "What is an aircraft-carrier?" and a naval database, the `identification` schema is selected and filled, and the corresponding filling content, expressed in English text, is shown in Fig. 3.3.

### 3.3.2 RST planners

RST (Rhetorical Structure Theory) was originally developed as an analytic tool for describing relationships between successive pieces of text [Mann & Thompson 87] and

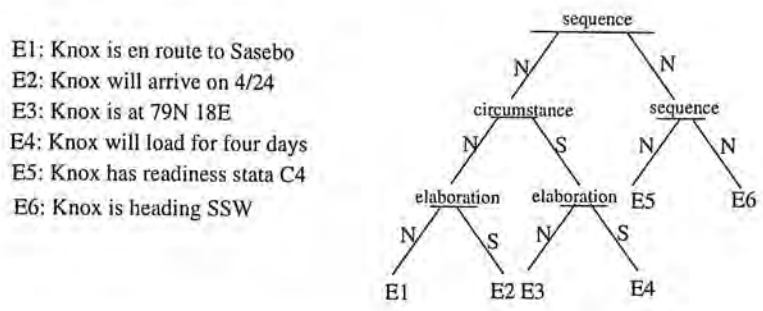



Figure 3.4: Input material and the corresponding RST tree.

later used as the basis of planning operators in a group of text planners [Hovy 90, Hovy 93, Moore & Paris 94]. It defines a number of relations to identify particular functional relationships between two non-overlapping spans of text: the nucleus (N) and the satellite (S). A relation definition contains constraint fields on N, S and on the combination of N and S and a field specifies the effect the writer intends to achieve in the relation. A relation can exist alone or together with other relations to form a schema. By applying schemas iteratively, a text can be analysed as a functionally dependent structure. As for goal-driven text planning, operators are constructed out of the RST relations to guide the inclusion of material into planning trees. The constraint fields in relations play the role of preconditions of planning operators; the effect field is the goal of the operator. When a communicative goal is given to the planner, it seeks an applicable operator for the goal to include material from the knowledge base. Each of the constraints on N and S of the operator in turn becomes a subgoal to be satisfied in the next round of tree expansion. This process continues until there is no more input material or the goal is satisfied. For example, given the input goal, `(BMB S H (POSITION-OF E105 ?NEXT)`, glossed as: *from the input units, tell the hearer the sequence of events of which E105 is a principal part* (i.e, the temporal position of Knox), and the glossed input material, the text planner will produce the structure, as shown in Fig. 3.4 [Hovy 90].

In [Moore & Paris 94], Moore and Paris focused on the use of RST theory in a dialogue environment. They argue that Hovy's operationalisation of RST relations into planning operators conflates the intentional and rhetorical structure, which is inadequate to handle dialogues. In their text planner, they preserve explicit representations for both

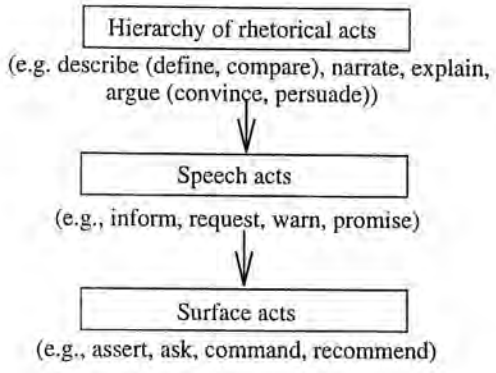



Figure 3.5: An integrated theory of communicative acts.

intentional and rhetorical structures. Another improvement to Hovy's planner is that their planner is able to do content selection in addition to text structuring. The choice of strategy to satisfy an intention must depend on what knowledge is available.

### 3.3.3 The text planner in the TEXPLAN system

In TEXPLAN [Maybury 90], the control of the text planner is similar to the one described in Section 3.2 and the planning operators are formulated on the basis of an integrated theory of communication acts, as shown in Fig. 3.5, which is motivated by observing the explanation types of text. In the figure, the top level is a hierarchy of rhetorical acts which characterise single or multiple sentences and correspond to the major types of explanation, including description, narration, exposition and argument. In this hierarchy, the rhetorical acts at a higher level can be decomposed into one or several acts at a lower level. For example, the hierarchy of rhetorical acts for description is shown in Fig. 3.6. According to the hierarchy, an entity can be described by defining it, by dividing it into its subtypes or sub-constituents, by providing details about it, by comparing it to or giving an analogy with something with which the reader is familiar. Similarly, down the hierarchy, we can define an entity logically, synonymically or antonymically.

In the middle of the hierarchy of communicative acts are speech acts which are concerned with illocutionary communication, such as `inform`, `request`, `warn` and `recommend`. The rhetorical acts at the lowest level are achieved by particular speech acts. In the

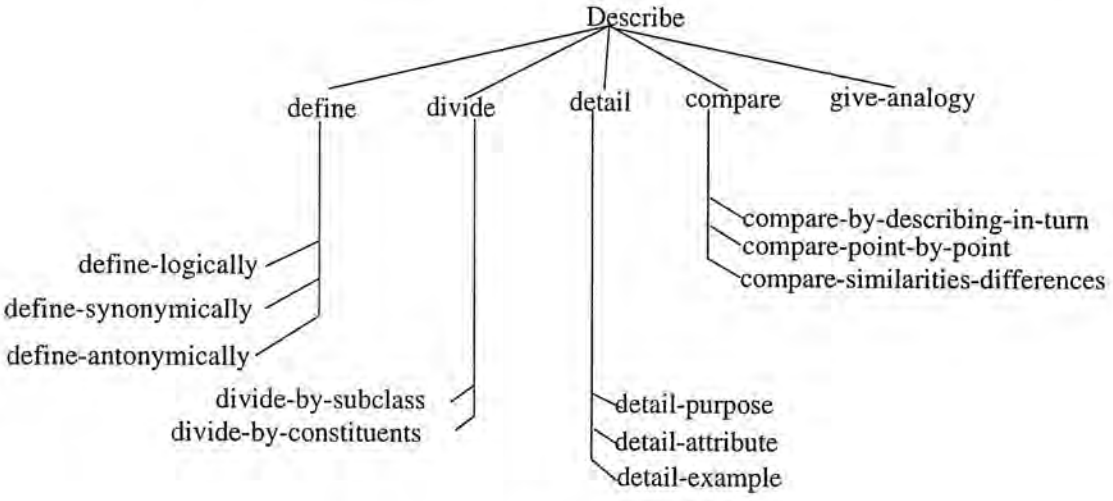



Figure 3.6: Hierarchy of rhetorical acts for description.

case of description, these acts are usually achieved by the speech act `inform`. For example, in Fig. 3.6, the rhetorical act `define-logically` is achieved by informing the logical definition of the entity being described. Other speech acts occur in particular types of texts. For example, in the exposition type of text, an instruction enabling the hearer to go from one place to another could first find and then describe a path between the two places. The instruction could then conclude by identifying the location of the destination. In the first step, a requesting act is used to ask the hearer to move along the path between the two places and an optional informing act is followed to describe the path. Then an informing act is used to identify the destination. For example, the following instruction is to enable a person in Rome, NY, to travel to a restaurant in another city, Syracuse, NY [Maybury 90].

> To get to the Country Inn from Rome take the New York State Thruway (Interstate 90) to Exit 39. Take a left onto Interstate 690. Travel a quarter mile to the Farrel Road exit. The Country Inn is located at 1615 State Fair Blvd.

At the lowest level of the communicative act hierarchy are the surface acts which guide the selection of appropriate sentence structures during the realization of the underlying message contents. For example, a message content having the agent John and Mary

and the location park of an action walk can be realized as the declarative form, as in (1), the imperative form, as in (2), etc.

(1) John and Mary walk to the park.
(2) (John and Mary) walk to the park.

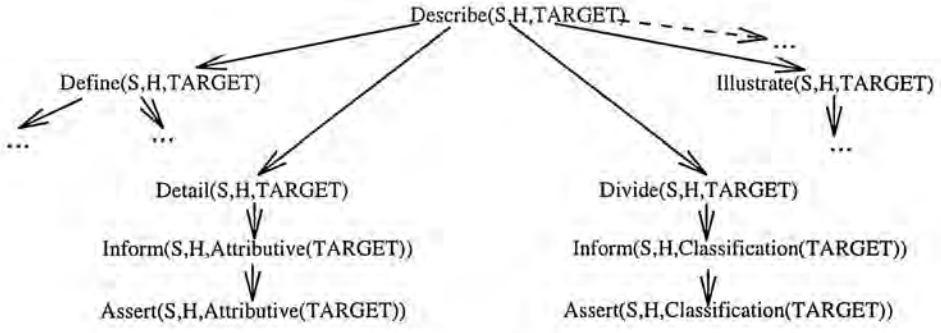
The surface acts characterise the communication of individual utterances in a discourse, such as `assert`, `ask`, `command`, `recommend`. Each surface act is related to a specific surface form, such as declarative, imperative, etc. A speech act can be accomplished by a surface act. For example, the speech act `inform` is achieved by the asserting surface act; in contrast, the requesting act can be achieved by the surface acts, `command`, `ask` or `recommend`, which correspond to imperative, interrogative and obligatory modal surface forms, respectively. The choice among various surface forms for a `request` is governed by constraints such as the relationship between the speaker and the hearer, etc. Since, in this paper, we generate descriptions for a tourist guide, all rhetorical acts are achieved by informing speech acts which are accomplished by asserting surface acts.

Composition of communicative acts in the hierarchy enables the creation of multisentential texts. For example, in Fig. 3.7a, the description of an entity can be achieved by first giving a definition, and optionally following this by providing detailed information, such as attributes and purpose, dividing the entity according to its classification or constituency, and providing illustrative or analogous samples. An example of an instantiation of this `describe` operator and its corresponding text are shown in the same figure.

As will be clarified, discourse segmentation structure is an important constraint in our anaphor generation rules. We employ Grosz and Sidner's discourse structure theory [Grosz & Sidner 86] in our empirical study. In our implementation, we needed a text planner that could produce discourse structure to meet the concept in [Grosz & Sidner 86], so that the anaphor generation rules could be implemented. TEXT's planner produces flat structures, for example, Fig. 3.3, that do not match the hierarchical structures we need. Both RST text planners and TEXPLAN's planner produces hierarchical discourse structures. In this thesis, we choose to employ TEXPLAN's planner as the

| | |
|---|---|
| **NAME** | extended-description |
| **HEADER** | Describe(*S, H entity*) |
| **CONSTRAINTS** | Entity?(*entity*) |
| **PRECONDITIONS** | |
| **ESSENTIAL** | KNOW-ABOUT(*S,entity*)∧ |
| | WANT(*S*,KNOW-ABOUT(*H, entity*)) |
| **DESIRABLE** | ¬KNOW-ABOUT(*H,entity*) |
| **EFFECTS** | KNOW-ABOUT(*H,entity*) |
| **DECOMPOSITION** | Define(*S,H,entity*) |
| | **optional**(Detail(*S,H,entity*)) |
| | **optional**(Divide(*S,H,entity*)) |
| | **optional**(Illustrate(*S,H,entity*))∨ |
| | Give-Analogy(*S,H,entity*)) |

(a)




**The corresponding text**

Targets are entities. They have a latitude/longitude, a cloud cover, a cloud height, a visibility, and a weather condition. There are five targets: passages, facilities, electronic hardware, weapons, and vehicles. Weapons, for example, are targets such as anti-aircraft missiles, surface-to-surface missiles, and enemy aircraft.

(b)

Figure 3.7: The `describe` operator and an instantiation.

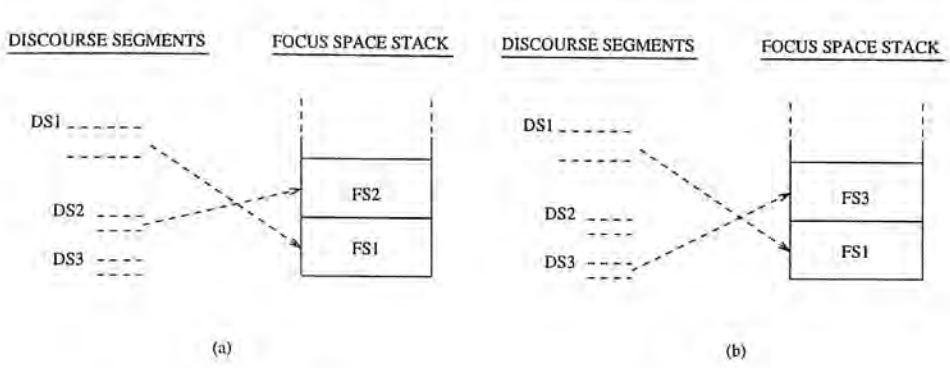



Figure 3.8: Discourse segmentation and the focus space stack.

framework of the text planner in our Chinese natural language generation system. We leave the use of an RST-type planner in our system in future study.

## 3.4 The Effect of Discourse Structure on Referring Expressions

In [Grosz & Sidner 86], Grosz and Sidner suggest that three structures can be identified within a discourse: *linguistic structure*, *intentional structure*, and *attentional state*. The first structure is the sequence of utterances that comprise the discourse. Underlying this is the intentional structure, which shows the relationship between the respective purposes of discourse segments. The third structure reflects the evolution of discourse entities as a discourse proceeds. A discourse is divided into hierarchical segments which have their respective discourse purposes, termed discourse segment purposes (DSPs). The dominance and satisfaction precedence relationship among DSPs in a discourse imposes a hierarchical structure on the discourse segments. The attentional state, modelled by a set of focus spaces, changes according to a set of transition rules of adding and deleting spaces. The structure of the set of focus spaces, in general, is stack-like. For example, in Fig. 3.8, there are three discourse segments DS1, DS2, and DS3, with discourse purposes DSP1, DSP2, and DSP3, respectively, where DSP1 dominates both DSP2 and DSP3. Initially, the focus space of DS1, FS1, is pushed into the stack; then as it proceeds to DS2, FS2 becomes the top entry, as shown in Fig. 3.8a. When segment DS3 is processed, FS2 is popped off from the stack and FS3 is pushed onto it, as shown in Fig. 3.8b.

In the theory, an idea related to referring expressions is the mutual effect between the linguistic expressions in utterances constituting the discourse and the discourse segment structure. On the one hand, linguistic expressions can be used to convey information about the discourse segment structure. For example, the explicit use of words and phrases, for example, *in the first place*, to indicate segment boundaries; referring expressions can be used to mark discourse boundaries [Grosz & Sidner 86]. On the other hand, the discourse segment structures constrain the use and interpretation of linguistic expressions. For example, there are different constraints on the use and interpretation of pronouns and reduced definite noun phrases within a segment than across segment boundaries [Grosz & Sidner 86]. What concerns us here is the interrelationship between the forms of referring expressions and the discourse segment structures. In a natural language generation system, the referring expression component is input a message structure which in a sense is similar to the discourse segment structure just described. Each discourse segment structure is accompanied with a focusing structure which becomes a focus space stack as the discourse proceeds. Referring expressions are then decided, based on the discourse segment structure and the focus space stack. In this thesis, we will employ Grosz and Sidner's theory as a condition in the course of investigating anaphor generation in Chinese.

## 3.5 Previous Work on Referring Expressions in Natural Language Generation

Studies on Chinese anaphora can be found in linguistic literature; however, no corresponding work has been done in computational linguistics, or natural language processing. In the literature, most of the computational work on referring expressions focuses on working out the problem for English [McDonald 80, Reiter 90, Reiter 91, Dale 92, Dale & Haddock 91, Reiter & Dale 92, Reiter & Dale 95]. Among them, MUMBLE [McDonald 80] and EPICURE [Dale 92] deal with the whole range of referring expressions. Other work is concerned with the decision about the form of nominal descriptions [Dale & Haddock 91, Reiter & Dale 92, Reiter & Dale 95], knowledge representation for referring expression generation [Dale & Haddock 91, Dale 92], or customising descriptions for different levels of users [Reiter 90, Reiter 91]. Since in this thesis, we

are not concerned with the generation of descriptions for different level of users, we will not look at Reiter's work. In our system, we take advantage of the knowledge representation in [Dale & Haddock 91, Dale 92]; we leave a discussion of this to the implementation chapter. In this section, we briefly survey the remaining work to investigate relevant theories and techniques.

We start with MUMBLE [McDonald 80]. MUMBLE discriminates between two kinds of referring expressions within a discourse: initial and subsequent. The decision as to what descriptions to use for initial referring expressions is a complicated process [McDonald 80, Dale 92]. In this thesis, we only consider subsequent referring expression. When an element referring to an object mentioned previously in the discourse is to be generated, the program first calls a procedure to determine whether it is pronominalizable. If it is, then a pronoun is used; otherwise, a non-pronominal referring expression is generated. In the pronominalization procedure, it first checks to see if it is already designated by the "strategic" component that a pronoun is required for the element. This occurs when the element is a cataphoric referring expression. If the element is not of this kind, then it checks whether the nature of the element allows the use of a pronoun. If the element plays a referential role, it is allowed to use a pronoun; otherwise, if it is for a descriptive purpose, then other anaphors, such as *one*, *such*, etc., must be used.

If a pronoun is allowed, some heuristic rules, which are based on linguistic features of English to figure out the relationship between the current element and its preceding referring expression, are employed to determine either for or against the use of pronoun. The nearby entities that might be taken by the reader as antecedents of the pronoun, if used, are subsequently considered to see whether they will cause ambiguity of the element. The same set of heuristic rules is applied again to evaluate the possible ambiguity of the pronoun. If the intended referent is still at the highest score, then pronominalization is carried out; otherwise, the pronoun is rejected.

In MUMBLE, when a non-pronominal referring expression is decided upon, a proper name is used if the referent has a proper name. Otherwise, if the initial referring expression has modifiers involved, MUMBLE omits all the modifiers initially used, leaving the head noun and a definite determiner.

EPICURE [Dale 92] is an natural language generation system specially designed for generating descriptions of cookery recipes in English. In EPICURE, the message content of a recipe to be produced is organised into a tree, where nonterminals are macro actions and terminals are primitive actions which can be realized as single sentences. The system maintains a discourse model that consists of two components, corresponding to Grosz's distinction between local and global focus [Grosz 77]. Global focus is concerned with a set of entities relevant to the overall discourse. Global focus can be decomposed into a collection of hierarchical focus spaces, corresponding to the hierarchical structure of the task-oriented dialogue [Grosz 77]. Local focus is concerned with identifying a single entity that is considered most centrally in a sentence. Local focus has greater effect on the process of pronominal expressions, whereas global focus is more relevant to definite noun phrases [Grosz *et al.* 83]. In EPICURE,the discourse model consists of a set of focus spaces which are hierarchically arranged to reflect the structure of the message tree. This part corresponds to the *global focus* of Grosz's model [Grosz 77]. A "cache memory" is maintained to contain the lexical, syntactic and semantic information of the current and preceding sentences, which corresponds to *local focus* in [Grosz 77]. Dale made use of the notion of *discourse centre* proposed by Grosz, *et al.* [Grosz *et al.* 83]: the result of the previous operation in a recipe, i.e., the object involved, is treated as the discourse centre.

The top-level algorithm or generation of referring expressions is roughly the same as in MUMBLE: the pronominalization decision is made first. A subsequent reference to an object is pronominalizable if it refers to the discourse centre. Following [Grosz *et al.* 83], a reference to other entities in "cache memory" may also be pronominalized, provided that the centre is pronominalized. Although, in EPICURE, Dale only dealt with pronouns with antecedents in the immediate preceding or current sentences, he nevertheless, in other papers [Dale 86, Dale 88], suggested a way to generate long-distance pronouns by means of Grosz and Sidner's discourse structure theory [Grosz & Sidner 86].

If a pronoun can not be used for the current element described, an appropriate description, i.e., a definite noun phrase is then built. In addition to EPICURE, other efforts have been paid on this particular problem [Dale 92, Reiter & Dale 92, Reiter & Dale 95].

This group of work aims at generating descriptions for a subsequent reference to distinguish it from the set of entities with which it might be confused. The main data structure in these algorithms is a *context set* which is the set of entities the hearer is currently assumed to be attending to except the intended referent. Basically their algorithms can be regarded as ruling out members of the *context set*. These algorithms pursue efficiency in producing an adequate description which can identity the intended referent unambiguously with a given context set. In his system [Dale 92], Dale used the global focus space [Grosz & Sidner 86], as the context set in his domain of small discourse. Following this idea, the context set grows as the discourse proceeds. Consider, for example, two nominal anaphors referring to the same entity occurring at different places in a discourse. According to the above algorithms, a single description would be produced for both anaphors, if the context sets at both places have the same elements. On the other hand, in general, a description with more distinguishing information is used for the second anaphor, if more distractors have entered into the context set. In [Grosz & Sidner 86], Grosz and Sidner claim that discourse segmentation is an important factor, though obviously not the only factor, governing the use of referring expressions. If the idea of context set is restricted to local focus space [Grosz & Sidner 86], then the resulting descriptions would be to some extent sensitive to discourse structure. Hence the algorithms would be refined to take account of discourse structure, but they would remain biased towards the purpose of distinguishing.

Existing work provides successful implementations for the generation of referring expressions in natural language generation, which can be employed as the framework of our system. However, most of them focus on referring expressions in English and hence can not account for problems specific to Chinese, namely, zero anaphora. Thus we need to develop a set of anaphor generation rules for our Chinese natural language generation system. Furthermore, no proper evaluation has been carried out for the referring expression generation rules used in previous systems. In this thesis, we will provide an evaluation for the referring expression component in our system.

## 3.6 An Approach to the Generation of Anaphors in Chinese

As shown in the preceding section, in the first step, algorithms for the generation of referring expressions try to pronominalise a reference. If this is inappropriate, then they nominalise it. In other words, the decision starts from thinking of an attenuated form, a pronoun here, for a reference. If the reference does not satisfy the constraints of the attenuated form, then a sophisticated form, a nominal description, is used. Following this idea, in our algorithm for the generation of Chinese anaphors, the sequence of decision becomes, in order, that zero is tried first, then pronoun and finally, nominal description.

Having determined a tripartite structure for the referring expression component in our Chinese natural language generation system, the remaining essential job is to obtain a rule as the basis of each stage in the component. Since whether the zero form is satisfied is determined at the beginning, our first effort is to establish a rule that can effectively distinguish the use of zeros from non-zeros. Then, in our second task of pursuing anaphor generation rules, we attempt to find out some constraints that can further characterise the use of pronouns and nominals out of the above non-zero cases. Finally, to complete the anaphor generation rules, we try to develop a rule for choosing an appropriate description for an anaphor if a nominal form is decided.

As surveyed previously, no efforts have been made regarding the generation of anaphors in Chinese. In other words, we have to develop a set of generation rules for the generation of anaphors in Chinese. In this thesis, we do not intend to account for phenomena of anaphora in general domains. Instead we focus our attention on a specific type of text, namely, descriptive text. Since the result of linguistic study can not be used directly to build up anaphor generation rules, we try to obtain regularities from a set of test texts written by human writers. The basic idea is as below. We compare by hand the anaphors occurring in the test texts and the ones generated by an algorithm using an anaphor generation rule, assuming the same semantic structures and context information. At the beginning, a simple constraint is included in the generation rule. Then more constraints are included and we repeat the above process

until the result is promising. After finishing this, we can see the effectiveness of the rules we have obtained from comparing the anaphors occurring in the human texts and the ones generated by the hypothetical algorithms. The rules obtained are then implemented in our Chinese natural language generation system to see the performance obtained.

## 3.7 Summary

In this chapter, we investigated from a computational point of view the generation of anaphors in Chinese. This includes identifying the position of the referring expression component in a natural language generation system, the framework of message that the referring expression component will operate on, the role of discourse structure affecting the use of anaphors and the previous work on the generation of anaphors in English. From the previous study, we propose a tripartite algorithm for the referring expression component in our Chinese natural language generation system. Since no rules can be employed directly from previous study, we further propose an empirical method to develop anaphor generation rules.

# Chapter 4

# The Decision Whether to Use a Zero Anaphor

## 4.1 Introduction

In this chapter, we start by establishing a rule for the generation of zero anaphors in Chinese. Although there are no clear rules delineated in previous linguistic work, we, nevertheless, can summarise a very simple rule, Rule 1 as shown below and in an associated decision tree in Fig. 4.1, for the generation of zero anaphors. Hereafter, we use non-zero anaphor to denote pronominal and nominal anaphors altogether.

> **Rule 1:** If an entity, $e$, in the current sentence was referred to in the immediately preceding sentence, then a zero anaphor is used for $e$; otherwise, a non-zero anaphor is used.

Although the above rule is able to account for almost all occurrences of zero anaphors, nevertheless, it is not clear whether it can be employed to generate zero anaphors appropriately. Thus, to this end, we performed an experiment by comparing the

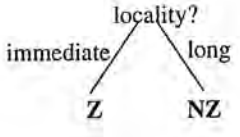



Figure 4.1: Decision tree for Rule 1.

zero anaphors generated by the algorithm employing this rule and those occurring in real text to see how well it works. The initial result showed that zero anaphors were over-generated to a large extent in the text produced by employing Rule 1. Consequently, we considered other well-known factors namely, syntactic constraints [Li & Thompson 81], discourse structure [Grosz & Sidner 86] and salience of objects in utterances [Sidner 83], to achieve better results. By taking into account syntactic constraints on zero anaphors in the rule, over-generations are reduced to some extent, but this has still not proven promising. The next experiment, using an enhanced rule taking advantage of discourse structure, showed that the over-generated zero anaphors could be reduced considerably. We further improved the rule by including the feature of prominence of topic in Chinese utterances, which results in an even better outcome. Through the above sequence of experiments, we worked out a reliable rule which considers a number of factors affecting the use of zero anaphors.

In Section 4.2, the first experiment will be described. The results of performing the experiment and the enhancements of the rule are shown in Section 4.3. Then some points about the results of the experiments and the extensions of the rule are investigated.

## 4.2 Experiment

A number of articles written by different authors were selected as the linguistic sources with which the text produced by employing the generation algorithms can be compared. In this chapter, the selected articles are restricted to the exposition type, namely, ones which explain an idea or discuss a problem. We selected three sets of articles from two books, with the details shown below.

- Sets 1 and 2:
  - **Source:** *Shiwange Weishenme* (One Hundred Thousand Whys), Vols. 1 and 4, Guojixueshe Chubanshe, Taipei, Taiwan, 1991.
  - **Description:** Each volume contains a number of scientific questions and answers for children, where answers are written by multiple authors and are of paragraph size.
  - **Number of paragraphs:** 42 for each volume.

Table 4.1: Occurrence of anaphors in the test data.

| Data | ZA | PA | NA | Total |
|---|---|---|---|---|
| Set 1 | 242 | 34 | 243 | 519 |
| % | 47 | 7 | 47 | 101 |
| Set 2 | 201 | 58 | 384 | 643 |
| % | 31 | 9 | 60 | 100 |
| Set 3 | 47 | 24 | 76 | 147 |
| % | 32 | 16 | 52 | 100 |

  - **Total number of sentences:** 701 and 673 in Vols. 1 and 4, respectively.
  - **Average number of sentences in a paragraph:** 16 and 17.

- Set 3:
  - **Source:** Li Wang, *Zhongguo Xiandai Yufa* (Modern Chinese Grammar), Zhonghua Shuju, Shanghai, China, 1947.
  - **Description:** We selected the introductory chapter from this book, the content of which is a brief introduction to grammar and modern Chinese grammar.
  - **Number of paragraphs:** 15
  - **Total number of sentences:** 230
  - **Average number of sentences in a paragraph:** 15

The numbers of anaphors occurring in the three sets of test data are shown in Table 4.1. In the table, ZA, PA and NA denote zero, pronominal and nominal anaphors, respectively. Note that we only deal with third person pronouns in Chinese; thus, in the table, and the following, pronominal anaphors, or pronouns, refer to third person cases. In this thesis, we treat the first and second person pronouns as nominal anaphors.

The experiment was executed in three steps. First, zero and non-zero anaphors within the selected articles were identified. The identification of zero anaphors can be achieved by finding out missing elements in sentences through their syntactic/semantic structures. For example, in (1b), shenshang (body) and gezhong butong de yanse (various kinds of colours) are the subject and object of the predicate you (have), respectively,

and **dou** (all) serves as the adverb. In this sentence, the topic, **xia, pangxie** (shrimps and crabs), is missing and thus is a zero anaphor. In the next sentence, the topic referring to the same entities is omitted again; besides this, the subject of the main verb **zhengzhu** (steaming cook) is omitted again, which is an intra-sentential zero anaphor. [1]

(1) a. xia, pangxie$^i$ de zhonglei hen duo,
shrimp crab GEN species very many
There are many species of shrimps and crabs.

b. $\phi_1^i$ shenshang dou you gezhong butong de yanse,
(they) body all have various different GEN colour
(Their) bodies have various colours.

c. keshi $\phi_2^i$ jingguo $\phi$ zhengzhu yiho,
however (they) through (one) steam after
However, after they are steamed,

d. $\phi_3^i$ shenti biaomian dou hui biancheng hongse le.
(they) body surface all will change-to red-colour ASP
(their) body surfaces will all become red.

Second, for each paragraph in the selected articles, we examined each sentence sequentially and recorded the occurrence of zero and non-zero anaphors that would be obtained by applying the algorithm using an anaphor generation rule, for example, Rule 1. Third, we noted down the differences between the results of steps 1 and 2.

In step 3, we categorised the differences between the results as: *correct*, *false* and *missing* types. If a reference created by the algorithm is the same as the one in the real text, then it belongs to the *correct* type. If a zero anaphor is created by the algorithm, while the corresponding position in the real text is non-zero anaphor, then it belongs to the *false* type. Conversely, if a zero anaphor is found in some position in the real text, while a non-zero anaphor is created by the algorithm, then it belongs to the *missing* type. The *false* and *missing* types correspond to the zero anaphors over- and under-generated by the algorithm, respectively. The task of step 3 is to count the

[1] See 2.2 for more detailed description about topic-comment structure in Chinese.

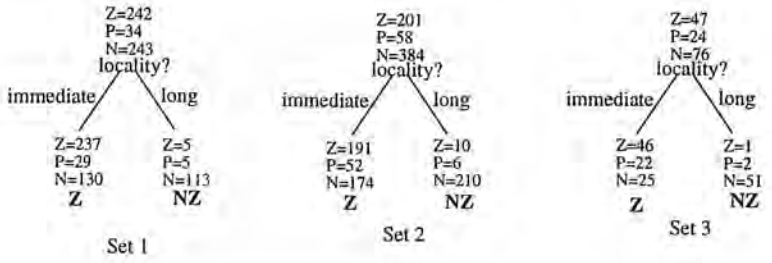



Figure 4.2: Classification trees for Rule 1.

number of cases in each type.

## 4.3 Results

In this section, we first show the result of employing Rule 1 for the algorithm on the selected articles. By inspecting the data, we add syntactic constraints to Rule 1, which becomes Rule 2, to improve the performance. Rule 2, though it achieves a better performance, still results in an unacceptable distance between the real and generated text. We therefore employ the notion of discourse structure to further enhance the above rule, which becomes Rule 3. In the round of experiments using the algorithms of Rule 3, we found that the distance between the real and generated text decreases greatly. Finally, we employ the notion of topic in Chinese to further refine the rules.

### 4.3.1 The effect of using Rule 1

For the convenience of illustration, in the following, we use a classification tree to represent the result of executing the first two steps in an experiment using the algorithm of a rule on a set of test data. Each classification tree is the decision tree of a rule, with each node annotated with the number of anaphors in the data satisfying the corresponding conditions in the rule. For example, the classification trees of Rule 1 for the sets of test data are shown in Fig. 4.2.

From the classification tree, the result of using an algorithm of a rule on a set of test data can be obtained as follows. The number of *correct* matches is the total numbers of zero and non-zero anaphors associated with zero and non-zero leaf nodes in the

Table 4.2: Result of using algorithm of Rule 1 on the test data.

| Data | Alg. | *Correct* | *False* | *Missing* | Total |
|---|---|---|---|---|---|
| Set 1 | Rule 1 | 355 | 159 | 5 | 519 |
| | % | 68 | 31 | 1 | 100 |
| Set 2 | Rule 1 | 407 | 226 | 10 | 643 |
| | % | 63 | 35 | 2 | 100 |
| Set 3 | Rule 1 | 99 | 47 | 1 | 147 |
| | % | 67 | 32 | 1 | 100 |

classification tree. The *false* and *missing* matches are the numbers of non-zero and zero anaphors associated with zero and non-zero leaf nodes in the tree. For example, the result of using Rule 1 on the test data is shown in Table 4.2. In the table, the *correct* rates of the test data are 68%, 63% and 67%, respectively. These figures obviously show an unpromising performance of the algorithm employing Rule 1. As for the types of mis-matches, the *false* rates are 31%, 35% and 32%; the *missing* rates are 1%, 2% and 1%. Apparently, what we need to do in the next step is to find constraints which can reduce the *false* rates and increase *correct* rates. As shown in the classification trees of the test data, the numbers of non-zeros are far greater than their counterparts, zeros, in the long distance cases of anaphors. Thus, in the following, we will not make any refinement to this condition because little progress would be obtained.

### 4.3.2 The effect of adding syntactic constraints

As introduced in Section 2.4, there are certain syntactic constraints on zero anaphora, regardless of discourse factors: the NP immediately after a coverb cannot be a zero anaphor, the pivotal NP in a serial verb construction cannot be zeroed, and, in written texts, the object NP in a sentence is normally not zeroed. These constraints can be used to distinguish the use of zero and non-zero anaphors. Therefore, we enhanced Rule 1 by adding the above syntactic constraints on zero anaphora, which becomes Rule 2 as below and the corresponding decision tree is shown in Fig. 4.3.

> **Rule 2:** If an entity, $e$, in the current sentence was referred to in the immediately preceding sentence and does not violate any syntactic constraint on zero anaphora, then a zero anaphor is used for $e$; otherwise a non-zero

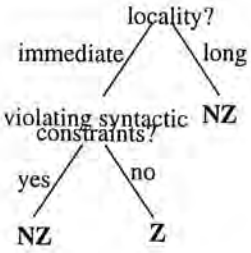



Figure 4.3: Decision tree for Rule 2.

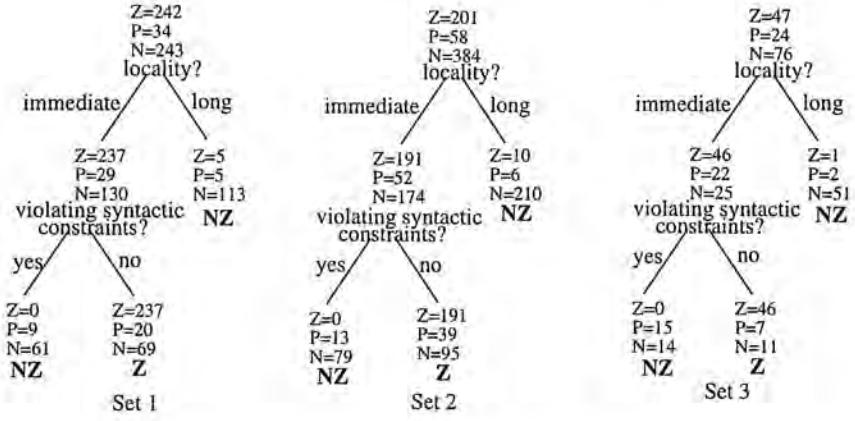



Figure 4.4: Classification trees for Rule 2.

one is used.

The classification trees of Rule 2 are shown in Fig. 4.4.

The result of using the algorithm of Rule 2 is shown in Table 4.3. The *correct* rates for the test data are 82%, 77% and 87%, respectively. Though Rule 2 improves its predecessor's performance, the result, however, still discourages us from using it for the generation of zero anaphors in Chinese.

As shown in [Grosz & Sidner 86], the structure of discourse is a significant factor affecting the use of anaphoric forms. Thus in the following, we employed the notion of discourse structure as the basis for enhancing the rule.

Table 4.3: Result of using algorithm of Rule 2 on the test data.

| Data | Alg. | *Correct* | *False* | *Missing* | Total |
|---|---|---|---|---|---|
| Set 1 | Rule 2 | 425 | 89 | 5 | 519 |
| | % | 82 | 17 | 1 | 100 |
| Set 2 | Rule 2 | 499 | 134 | 10 | 643 |
| | % | 77 | 21 | 2 | 100 |
| Set 3 | Rule 2 | 128 | 18 | 1 | 147 |
| | % | 87 | 12 | 1 | 100 |

### 4.3.3 The effect of adding discourse structure

An important idea in Grosz and Sidner's discourse structure theory, as introduced in Section 3.4, is the mutual effect between the linguistic expressions in utterances constituting the discourse and the discourse segment structure. In natural language generation systems, the semantic structures of messages to be produced are usually organised according to their hierarchical intentional structures; then, based on the structures, referring expressions are decided [Dale 92, Hovy 93]. Hence, here, we employ the idea of discourse structure to improve our algorithm for the generation of zero anaphors.

Li and Thompson, in their study [Li & Thompson 79], introduced in Section 2.3, proposed the idea that the use of non-zero anaphors has to do with the segment boundaries in a discourse. For example, in (2), sentences *b* to *d* describe the appearance of a person who enters a room, while the following sentences, *e* to *g*, provide background information about that person. Thus a pronoun ta (he/she) rather than a zero anaphor is used to "highlight" the shift of intention [Li & Thompson 81], and hence the beginning of a discourse segment.

(2) a. waibian jinlai le yige ren$^i$,
outside enter ASPECT one person
From outside came a person.

b. $\phi_1^i$ liangge hong yanjing,
(he) two red eye
(He) had two red eyes.

c. $\phi_2^i$ yifu da yuan lian,

(he) a big round face

(He) had one big round face.

d. $\phi_3^i$ dai zhe yiding xiao maozi,

(he) wear ASPECT a small hat

(He) was wearing a small hat.

e. ta$^i$ xing Xia,

he surname Xia

He had the surname Xia.

f. $\phi_4^i$ ershiwu sui,

(he) this-year 25 year

(He) (was) 25 years old.

g. $\phi_5^i$ zhu taibei.

(he) live-in taibei

(He) lived in Taipei.

In general, a zero anaphor used to refer to some entity in the previous sentence might be expected to indicate the continuation of a discourse segment, while a non-zero anaphor occurring in the same situation signals a boundary of a discourse segment. From the generator's perspective, when the decision about the anaphoric form for a phrase referring to some entity in the previous sentence is to be made, the factor of discourse segment boundaries must be taken into consideration. Therefore, based on this idea, we improve the previous rules for generation of zero anaphors, to make the following rule. The corresponding decision tree is shown in Fig. 4.5.

> **Rule 3:** If an entity, $e$, in the current sentence $u$ was referred to in the immediately preceding sentence and does not violate any syntactic constraints on zero anaphora, then if $u$ is not the beginning of a discourse segment, then a zero anaphor is used for $e$; otherwise, a non-zero anaphor is used.

To perform the experiments for the new rule, we have to access the discourse segment structures of the test data. Therefore, we annotated the boundaries between discourse segments in the test data and the hierarchical discourse structures according to the discourse segment intentions. Since the above annotations were based on intuition,

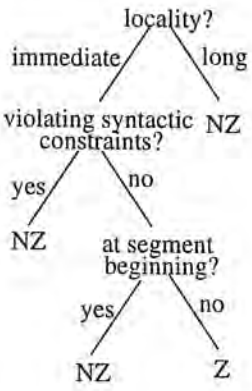



Figure 4.5: Decision tree for Rule 3.

we further carried out a test by comparing our annotations with those of other native speakers of Chinese to see whether our intuitions about the discourse structures of the test data were reliable for the purpose of the experiments. In the test, four native speakers of Chinese were asked to annotate discourse segment boundaries for five articles selected from the test data. Each speaker was given a short description in Chinese about the idea of discourse structure and the task to be done, namely annotate the discourse segment boundaries and the hierarchical structures according to the intentions of the discourse segments. Compared with the speakers' results, 84% of our annotation markers on average match those of the speakers. There are 16% of our markers not occurring in the speakers' results; on the other hand, 8% of the speakers' markers do not occur in our annotations. From the above comparisons the annotations we made were highly reliable for the purpose of the experiment. For a detailed description and the results of the test, please refer to Appendix A. The result in Appendix A shows that sentential marks in the test texts closely correlate to the boundaries between discourse segments. In analysing the test data, we have assumed the beginning of a discourse segment to be where a sentential mark occurs. To illustrate the discourse segmentation, we show in Fig. 4.6 an example of a discourse segment structure selected from the test data.

We then performed the experiment by employing the algorithm of Rule 3. The classification trees and results of the experiment are shown in Fig. 4.7 and Table 4.4, respectively. By taking into account the effect of discourse segment structure, the al-

a. xia, pangxie[1] de zhonglei hen duo,
b. $\phi^1$ shenshang dou you gezhong butong de yanse,
c. keshi $\phi^1$ jingguo $\phi$ zhengzhu yihou,
d. $\phi^1$ shenti biaomian dou hui biancheng hongse le.
e. zhe shi weishenme ne?
f. yuanlai zai xia, xie jiake xiamian de biaopizhong you yizhong yu re hui bei fenjie chulai de dongxi[2],
g. $\phi^2$ ming jiao "xiahongsu"[3],
h. ta[2] yu qita sesu buyiyang,
i. $\phi^2$ bu pa gaowen,
j. $\phi^2$ yudao gaowen,
k. ta[2] jiu cong $\phi^1$ jiakezhong chendian chulai,
l. $\phi^2$ bing xianchu feichang xianyan de hongse.
m. youyu "xiahongsu"[3] neng rongjie zai youzhong,
n. ru $\phi$ yong you[4] peng xia she,
o. you[4] jiu biancheng yin ren xiai de chenghongse le.

**Translation**

a. Both shrimps and crabs have many species.
b. Their bodies have various colours.
c. However, after they are cooked,
d. their body surfaces will all become red.
e. Why is this?
f. On the underside of the shell of both shrimps and crabs, there is a red stuff which will permeate to the surface if heated.
g. It is named "Xiahongsu".
h. It is different from other sorts of pigment.
i. It is able to endure high temperatures.
j. Upon being exposed to high temperatures,
k. it then exudes from the shell,
l. and displays a very bright red colour.
m. Since "Xiahongsu" can be dissolved in oil,
n. if we use oil to cook shrimps,
o. the oil will also become an attractively red colour.

Figure 4.6: An example of discourse segment structure from the test data.

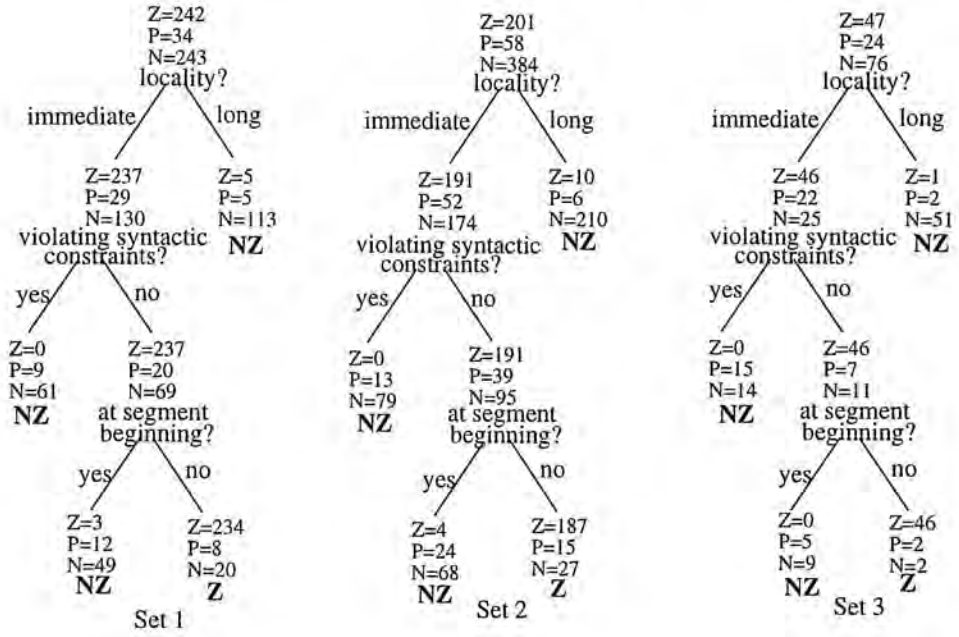



Figure 4.7: Classification trees of Rule 3.

Table 4.4: Result of using algorithm of Rule 3 on the test data.

| Data | Alg. | *Correct* | *False* | *Missing* | Total |
|---|---|---|---|---|---|
| Set 1 | Rule 3 | 483 | 28 | 8 | 519 |
| | % | 93 | 5 | 2 | 100 |
| Set 2 | Rule 3 | 587 | 42 | 14 | 643 |
| | % | 91 | 7 | 2 | 100 |
| Set 3 | Rule 3 | 142 | 4 | 1 | 147 |
| | % | 97 | 3 | 1 | 101 |

gorithm of Rule 3 obtained 93%, 91% and 97% *correct* matches in the test data. The result shows that Rule 3 is helpful for the decision as to whether to use a zero anaphor. These figures also reveal the fact that discourse segment structure is an important factor in deciding the use of zero anaphors in Chinese. Also as shown in Fig. 4.7, 3, 4 and 0 cases were under-generated by the new algorithm. This phenomenon indicates that discourse structure is not an absolute factor affecting the use of zero anaphors.

### 4.3.4 The effect of topic

Although the zero anaphors generated by the algorithm of Rule 3 look considerably similar to those in the test data, there are, nevertheless, still a number of over-generations for Sets 1 and 2 of the test data, namely, the non-zero anaphors associated with the zero leaf nodes in their classification trees. Here, we use the notion of topic in Chinese, introduced in Section 2.2, to further refine the previous rule.

The basic idea here is to investigate the positions of antecedent and anaphor in their respective sentences. Then we observe the occurrence of both the antecedent and anaphor in the topic position to see the effect of topic on zero anaphora. In the following, we divided the position of anaphors in their respective sentences into topic and non-topic cases. For each anaphor, its antecedent's position is classified as one of the following categories: either an entire or part of a topic phrase, for example, the antecedent $\phi^1$ in sentence b in Fig. 4.6 is part of the topic phrase in sentence a, a direct object or the NP following a presentative verb, such as yige ren (a person) in (2) and others. Thus we have the following types of antecedent-anaphor pairs.

| | **Position of antecedent** | **Position of anaphor** |
|---|---|---|
| **Type A:** | whole or part of topic | topic |
| **Type B:** | object-1 or NP following pres. verb | topic |
| **Type C:** | whole or part of topic | non-topic |
| **Type D:** | object-1 or NP following pres. verb | non-topic |
| **Type E:** | others | topic |
| **Type F:** | others | non-topic |

Since in the new rule the condition of topic in utterance will be considered to refine the zero leaf node in the decision tree of Rule 3, we focus on investigating the corresponding anaphors in the classification trees. According to the above classification, the numbers of various antecedent-anaphor pairs in the test data are shown in Table 4.5. Obviously, for columns A and B in the table, non-zero cases, namely the sums of pronouns and nominals, are in the minority for all sets of data. Chen [Chen 87] found a higher percentage of zero anaphors occurring in the topic position with their antecedent most frequently in the topic or object positions of the immediately previ-

Table 4.5: Occurrence of types of antecedent-anaphor pairs in the test data.

| Data | Type | A | B | C | D | E | F | Total |
|---|---|---|---|---|---|---|---|---|
| Set 1 | Z | 200 | 24 | 1 | 3 | 0 | 6 | 234 |
| | P | 2 | 1 | 5 | 0 | 0 | 0 | 8 |
| | N | 6 | 3 | 1 | 1 | 9 | 0 | 20 |
| Set 2 | Z | 159 | 23 | 3 | 0 | 0 | 2 | 187 |
| | P | 6 | 2 | 6 | 0 | 1 | 0 | 15 |
| | N | 2 | 8 | 1 | 0 | 16 | 0 | 27 |
| Set 3 | Z | 44 | 0 | 1 | 0 | 0 | 1 | 46 |
| | P | 2 | 0 | 0 | 0 | 0 | 0 | 2 |
| | N | 2 | 0 | 0 | 0 | 0 | 0 | 2 |

ous sentence, which strongly supports the idea of letting anaphors of Types A and B be zero. Zero anaphors of Types A and B are generally understood because they are salient [Li & Thompson 81]. Types C to F of anaphors are not so salient as Types A and B; thus we group Types C to F as being not salient. The total numbers of zero cases for the non-salient type are 10, 5 and 2 in Sets 1, 2 and 3 data, respectively; the total numbers of non-zeros for the same type are 16, 24 and 2. Thus we let anaphors of the non-salient type be non-zero. We chose zero for Types A and B and non-zero for other types, which resulted in a new rule, Rule 4, as below and its decision tree is shown in Fig. 4.8.

> **Rule 4:** If an entity, $e$, in the current sentence $u$, was referred to in the immediately preceding sentence, then if $e$ violates some syntactic constraint for zero anaphora, then a nominal or pronominal anaphor is used for $e$; otherwise, if $u$ is at the beginning of a discourse segment, then a pronominal or nominal anaphor is used; otherwise, if $e$ is salient then a zero anaphor is used for $e$; otherwise, a non-zero anaphor is used.

The classification trees of Rule 4 are shown in Fig. 4.9 and the result along with previous ones are shown in Table 4.6. As shown in the table, the *correct* rates of Sets 1 and 2 increased from 93% and 91% to 94% and 94%, while Set 3 decreased from 97% to 95%.


Figure 4.8: Decision tree for Rule 4.

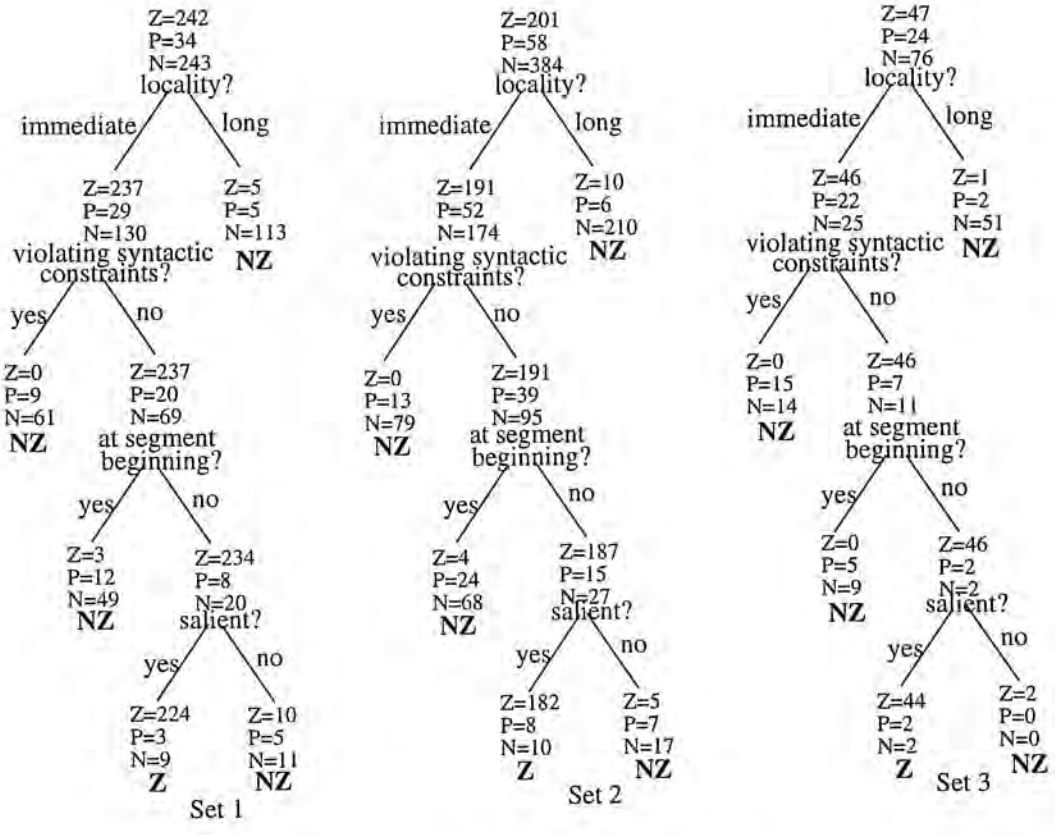



Figure 4.9: Classification trees for Rule 4.

Table 4.6: Results of using algorithms of Rules 1, 2, 3 and 4 on the test data.

| Data | Alg. | *Correct* | *False* | *Missing* | Total |
|---|---|---|---|---|---|
| Set 1 | Rule 1 | 355 | 159 | 5 | 519 |
| | % | 68 | 31 | 1 | 100 |
| | Rule 2 | 425 | 89 | 5 | 519 |
| | % | 82 | 17 | 1 | 100 |
| | Rule 3 | 483 | 28 | 8 | 519 |
| | % | 93 | 5 | 2 | 100 |
| | Rule 4 | 489 | 12 | 18 | 519 |
| | % | 94 | 2 | 3 | 99 |
| Set 2 | Rule 1 | 407 | 226 | 10 | 643 |
| | % | 63 | 35 | 2 | 100 |
| | Rule 2 | 499 | 134 | 10 | 643 |
| | % | 77 | 21 | 2 | 100 |
| | Rule 3 | 587 | 42 | 14 | 643 |
| | % | 91 | 7 | 2 | 100 |
| | Rule 4 | 606 | 18 | 19 | 643 |
| | % | 94 | 3 | 3 | 100 |
| Set 3 | Rule 1 | 99 | 47 | 1 | 147 |
| | % | 67 | 32 | 1 | 100 |
| | Rule 2 | 128 | 18 | 1 | 147 |
| | % | 87 | 12 | 1 | 100 |
| | Rule 3 | 142 | 4 | 1 | 147 |
| | % | 97 | 3 | 1 | 101 |
| | Rule 4 | 140 | 4 | 3 | 147 |
| | % | 95 | 3 | 2 | 100 |

## 4.4 Discussion

The experimental result shows a promising performance of Rule 4 for generating zero anaphors. Although this rule contains various conditions, they, however, do not account for the general principle, as stated at the beginning of this paper, that zero anaphors can occur in any grammatical slot with an antecedent that may occur in any grammatical slot, regardless of the distance between them [Li & Thompson 79]. Specifically, according to the rule, the following kinds of zero anaphors would not be generated, although they could possibly occur in real texts. First, long distance zero anaphors are missing. Long distance zero anaphors can be explained linguistically by employing some sort of discourse structure theory, like Grosz and Sidner's discourse structure theory [Grosz & Sidner 86] and Rhetorical Structure Theory [Mann & Thompson 87, Chen 87]. We need further investigation to have the above idea realized in natural language generation systems [Dale 92, Hovy 93].

Second, by using Rule 4, zero anaphors would not occur at the beginning of discourse segments. From the experiment in Section 4.3, just a few cases of zero anaphors in the test data occur at the beginning of discourse segments. These cases may result from a mismatch between our discourse segment boundaries and those of the authors.

Finally, by using Rule 4, zero anaphors would not occur in non-salient positions within sentences. As described in Section 4.4, the difference between the number of zero and non-zero anaphors in this case is not so significant as in other cases. The Set 1 data, in particular, had 10 and 16 zero and non-zero anaphors satisfying the non-salience condition, respectively. We therefore, carried out a test to investigate the preference for zero or non-zero forms in this situation. We asked five native speakers of Chinese, on the basis of their intuition, to give their preference on these cases. Basically, for each case, they were asked to choose one of zero, non-zero and both equally good. The result of the test showed that the speakers to some extent agree with Rule 4. The detailed description and results are shown in Appendix B.

In addition to the under-generation on zero anaphors described above, Rule 4 would result in over-generations of zero anaphors as well. As stated in Rule 4, salient anaphors will be zeroed. The experiment in Section 4.4 showed that there were some non-zero

cases in the test data against the salience condition, for example, ta (it) of sentence $k$ in Fig. 4.6, repeated below.

...

h.ta$^2$ yu qita sesu buyiyang,
  it and other pigment different
  It is different from other pigments.
i. $\phi^2$ bu pa gaowen,
  (it) not afraid high-temperature
  (It) is is able to endure high temperature.
j. $\phi^2$ yudao gaowen,
  (it) meet high temperature
  Upon being exposed to high temperature
k.ta$^2$ jiu cong $\phi^1$ jiakezhong chendian chulai,
  it then from (their) shell-in exude come
  It then exudes from the shells,
l. $\phi^2$ bing xianchu feichang xianyan de hongse.
  (it) and display very bright NOM red-colour
  and (it) displays a very bright red colour.
  ...

We are not certain what causes these non-zero cases to occur. However, from the ratios between zero and non-zero anaphors in the test data, zero anaphors are the normal cases under the salience condition.

## 4.5 Summary

By doing experiments on a number of descriptive articles, we obtained a rule for the generation of zero anaphors, which incorporates the ideas of locality between anaphor and antecedent, syntactic constraints, discourse segment structure and salience of objects in discourse. In the text generated by hand employing the algorithm of the above rule, assuming the same semantic structure and discourse segment structure as the real text, the use of zero anaphors is very close to those occurring in the real text.

In a stepwise empirical study, the algorithms are improved through considering real data, which provides an assessment of the effectiveness of the rule. The result of the assessment thus encourages us to employ the rule as a part of the referring expression component in the Chinese natural language generation system we are developing.

# Chapter 5

# Towards the Generation of All Kinds of Anaphors

## 5.1 Introduction

In the preceding chapter, we established a rule for the use of zero anaphors. In this chapter, we attempt to refine the non-zero parts of the rule to account for pronouns and nominal anaphors. What we intend to do is find constraints which can distinguish between the use of pronominal and nominal anaphors. We carried out experiments using the new rules to investigate their performance. According to our survey, as shown in Section 3.5, previous natural language generation work on the use of pronouns can not be applied for our purpose because it is either language- or domain-specific [McDonald 80, Dale 92]. Former linguistic studies, as surveyed in Section 2.5, provides some idea for the selection of pronouns in discourse. In the following, we will encode some constraints based on the idea to refine our rule.

## 5.2 Refinements on the Zero Anaphor Generation Rule

To distinguish between pronouns and nominal anaphors, we carried out a similar experiment to that described in the preceding chapter. We first performed an experiment with a rule letting non-zeros in the previous rule always be nominal. Then we repeated the experiment by appending an animacy constraint to the non-zero leaf nodes in the previous rule. Finally, we discuss the effectiveness of the new rule.

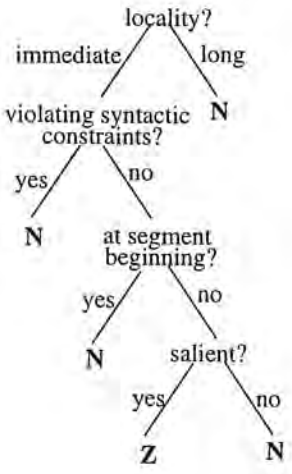



Figure 5.1: Decision tree for Rule 5.

### 5.2.1 Using nominal forms for non-zeros

As shown in the classification trees of Rule 4 in Fig. 4.9, pronouns are in the minority of the non-zeros in the test data. A simple way to refine the previous anaphor generation rule is to let the non-zero parts in the rule be nominal, which resulted in a new rule, Rule 5, with decision tree shown in Fig. 5.1. The classification trees of Rule 5 are shown in Fig. 5.2.

To demonstrate the result of using the new rule, we extended the definition of *correct*, *false* and *missing* matches used previously for zero and non-zero anaphors to zero, pronominal and nominal anaphors. The numbers of *correct* matches for zero, pronoun and nominal in the test data can be obtained by summing up anaphors associated with the leaf nodes labelled $Z$, $P$ and $N$ in the classification trees, respectively. The *false* matches of zero anaphors, for instance, are the sum of non-zero anaphors associated with the leaf nodes labelled Z in the classification trees. Conversely, the *missing* matches of zero anaphors, for instance, are the sum of zero anaphors associated with the leaf nodes labelled with non-zeros. The *false* and *missing* matches of pronouns and nominals can be obtained in a similar way. For distinguishing purpose, hereafter, we use *overall correct*, *overall false* and *overall missing* matches to denote the total numbers of *correct*, *false* and *missing* matches for all anaphors. According to the new definition, the result from using Rule 5 is shown in Table 5.1. Note that both the num-

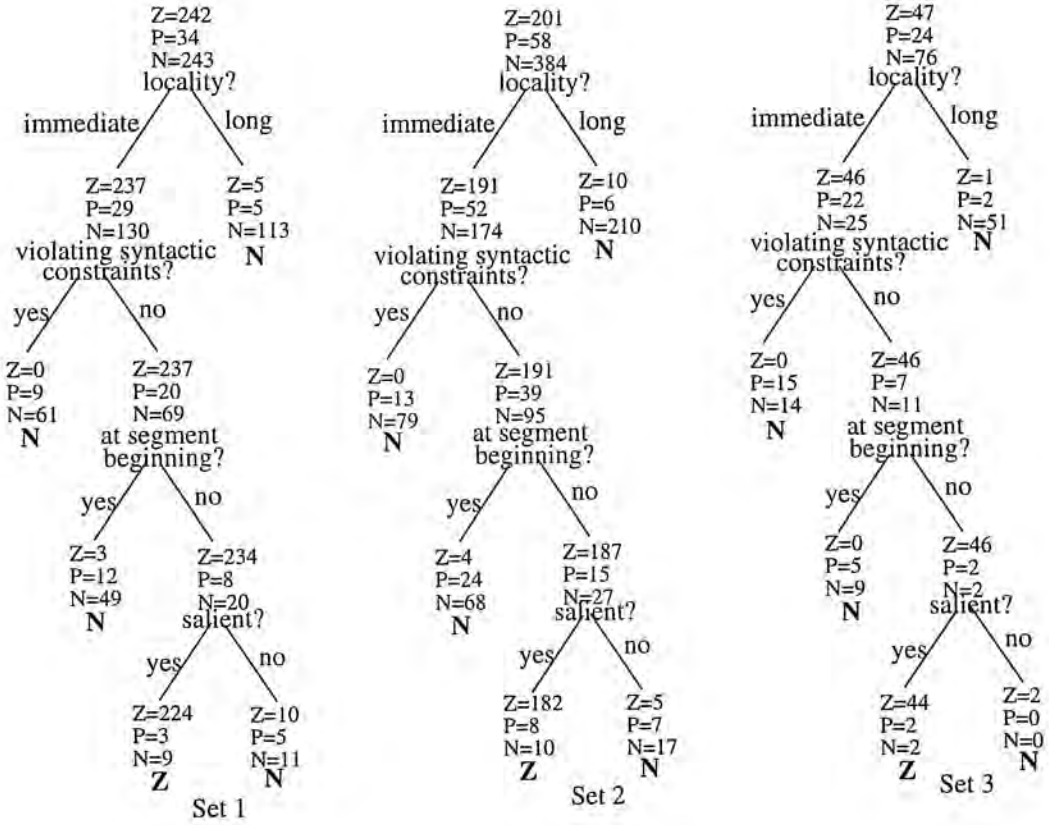



Figure 5.2: Classification trees for Rule 5.

bers of matches and the percentages with respect to the total numbers of anaphors in the test data are given in the table. For example, there are 224 *correct* matches of zero anaphors for the Set 1 data, and the total number of anaphors in the same set of data is 519. Thus the percentage of the *correct* matches is 43% (224/519). The percentage of *overall correct* matches for the Set 1 data, for example, is the total percentages of all anaphors which is 88% (43%+45%).

Table 5.1: Result of using Rule 5 on the test data.

| | | *Correct* | | | *False* | | | *Missing* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Data | Alg. | Z | P | N | Z | P | N | Z | P | N | Total anaphors |
| Set 1 | Rule 5 | 224 | 0 | 234 | 12 | 0 | 49 | 18 | 34 | 9 | 519 |
| | % | 43 | 0 | 45 | 2 | 0 | 9 | 3 | 7 | 2 | |
| Set 2 | Rule 5 | 182 | 0 | 374 | 18 | 0 | 69 | 19 | 58 | 10 | 643 |
| | % | 28 | 0 | 58 | 3 | 0 | 11 | 3 | 9 | 2 | |
| Set 3 | Rule 5 | 44 | 0 | 74 | 4 | 0 | 25 | 3 | 24 | 2 | 147 |
| | % | 30 | 0 | 50 | 3 | 0 | 17 | 2 | 26 | 2 | |

The percentages of overall *correct* matches are 88%, 86% and 80%, respectively, which is quite promising. However, according to this rule, no pronouns would occur in the generated texts at all. Although pronouns are referred to as a special cases of zero anaphors and occur rarely as shown in the test data, they should, however, not be ignored from discourse. For example, in (1), if pronouns in b, c and g are all replaced by nominal anaphors, such as nage yingguoren (that Briton), then it would result in incoherent discourse.

(1)a.jiaru you yige yingguoren$^i$ gen ni xue zhongguo yufa,
if have a Briton from you learn Chinese grammar
Suppose that a Briton learns Chinese grammar from you.
b.ni gaosu ta$^i$, "ma" shi mingci, "pao" shi dongci,
you tell he horse is noun run is verb
You tell him that "horse" is a noun, and "run" is a verb.
c.$\phi$ you gaosu ta$^i$,
(you) furthermore tell he
Furthermore (you) tell him that
d.zai "gou yao Lu Dongbin" zheju hua li,
in "dog bite Ludongbin" this word inside

inside the sentence "dog bite Ludongbin",

e. "gou" shi zhuyu, "yao" shi dongci, "Ludongbin" shi mudiwei,
dog is subject bite is verb Ludongbin is object
"dog" is the subject, "bite" the verb, and "Ludongbin" the object.

f. nage yingguoren$^i$ yiding hen shiwang,
that Briton certainly very disappointed
That Briton must be very disappointed.

...

g.ni bingmeiyou ba zhongguoyu de jiegou fangshi gausu ta$^i$.
you do-not BA Chinese-language ASSOC structure pattern tell he
You didn't tell him the structure pattern of Chinese.

### 5.2.2 The effect of animacy on pronominal encoding

As described in Section 2.5, previous linguistic studies [Li & Thompson 79, Chen 86] showed that pronouns are frequently used when the anaphors occur at places marked as minor discontinuities and when referring to things that are highly noteworthy. The conditions of minor discontinuity were not clearly stated, and individual judgements on this are likely to vary. Thus we will not take it as a constraint to further refine our rule. As for the other discourse factor, high noteworthiness, the animacy condition can be determined according to the features of the referent. In an examination on inanimate anaphors in his test data, Chen [Chen 86] found that there were only a few instances of pronouns. On the other hand, the percentage of inanimate anaphors being encoded in nominal forms is higher than that of pronouns. Thus, according to Chen's results, we employ the animacy of the referent as a constraint, as shown below, to refine our rule.

> If a non-zero anaphor is animate, then it is pronominalised; otherwise, it is nominalised.

In the classification trees in Fig. 4.9, nominal anaphors obviously dominate others in the long distance cases; thus we chose nominal forms for all long distance anaphors. Therefore we attached the animacy constraint to the remaining NZ leaf nodes in Rule 4,

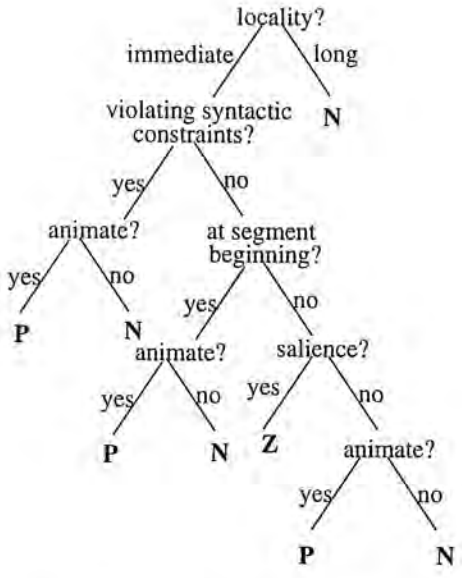



Figure 5.3: Decision tree for Rule 6.

as shown in Fig. 4.8, and obtained a new rule, Rule 6, as shown in Fig. 5.3. In general, animate objects characterise living things, especially animal life. In this experiment, we adopted this concept to determine the animacy of anaphors. The resulting classification trees for the test data are shown in Fig. 5.4.

The result of using Rule 6 is shown in Table 5.2. The overall *correct* rates increased from 88%, 86% and 80% to 89%, 88% and 84%. Although the increase of overall *correct* rates was not significant, 44 (15/34), 41 (24/58) and 17% (4/24) of the pronouns in the test data, however, were correctly matched by using the new rule. More specifically, among the 26, 44 and 20 pronouns we examined in this experiment, 15, 24 and 4 were *correctly* matched by involving the animacy constraint; in other words, 11, 20 and 16 cases were missed out by the same constraint. On the other hand, 10, 15 and 1 pronouns would be over-generated by using the animacy constraint, namely, the *false* matches of pronouns. The above figures show that the animacy constraint, though effective to some extent, still malfunctions in some situations.

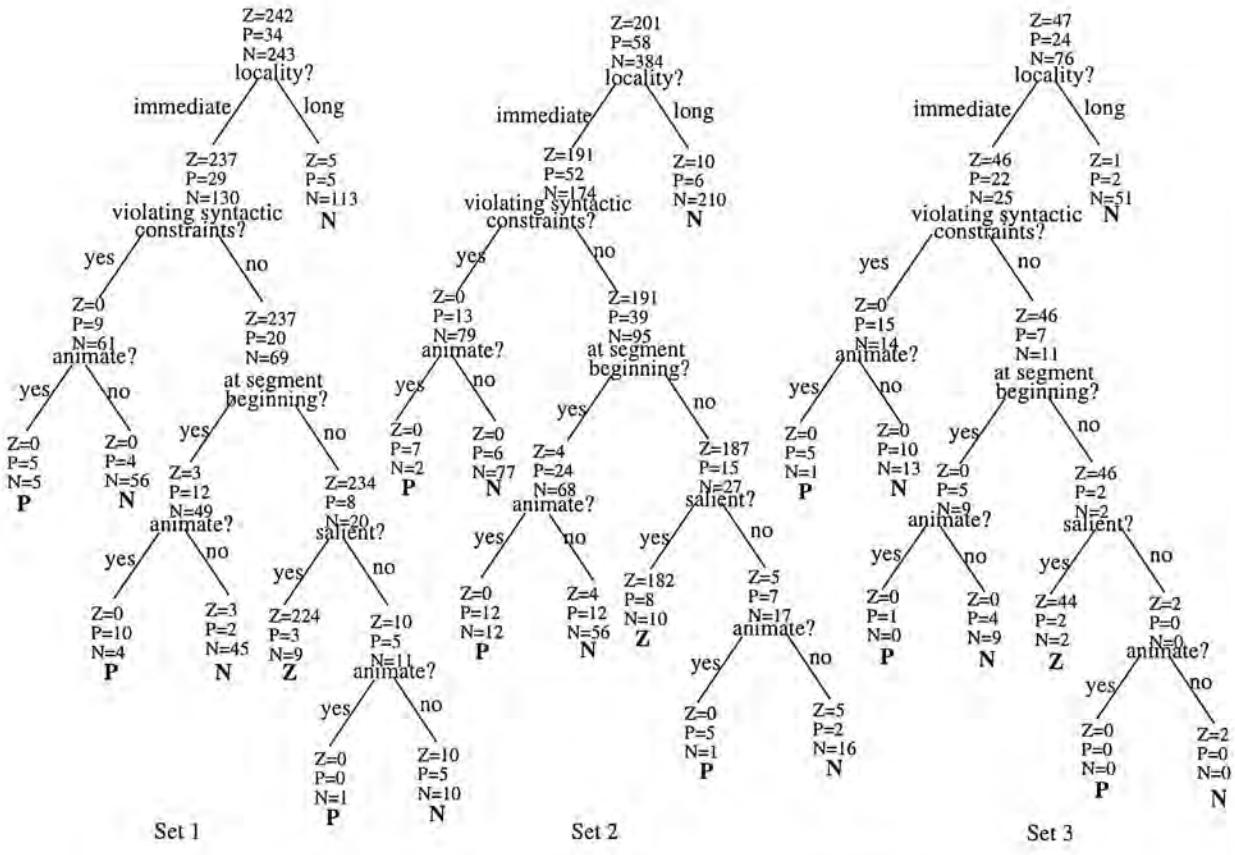



Figure 5.4: Classification trees of the test data using Rule 6.

Table 5.2: Result of using Rule 6 on the test data.

| | | *Correct* | | | *False* | | | *Missing* | | | Total anaphors |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Data | Alg. | Z | P | N | Z | P | N | Z | P | N | |
| Set 1 | Rule 6 | 224 | 15 | 224 | 12 | 10 | 34 | 18 | 19 | 19 | 519 |
| | % | 43 | 3 | 43 | 2 | 2 | 6 | 3 | 4 | 4 | |
| Set 2 | Rule 6 | 182 | 24 | 359 | 18 | 15 | 45 | 19 | 34 | 25 | 643 |
| | % | 28 | 4 | 56 | 3 | 2 | 7 | 3 | 5 | 4 | |
| Set 3 | Rule 6 | 44 | 6 | 73 | 4 | 1 | 19 | 3 | 18 | 3 | 147 |
| | % | 30 | 4 | 50 | 3 | 1 | 13 | 2 | 12 | 2 | |

### 5.2.3 Problems with using the animacy constraint

Here we focus on examining the pronouns and nominal anaphors in the test data that the animacy constraint in Rule 6 fails to account for, namely, the nominal anaphors under the animate nodes in the classification trees and the pronouns under the inanimate nodes in the trees.

First, we investigate the *false* matches of pronouns, namely, the over-generation of pronouns. There were 10, 15 and 1 cases of over-generation of pronouns resulting from the animacy constraint all of which were nominal anaphors in the test data. These cases are animate objects according to our criteria; the authors of the test data, however, chose to use nominal forms instead of pronouns as the animacy constraint expects. Observing the test data, we found that these nominal anaphors occur in sequences of anaphoric references where pronominal and nominal forms are mixed up together. For example, in the sequence of references to pangxie (crab) in (2), it occurs as pronouns in b, e and i and as nominal forms in d, g and l.

(2)a.ba hepangxie$^i$ fangzai penzili,
BA river-crab put-in basin-in
Put river-crabs in a basin.

b.dai yihuir, $\phi^i$ jiu hui cong ta$^i$ de zuibian tuchu xuduo xiao paopao.
wait a-while (it) then will from it ASSOC mouth-side vomit-out much small bubble
After a while, (it) will expel lots of bubbles out of a side of its mouth.

c.zhe shi zenme huishi ne?
this is why matter
Why does this happen?

d.hepangxie$^i$ shi shenghuo zai shuili de dongwu,
river-crab is live in water-in ASSOC animal
River-crab is an animal which lives in water.

e.ta$^i$ han yu yiyang,
it and fish similar
It is similar to fish.

f. $\phi^i$ yong sai zai shuizhong huxi.

(it) use gill in water-in breathe

(it) uses gill to breathe in water.

g.hepangxie$^i$ bu guang neng zai shuili shenghuo,

river-crab not only can in water-in live

River-crab not only can live in water,

h.$\phi^i$ hai keyi padao ludishang laizhao shiwu chi.

(it) also can crawl-to land-above come-find food eat

but also can crawl onto land to find food to eat.

i. buguo, ta$^i$ shang an yihou,

however it up shore after

However, after it reaches the shore,

j. $\phi^i$ sai jiu hui bian ganzao,

(it) gill then will become dry

(its) gills become dry.

k.$\phi^i$ huxi bu fangbian.

(it) breathe not convenient

(It) makes breathing inconvenient.

l. zheyang, hepangxie$^i$ jiu ba cunzai saili de shui han huchu de kongqi$^j$

so river-crab then BA exist gill-in ASSOC water and breathe-out NOM air

So river-crab expels the water existing in its gill and the air breathing out

cong $\phi^i$ zui de liangbian tu chulai,

from (it) mouth ASSOC two-side vomit out-come

from both sides of its mouth.

m$\phi^j$ xingcheng xuduo xiao paopao$^k$,

(they) form many small bubbles

(They) form lots of bubbles.

n.$\phi^k$ dui $\phi^i$ zai zui de mianqian.

(they) heap-up in mouth ASSOC front

(They) gather in front of its mouth.

A similar situation occurs in the *false* matches, namely, in the under-generation of pronouns. For example, in (3), both pronouns in b and c are *missing* matches; in

other words, they are nominalised by using Rule 6. In f, the referent yueliang occurs again, with its antecedent occurring immediately in e. However, the author chooses the nominal form for this reference instead of pronoun like the similar reference in c.

(3)a.yueliang$^i$ gaogao gua zai tianshang,
moon high hang at sky
The moon hangs high in the sky.

b.ta$^i$ li dimian henyuan henyuan,
it away ground very-far very-far
It is far far away from the earth.

c.meiyou shenme dongxi keyi dangzhu ta$^i$.
no-have what thing can block it
Nothing can block it.

d.suoyi $\phi$ buguan zou dao nar,
therefore (we) no-matter walk to where
Therefore, no matter where (we) go,

e.$\phi$ zong neng kanjian yueliang$^i$
(we) always can see moon
(we) can always see the moon.

f. zheyang women jiu hui juede yueliang$^i$ zai genzhe ren zou.
this-way we then will feel moon at follow person go
Thus we feel that moon is following us.

From the above investigation, we found that in these *missing* and *false* matches of pronouns either pronominal or nominal form is acceptable. The use of pronouns in these places to some extent has to do with the writer's own judgement and style. Following our previous work, the next step is to find more computationally tractable constraints to refine the rule so as to make the generated text as close as possible to the real text. We are, however, not clear how to obtain the appropriate constraints because they seem to involve the writer's intention, world knowledge, etc., which are beyond the scope of our study. Furthermore, because of divergences between different writers, it makes little sense to establish a rule to serve as a general account for all writers. Therefore, we halted the pursuit of further refinements to the rule.

The rule developed here, Rule 6, as shown in Fig. 5.3, is to serve as the decision part of the referring expression component in Chinese natural language generation systems which in a sense plays the role of a writer. When an anaphor is to be created in the generation of text, the system consults the rule to decide an appropriate form for it. As the rule suggests, certain kinds of forms are used for anaphors satisfying specific constraints. Though it generates most cases appropriately, however, the resulting forms in some situations are not appropriate. For example, in (4), the pronoun in d is preferred over the nominal form generated by the rule, even though it is inanimate in nature.

(4)a.you yizhong mianji jiao da de huoshankou$^i$ jiao pohuoshankou,
have a-kind area more large ASS crater name broken-crater
There is a kind of crater with a larger area called broken crater.

b.$\phi^i$ zhuang ru pendi,
(it) shape like basin
(It) is shaped like a basin.

c.$\phi^i$ lue cheng yuanxing,
(it) about present round-shape
(It) presents round shape.

d.ta$^i$ de banjing bi yiban huoshankou da shubei zhi shushibei,
it ASS radius compare ordinary crater large several-times to several-ten-times
Its radius is several to tens of times larger than ordinary craters.

...

Thus, to make the generated text more natural, we propose that Chinese generation systems have their own extra judgement and style in addition to the basic rule described previously in the decision of anaphoric forms. In the following section, we propose a way to organise these extra heuristics.

## 5.3 Approach to Accounting for Personal Style

The personal style that a writer attempts to impose on the decision of anaphoric forms may originate from various sources. In this thesis, we do not aim at a full account for

the above phenomenon; what we are concerned with is to obtain additional constraints to feed Rule 6, so that variety of resulting anaphoric forms in the generated text is gained, while appropriateness is retained.

Chao, in [Chao 68], has noted that

> "··· the difference between ta (s/he, it) as a pronoun for animate objects and inanimate objects (···) is a matter of the speaker's use of the word as having animate or inanimate reference and not a matter of the biological or physical nature of the referent."

This suggests that the animacy constraint we used in the experiment could be further extended to involve our own idea on animacy. After re-examining the criterion for determining the animacy of objects described previously, the criterion involving personal judgement could be easily obtained, if the investigation rests on a specific domain and genre. In fact, most of the existing natural language generation systems were built on specific domains [McKeown 85, Dale 92, Hovy 93]; the Chinese natural generation system we are developing is also domain-specific. Therefore, we propose to extend the above criterion by adding some domain-specific constraints. In our generation system, described later in the chapter on implementation , for example, the domain knowledge base consists of inanimate objects in a national park, including plants, trails, lakes, craters, mountains, etc. When a new inanimate object is entered into the domain knowledge base, we annotate it as animate or inanimate. This requires some heuristics to aid the determination of animacy of objects. One heuristic, for example, is the personification of an object [Chao 68]. In real text, we can find evidence of similar ideas. For example, some features of a tube rose are described in (5); the writer used pronouns in c, g, h and i to refer to it.

(5)a.huaer chabuduo dou zai baitian kaifang,
flower almost all at daytime open
Almost all flowers blossom in the daytime.
b.yelaixiang$^{i}$ que butong,
tube-rose but different
But tube rose is different.

c.ta$^i$ zai bangwan kai hua,

it at evening open flower

It blossoms in the evening.

d.yue dao shenye $\phi^i$ fachu de xiangwei yue nong.

more till deep-night it spread NOM perfume more strong

The deeper in the night, the stronger perfume it spreads.

...

e.bangwan yelaixiang$^i$ kaifang,

evening tube rose open

Tube rose blossoms in the evening.

f. keshi hudie mifeng dou bu zai chulai le,

but butterfly bee all not again come-out ASP

But butterfly and bee don't come out.

g.ta$^i$ zhiyou shanfachu qianglie de xiangwei,

it only spread-out stronger NOM perfume

It only spreads out stronger perfume,

h.caineng jingyou naxie wanshang chulai huodong de feie dao ta$^i$ zheli lai,

can attract those evening come-out work NOM moths come it here come

so that it can attract those moth coming out in the evening close to it,

i. bang ta$^i$ chuanshou huafen.

help it propagate pollen

and help it to propagate pollen.

j. zheyang yelaixiang$^i$ ...

this-way tube rose ...

Thus tube rose ...

As shown in Chapter 4, non-zero anaphors, namely, pronouns and nominal anaphors, are commonly used to indicate the beginning of discourse segments in a sequence of references to the same object in the topic positions. Furthermore, as shown in [Chen 86, Grosz & Sidner 86], pronouns and nominal anaphors in turn are used to signal different types of discourse segments. For example, in (6), nominal anaphors referring to crab are used in d and f, while pronouns are used in c, e and h. The above

nominal anaphors in (6) all occur at the beginning of the "sentences", while pronouns occur at the beginning of discourse segments within "sentences".

(6)a.pangxie$^i$ ping zhe $\phi^i$ diyidui chujiaoshang de ganjuemao han
crab rely-on ASP (it) first-pair feeler-on ASSOC sensing-hair and
Crab rely on the sensing-hair on the first pair of feelers and a pair of
yidui yanjing zai shuizhong zhao yu, xia chi,
a-pair eyes in water-in find fish prawn eat
eyes to find fish and prawns to eat in the water .

b.anshang ru you shiwu,
shore-above if have food
If there is food on shore,

c.ta$^i$ yeyao qiang,
(it) also-want rob
(it) robs too.
... (omitted text terminating in a full stop)

d.pangxie$^i$ weishenme yao hengzhe pa ne?
crab why want cross-wise crawl
Why does crab crawl cross-wise?

e.zhe han ta$^i$ na sidui buzu de jiegou you guanxi.
this and it that four-pair foot ASSOC structure have relation
This has to with the structures of its four pairs of feet.

f. pangxie$^i$ you sidui buzu$^j$
crab has four-pair foot
Crab has four pairs of feet.

g.$\phi^j$ sucheng "tuier",
(they) commonly-call "tuier"
(They) are commonly called "tuier".
... (omitted text terminating in a full stop)

h.$\phi^i$ yong $\phi^i$ ling yibian de buzu sheng qilai,
(it) use (it) other one-side ASSOC foot stretch up
(It) uses the foot on the other side to lift.

i. $\phi^i$ ba $\phi^i$ shenti tui guoqu,

(it) BA (it) body push get-through

(It) pushes (its) body through.

j. suoyi, ta$^i$ zhi hui hengzhe pa.

therefore it only can cross-wise crawl

Therefore, it can only crawl cross-wise.

The above example in a sense motivates the idea of using pronominal and nominal anaphors for different types of discourse segments. Chen in his study [Chen 86] found that nominal forms are usually chosen to encode anaphors at the beginning of a kind of discourse segment, called major breaks which correspond to paragraph boundaries, while pronouns are usually used at the beginning of another kind of discourse segments, called minor breaks which are associated with "sentence" boundaries. Obviously, Chen's observation about both kinds of breaks in a discourse is different from the types of discourse segment revealed in (6). Although there exist different types of discourse segments [Chen 86, Grosz & Sidner 86], however, there seem to be no consensual criteria for determining the type of a discourse segment.

In written text, some discourse segments are explicitly marked, such as by punctuation marks or paragraph boundaries, and the marks correspond to some extent to the types of discourse segments. However, on the contrary, in natural language generation systems, no such marks exist in the discourse structures produced by text planners to indicate the different types of breaks. Thus, the system designer must annotate the boundaries of breaks in discourse structures to facilitate the determination of anaphoric forms. Although there are no clear criteria for determining the types of discourse segments, we can borrow the idea from the study of information flow that, for example, in [Chafe 79], major breaks often occur at the points where coherence of space, time, topic, event or world are broken. The discourse structure produced by a text planner is itself a record of the coherence of the discourse [Hovy 93]. Thus following the above idea, we can refer to the beginning of a discourse as a major break if it signals a discontinuity of one of the above kinds of coherence; otherwise, it is a minor one. In our generation system, however, the text planner, as described in Chapter 7 is not equipped with a strong inference component; the types for discourse segments are

decided by hand at the design stage of the planning facilities.

## 5.4 Summary

In this chapter, we refined the non-zero parts in Rule 4 to obtain a rule, Rule 6 that is able to account for the generation of all kinds of anaphors. As before, we performed a sequence of experiments to achieve the rule. The experimental result shows that on average 87% of all kinds of anaphors in the test data are correctly matched by using Rule 6. Since the factors affecting the use of pronouns in Chinese are complicated, we also propose some heuristics to make the rule creating anaphors closer to human performance.

# Chapter 6

# Choosing Descriptions for Nominal Anaphors

## 6.1 Introduction

We have established a rule that includes a set of syntactic, semantic and discourse-oriented constraints to decide between the generation of zero, pronominal and nominal anaphors. Nominal anaphors do not have unique forms like their zero and pronominal counterparts. The description can be the same as the initial reference, parts of the information in the initial reference can be removed, new information can be added to the initial reference, or even a different lexical item can be used for a nominal anaphor. In this chapter, we investigate the choice of appropriate descriptions for nominal anaphors in Chinese natural language generation.

Former related research in natural language generation [Dale & Haddock 91, Dale 92, Reiter & Dale 92, Tutin & Kittredge 92, Horacek 95] focused on creating referring expressions for entities to distinguish them from a set of objects that the reader is assumed to be attending to. These algorithms can efficiently create descriptions to identify the intended referent unambiguously. The resulting descriptions, however, only reflect the attentional aspect of discourse [Grosz & Sidner 86]. In this chapter, we attempt to investigate the role of descriptions for nominal anaphors in another aspect of discourse, namely, intention [Grosz & Sidner 86]. We propose a preference rule for choosing different descriptions for nominal anaphors to reflect shift of intention in a discourse. To investigate the effectiveness of the rule, we performed two experiments using the three

sets of Chinese text as the test data. Basically, the experiments were carried out by comparing the nominal descriptions in the test data with the corresponding ones created by using a simple rule and then the preference rule, assuming the same semantic structures and context. The comparison of the results shows that the preference rule is effective.

## 6.2 Analysis of Nominal Anaphors in the Test Data

The surface structure of a Chinese nominal anaphor is a noun phrase which consists of a head noun optionally preceded by associative phrase, articles, relative clauses and adjectives [Li & Thompson 81]. In Chinese whether one chooses articles for nominal descriptions depends on complicated factors [Teng 75, Li & Thompson 81]. Observing the test data, we found that nominal anaphors are not commonly marked with articles. Thus we choose not to use articles for descriptions of nominal anaphors in our Chinese natural language generation system, as motivated by the data in Sec. 6.5. The nominal descriptions investigated in the remainder of this chapter are thought of as noun phrases of the above scheme without articles. A nominal anaphor is referred to as a *reduced* form, or a *reduction*, of the initial reference if its head noun is the same as the initial reference, and its modification part is a strict subset of the optional part in the initial reference; otherwise, if it is identical to the initial reference, then it is a *full* description.

Observing nominal anaphors occurring in the test data, we can classify nominal descriptions as below, with examples shown in Table 6.1.

**A** The initial reference is a bare noun, and the subsequent reference is the same as the initial reference.

**B** The initial reference is reducible, and the subsequent reference is the same as the initial reference.

**C** The initial reference is reducible and the subsequent reference is a reduced form of the initial reference without new information.

**D** The subsequent reference has new information in addition to the initial reference.

**E** Otherwise.

Table 6.1: Examples of descriptions of nominal anaphors.

| | **Initial references** | **Nominal anaphors** |
|---|---|---|
| A | zuqiu (football) | zuqiu (football) |
| B | tie-tong (iron barrel) | tie-tong (iron barrel) |
| C | *tie*-tong (*iron* barrel) | tong (barrel) |
| D | shui (water) | *yuan-wan-zhong de* shui (water *in the round bowl*) |
| E | *qian* (*money*) | *neixie chaopiao* (*those notes*) |

The occurrence of the types of nominal anaphors in the test data, in terms of the above classification, is shown in Table 6.2. Note that, first and second person pronouns in the test data are excluded in this table, which explains the differences between the total nominal anaphors in this chapter and in the previous two chapters.

Table 6.2: Occurrence of nominal anaphors in the test data.

| Data | A | B | C | D | E | Total |
|---|---|---|---|---|---|---|
| Set1 | 147 | 38 | 33 | 6 | 10 | 234 |
| | 63% | 16% | 14% | 3% | 4% | 100% |
| Set2 | 248 | 35 | 39 | 25 | 13 | 360 |
| | 67% | 10% | 14% | 6% | 4% | 100% |
| Set3 | 46 | 12 | 0 | 0 | 0 | 58 |
| | 79% | 21% | 0% | 0% | 0% | 100% |

## 6.3 A Preference Rule for Nominal Descriptions

The decision about what descriptions to use for initial references is a complicated process [McDonald 80, Dale 92]. In this chapter, we only consider subsequent reference. Previous work on the generation of referring expressions focused on producing minimal distinguishing descriptions [Dale & Haddock 91, Dale 92, Reiter & Dale 92] or descriptions customised for different level of hearers [Reiter 90]. Since we are not concerned with the generation of descriptions for different level of users, we look only at the former group of work, which is concerned with generating discriminating descriptions for subsequent references. As described in Sec. 3.5, if more distractors have entered into the *context set* which contains the entities occurring so far in the discourse except the intended referent, then more distinguishing information is used for a subsequent reference. Two entities are said to be distractors to each other if they are of the same category. For example, *the black dog* and *the brown dog* are distractors to each other

because they are of the same category, *dog*; yige yuanwan (a round bowl) and yige fangwan (a square bowl) in (1b) is another example of distractors. The entity, *the big cat*, is not a distractor to the above example because it is of different category *cat*. In [Grosz & Sidner 86], Grosz and Sidner claim that discourse segmentation is an important factor, obviously not the only one, governing the use of referring expressions. If, the idea of *context set* were restricted to local focus space [Grosz & Sidner 86], then the resulting descriptions would be to some extent sensitive in dealing with the local aspect of discourse structure. Although the algorithms would be refined due to the introduction of discourse structure, they would essentially still serve the distinguishing purpose.

The beginnings of discourse segments in a sense indicate shifts of intentions in a discourse [Grosz & Sidner 86]. In this situation, subsequent references may be preferred to be full descriptions rather than reduced ones or pronouns to emphasise the beginning of discourse segments, even if the referents have just been mentioned in the immediately previous sentence. Some examples were used to illustrate this idea, for example, in [Grosz & Sidner 86] and [Dale 92]. A similar situation happens in Chinese discourse. First of all, let's look at a characteristic of discourse segment structure in Chinese written text. In Chinese written text, a sentential mark, ".", is normally inserted at the end of a "sentence", which is a meaning-complete unit in a discourse, such as a to d, e to j, k to n, o to p and q to s in Fig. 6.1; [1] on the other hand, commas are inserted between sentences within a "sentence" as separators [Liu 84]. As shown in Chap. 4, a "sentence" to a large extent corresponds to a discourse segment. A Chinese discourse, say a paragraph of written text, therefore consists of a sequence of "sentences" and the corresponding intentions altogether form the intention of the discourse.

Among the groups of initial and subsequent references in Fig. 6.1, we focus on the one indexed *j*, la fengzheng de xian (the string pulling the kite). After it is initially introduced in b, it then appears in zero and nominal forms alternatively in the rest of the discourse, as shown schematically in Fig. 6.2. At the beginning of the second "sentence," it appears in a full description and then in four reduced descriptions in

[1] This is obtained from the Set 1 test data.

a. fengzheng$^i$ $\phi$ fangdao gaokong shangqu yihou,
b. la fengzheng$^i$ de xian$^j$ zhenme ye la bu zhi,
c. $\phi^j$ zongshi xiang xia wan,
d. zhe shi weishenme ne?
e. yuanlai, buguan fang fengzheng$^i$ de xian$^j$ you duome xi,
f. $\phi^j$ dou shi you zhongliang de,
g. xian$^j$ de zhongliang shi youyu diqiu dui xian$^j$ you xiyin de liliang$^l$ er chansheng de,
h. zhege liliang$^l$ haoxiang wuxing de shou,
i. $\phi^k$ ba xian$^j$ xiangxi zhuai,
j. xian$^j$ $\phi$ jiu la bu zhi le.
k. qishi, fengzheng$^i$ ye you zhongliang,
l. yinwei feng$^m$ chui zhe fengzheng$^i$,
m. $\phi^m$ shi fengzheng$^i$ xiang shang sheng,
n. suoyi fengzheng$^i$ bingbu xiang xia chen.
o. zheyang, $\phi$ zai fang fengzheng$^i$ shi,
p. piao zai kongzhong de xian$^j$ xingcheng yige wanqu de huxing.
q. piao zai kongzhong de xian$^j$ yue chang,
r. xian$^j$ wanqu de yue lihai,
s. $\phi^j$ yue la bu zhi.

**Translation:**

a. When flying a kite$^i$ in the sky,
b. the string pulling the kite$^{ij}$ can't be pulled straight.
c. It$^j$ is always bent downwards.
d. Why is that?
e. However thin the string pulling the kite$^{ij}$ is,
f. (it)$^j$ all has weight.
g. The weight of the string$^j$ is due to the attracting power of the earth on the string$^{jl}$.
h. This power$^l$ is like a invisible hand.
i. (It)$^l$ pulls the string$^j$ down.
j. The string$^j$ then can not be pulled straight.
k. However, the kite$^i$ also has weight.
l. Since the wind$^m$ blows the kite$^i$,
m. (it)$^m$ makes the kite$^i$ rise.
n. Therefore, the kite$^i$ does not fall down.
o. So when flying a kite$^i$,
p. the string fluttering in the sky$^j$ forms a curved arc.
q. The longer the string fluttering in the sky$^j$,
r. the more curved the string$^j$ is,
s. and the more difficult (it)$^j$ is to pull straight.

Figure 6.1: A sample Chinese written text.

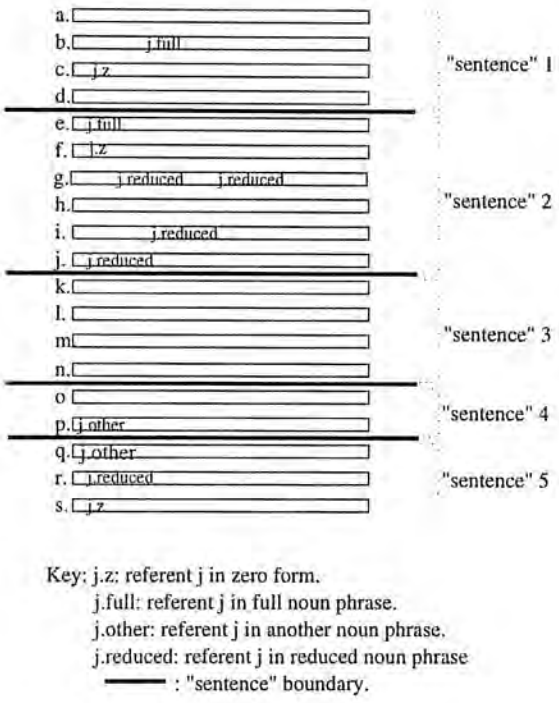



Figure 6.2: Occurrence of referent *j* in the discourse in Fig. 6.1.

the rest of the "sentence." It is not mentioned in the third "sentence." When it is reintroduced into the fourth "sentence," it appears in another noun phrase, piao zai kongzhong de xian (the string fluttering in the sky), which is not reduced. Then, in the last "sentence," it repeats the same patterns as in the second "sentence." Since there are no distracting elements for the string in the discourse, the use of full descriptions at the beginning of "sentences," e and q, can be interpreted as emphasising that a new discourse segment, "sentence", has begun. The accompanying reduced descriptions can then be explained as being intended to contrast with the emphasis at the beginning of "sentences." Note that a full description is used for the subsequent reference in p that is not the beginning of "sentence" because it is the first mention in the "sentence." Thus, we would generalise the above interpretation to be that a full description is preferred for a subsequent reference if it is at the beginning of a "sentence" or the first mention in the "sentence"; otherwise, a reduced description is preferred.

Should distracting elements occur in a "sentence," a sufficiently distinguishable description is required for a subsequent reference within the "sentence" instead of a reduced

one, even if it has been mentioned previously in the "sentence," for example, yuanwan (the round bowl) in (1d) and fangwan (the square bowl) in (1e). [2]

(1)a.zhaolai tongyang daxiao de liangkuai tiepi,
get same big-small NOM two iron-piece
Get two pieces of iron of the same size.

b.zuocheng yige yuanwan$^i$ he yige fangwan$^j$.
make one round-bowl and one square-bowl
Make a round and a square bowl.

c.ba yuanwan$^i$li zhuangman le shui,
BA round-bowl-in fill-full ASPECT water
Fill the round bowl full of water.

d.ranhou ba yuanwan$^i$zhong de shui manman daojin fangwan$^j$li,
then BA round-bowl-in GEN water slowly fill-in square-bowl-in
Then slowly pour the water in the round bowl into the square bowl.

e.ni hui faxian fangwan$^j$ zhuangbuxia zhexie shui,
you will find square-bowl fill-not-in these water
You will find that the square bowl can't hold this water.

f. youxie shui hui liu chulai.
have-some water will flow out-come
Some water will overflow.

On the basis of the above observations, we propose the following preference rule for the generation of descriptions for nominal anaphors in Chinese.

> If a nominal anaphor, $n$, is the first mention in a "sentence," then a full description is preferred; otherwise, if $n$ is within a "sentence" and has been mentioned previously in the same "sentence" without distracting elements, then a reduced description is preferred; otherwise a full description is preferred.

In the following we show the effect of using this preference rule.

[2] This is also obtained from Set 1 test data.

Table 6.3: Result of using the simple rule on the test data.

| Data | Matched | A | B | C | D | E | Total | % |
|---|---|---|---|---|---|---|---|---|
| Set1 | yes | 137 | 35 | 0 | 0 | 0 | 172 | 79 |
| | no | 0 | 0 | 30 | 6 | 9 | 45 | 21 |
| Set2 | yes | 232 | 32 | 0 | 0 | 0 | 264 | 78 |
| | no | 0 | 0 | 37 | 25 | 11 | 73 | 22 |
| Set3 | yes | 46 | 12 | 0 | 0 | 0 | 58 | 100 |
| | no | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 6.4: Result of using the preference rule on the test data.

| Data | Matched | A | B | C | D | E | Total | % |
|---|---|---|---|---|---|---|---|---|
| Set1 | yes | 137 | 28 | 26 | 0 | 0 | 191 | 88 |
| | no | 0 | 7 | 4 | 6 | 9 | 26 | 12 |
| Set2 | yes | 232 | 27 | 27 | 0 | 0 | 286 | 85 |
| | no | 0 | 6 | 9 | 25 | 11 | 51 | 16 |
| Set3 | yes | 46 | 12 | 0 | 0 | 0 | 58 | 100 |
| | no | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

## 6.4 Experimental Results

In this section, we describe experiments on employing rules for the generation of nominal anaphor descriptions to show their effects. First we use a simple rule and then the preference rule described previously.

### 6.4.1 Effect of using a simple rule

The experiment is described below.

- For each nominal anaphor with which we are concerned in a set of test data written by humans, repeat the following steps.
  - A nominal description is generated by using *a rule*, assuming the same semantic structure and context.
  - Then the resulting description is compared with the corresponding description in the text.
  - The comparison is matched if both sides are of either full or reduced descriptions of the initial reference; otherwise it is unmatched.

At the end of the experiment, the numbers of matches were collected to show the effect of the rule. Following our work in the previous chapters, we employed the same sets of text, Sets 1, 2 and 3, as the test data in this chapter. Since the aim of this work is to refine its predecessors, Rule 6 in Chap. 5, in the following, we focus on nominal anaphors which were correctly matched by using Rule 6. The differences between the total nominal anaphors in Table 6.2 and the following are the unmatched cases by using the same rule. Note that first and second person pronouns in the test data, which are treated as special cases of nominal anaphors in the previous two chapters, are excluded in this chapter. Therefore, the exclusion of first and second person pronouns should be taken into account, when comparing the total number of nominal anaphors in this chapter and in the previous chapters.

We started with a simple rule for the generation of nominal anaphor descriptions as below.

> Leave the description of the initial reference unchanged for nominal anaphors throughout the discourse.

In other words, according to this rule, only full descriptions of the initial references would be produced. The result of the experiment using this rule is shown in Table 6.3. The types A to E in this table, and in the following Table 6.4, are described in Sec. 6.2. The result summarises the fact that all of the nominal anaphors having full descriptions are correctly matched by using the simple rule, which amounts to 79, 78 and 100% of the nominal anaphors concerned. However, reduced descriptions and the other two types of descriptions, D and E, would not occur in the generated texts.

### 6.4.2 The effect of using the preference rule

We repeated the previous experiment by using the preference rule described formerly and obtained statistics as shown in Table 6.4. As shown in the table, by using the new rule, in addition to the fact that the majority of the nominal anaphors using full descriptions are correctly matched, a considerable number of reduced descriptions are matched as well, giving overall matches of 88, 85 and 100%. If we only consider Types A, B and C, namely, full and reduced descriptions in the test data, the match rates

become 94% (190/202), 94% (284/301) and 100% (58/58). Both groups of figures show that the preference rule is promising in the choice of reduced descriptions for nominal anaphors.

### 6.4.3 Discussion

What we are concerned with here is first to investigate the unmatched cases of Types B and C anaphors resulting from using the preference rule. Then, we investigate the implementation of the preference rule in Chinese natural language generation systems.

As discussed previously, the preference rule relies heavily on the notion of "sentence" boundary. In our experiment, we referred to the points right after sentential marks as the beginnings of "sentences". The beginning of a "sentence" generally signals a shift of intention in the discourse. In the test data, we found some cases which look like shifts of intentions that are not preceded by sentential marks, for example, d in (2) and c in (3). In (2d) and (3c), full descriptions are used for the anaphors, zhege xiaodong (this small hole) and suan de shiwu (sour food), not reduced ones as the preference rule expects. Note that, in these examples, (2d) and (3c) are preceded by adverbial phrases, *therefore* and *every time*. As shown in [Li & Thompson 79, Grosz & Sidner 86], the use of adverbial phrases is an important way to indicate shifts of intentions in a discourse. Therefore, the full descriptions in (2d) and (3c) in a sense do not violate the preference rule, if the determination of "sentence" boundaries is extended to include the condition of having preceding adverbial phrases.

(2)a.jiangluosan$^i$-dingshang you le yige xiaodong$^j$,
parachute-above have ASPECT a small-hole
There is a small hole at the top of the parachute.

b.san$^i$nei yibufen kongqi jiu cong zhege xiaodong$^j$li pai le chuqu,
parachute-in a-part air then from this small-hole-in push ASPECT out
A part of the air in the parachute then comes out from this small hole.

c.waimian de kongqi ye neng jishi jinru san$^i$nei bucong,
outside ASSOC air also can in-time come-in parachute-in replenish
The outside air can also come in the parachute in time to replenish it.

d.suoyi, zhege xiaodong$^j$ jiao paiqikong.

therefore this small-hole call push-out-air-hole

Therefore this small hole is called an air-hole.

(3)a.tuoye liulian de duoshao quan kao danao$^i$ zhihui,

saliva flow-amount NOM more-less completely rely brain command

The flow rate of saliva completely relies on commands from the brain.

b.suan de shiwu$^j$ dui tuoyexian$^k$ ciji zui qiang,

sour NOM food to salivary-gland stimulate most big

Sour food engenders the strongest stimulation from salivary gland.

c.meici $\phi$ chi suan de shiwu$^j$ shi,

every-time (one) eat sour NOM food ASPECT,

Every time when (one) eats sour food,

d.danao$^i$ jiu fachu xinhao$^l$,

the brain then send-out signal

the brain then sends out signals

e.$\phi^l$ rang tuoyexian$^k$ liuchu haoduo koushui lai.

(it) let salivary-gland flow-out a-lot-of saliva come

(which) makes salivary gland release a lot of saliva.

(4)a.1851 nian, ta$^i$ zai Bali zuo le yige shiyan,

1851 year he in Paris make ASPECT a experiment

In 1851, he conducted an experiment in Paris.

b.ta$^i$ yong yitiao chang 67 gongchi de xisheng$^j$,

he use a length 67 meter ASSOC thin-string

He used a 67-meter long string.

c.$\phi^j$ xiamian shuan zhe yige zhongqiu$^k$,

(it) under tie ASPECT a heavy-ball

Under (the string) (he) tied a heavy ball.

d.$\phi^j$ shangduan xi zai gaogao de difang,

(it) above tie at high-high NOM place

The upper part of (the string) was tied at a high position.

e.rang zhongqiu$^k$ zidongdi baidong qilai.

let heavy-ball automatically swing

Let the heavy ball swing automatically.

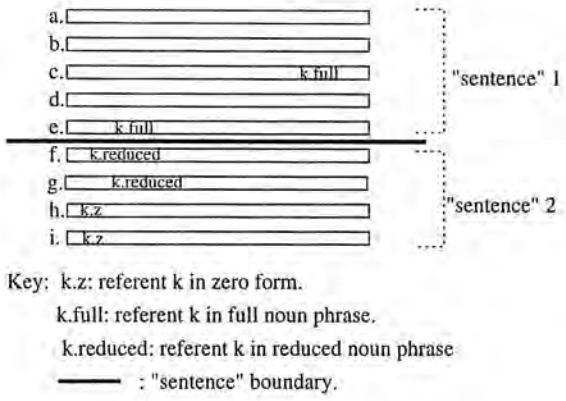



Figure 6.3: Occurrence of referent $k$ in the discourse in (4).

f. qiu$^k$ de xiamian fang yige shapan$^l$,
ball ASSOC under-place place a sand-pan
(He) placed a sand-pan under the ball

g.dang qiu$^k$ baidong shi,
let when ball swing ASP
When the ball swung,

h.$\phi^k$ qingqingde cong shapan$^l$shang huaguo,
(it) gently from sand-pan above cut-ASP
(it) gently cut the surface of the sand-pan.

i. $\phi^k$ liuxia huahen.
(it) leave cut-mark
(It) left a cutting mark.

We also found some cases of reduced descriptions being used at the beginning of "sentences". For example, in (4), f is at the beginning of a "sentence"; a reduced description, qiu (ball) instead of the full description zhongqiu (heavy ball) expected by using the preference rule is used for the nominal anaphor. Furthermore, as shown in Fig. 6.3, the writer used a full description for the anaphor in e, where the initial reference occurred locally in the same "sentence". This example shows that the writer did not use changes of descriptions as a way to signal shift of major intention.

In brief, the above observations suggest that the boundary condition in the preference

rule should allow for more than "sentence" boundaries. Furthermore, the use of descriptions for nominal anaphors may be dependent to some extent upon a writer's own style.

In the Chinese generation system we are developing, the "sentence" boundaries in a generated text are determined on the basis of the characteristics of the hierarchy of discourse structure produced by the text planner. [3] The condition of "sentence" boundary in the preference rule is determined when traversing the discourse structure. Since, in the current implementation of our system, we do not consider using adverbial phrases as the indicator of shift of intention, the extension of using adverbial phrase to signal shift of intention for the preference rule is left to a future study. In our development of a rule for nominal descriptions, we also did not take personal style into consideration. Observing the test data, those cases related to personal style are the minority, and thus negligible. It is impractical to expend a great deal of effort on minor cases. Furthermore, the descriptions resulting from using the preference rule are no less appropriate than the ones imposing personal style. Thus we do not intend to extend the preference rule by considering personal style.

As for the implementation of reduced descriptions, we basically employ the representation of entities in [Dale 92]. In other words, an entity is represented as a substance, that will be realised as the head noun of a noun phrase, and a property list, that will be the modification of the head noun. A straightforward method is to take both parts above to form a full description and to use the former part only as a reduced description. However, from the observation of the test data, some reduced descriptions can not be obtained by the above method. For example, in (5), a full description **jiangluosan** (parachute) occurs in a, and later on, in b, c and d, descriptions **san** (parachute) are used. Here a superordinate/subordinate relationship exists between the anaphor and antecedent. In Chinese, an anaphor of this kind can use the head of its antecedent, for example, **san** here, which makes it look like a "reduced" description. In our experiment, we treated it as a reduced description. Here the missing part **jiangluo** (descending) in the full description would normally not be represented in the property list because it

[3] See Chap. 7 for details.

is bound to san to form a compound. [4] To account for noun phrases like this, in the implementation we indicate in the representation of an entity the default reduced form. Then, when a reduced description is decided for a nominal anaphor, the system first tries to use the default reduced form. Otherwise, if there is none, it takes the substance part as the reduced form.

(5)a.yaoshi jiangluosan$^i$-dingshang mei you xiaodong,
if parachute-top-above no have small-hole
If at the top of a parachute has no small hole,
b.cong sanyi-bianshang liu jinqu de kongqi$^j$ jiu dou jizai san$^i$ding,
from parachute-cloth-edge-above flow come-in NOM air then all
concentrate-in parachute-top
then the air flow in the parachute would concentrate in the top part of the parachute.
c.$\phi^j$ bu rongyi paodao san$^i$wai qu,
(it) not easy run-in parachute-out go
(It) is not easy to escape the parachute.
d.waimian liudong de kongqi ye bu rongyi jinru san$^i$nei,
outside flow-move NOM air also not easy get-in parachute-in
The outside air could not easily enter the parachute.
...

## 6.5 Articles in Nominal Descriptions

Observing our test data, we classified the following types of articles used for nominal anaphors.

**Type A** The initial reference, IR, has an indefinite marker and the subsequent reference, SR, has a definite marker.

**Type B** The IR has an indefinite marker but the SR has no marker.

**Type C** The IR has no marker and the SR has no marker either.

**Type D** The IR has no marker but the SR has a definite marker.

[4] This has to do with the characteristics of compounds in Chinese. See [Chao 68] for detail discussion about compounds in Chinese.

Table 6.5: Occurrence of articles in nominal anaphors in the test data.

| | A | B | C | D | Total |
|---|---|---|---|---|---|
| Set 1 | 7 | 21 | 196 | 10 | 234 |
| % | 3 | 9 | 84 | 4 | 100 |
| Set 2 | 12 | 23 | 317 | 8 | 360 |
| % | 3 | 6 | 88 | 2 | 99 |
| Set 3 | 2 | 0 | 55 | 1 | 58 |
| % | 3 | 0 | 95 | 2 | 100 |

The occurrence of each type of nominal anaphor in the test data is shown in Table 6.5. As shown in the table, Type B and C nominal anaphors altogether occupy 93%, 94% and 95%, which show that articles are rarely used in nominal descriptions in descriptive texts. The figures show that nominal descriptions without articles would normally cause little trouble for the reader in understanding the text. As pointed out in [Teng 75], references do not have to be marked in any specific way, although explicit markers in noun phrases would help clear interpretation. The use of articles for nominal anaphors involves very complicated factors. Unfortunately, from linguistic studies we are not sure how to account for this problem [Teng 75, Li & Thompson 81]. In this chapter, we do not intend to examine the sophisticated behaviour used by human writers to cope with this problem. Instead, we aim at getting comprehensible descriptions for nominal anaphors produced by our generation system, in other words, descriptions where the reader can grasp the referent easily. The descriptions produced by using the preference rule investigated previously without using articles largely meets the above requirement. Furthermore, the figures shown in Table 6.5 also suggest using the plain form for nominal descriptions. Thus we choose not to use articles in the descriptions of nominal anaphors produced by our system.

## 6.6 Summary

A rule for the generation of nominal descriptions based on empirical study is presented. The rule uses full and reduced descriptions to characterise shifts of intention in the generated discourse. The experimental results show that 88, 85 and 100% of the ones predicted by the previous rule can be captured by using this rule. We also investigate

the use of articles in the descriptions of nominal anaphors. Since 94% in average of nominal anaphors have no articles in the test data, we choose not to use articles for nominal descriptions.

# Chapter 7

# Implementation

## 7.1 Introduction

In previous chapters, we explored the usage of anaphors in Chinese and obtained some rules. To show how these rules work in a real system, we have implemented them in a Chinese natural language generation system. Since we aim at descriptive text, we have built up a generation system which is able to generate descriptive texts, similar to the ones produced by TEXT [McKeown 85] and TEXPLAN [Maybury 90]. The domain we chose for the system is descriptions for plants, animals, scenic locations, etc., in a National Park. [1] The resulting system could be further extended as the text generation component in, for example, a tour-guide explanation system.

The system, like conventional ones [McKeown 85, Maybury 90, Hovy 93, Reiter 94], is divided into strategic and tactical components, as shown in Fig. 7.1. The strategic component arranges message contents in response to the input goal into a well-organised hierarchical discourse structure by using a text planner. Accompanying the text planner are three knowledge bases providing the contents of messages, ways of organising the message contents, and the characteristics of various levels of users, with which the planner can adapt the message contents for specific users. The tactical component takes the hierarchical discourse structure as input and produces surface sentences with punctuation marks inserted appropriately. Within the tactical component, the first task is linearising in depth-first order the message units in the discourse structure and

[1] We choose the Yangming Mountain National Park which is 20 km north of Taipei.

mapping them into syntactically-oriented representations. Referring expressions, the main concern of this thesis, are generated within the mapping process. Then a linguistic realisation program is invoked to convert the syntactic representation into surface strings in Chinese. The discourse model maintains local and global focus spaces corresponding to the attentional states in Grosz and Sidner's theory of discourse structure [Grosz & Sidner 86].

In this thesis, we do not intend to invent a new generation system. What concerns us most is how the referring expression component works in a Chinese natural language generation system. Thus, instead of designing a new system, we borrowed ideas from well-developed generation systems and simplified them to some extent as the basis on which to construct our system. Among them, we employ the idea of a domain knowledge base as in the TEXT system [McKeown 85], text planning as in the TEXPLAN system [Maybury 90], and semantic and syntactic representations as in the Epicure system [Dale 92]. We developed a simple linguistic realisation program based on the PATR-II formalism [Shieber 86].

The remainder of the chapter is organised as follows. In Section 7.2, we describe the representation and content of the domain knowledge base. In Section 7.3, we describe the operation of the planner and its accompanying knowledge bases. In Section 7.4, we describe the linearisation program, including the implementation of our anaphor generation rules. Finally, in Section 7.5, we describe the realisation program and the accompanying syntactic rules and lexicon.

## 7.2 Domain Knowledge Base

The domain knowledge base contains information about entities in the National Park. The knowledge base can be divided into two parts: (a) descriptive knowledge about entities in the domain; and (b) essential characteristics of entities mentioned in part (a). Part (a) includes part or all of the following kind of information to describe an entity [McKeown 85, Maybury 90]:

- *definitional information* providing classification and characteristics to identify an entity;

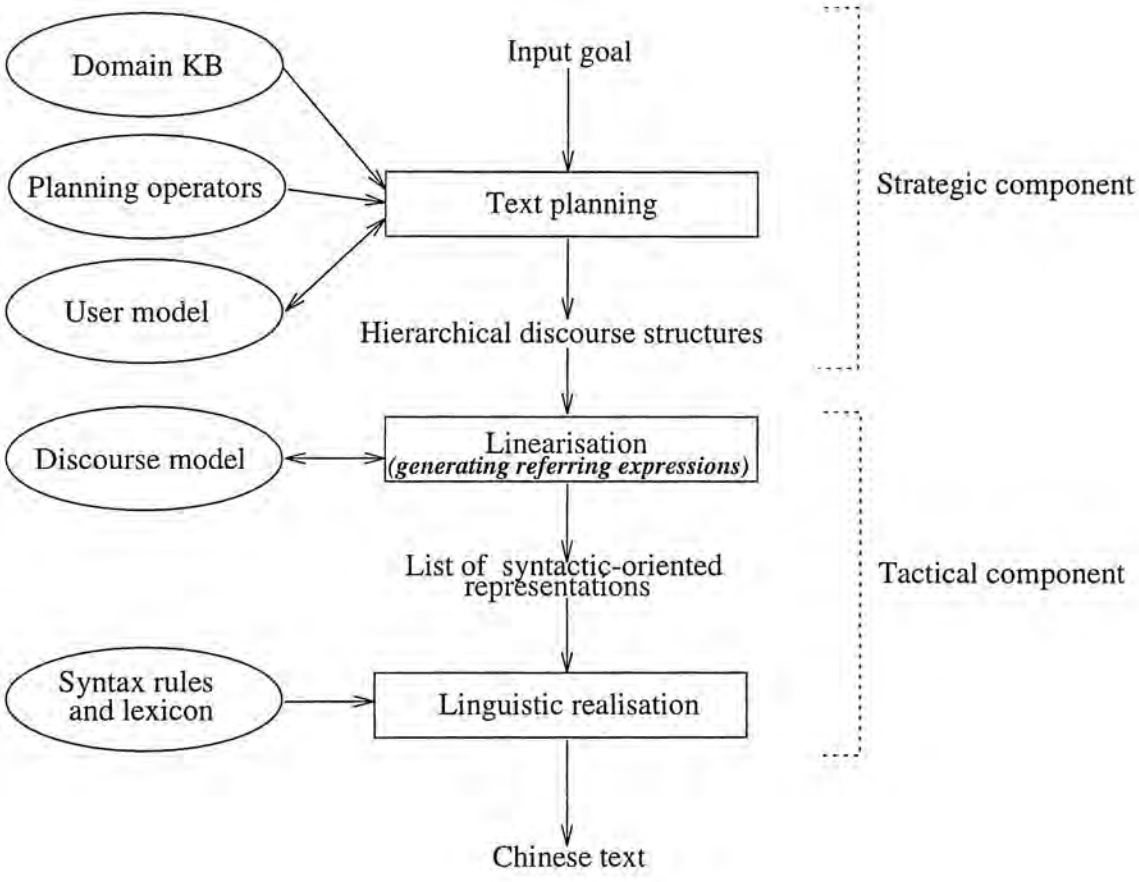



Figure 7.1: Diagram of our Chinese natural language generation system.

- *attributive information* indicating properties associated with an entity;
- *divisional information* specifying components and subtypes of an entity; and
- *functional information* indicating the purpose of an entity.

The knowledge base maintains a set of nodes to represent entities and their associated information and a number of links between nodes. Each node is labelled by an entity marker or contains information connected to the corresponding entity node by a link labelling the kind of information. For example, the part of the knowledge base corresponding to the plant *Isoetes taiwanensis* is shown in Fig. 7.2. In the figure, entity nodes are represented by oval boxes, information corresponding to entities is represented by round square boxes and types of links between nodes are indicated in boldface. The dashed links in the figure indicate the coreference relationships between entities. Since in our system we rely on unification as the basic operation to manipulate information, we implemented the above knowledge base in feature structures [Gazdar & Mellish 89], as shown in Fig 7.3. [2] We used a mixture of English, Chinese and shorthand notation in the feature structure. For example, we use "it" to denote *Isoetes taiwanensis*. To simplify the implementation of later components, the text planner and the linguistic realisation program, we adopted some linguistically-oriented representations in the domain knowledge base, as described below.

Locative particles are specific to Chinese, which occur in locative phrases, as described in Sec. 2.3. We used Chinese words, instead of a neutral representation, to indicate this kind of information in the domain knowledge base. For example, in Fig. 7.3, the attribute of leaves, `growing(zai):shang(p11)`, means that leaves grow on top of the stalk, where zai(at) is a preposition, shang(top) is a locative particle and `p11` is the index of the entity stalk. This attribute is realised as the following sentence,

yezi zhang zai duan-qiujing-shang
leaf grow at short stalk-top
Leaves grow on top of the short stalk

[2] In fact, this is not quite standard, as we have included in the feature values, `about(2,cm)`, for example, in Fig. 7.3.

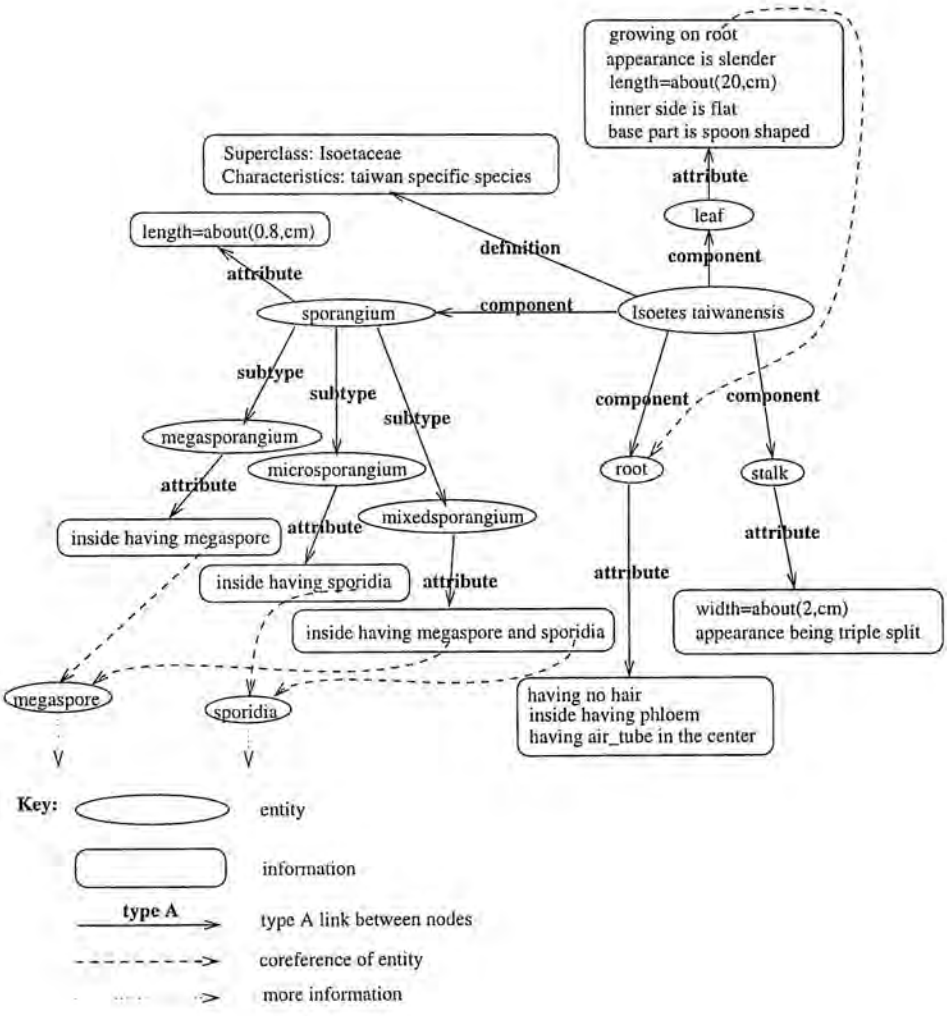



Figure 7.2: Diagram of knowledge base about *Isoetes taiwanensis*.

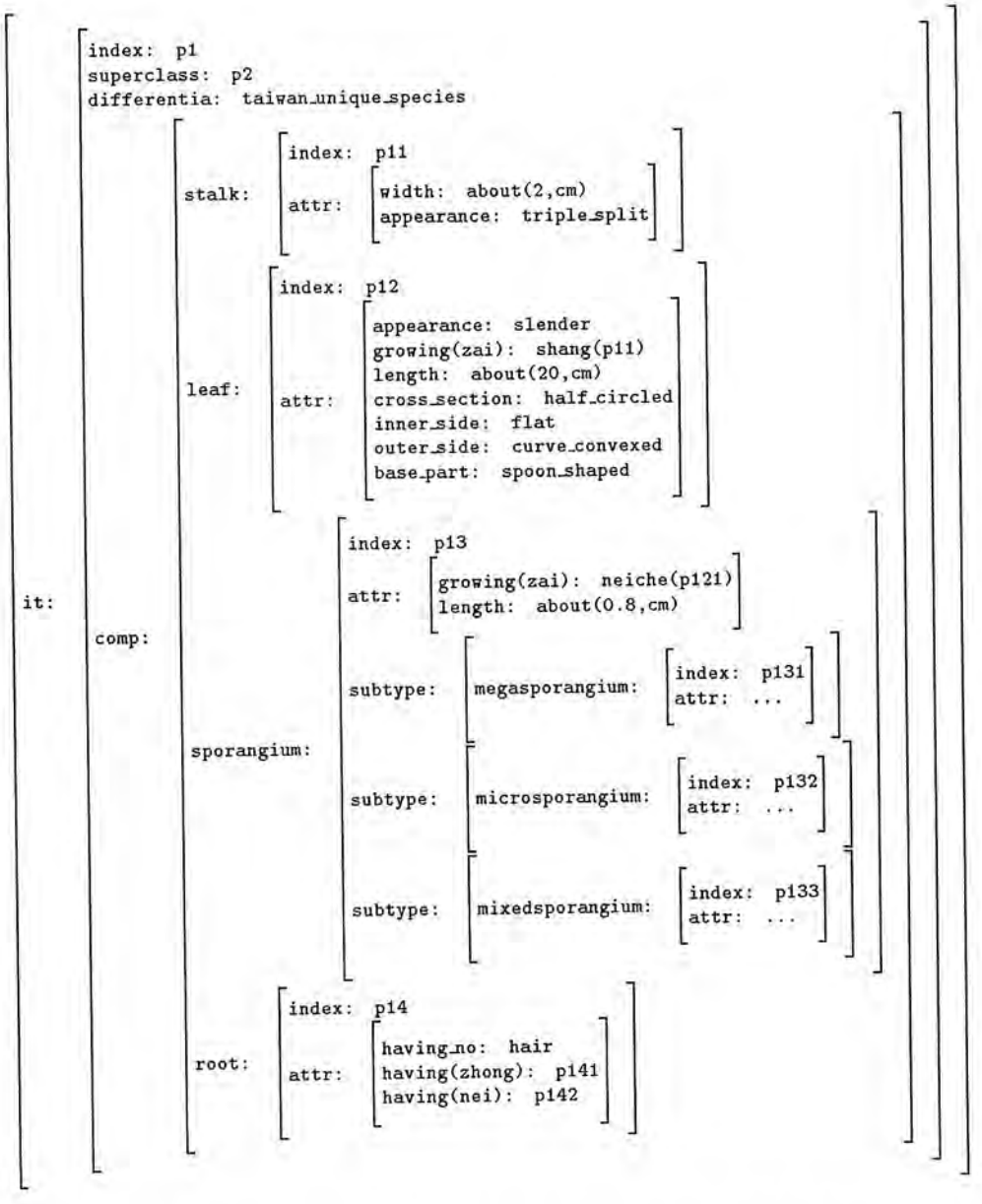



Figure 7.3: knowledge base about *Isoetes taiwanensis* in feature structure.

where zai duan-quijing-shang is a locative phrase. In addition to the object position in a sentence, a locative phrase may occur in the sentence-initial position. In our domain, this kind of locative phrase generally occurs in presentative sentences; [3] the coverb zai can be omitted in this situation. [4] In the domain knowledge base, this kind of information is represented as, for example, the attribute of root in Fig. 7.3, `having(nei):p142`, where nei (in) is a locative particle and `p142` is the index of `wei_guan_shu_dan_tiao` (phloem). [5] It is realised as the following sentence,

(zai) gen-nei you wei_guan_shu_dan_tiao.
(at) root-in have phloem
In the root, there is a phloem.

The above representation of information about locative phrases largely simplifies the creation of semantic structures described in Sec. 7.3.2. [6] In semantic representation, the above kinds of information will be mapped into the location and patient roles, respectively. The value of each role is a feature structure which contains the `locative`, `location` and `obj` features. In creating the semantic structure for the above kinds of feature structures, the coverb is mapped to the `locative` feature, the locative particle is mapped to the `location`, and the value part goes to the `obj` feature. For example, the attributes, `growing(zai):shang(p11)` and `having(nei):p142`, correspond to the following `location` and `patient` roles, respectively.

$$\begin{bmatrix} \texttt{location:} & \begin{bmatrix} \texttt{locative:} & \texttt{zai} \\ \texttt{location:} & \texttt{shang} \\ \texttt{obj:} & \texttt{p11} \end{bmatrix} \end{bmatrix} \begin{bmatrix} \texttt{patient:} & \begin{bmatrix} \texttt{locative:} & \texttt{null} \\ \texttt{location:} & \texttt{nei} \\ \texttt{obj:} & \texttt{p142} \end{bmatrix} \end{bmatrix}$$

The negation of a feature is represented as an individual feature like ordinary ones. For example, "(inside) the air tube of root has no hair" is represented as `having_no(li):hair`, where li(inside) is a locative particle.

[3] This kind of sentence presents a new entity in the discourse. See Sec. 2.2 for details.

[4] See Sec. 2.3 and [Li & Thompson 81] for details.

[5] Note that we deliberately omit the coverb zai in the notation.

[6] Of course, the representation of the information about locative phrases is limited to the above types.

Part (b) of the knowledge base contains essential characteristics about entities mentioned in part (a), each of which is represented as a feature structure of the following format.

```
[ index:  ...
  spec:   [ substance:  ...
            prop:  ...
            animate:  ...
            article:  ... ] ]
```

Each entity has a unique index as its identifier in the domain knowledge base. The `spec` feature specifies the essential characteristics of the entity. The first feature specifies the substance of the entity in nature that will be realised as the head noun in the noun phrase. The second lists the properties associated with the entity, such as size, colour, etc. Note that the properties here are realised as the adjectives of the noun phrase, in contrast with the ones in the attribute feature in part (a) which are realised as sentences. The `animate` feature indicates the animacy of the entity. This is used in the decision whether to use a pronoun or nominal description. [7] The last feature, `article`, indicates the article used in the initial reference. [8] In natural language generation systems, the decision of the description for an initial reference is quite a complicated task [McDonald 80, Dale 92]. Since in this thesis we focus on the work of subsequent reference, our system employs a simple way to deal with initial references: using full descriptions for initial references. In other words, all the information in the `property` and `article` features will be realised in the initial reference. To allow for appropriate descriptions for initial references, the builder of the domain knowledge therefore has to take care to include properties and articles for entity specifications. For example, the following specification is realised as **yi ge duan qujing** (a short stalk) for the initial reference, where **yi_ ge** is the article.

[7] See Chap. 5 for details.

[8] Note that, as described in Sec. 6.5, we choose not to use articles in the descriptions of nominal anaphors. Thus the `article` feature will not affect the nominal descriptions.

```
[ index:  p11
  spec:   [ substance:  it/stalk
            prop:  [size/short]
            animate:  -
            article:  yi_ge ]
]
```

One should recall that in the decision rule for nominal anaphora, reduced descriptions will be generated in certain situations. This reduction is made by omitting the prop from the final noun phrase, for example, using qujing for the above case.

## 7.3 Text Planning

The preceding section describes what the system knows about the domain in question. To have what the system knows about presented appropriately to the user, a natural language generation system needs a text planner to extract and organise information stored in the knowledge base to fit the user's requirements. In our system, the plan operator library is the main knowledge source guiding the organisation of message contents to be presented to the user. The idea of text planning here basically comes from TEXPLAN [Maybury 90]. In the following, we first introduce a plan operator library based on the integrated theory of communicative acts described in Sec. 3.3.3. Then we describe the control part of the planner and give examples to illustrate its operation.

### 7.3.1 A plan library

In our system, the plan library contains a set of operators which are encoded to achieve communicative acts in the hierarchy in Fig. 3.6. An operator has in order **header**, **constraints**, **effects** and **decomposition** fields. An instantiated operator can be selected to achieve a subgoal only if the constraints are satisfied. Propositions in the constraints field encode restrictions on the domain knowledge base or the user model. Listed in the effects field are the intended results after executing the selected operator. The subgoal in the decomposition field can be either optional, or universally quantified, or both. An instantiated operator is decomposed according to the availability

of subgoals specified in the decomposition field. For example, the *describe-by-defining* operator is shown as below

| | |
|---|---|
| **Name:** | describe-by-defining |
| **Header:** | describe(s,h,E). |
| **Constraints:** | is_entity(E),not(know_about(h,E)), (haste(s);haste(h)). |
| **Effects:** | know_about(h,E). |
| **Decomposition:** | define(s,h,E). |

The header indicates the name of the communicative acts; the arguments $s$, $h$ and $E$ correspond to the speaker, hearer and the entity to be described. The propositions in the constraints field specify that $E$ must be an entity, the hearer must not know about the entity and either the speaker or the hearer must be in haste to get the result, which leads to the choice of a terse description. The effect field indicates that the hearer will know about the entity after the operator is executed. Finally, the operator specifies that it would be decomposed into a subgoal `define` in the next round of execution.

The `define` act can be achieved by defining logical, synonymic and antonymic definitions. [9] In addition to the above, an entity can also be described by stating constituency and attributes, as shown in Fig. 7.4. In our system, we only implemented one of the speech acts, `inform`, as shown in Fig. 7.4. The `inform` operator constrains $P$ to be a rhetorical proposition in the system, and its effect is that the hearer believes that the speaker believes the proposition $P$. It is decomposed into a single `assert` act.

The rhetorical propositions guide the extraction of information from the domain knowledge base. In our system, in addition to extracting knowledge, each rhetorical proposition also manipulates the conversion of the extracted knowledge into the corresponding semantic structure which then becomes the input of the linearisation component.

An extended description can be made by combining the above techniques, giving definitional and then detailed information, followed by divisional information as shown below.

[9] At the current stage, we have only implemented the *define-by-logical-definition* operator

| | |
|---|---|
| **Name:** | define-by-logical-definition |
| **Header:** | define(s,h,E). |
| **Constraints:** | has_superclass(E,C). |
| **Effects:** | know(h,superclass(E,C)),<br>know(h,differentia(E)). |
| **Decomposition:** | inform(s,h,logical_definition(E)). |

| | |
|---|---|
| **Name:** | describe-by-constituency |
| **Header:** | describe(s,h,E). |
| **Constraints:** | has_subpart(E,P),<br>not(know_about(h,E)),<br>(haste(s);haste(h)). |
| **Effects:** | know_about(h,E),<br>know(h,subpart(E,P)). |
| **Decomposition:** | inform(s,h,constituency(E)). |

| | |
|---|---|
| **Name:** | describe-by-attribution |
| **Header:** | describe(s,h,E). |
| **Constraints:** | has_attribute(E,A),<br>not(know_about(h,E)),<br>(haste(s);haste(h)). |
| **Effects:** | know_about(h,E),<br>know(h,attribute(E,A)). |
| **Decomposition:** | forall(X^member(X,A),<br>inform(s,h,attribution(E,X))). |

| | |
|---|---|
| **Name:** | inform-by-assertion |
| **Header:** | inform(s,h,P). |
| **Constraints:** | proposition(P) |
| **Effects:** | believe(h,believe(s,P)). |
| **Decomposition:** | assert(s,h,P). |

Figure 7.4: Operators for terse descriptions.

| | |
|---|---|
| **Name:** | Extended-description |
| **Header:** | describe(s,h,E) |
| **Constraints:** | is_entity(E),not(know_about(h,E)),definable(E) |
| **Effects:** | know_about(h,E) |
| **Decomposition:** | define(s,h,E),<br>optional(detail(s,h,E)),<br>optional(divide(s,h,E)). |

In the decomposition field, the `detail` can be achieved by giving attributive, or purpose information, or both. In the attributes of an entity, new entities may occur. For example, in Fig. 7.2, inside the *root* of *Isoetes taiwanensis* there is the *phloem*; inside the *megasporangium* there is the *megaspore*, etc. Thus in addition to the above `detail` operators we have an extended `detail` operator, *detail-by-attribution with new elements*, as shown in Fig. 7.5.

Divisional information includes both subparts and subtypes of an entity. For example, as shown in Fig. 7.2, *Isoetes taiwanensis* has subparts, *leaf*, *stalk*, *sporangium* and *root*; the *sporangium* of *Isoetes taiwanensis* has three subtypes, *megasporangium*, *microsporangium* and *mixed sporangium*. To account for such information, we have two `divide` operators in the library, i.e., *divide-by-constituency* and *divide-by-classification*. As shown in Fig. 7.3, the first `divide` operator is achieved by giving all subparts and the accompanying detailed information; the other is achieved by first informing of the subtypes of the entity and then following by detail on each subtype. Subparts of an

entity may have divisional information. For example, as shown in Fig. 7.2, the subpart *sporangium* of *Isoetes taiwanensis* is classified as three subtypes. To present this sort of divisional information in the plan tree, we therefore established an alternative *detail-by-attribution* operator with optional divisional information. Note that the `divide` in the decomposition field is optional. This operator is activated if the user wants the divisional information. [10] An entity may be described in various aspects. For example, we can describe a lake in the National Park, Menghuan Lake in both geographical and ecological aspects. The complete list of operators for extended descriptions is shown in Fig.7.5.

### 7.3.2 The planner

The algorithm for the planner, adopted from [Maybury 90], is shown in Fig. 7.6. The planner takes a list of subgoals, $[g_1, g_2, \cdots, g_n]$, as the input, where the first subgoal, $g_1$, is denoted as FIRST and the rest of subgoals is denoted as a list REST. When the planner is first invoked, the user's goal is the only subgoal in the input list. The plan structure resulted by the algorithm is a tree where the internal nodes are decomposed subgoals of the user's goal and the terminal nodes are attached with the semantic structures of the message contents.

The planner first checks whether the input list is empty. If it is, then the plan structure is returned. Otherwise, in Step P1, if it is a surface act, the algorithm performs the following tasks:(1) concatenating the current subgoal, FIRST, to the plan structure; (2) extracting the associated message content from the domain knowledge base according to the rhetorical predicates; (3) converting the extracted content into a semantic structure as shown in Fig 7.7; and (4) continuing the recursion. The second task is based on the definition of the corresponding rhetorical predicate. For example, the *logical definition* predicate gets the superclass and, if available, the distinguishing characteristics of the entity from the knowledge base. Details about rhetorical predicates are shown in Appendix C. The third task is described in the next subsection.

If, on the other hand, a rhetorical act operator or a speech act operator is met in Step P2, the program finds the effects field in operators that achieve the current subgoal.

[10] See Sec. 7.3.4 for examples.

**Name:** Extended-description
**Header:** describe(s,h,E).
**Constraints:** is_entity(E),definable(E), not(know_about(h,E)).
**Effects:** know_about(h,E).
**Decomposition:** define(s,h,E), optional(detail(s,h,E)), optional(divide(s,h,E)).

**Name:** Extended-description in aspects
**Header:** describe(s,h,E).
**Constraints:** is_entity(E),has_aspect(E,As), not(know_about(h,E)).
**Effects:** know_about(h,E).
**Decomposition:** forall(X^member(X,As), describe(s,h,X)).

**Name:** Detail-by-attribution
**Header:** detail(s,h,E).
**Constraints:** has_attribute(E,A).
**Effects:** know_about(h,E), know(h,attribution(E,A)).
**Decomposition:** forall(X^member(X,A), inform(s,h,attribute(E,X))).

**Name:** Detail-by-attribution with new elements
**Header:** detail(s,h,E).
**Constraints:** has_attribute(E,A), new_elements(A,New), not(know_about(h,New)), want_new_info(h,E).
**Effects:** know_about(h,E), know(h,attribution(E,A)), know_about(h,New).
**Decomposition:** forall(X^member(X,A), inform(s,h,attribute(E,X)), N1^new_element(X,N1), optional(detail(s,h,N1))).

**Name:** Detail-by-indicating-purpose
**Header:** detail(s,h,E).
**Constraints:** has_purpose(E).
**Effect:** know(h,purpose(E)).
**Decomposition:** inform(s,h,purpose(E)).

**Name:** Detail-by-attribution, with division.
**Header:** detail(s,h,E).
**Constraints:** has_attribute(E,A), want_division_info(h,E).
**Effects:** know_about(h,E), know(h,attribution(E,A)).
**Decomposition:** forall(X^member(X,A), inform(s,h,attribute(E,X))), optional(division(s,h,E)).

**Name:** Detail-by-indicating-purpose and-attributes
**Header:** detail(s,h,E).
**Constraints:** has_attribute(E,A), has_purpose(E).
**Effects:** know_about(h,E), know(h,purpose(E)), know(h,attribute(E,A)).
**Decomposition:** inform(s,h,purpose(E)), inform(s,h,attribution(E)).

**Name:** Divide-by-constituency
**Header:** divide(s,h,E).
**Constraints:** has_subpart(E,Ps).
**Effects:** know(h,subpart(E,P)).
**Decomposition:** forall(X^member(X,Ps), inform(s,h,has_subpart(E,X)), detail(s,h,X)).

**Name:** Divide-by-classification
**Header:** divide(s,h,E).
**Constraints:** has_subtype(E,Ts).
**Effects:** know(h,subtype(E,Ts)).
**Decomposition:** inform(s,h,classification(E,Ts)), forall(X^member(X,Ts), detail(s,h,X)).

**Name:** divide-according-to-aspects
**Header:** divide(s,h,E).
**Constraints:** has_aspect(E,As).
**Effects:** know_about(h,E), know(h,aspect(E,As)).
**Decomposition:** forall(X^member(X,As), describe(s,h,X)).

Figure 7.5: Operators for extended descriptions.

```
Initialisation: The planner is given a list of subgoals, LIST, as
                the input; the first subgoal in LIST is represented
                as FIRST, and the rest of subgoals in LIST is
                represented as a list, REST.
                When the planner is first invoked, the user's goal is
                the only subgoal in LIST.
                The output is a plan tree which is initially empty.

P0: if LIST is empty
    then return the plan structure;
P1: else if FIRST is a surface act
         then concatenate FIRST to the plan structure,
              acquire message content from the domain KB,
              convert it into semantic structure,
              attach the semantic structure to the plan structure,
              and recurse P0 with REST as the input list;
P2:      else find the effects field in operators that achieve FIRST;
P3:           if no operators achieve FIRST
              then report error and halt;
              else
P4:               select those operators with constraints satisfied,
                  select the first priority operator and decompose it,
                  (the result of decomposition is a new list of
                   subgoals, LIST1.)
                   concatenate FIRST to the plan structure,
                   recurse P0 with LIST1 as the input list, and
                   recurse P0 with REST as the input list
```

Figure 7.6: Algorithm for planner.

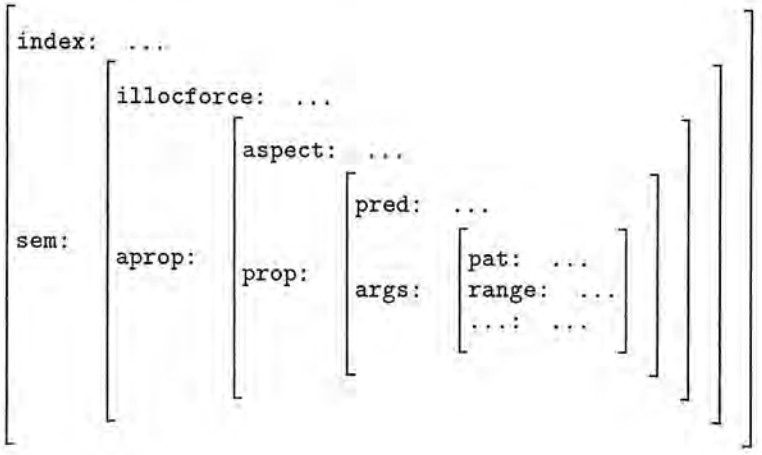



Figure 7.7: Semantic structure of message units.

If it cannot find any operator to achieve the subgoal, it reports an error and halts. Otherwise, in Step P4, the program selects the operators with the constraints satisfied, which may produce multiple results. The satisfied operators are prioritised according to the number of constraints in their **Constraints** field. The operator with the highest number of constraints is prioritised first. Then the **Decomposition** field of the selected operator is examined to obtain the subgoals of the next recursion. Before recursing on the decomposed subgoals and the rest of the subgoals, the selected operator is concatenated to the plan structure.

### 7.3.3 Building semantic structure

As described previously in the text planning algorithm, extracted message contents attached to surface act nodes are converted into semantic structures. The semantic structure is in a feature structure format slightly adapted from [Dale 92], as shown in Fig. 7.7. In the structure, first comes the index field indicating the sequential order in the discourse to be generated. The `sem` feature specifies the illocutionary force, here `assert`, and aspect and proposition. The `aspect` feature indicates the aspect marker associated with the sentence of the message unit, here `null`. [11] The `prop` field specifies the predicate and arguments.

The conversion is mainly achieved through a set of mapping rules. In the semantic

[11] Associated with the assertions are state verbs which do not assume aspectual inflection [Teng 75].

structure, the value of `index` is obtained from an index counter which automatically increases by one after creating a semantic structure. The `illocforce` value comes from the type of surface act, here assert. The `aspect` is nulled for sentences in descriptive text. [12] The remainder of the semantic structure is taken from the extracted content. The extracted items are generally entities to be assigned to the `args` feature. For example, the resulting content of the *logical definition* predicate is the superclass of the entity described; the entity and its superclass are assigned to the `patient` and `range` in the `args`, respectively. For each rhetorical proposition we chose an appropriate predicate value; for example, we use "belonging" as the predicate for superclass information. The extracted content of `attribution` is different from others in that it is a feature-value pair. The features in the pairs can be verbal, such as `growing(zai)`: `shang(p11)` or nominal, such as `length: about(20,cm)` for the leaves of *Isoetes taiwanensis*. For the former case, the feature is used as the predicate and the value is assigned as one of the arguments. On the other hand, the feature in the other case is assigned as one of the arguments, while the value is used as the predicate. The above two types of predicates are realised as verbal and nominal phrases, respectively, in Chinese sentences [Chao 68]. The mapping rules can be found in Appendix C.

### 7.3.4 Worked examples

Given the domain knowledge base and planning library described previously, the user model without any proposition in it, and a input goal

`know_about(h,p1)`, where `p1` is *Isoetes taiwanensis*,

the system invokes the planner to proceed as follows. The planner starts with matching the input goal with the effects fields of operators. In this case, all of the `describe` operators are candidates because their effects fields all match the goal. The variable `E` in both operators is therefore instantiated as `p1`. Further considering the constraints fields, the terse one is ruled out because no `haste(h)` is stored in the user model. The extended `describe` in aspect is also ruled out because the constraint `has_aspect(E,A)` is not satisfied. The surviving operator, instantiated as below, is then decomposed.

[12] Note, however, we are not certain, at the moment, on the means to establish the assignment of aspect markers for other types of text. We leave it to future study.

**Name:** Extended-description
**Header:** describe(s,h,p1)
**Constraints:** is_entity(p1),not(know_about(h,p1)),definable(p1).
**Effects:** know_about(h,p1)
**Decomposition:** define(s,h,p1),
optional(detail(s,h,p1)),
optional(divide(s,h,p1)).

As shown in Fig. 7.2, no attribute and purpose information of *Isoetes taiwanensis* exist in the domain knowledge base. Thus none of the `detail` operators in the library, as shown in Fig. 7.5, are available. On the other hand, there is information about the components of *Isoetes taiwanensis* in the knowledge base, which makes one of the `divide` operators, *divide-by-constituency*, available. Thus the above operator is decomposed in order into `define` and `divide`, which become subgoals in the next round of execution. The first subgoal, `define`, is achieved by `informing` *logical definition* which is then decomposed into an `assert` node. Upon detecting a surface act, here `assert`, the associated message content is extracted and converted into the two semantic structures shown in Fig 7.8. The first structure states that `p1`, *Isoetes taiwanensis*, belongs to `p2`, *Isoetes* and the other indicates that *Isoetes taiwanensis* is a Taiwan-unique species.

The remaining `divide` subgoal is decomposed into four `inform-detail` pairs which provide attributive information about *Isoetes taiwanensis*' subparts, leaves, stalk, sporangium and root. Each `inform` is achieved by an `assert(s,h,has_ subpart(p1,p11))`, for example, which means *Isoetes taiwanensis* has subpart leaves. Since there exists no purpose information about the subparts of *Isoetes taiwanensis*, purpose-related `detail` operators are ruled out. Similarly, neither `want_division_information` nor `want_new_info` are indicated in the user model; the two `details` with division and new elements are ruled out as well. As shown in Fig. 7.2, all the subparts have attributive information; only the plain `detail` operator is satisfied. Each instantiated plain `detail` operator is then decomposed into a sequence of informing attributes. The `informs` are then decomposed into `asserts` and the corresponding semantic structures are attached. The resulting plan tree is shown in indented form in Fig. 7.9, with the semantic structures omitted. Note that, for convenience of reference, each `assert` node is annotated with a sequential index.

The second example has the same input goal as the previous one and the following propositions in the user model

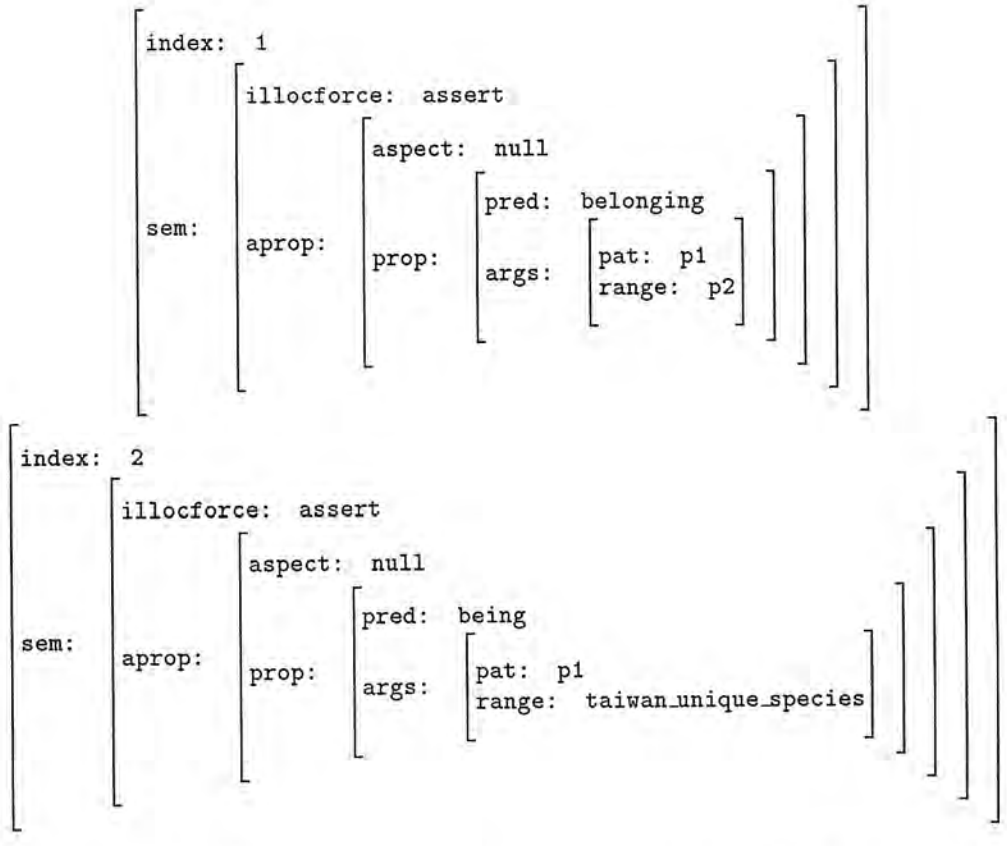



Figure 7.8: Semantic structures associated with *logical definition*.

`want_new_info(h,X)` and `want_division_info(h,X)`,
where X is in {`p11,p12,p13,p14`}.

Note that the elements in the above set are the indices of components of *Isoetes taiwanensis*. The processing of the input goal is similar to the preceding example except that `details` of subparts sporangium and root are changed to *detail-by-attribution, with division* and *detail-by-attribution, with new elements*, respectively. This is because they have information about division, subtypes of *sporangium*, and new elements, *phloem* and *air tube* of *root*, in the domain knowledge base. The resulting plan tree is shown in Fig. 7.10. Note that, since the part indexed from (1) to (13) is the same as Fig. 7.9, we omit this part for convenience in the new plan tree.

## 7.4 Linearisation

The text planner arranges message content into a hierarchical structure with leaf nodes annotated with semantic structures of sentence units. In this section, we describe the linearisation program which walks through the discourse structures created by the planner in depth-first order and linearises the message content. Within the traversal of the discourse structures, the program also performs the following tasks: (1) inserting appropriate punctuation marks between sentence units; (2) converting each semantic structure into a deep syntactic structure; and (3) deciding the forms of anaphors occurring in the semantic structures. In the following, we first introduce the representation of deep syntactic structure. Since both tasks (1) and (3) rely on proper segmentation of discourse structures, we then describe the method of discourse segmentation used in our system. Finally we describe the discourse model and the linearisation program.

### 7.4.1 Representation of deep syntactic structure

To aid the realisation of message units, we need a representation closer to the syntax of linguistic items, termed deep syntactic structure, in which ordering information of arguments is specified. Due to the topic-comment structure of Chinese, as described in Sec. 2.2, we propose the following structure as the framework for deep syntactic structure.

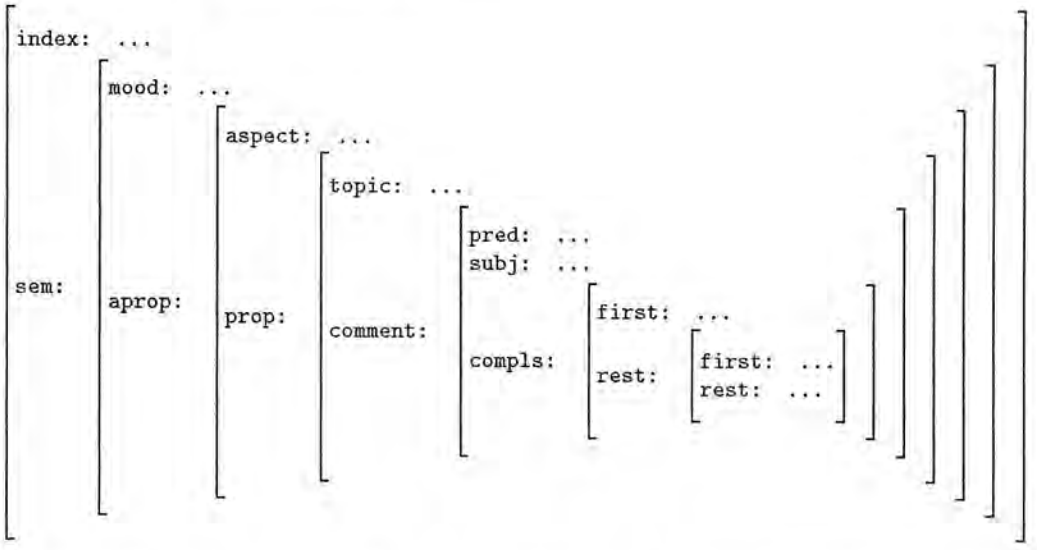
```
describe(s,h,p1)
   define(s,h,p1)
      inform(s,h,logical_definition(p1))
         assert(s,h,logical_definition(p1))..........................(1,2)
   divide(s,h,p1)
      inform(s,h,has_subpart(p1,p11))
         assert(s,h,has_subpart(p1,p11))............................(3)
      detail(s,h,p11)
         inform(s,h,attribution(p11,[appearance,triple_split]))
            assert(s,h,attribution(p11,[appearance,triple_split]))....(4)
         inform(s,h,attribution(p11,[width,about(2,cm)]))
            assert(s,h,attribution(p11,[width,about(2,cm)]))..........(5)
      inform(s,h,has_subpart(p1,p12))
         assert(s,h,has_subpart(p1,p12))............................(6)
      detail(s,h,p12)
         inform(s,h,attribution(p12,[appearance,slender]))
            assert(s,h,attribution(p12,[appearance,slender]))........(7)
         inform(s,h,attribution(p12,[base_part,spoon_shaped]))
            assert(s,h,attribution(p12,[base_part,spoon_shaped]))....(8)
         inform(s,h,attribution(p12,[cross_section,half_circled]))
            assert(s,h,attribution(p12,[cross_section,half_circled])).(9)
         inform(s,h,attribution(p12,[inner_side,flat]))
            assert(s,h,attribution(p12,[inner_side,flat]))...........(10)
         inform(s,h,attribution(p12,[length,about(20,cm)]))
            assert(s,h,attribution(p12,[length,about(20,cm)]))........(11)
         inform(s,h,attribution(p12,[outer_side,curve_]))
            assert(s,h,attribution(p12,[outer_side,curve_convexed]))..(12)
         inform(s,h,attribution(p12,[growing(zai),shang(p11)]))
            assert(s,h,attribution(p12,[growing(zai),shang(p11)]))....(13)
      inform(s,h,has_subpart(p1,p13))
         assert(s,h,has_subpart(p1,p13))............................(14)
      detail(s,h,p13)
         inform(s,h,attribution(p13,[length,about(0.8,cm)]))
            assert(s,h,attribution(p13,[length,about(0.8,cm)]))......(15)
         inform(s,h,attribution(p13,[growing(zai),neiche(p121)]))
            assert(s,h,attribution(p13,[growing(zai),neiche(p121)]))..(16)
      inform(s,h,has_subpart(p1,p14))
         assert(s,h,has_subpart(p1,p14))............................(17)
      detail(s,h,p14)
         inform(s,h,attribution(p14,[having_no,hair]))
            assert(s,h,attribution(p14,[having_no,hair]))...........(18)
         inform(s,h,attribution(p14,[organisation,simple]))
            assert(s,h,attribution(p14,[organisation,simple]))........(19)
         inform(s,h,attribution(p14,[having(nei),[p142]]))
            assert(s,h,attribution(p14,[having(nei),[p142]]))........(20)
         inform(s,h,attribution(p14,[having(zhong),[p141]]))
            assert(s,h,attribution(p14,[having(zhong),[p141]]))......(21)
```

Figure 7.9: Plan tree of extended description for `know_about(h,p1)`.

(This part is the same as (1) to (13) in Fig. 7.9.)
...

```
inform(s,h,has_subpart(p1,p13))
   assert(s,h,has_subpart(p1,p13))
detail(s,h,p13)
   inform(s,h,attribution(p13,[length,about(0.8,cm)]))
      assert(s,h,attribution(p13,[length,about(0.8,cm)]))
   inform(s,h,attribution(p13,[growing(zai),neiche(p121)]))
      assert(s,h,attribution(p13,[growing(zai),neiche(p121)]))
   divide(s,h,p13)
      inform(s,h,classification(p13,[p131,p132,p133]))
         assert(s,h,classification(p13,[p131,p132,p133]))
      detail(s,h,p131)
         inform(s,h,attribution(p131,[having(nei),p1311]))
            assert(s,h,attribution(p131,[having(nei),p1311]))
         detail(s,h,p1311)
            inform(s,h,attribution(p1311,[quan,about([300,500],ge)]))
               assert(s,h,attribution(p1311,[quan,about([300,500],ge)]))
            inform(s,h,attribution(p1311,[diameter,about([310,390],um)]))
               assert(s,h,attribution(p1311,[diameter,about([310,390],um)]))
            inform(s,h,attribution(p1311,[appearance,tetrahedron]))
               assert(s,h,attribution(p1311,[appearance,tetrahedron]))
            inform(s,h,attribution(p1311,[having(shang),p13111]))
               assert(s,h,attribution(p1311,[having(shang),p13111]))
         detail(s,h,p132)
            inform(s,h,attribution(p132,[having(nei),p1321]))
               assert(s,h,attribution(p132,[having(nei),p1321]))
             detail(s,h,p1321)
                inform(s,h,attribution(p1321,[diameter,about([15,25],um)]))
                   assert(s,h,attribution(p1321,[diameter,about([15,25],um)]))
                inform(s,h,attribution(p1321,[appearance,bihedran]))
                   assert(s,h,attribution(p1321,[appearance,bihedran]))
                inform(s,h,attribution(p1321,[having(shang),p13211]))
                   assert(s,h,attribution(p1321,[having(shang),p13211]))
         detail(s,h,p133)
            inform(s,h,attribution(p133,[having(nei),[p1311,p1321]]))
               assert(s,h,attribution(p133,[having(nei),[p1311,p1321]]))
inform(s,h,has_subpart(p1,p14))
   assert(s,h,has_subpart(p1,p14))
detail(s,h,p14)
   inform(s,h,attribution(p14,[organisation,simple]))
      assert(s,h,attribution(p14,[organisation,simple]))
   inform(s,h,attribution(p14,[having_no,hair]))
      assert(s,h,attribution(p14,[having_no,hair]))
   inform(s,h,attribution(p14,[having(zhong),p141]))
      assert(s,h,attribution(p14,[having(zhong),p141]))
   inform(s,h,attribution(p14,[having(nei),p142]))
      assert(s,h,attribution(p14,[having(nei),p142]))
```

Figure 7.10: Plan tree of extended description for `know_about(h,p1)`, with additional user's requirements.




The index plays the same role as its counterpart in semantic structure. The mood affects the type of sentence to be produced. The aspect specifies what kind of aspect marker should be used. Both index and aspect get their values from the corresponding positions in the semantic structures. The value of mood is declarative for assert. Then comes the topic-comment structure. The arguments, topic, subj and compls, are filled according to a set of mapping rules from the argument list in the semantic structure. At the current stage, four roles are used in the argument list: *experiencer*, *patient*, *range* and *location*. [13] The mapping rules specify the linking between semantic roles and syntactic argument, as below.

- If a semantic structure contains both *experiencer* and *patient*, then take the former as the topic and the other as the subject. [14]
- Otherwise, if a semantic structure only contains *patient*, then take it as the topic and set topic and subject identical.
- If *range* exists, set the first of compls to it.
- If *location* exists, set the first of compls to it.

Note that in the current domain, the pred in a semantic structure can have either a

[13] A detailed list of semantic roles can be found in [Teng 75]. The precise choice of thematic roles is not crucial to the workings of the implementation.

[14] This corresponds to the type of sentences having both topic and subject.

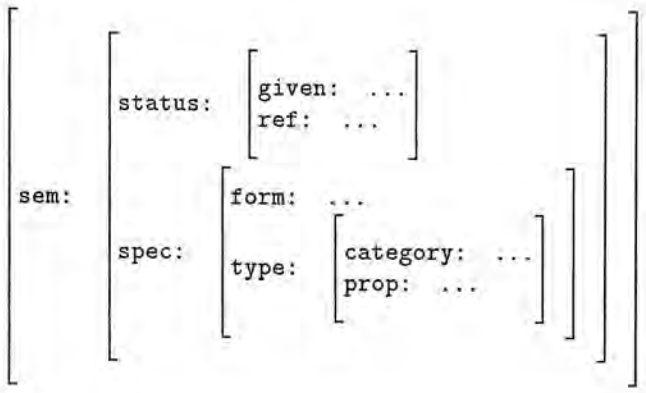



Figure 7.11: The deep syntactic structure for arguments.

`range` or a `location` as its complement. Consequently, `range` and `location` can not occur simultaneously in semantic structures. [15]

In semantic structures, referential entities are represented by indices linked to entries in the domain knowledge base. There is no information about the anaphoric form for referential entities; moreover, the content of descriptions, either in full or reduced form, are not specified if nominal anaphors are decided. The decision on whether to use a determiner relies on the status information of referential entities in the discourse where they occur. Consequently, we add features, as shown in Fig. 7.11, to encode the above information for noun phrases. This structure is inserted in the previous deep syntactic structure, where noun phrases may occur, for example, `topic`, `subj`, etc.

The `given` of `status` indicates whether the entity occurs previously in the discourse and `ref` whether the entity is referential or not. In our implementation, an entity is referred to as non-referential if it has no entry in Part b of the domain knowledge base; otherwise, it is referential. [16] The `given` of `status` is set to + if it occurs previously in the list of referential entities. Otherwise, it is set to - if it is referential but did not occur previously. In `spec`, `form` indicates the surface form of the referential entity, `zero`, `pronoun` or `nominal`. Corresponding to the decision of the nominal description rule as described in Chap. 6, `prop` either copies all or none of the corresponding value in the domain knowledge base, which results in a full or a reduced description for a

[15] The set of rules could be extended to produce other types of sentences.

[16] See Sec. 7.2 for descriptions of Part b.

nominal anaphor, respectively.

### 7.4.2 Discourse segmentation

As described in Sec. 3.4,Grosz and Sidner [Grosz & Sidner 86] suggest using cue phrases, such as *however*, *for example*, etc., and anaphoric forms, as explicit clues for judging discourse segment boundaries in language analysis. In their theory, two basic relations of intentions between discourse segments, *dominance* and *satisfaction precedence*, are used to construct the hierarchical structure of discourse segments. In the case of generation, we find ourselves in the opposite position: there exist hierarchical structures created by a planner; what we need to do is choose to generate cue phrases, anaphors and punctuation marks along with the sentences in the plan structures so that the generated text is coherent. In this thesis, we focus on the generation of anaphors and punctuation marks and leave the others to future study.

The trees produced by the planner as described in the preceding section can be referred to as discourse structures as defined in Grosz and Sidner's theory, where each internal node corresponds to a discourse segment [Hovy 93] and the hierarchical and sibling relations between internal nodes correspond to the above two intentional relations, respectively. The sequence of message units associated with leaf nodes corresponds to this linguistic structure. For illustration, the plan structure shown in Fig. 7.9 can be redrawn as the embedded structure shown in Fig. 7.12a.

As mentioned previously, there are two kinds of internal nodes in plan trees: rhetorical act and speech act nodes. In terms of communication level, the former is higher than the latter; a rhetorical act can be decomposed into speech acts, but not vice versa. In the integrated theory of communicative acts, speech act nodes deal with the illocutionary forms of individual utterances. In the case of descriptive text, the usual form of illocutionary form of utterance is informing which is ultimately achieved by an asserting surface act. A speech act node covers a single rhetorical proposition. It is not possible to decide whether the intention of a discourse segment is finished just by looking at individual speech act nodes. For example, in Fig. 7.12a, we are not able to discern the purpose of informing the attributes of the entity *p11* at the level of `inform` nodes, unless we look at a higher level. Rhetorical acts, in contrast to speech acts, have

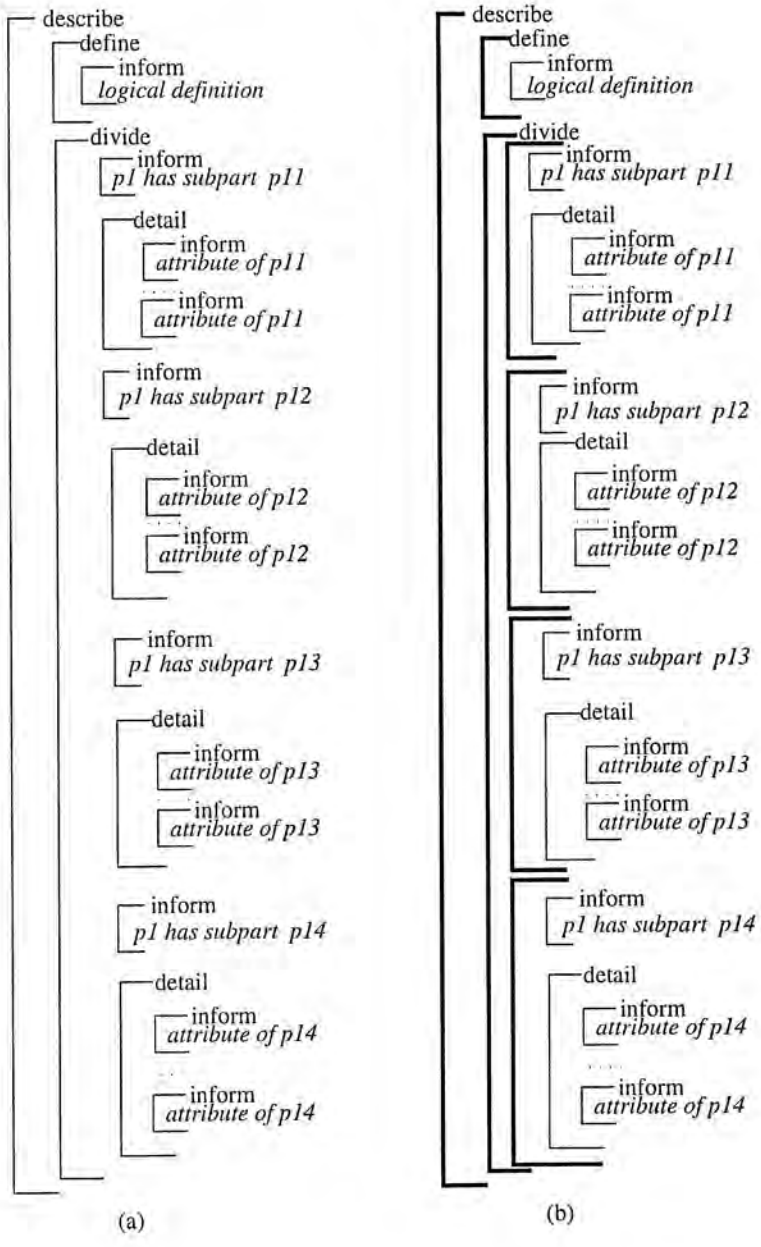



Figure 7.12: Discourse segment structure of the plan in Fig.7.9.

to do with a more abstract level of communication; they provide complete meanings of a larger size. Therefore, we choose rhetorical act nodes in plan trees as the level for discourse segmentation.

A straightforward method for segmenting the plan tree is adopting a rhetorical act node as a discourse segment. For example, the extended description operator in the plan library, as shown in Fig. 7.5 can be maximally decomposed into three rhetorical acts; therefore, it contains three embedded discourse segments. A rhetorical act, on the other hand, may be decomposed into a mixture of speech acts and rhetorical acts, such as the *divide-by-constituency* operator in the library, repeated below.

| | |
|---|---|
| **Name:** | Divide-by-constituency |
| **Header:** | divide(s,h,E). |
| **Constraints:** | has_subpart(E,). |
| **Effects:** | know(h,subpart(E,P)). |
| **Decomposition:** | forall(X^member(X,Ps), |
| | inform(s,h,has_subpart(E,X)), |
| | detail(s,h,X)). |

In the decomposition field, an `inform` is followed by a `detail` act, where the latter is used to provide more detailed information for the entity given in the former `inform` act. Obviously, it would be better to consider such pairs as integrated units; in other words, the `inform` and the following `detail` are in a discourse segment. Examining the plan library, a similar case occurs in the other `divide` operator, *divide-by-classification*. In the decomposition field of *extended-description*, the `define` is decomposed into an `inform` which gives the definition and distinguishing characteristics of the entity in question; the following optional `detail` then provides more information for the entity. The `define-detail` pair, though the former exists on a rhetorical act level, is in a sense similar to the previous `inform-detail` pairs. Therefore, the method for discourse segmentation is modified as below.

> Examining the decomposition fields of operators, if a rhetorical act subgoal is in close relation with its preceding subgoal, either rhetorical act or speech act, then the rhetorical act along with its preceding one form a discourse segment; otherwise, a rhetorical act is taken as a discourse segment.

Note that the extent of *close relation* mentioned above may vary according to the

designer's judgement on the relationship between subgoals in the decomposition fields of operators. In our implementation, we examined the subgoals in the decomposition fields of planning operators and checked the relationships between the subgoals. The *divide-by-constituency* as shown in Fig. 7.5, for example, is decomposed into a sequence of `inform-detail` pairs. Each `inform` informs that the entity in question has a subpart and the following `detail` gives detailed information about the subpart. The message contents of the `detail` would not make sense if separated from its preceding `inform`. Thus we take such a pair as a discourse segment. Similar cases can be found in *divide-by-classification*, and two sophisticated *detail-by-attribution* operators as shown in Fig. 7.5. Based on the above method, the resulting segmentation on the plan tree in Fig. 7.9 becomes the dark lines in Fig. 7.12b.

### 7.4.3 Generation of punctuation marks

From the description in Sec. 2.4, size is not an appropriate measure to judge "sentences." Therefore, we employ the idea of meaning completeness as the factor to characterise "sentences."

The idea of meaning completeness may not cause too much trouble for human writers because they more or less have the ability to infer satisfying conditions. However, it is not a straightforward process to be encoded into natural language generation by computer. A proper mapping of meaning completeness into a span of text in a plan tree is therefore an essential step towards a successful generation of punctuation marks.

From the discussion in the preceding section, in terms of meaning completeness, discourse segments are more appropriate than other levels of nodes in plan trees, i.e., speech acts and rhetorical acts, as the scope of "sentences." However, simply choosing a discourse segment as the scope of a "sentence" would result in the problem of recursion. In plan trees, discourse segments may occur at different levels. In other words, discourse segments may be embedded in others; for example, in Fig. 7.12b, the *define* segment and the *divide* segment are embedded in the *describe* segment, and the *inform-detail* pairs are embedded in the *divide*. If a discourse segment, no matter its level, is taken as a "sentence," then it may contain or be embedded in other "sentences," which is not allowed in Chinese text. For example, the corresponding embedded "sentence"

structure to Fig. 7.12b, would be as shown in Fig.7.13a, if applying the above method.

In Chinese, several "sentences" are concatenated to form a larger unit, a paragraph. Considering, for example, the discourse structure shown in Fig. 7.12b, the corresponding paragraph is formed by concatenating the lowest level segments. In implementation, we have to ignore the higher level segments to eliminate the recursion problem, which results in, for example, the structure as shown in Fig. 7.13b. In the linear "sentence" structure, a sentential mark is inserted at the end of a "sentence" and, otherwise, a comma is inserted.

### 7.4.4 The linearisation program

We first describe the framework of the linearisation program which is used to achieve the three tasks mentioned at the beginning of this section. Then we provide a detailed description of the respective tasks which happen at certain points in the framework and the data structures used in the program.

The framework of the linearisation program is a depth-first traversal, as shown in Fig. 7.15, of the plan tree produced by the planner. The plan tree is represented as the following recursive list, where parent is the root node and $\text{child}_i$'s are its child nodes.

$$[\text{parent},[\text{child}_1,\text{child}_2, \cdots, \text{child}_n]].$$

Each $\text{child}_i$ can either be a terminal node that contains a list of semantic structures or a non-terminal node that has the above structure. For example, in the plan tree shown in Fig. 7.14, the root `describe( ...)` is a parent node which has two child nodes, `define(...),[inform ...]` and `detail(...),[inform ...]`; `define(...)` and `detail(...)` are parent nodes which have one and four `informs` as their child nodes, respectively. Each terminal node contains a list of semantic structures, for example, `[[index:1, ...]]`.

For convenience of illustration, we employ the notation for list structures used in Prolog [Clocksin & Mellish 94] in the algorithm. The above list structure is represented in a list of Prolog variables, `[ParentNode,ChildNodes|More]`. Initially, in the depth-first traversal, the root of the input plan tree becomes the first `ParentNode`, `ChildNodes`

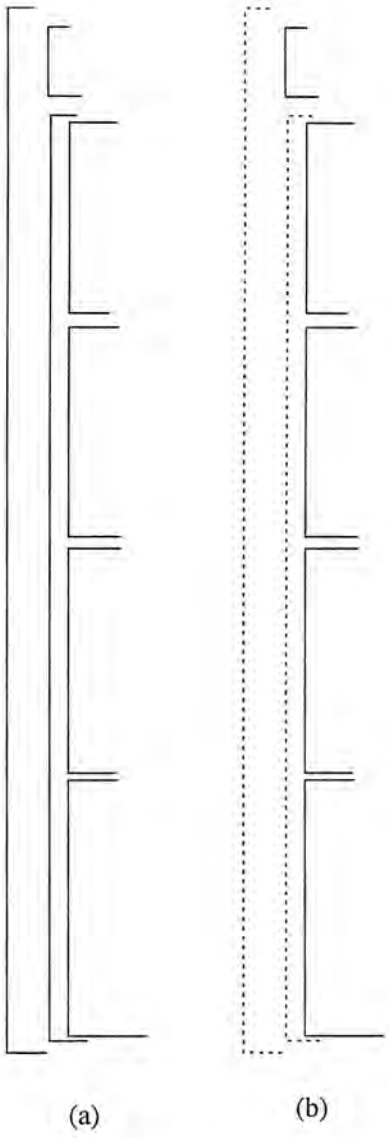



Figure 7.13: (a) The embedded "sentence" structure and (b) the resulting linear "sentence" structure for the plan tree in Fig. 7.11b by ignoring the higher level segments.

```
[describe(s,h,m11),
   [define(s,h,m11),
      [inform(s,h,logical_definition(m11)),
         [assert(s,h,logical_definition(m11)),
            [[index:1, ...]]
         ]
      ],
    detail(s,h,m11),
      [inform(s,h,attribution(m11,[height,null(977,m)])),
         [assert(s,h,attribution(m11,[height,null(977,m)])),
            [[index:2, ...]]
         ],
       inform(s,h,attribution(m11,[mt_shape,hun_yuan_zheng_qi])),
         [assert(s,h,attribution(m11,[mt_shape,hun_yuan_zheng_qi])),
            [[index:3, ...]]
         ],
       inform(s,h,attribution(m11,[having(ding_bu),m111])),
         [assert(s,h,attribution(m11,[having(ding_bu),m111])),
            [[index:4, ...]]
         ],
       inform(s,h,attribution(m11,[having(ding_bu_xi_che),13])),
         [assert(s,h,attribution(m11,[having(ding_bu_xi_che),13])),
            [[index:5, ...]]
         ]
      ]
   ]
]
```

Figure 7.14: An example of a plan tree as a list structure.

is instantiated as the child nodes list of the root, and `More` is an empty list. At later stages, `More` is used to denote remaining nodes to be processed.

The discourse model, implemented by `Stack` in the algorithm, contains information about the discourse and is used by the linearisation program to determine how to refer to entities mentioned previously in the discourse. In the discourse model, we mainly distinguish local and global focus, as in [Grosz & Sidner 86, Dale 92]. The former consists of two parts: the semantic structures of the current and immediately preceding sentences which are directly taken from the leaf nodes in the plan tree. To facilitate the decision of anaphoric form, each semantic structure is appended with a list of entities occurring in the structure in which the first and second elements are topic and object-1 in the sentence and the rest are the others. The global focus consists of (1) a stack of focus spaces for discourse segments that have not been closed; and (2) a list of focus spaces corresponding to those discourse segments already closed [Dale 92]. A new focus space is pushed onto the stack when encountering a new discourse segment. The new focus space consists of a list of entities, initially empty, occurring so far in the discourse segment. The top of the stack is popped off and appended to the list of closed discourse segments when a discourse segment is finished.

The program also maintains the text plan path for the current and previous sentence, `Curr_path` and `Prev_path`. Each path is a sequence of rhetorical act node names from the root to the embedding node in the plan tree. For example, the path corresponding to the sentences of `inform-`*logical-definition* in Fig. 7.12 is `describe|define`. [17] When the traversal enters a new rhetorical act node, the node name is appended to the `Curr_path`. On the other hand, when a rhetorical act node is finished, the corresponding node name is removed. These two paths are used to facilitate the detection of discourse segment boundaries that will be used in the generation of punctuation marks and nominal descriptions.

When the program in Fig 7.15 is invoked, it first checks the base condition in L1. If the input is an empty list, then it returns. Otherwise, if the current parent node is a rhetorical act, then the program prepares a new focus space if it is at the beginning of a discourse segment. After recursing on the child nodes of the current node in Step

[17] We use | as the separator between names in a path.

```
Initialisation: the input to the algorithm is a list of parent-children pairs
                represented in Prolog, [ParentNode,ChildNodes|More], where
                ParentNode and ChildNodes is the first pair and More is the
                list of the rest pairs. When dft is invoked, the root and its
                child nodes in the input plan tree are the only pair in the
                initial input list and More is empty initially.
                We use Curr_path and Prev_path to denote the plan
                tree paths of the current and previous encountered
                leaf nodes in the plan tree.
                We use Stack to denote the focus space stack.

L0: dft([ParentNode,ChildNodes|More]);
L1      if the input list is empty
        then return;
L2:     else if ParentNode is a rhetorical act
L3:          then if it is at the beginning of a discourse segment
                  then push a new focus space onto Stack;
L4:               append a node to Curr_path;
L5:               push a focus space onto Stack;
L6:               call dft(ChildNodes);
L7:               pop a focus space off Stack;
L8:               remove a node from Curr_path;
L9:               call dft(More);
L10:         else if it is a speech act
L11:              then call dft(ChildNodes);
L12:                   call dft(More);
L13:              else %it is a surface act
L14:                   insert a punctuation mark;
L15:                   output the deep syntactic structure of the
                       associated message;
L16:                   update Prev_path as Curr_path;
```

Figure 7.15: Algorithm for linearisation.

L6, a focus space is popped off and the last node is removed from the current path. Then the program recurs on the remaining parent-children pairs in Step L9. If the current node is a speech act, the program recurs on the child nodes and the remaining parent-children pairs in Steps L11 and L12.

Finally, the current node is a surface act; the program performs the three tasks of linearisation mentioned earlier. In the following, we explicate the task of generating punctuation marks, i.e., `L14`. The remaining two tasks, implemented in step L15, are described in the next subsection. We established a set of rules which take into account values of `Curr_path` and `Prev_path` for the determination of using sentential mark or comma. We examined the planning operators to get all possible expansions of extended descriptions and then noted down the conditions under which sentential marks and commas are used. Each condition can be characterised as a pair of paths of leaf nodes. The rules for the determination of whether to use a sentential mark or a comma are then created by listing the conditions as pairs of paths. In our work, we found that the number of cases corresponding to sentential mark is far less than its counterpart. Since the two punctuation marks in our system are mutually exclusive to each other, instead of enumerating all rules for both cases, we only list the sentential mark cases, as shown in Fig. 7.16. Any pair of paths not satisfying the `sentential_rule` is taken to indicate a comma. For example, in Fig. 7.12b, when the traversal reaches the leaf node at the beginning of `divide`, the `Curr_path` is `describe|divide` and the `Prev_path` is `describe|define`. The program takes these two values to consult the set of rules and gets a sentential mark as the result. Next, when the traversal visits the first leaf node under the succeeding `detail`, the `Curr_path` is `describe|divide|detail` and the `Prev_path` becomes `describe|divide`. Taking these two values to consult the set of rules, a comma is obtained.

The current rule base for the determination of punctuation marks, though it is able to account for the plan trees based on the current planning operators, has deficiencies in dealing with recursive operators, for example. Also, when new operators are entered into the library, new rules are needed. In the future, we will develop a more general way for the determination of punctuation marks.

```
if sentential_rule(Curr_path,Prev_path) then
   return('.');
else
   return(',').

sentential_rule(describe|describe|define,
                describe|describe|detail).
sentential_rule(describe|divide,
                describe|define).
sentential_rule(describe|divide,
                describe|divide|detail).
sentential_rule(describe|describe|divide,
                describe|describe|divide|detail).
sentential_rule(describe|divide,
                describe|detail).
sentential_rule(describe|describe|divide,
                describe|describe|define).
sentential_rule(describe|describe|divide,
                describe|describe|detail).
```

Figure 7.16: Rules for determination of punctuation marks.

### 7.4.5 Choosing anaphoric forms

The decision of anaphoric forms is carried out when building up deep syntactic structures for the arguments in the corresponding semantic structures. The conversion of semantic structures into deep syntactic structures is based on the representation and set of mapping rules described in Sec. 7.4.1. In the following, we describe the implementation of the decision of anaphoric forms. The algorithm for the decision of anaphoric forms, as shown in Fig. 7.17 is invoked at the beginning when the deep syntactic structure for an argument, as shown in Fig. 7.11, is to be constructed. [18] In the algorithm, the main procedure first gets the referential status of the argument in question. Note that in our system an argument is referred to as referential if it has an entry in the domain knowledge base; otherwise, it is non-referential. The referential status of the argument is established by checking the entities occurring so far. An argument is an initial reference if it is referential and does not occur previously, while it is a subsequent reference if it is referential and was mentioned previously. If

[18] In the current implementation, we do not take into account the situation of distracting elements in discourse. We only choose a full or reduced description for a nominal anaphor.

the argument is a subsequent reference, then `decide_zpn` is called upon to decide the anaphoric form. It first gets the values of the conditions for the decision of whether using a zero or non-zero form.

Then it consults the zero anaphor rules given the values just obtained. In `decide_zpn`, the anaphoric form is set according to the result of consulting the zero anaphor rules. If a non-zero form is decided, then the animacy condition of the argument is further checked. This is done by simply checking the `animate` field of the corresponding entry in the domain knowledge base, as described in Sec. 7.2.

If it is animate then the `form` is set to `p`. Otherwise, the procedure gets the latest occurrence of the entity. If it is in the current discourse segment, then it sets the `prop` field to `null`; otherwise, it copies the `prop` from the corresponding entry in the domain knowledge base. [19]

Considering the plan tree shown in Fig.7.12, for example, the resulting deep syntactic structures for the `assert`-*logical-definition* are shown in Fig. 7.18 with indices 1 and 2. Both structures have their topic and subject identical. In the first structure, the topic is an initial reference; thus it receives "-" as the value of `given`. The topic in the other structure occurred in 1 and it satisfies the condition of the zero anaphor rule; thus, `given` is set to `+` and `form` is set to `z`. Another deep structure, shown in Fig. 7.19, also has identical topic and subject. The topic satisfies the condition for a zero anaphor; it is zeroed. The first element in `compls`, which will fill in the object position in the syntactic structure, is expressed as a nominal anaphor and its latest occurrence lies in a different discourse segment; thus a full noun phrase is used.

## 7.5 Realisation

The realisation program takes the form of a sequence of deep syntactic structures intermixed with punctuation marks as input and produces multisentential text with punctuation marks inserted. The task of generating punctuation marks has been achieved in the linearisation program. Here, we focus on the realisation of surface strings from

[19] Note that, as described in Sec. 5.3, the pronominalisation rule can be further improved by taking some heuristic rules. In this chapter, we do not implement these heuristic rules.

```
Initialisation: Input is an entity, E and output is the deep syntactic
                structure, F, of the entity as a feature structure.
                All the set commands below are assigning values to
                features in F.

A1: get the referential status, S, of E;
A2: case S of:
       initial reference ->
          set given as -;
          use the full description for E;
       subsequent reference ->
          set given as +;
          call decide_zpn(E);
       non-referential ->
          set ref as -;

decide_zpn(E)
D1: call locality(E,L), syn_const(E,C), ds_boundary(D), and
    salience(E,S), to get the values of locality, syntactic
    constraints, discourse segment boundaries and salience
    for E in variables L, C, D, S, and call z(L,C,D,S,Anaphor).
    The variable Anaphor returns either zero or non-zero.
D2: if Anaphor is zero then
       set form as z;
D3: else if Anaphor is non-zero then
            if E is animate then
D4:            set form as p; % a pronoun
D5:       else set form as n; % a nominal form
D6:            find the latest occurrence of E, T;
D7:            if E and T belong to the same discourse segment then
D8:               set property as null;
D9:            else
                  set property as full.
D10:           set the category from the substance in the domain
               knowledge base;
```

Figure 7.17: Algorithm for the decision of anaphoric forms, Part 1.

```
locality(E,L):
   get the list of entities, List, in the immediately preceding sentence
   from the discourse model;
   if E is in List, then return i for L; % means immediate case
   else return 1;

syn_const(E,C):
   if E is being used as a range or location then return y for S;
   else if the current sentence has both experiencer and patient
           and E is used as a patient then return y for S; % means violating
                                                           % constraints.
        else return n for S;

ds_boundary(D):
   get the values of Curr_path and Prev_path; % current and previous
                                              % plan tree paths;
                                              % see Sec. 7.4.4 for details.
   consult sentential rules in Fig. 7.16 by giving
   Curr_path and Prev_path ;
   if '.' is obtained then return y for D; % means segment beginning
   else return n for D;

salience(E,S):
   get the semantic structure of the current and previous sentences
   from the discourse model;
   if E occurs as the topic or a part of the topic of current and previous
   sentence then return y for S; % means salient
   else return n for S;

% Decision whether to use zero or non-zero form.
% An anaphor is zeroed if it is local, not violate any syntactic
% constraint, not at the beginning of discourse segment and salient;
% otherwise, it is non-zeroed.
z(L,C,D,S,Anaphor)
   if L=i, C=n, D=n and S=y, then return zero for Anaphor;
    else return non-zero for Anaphor.
```

Figure 7.17: Algorithm for the decision of anaphoric forms, Part 2.

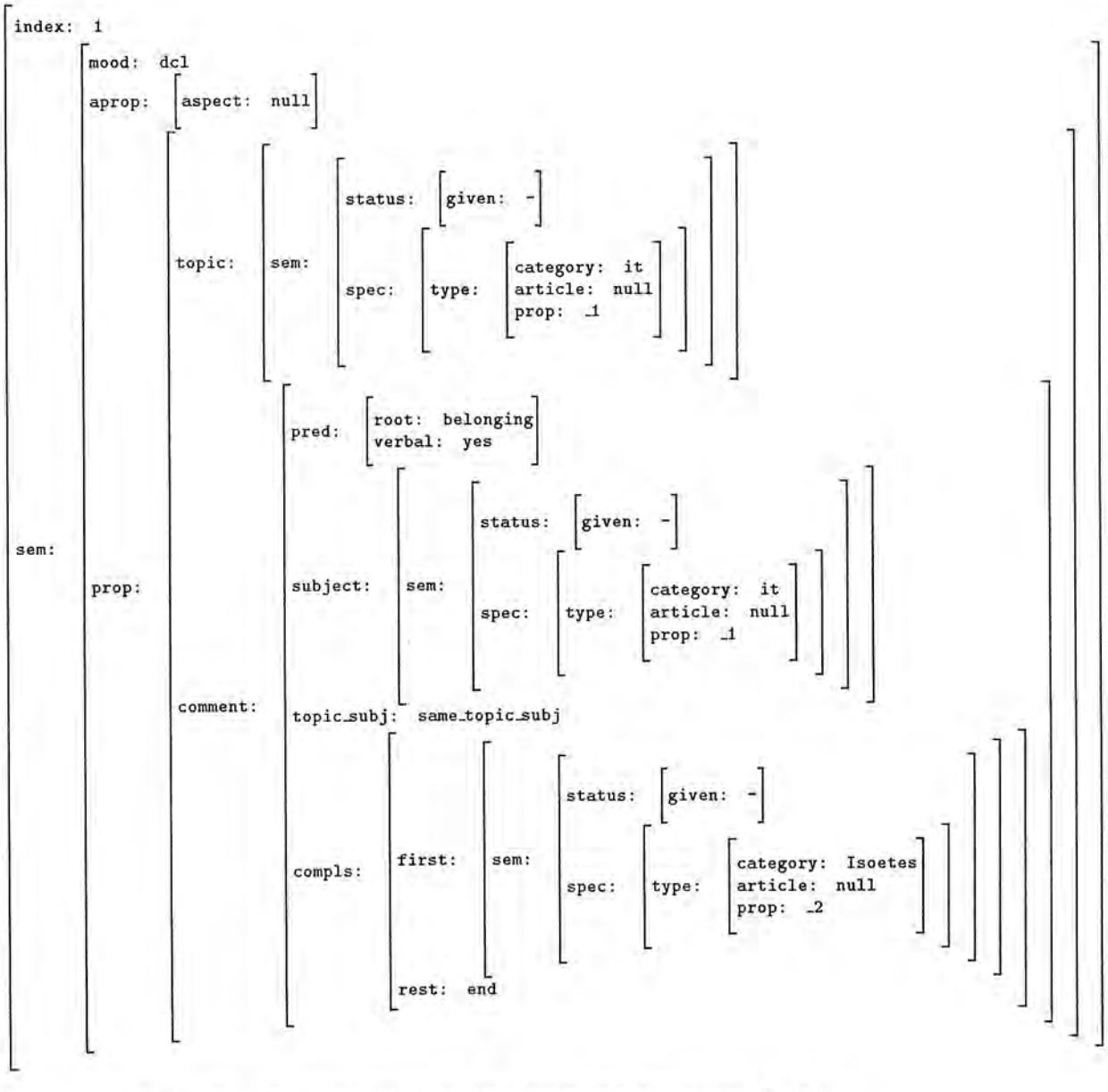



Figure 7.18: Two examples of deep syntactic structures, Part 1.

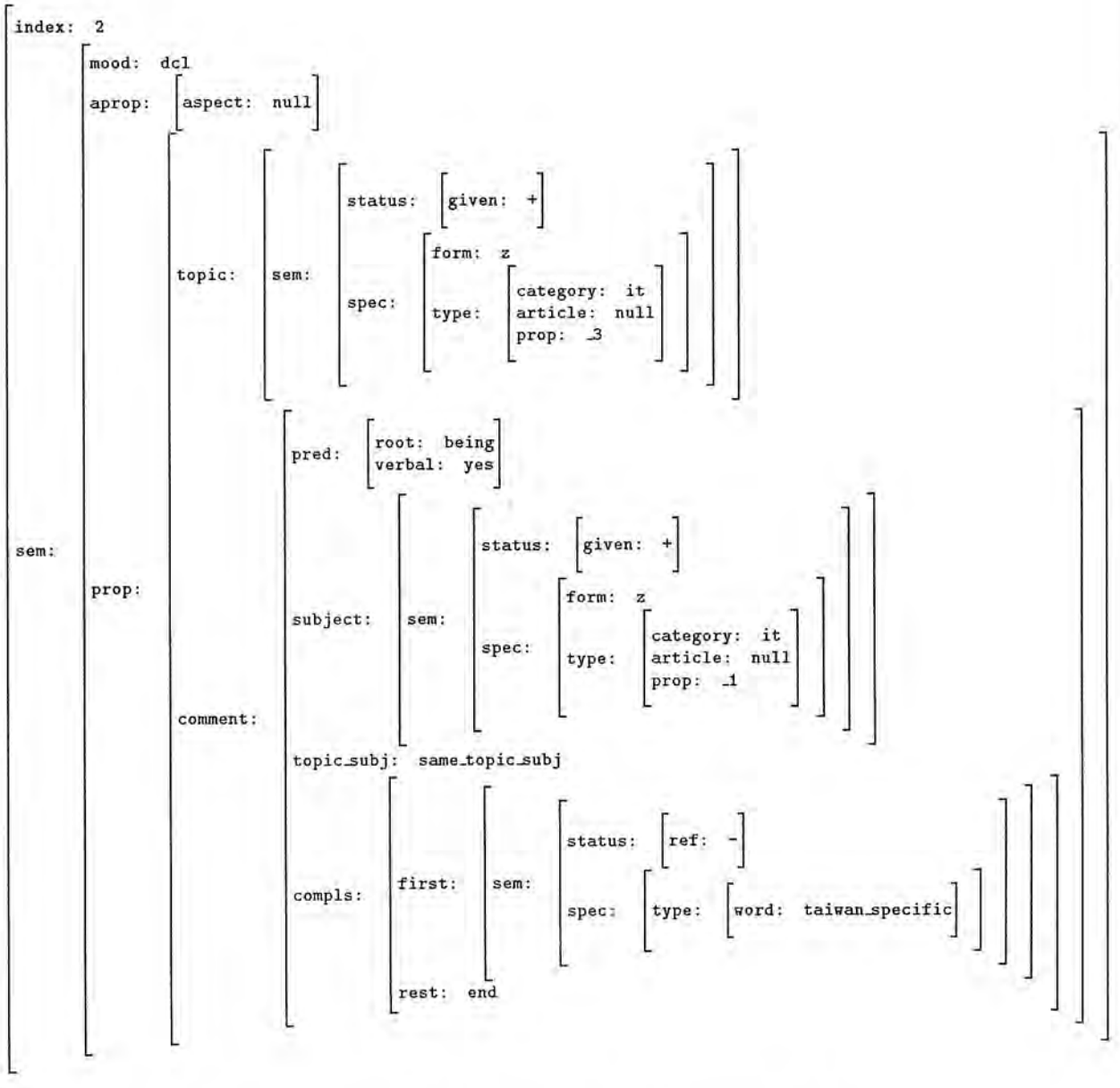



Figure 7.18: Two examples of deep syntactic structures, Part 2.

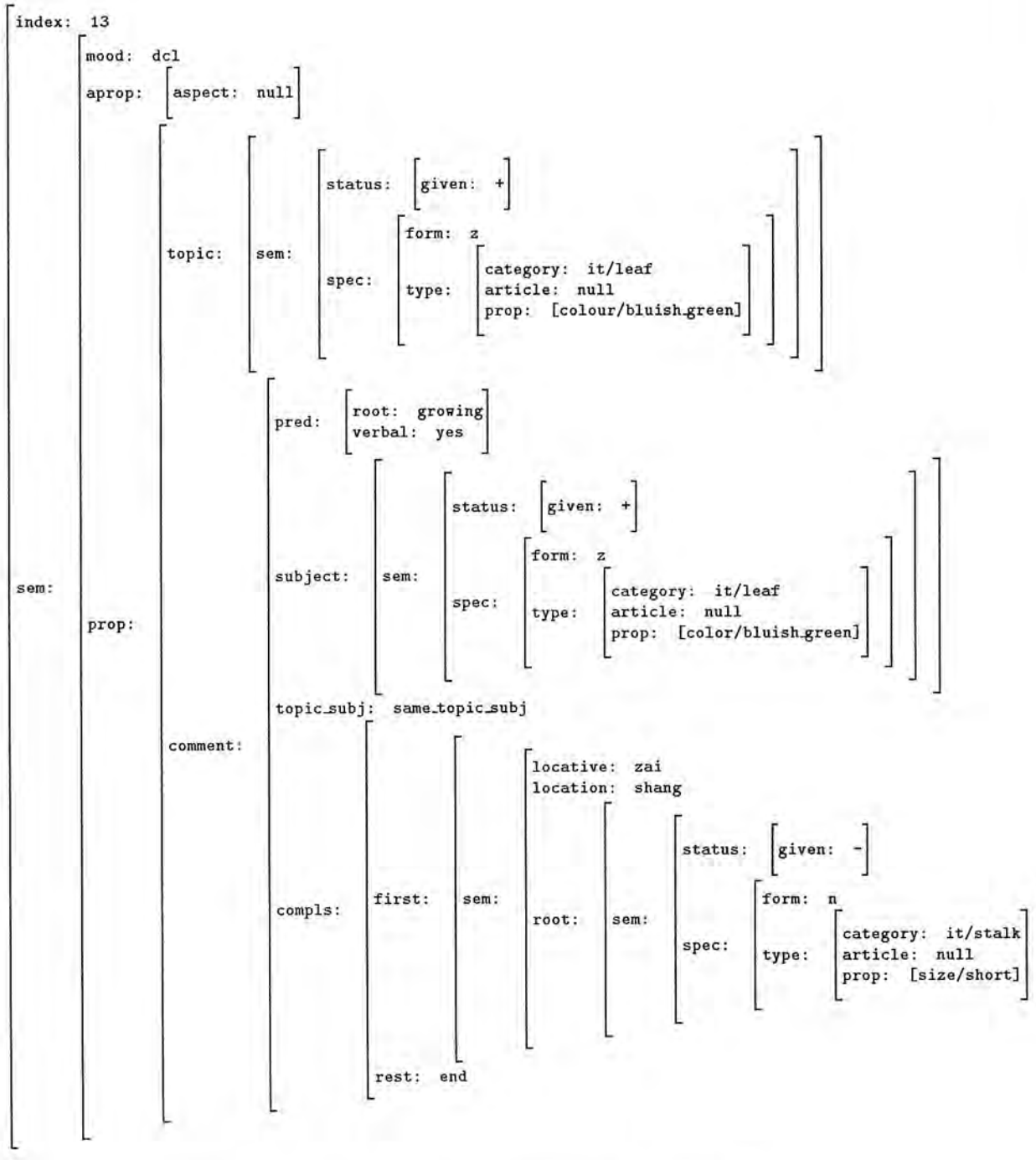



Figure 7.19: Another example of deep syntactic structure.

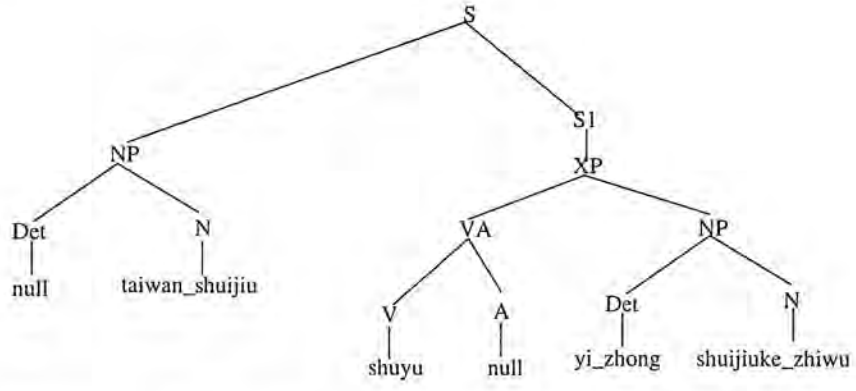



Figure 7.20: Syntax tree of the deep syntactic structure in Fig. 7.17.

deep syntactic structures. The essential task for realisation is to take elements in a deep syntactic structure, map them to syntactic roles according to a set of syntactic rules, and get the surface words from a lexicon. In this section, we first describe the set of syntactic rules and lexicon used in our system. Then we describe the realisation program with some illustrative examples.

### 7.5.1 Syntax rules and lexicon

The purpose of syntactic rules and lexicon is to map the deep syntactic structure, as shown in Fig. 7.19, for example, to the syntactic tree as shown in Fig. 7.20. The syntactic rules used here are adapted from [Dale 92]. The representation of the syntactic rules and lexical entries is in the PATR-II formalism [Shieber 86]. At the beginning, an `S`-rule, as shown below, takes up the topic-comment structure where the `NP` and `S1` get the topic and comment in `S`, respectively. [20] This rule is selected only when the input is a complete deep syntactic structure, for example, Figs. 7.18 and 7.19. The information specified in deep syntactic structures guides the generator to avoid over-generation. For example, the `index` feature in a deep syntactic structure indicates that it is a complete structure; the `topic_subj` in a deep syntactic structure guides the generator to choose one of the `S1` rules as shown later in Fig 7.22.

```
S ---> [NP,S1]
```

[20] As in [Dale 92], some constraints which are used to improve the efficiency of the program and limit over-generation by the syntax rules are not shown here.

```
(1) NP ---> Word
        NP:sem:status:ref = -,
        NP:sem:spec:type:word = Word.

(2) NP ---> Name
        NP:sem:spec:type:name = Name.

(3) NP ---> null
        NP:sem:spec:form = z.

(4) NP ---> pronoun
        NP:sem:spec:form = p.

(5) NP ---> [Det,NP1]
        Det:sem:given = NP:sem:status:given,
        Det:sem = NP:sem:spec:type:article,
        NP1:sem:spec:type:category = NP:sem:spec:type:category,
        NP1:sem:spec:type:prop = NP:sem:spec:type:prop.
(6) NP1 ---> [Adjs,N]
        Adjs:sem:prop === NP1:sem:prop,
        N:category = noun,
        N:content = NP1:sem:spec:type:category.
(7) Adjs ---> Adj_words
        Adj:sem:prop = P,
        get_adj_word(P,Adj_words).
```

Figure 7.21: Syntax rules for noun phrases.

```
NP:sem = S:sem:aprop:prop:topic:sem,

S1:sem = S:sem:aprop:prop:comment,

S1:sem:aspect = S:sem:aprop:aspect.
```

We have a set of NP rules, as shown in Fig. 7.21, to deal with non-referential entities, proper names, null word, i.e., zero anaphor, pronouns, and noun phrases. The comment structure itself, S1 in the S-rule, may be a subject followed by a predicate or just a predicate, where the predicate can be verbal or nominal. Consequently, we establish a set of rules, as shown in Fig. 7.22, to cope with the comment structure. The two S1-rules are for sentential and non-sentential comment, respectively, where XP denotes verbal or nominal predicate. The first S1 is selected only when the topic and comment are different. If they are the same, the second S1 is selected. Note that the topic_subj of the comment in a deep syntactic structure, for example, Fig. 7.18, provides information

to choose between the S1 rules. Rules 3 and 4 represent intransitive and transitive verb phrases, where VA is then analysed as a verb and an aspect marker, as in Rule 7. Rules 5 and 6 are for nominal predicates, such as the underlined in (1) and (2).

(1)yezi jibu tangcizhuang.

leaf base-part spoon-shape

The base part of leaf is spoon-shaped.

(2)yezi chang yue ershi gongfen.

leaf length about 20 cm

The leaf length is about 20 cm.

Rule 6 is specific for nominal predicates like yue ershi gongfen, where yue (about) is a range adverb [Chao 68], ershi (20) is a number and gongfen (cm) is a standard measure. Note that "=.." in Rule 6 is an operator in Prolog which matches the functor and arguments in a predicate into a list [Clocksin & Mellish 94], for example, `about(20,cm)=..[about,20,cm]`.

When the construction of the parse tree reaches leaf nodes, the parser obtains the corresponding surface words from the lexicon by matching the category and content. Basically, a lexical entry contains `category` and `content` as the constraints to guide the selection of appropriate words. For example, typical lexical entries are shown in Fig. 7.23. Note that `it` in (1) is a shorthand notation for *Isoetes taiwanensis*. The arguments for the `word` predicate are the surface expression and feature structure of the corresponding word, respectively.

### 7.5.2 Realisation program

The realisation program is coded based on a top-down parser [Gazdar & Mellish 89], as shown in Fig. 7.24. For example, taking the first deep syntactic structure in Fig. 7.18 as the input to the algorithm, the S-rule is selected, where the NP matches the topic and S1 takes the comment part. The S1 is decomposed according to Rules 2 and then 4 to obtain the verb phrase. It is decomposed further and a syntax tree is finally obtained as shown in Fig. 7.20. During the construction of the parse tree, surface words are printed out, which results in the surface string

```
(1) S1 ---> [NP,XP]
        NP:sem = S1:sem:subj:sem,
        XP:sem:pred = S1:sem:pred,
        XP:sem:aspect = S1:sem:aspect,
        XP:sem:compls = S1:sem:compls.
(2) S1 ---> [XP]
        XP:sem:pred = S1:sem:pred,
        XP:sem:aspect = S1:sem:aspect,
        XP:sem:compls = S1:sem:compls.
(3) XP ---> [VA]
        XP:sem:pred:verbal = +,
        VA:pred:root = XP:sem:pred:root,
        VA:sem:aspect = XP:sem:aspect.
(4) XP ---> [VA,NP]
        XP:sem:pred:verbal = +,
        VA:pred:root = XP:sem:pred:root,
        VA:sem:aspect = XP:sem:aspect,
        NP:sem = XP:sem:compls:first:sem.
(5) XP ---> [N]
        XP:sem:pred:verbal = -,
        N:category = noun,
        N:content = XP:sem:pred:root.
(6) XP ---> [Adv,Num,Unit]
        XP:sem:pres:verbal = -,
        XP:sem:pred:root = Root,
        Root=..[A,N,U],
        Adv:category = adv_range,
        Adv:content = A,
        Num:category = number,
        Num:content = N,
        Unit:category = stand_measure,
        Unit:content = U.
(7) VA ---> [V,A]
        V:category = verb,
        V:content = VA:sem:pred:root,
        A:category = aspect,
        A:content = VA:sem:aspect.
```

Figure 7.22: Syntax rules for the comment structure.

```
(1) word(taiwan_shuijiu,W):-
        W:category = noun,
        W:content = it.
(2) word(gen,W):-
        W:category = noun,
        W:content = it/root.
(3) word(shuyu,W):-
        W:category = verb,
        W:content = belonging.
(4) word(gongfen,W):-
        W:category = stand_measure,
        W:content = cm.
(5) word(yue,W):-
        W:category = range_adv,
        W:content = about.
```

Figure 7.23: Typical lexical entries.

```
Initialisation: input is a deep syntactic structure and output is the
                corresponding surface string.
                In each iteration, Deep is the current deep structure
                to be matched with the syntactic rules; initially, it is
                set to the input.

R1: find a rule, Mother ---> Daughters, instantiated via satisfying the
    constraints, to match the current deep syntactic structure;
R2: case Daughters of:
R3:    empty ->
          return;
R4:    a word ->
          print it out;
R5:    a list, [Head|Tail] ->
          recurse R1 on Head,
          recurse R2 on Tail.
```

Figure 7.24: Algorithm for realisation program.

taiwan_shuijiu shuyu yi_zhong shuijiuke_zhiwu

# Chapter 8

# Evaluation

## 8.1 Introduction

We have established several rules for the generation of anaphors in Chinese, including the decision of zero, pronominal and nominal anaphors, and a rule for the choice of a description if a nominal anaphor is decided upon. In the preceding chapter, these rules were implemented in our Chinese natural language generation system and a number of texts for describing entities in a national park were generated. As shown in Chaps. 4 to 6, these rules were obtained from empirical studies. The experimental results show that the anaphors generated by using these rules largely match the ones in the test texts, assuming the same semantic structures and contextual information. This shows the performance of the rules. Previous test data was the training data which has more complicated semantic structures than the ones in our system. Furthermore, the assumed contextual information, for example, discourse structures, may be different to the one implemented in our system. Thus, the performance of the anaphor generation algorithm based on the previous rules may be different to the experimental results. In this chapter, we attempt a post-evaluation by asking some native speakers of Chinese to judge the result of the generated anaphors.

## 8.2 Previous Work and Our Approach

Though the field of natural language generation has progressed towards composing complex texts, the evaluation of natural language generation systems has remained

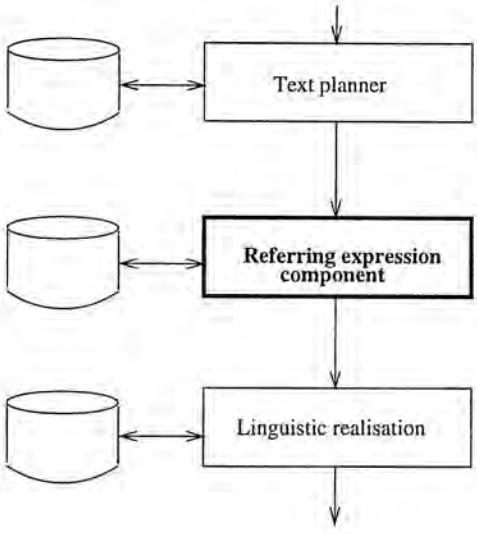



Figure 8.1: Referring expression component in the Chinese natural language system.

at the discussion stage [Maybury 90, Meteer & McDonald 91]. Two broad methods have been identified for evaluating natural language generation systems: *glass box* and *black box* evaluation [Meteer & McDonald 91]. The *glass box* method is concerned with examining the internal working of individual components in a system, while the latter looks at the behaviour of input and output to the generation systems. The difficulty of the *glass box* method is the lack of a clear division between components in generation systems. The problem of the *black box* method, on the one hand, is determining the appropriate input for generation. On the other hand, it is very difficult to be objective in evaluating the output text.

In this chapter, we aim to investigate the quality of anaphors generated by the referring expression component in our Chinese natural language generation system. As shown previously, the referring expression component lies between the text planner and the linguistic realisation component in the system, as briefly repeated in Fig. 8.1. As is discussed in Chaps. 4 to 6, the algorithm of the referring expression component first determines an appropriate form for an anaphor to be generated and then chooses a description if a nominal form is decided upon. Suppose that the referring expression components we wish to compare all adopt the above basic algorithm. Then the essential characteristic to distinguish them from each other becomes the rules used in the

components and how these rules are implemented. If all of these referring expression components are embedded in the same Chinese natural language generation system, as in Fig. 8.1, for example, then, given an input to the system, anaphors in the resulting texts can be characterised as the rules used in the referring expression component and their implementation. By adopting this approach, we need not worry about the problems of both evaluation methods stated above, except the objective evaluation of output text. Since there is no machine that can read the generated texts and give an impartial judgement about them, we rely on the opinions of human readers who are native speakers of Chinese to investigate the quality of the generated anaphors. This is an easier task than assessing the quality of whole texts. To compensate for possible bias among the individual readers, we sent the output texts to a group of readers for viewing and took the average of their outcomes as the measurement. In brief, each object system in our evaluation work is thought of as having the same individual components, including control and knowledge bases, except that the anaphor generation rules used in the referring expression components are different to each other. In practice, we employ our Chinese natural language generation described in the preceding chapter as the backbone of the evaluation work because it is easy for us to control and maintain. What we have to do for each generation system is simply to replace the corresponding generation rule.

## 8.3 Systems to Compare and the Test Task

Having described the framework of evaluation, in this section, we give details about the object systems to be compared in the evaluation work and the tasks to be performed in the evaluation work.

### 8.3.1 Systems to compare

In the existing literature, we cannot find other work on the generation of Chinese referring expressions, which means that we have no real working systems to compare with. Thus, we turn our attention to working out some anaphor generation rules that may be used in Chinese generation systems. The anaphor generation rule we obtained, as repeated in Fig. 8.2, consists of a number of constraints that, as investigated in

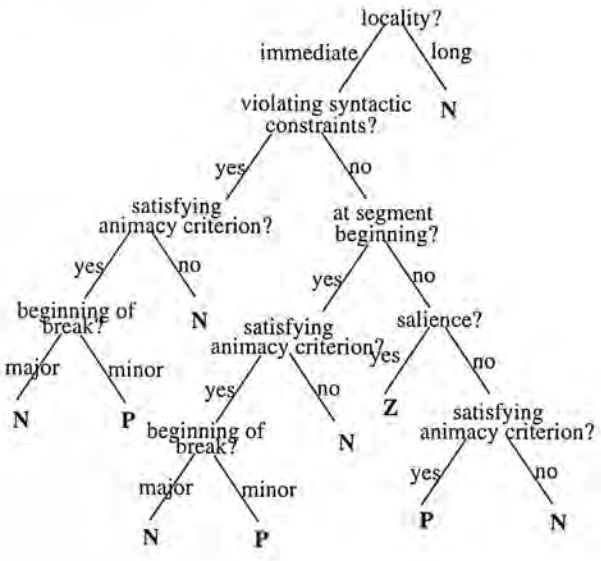



Figure 8.2: A Chinese anaphor generation rule.

Chaps. 4 to 6, were obtained by consulting relevant linguistic studies. Consequently, subsets of constraints in the above rule can be thought of as possible rules, if not complete, for the generation of anaphors in Chinese. In this chapter, we equipped each system with such a possible anaphor generation rule.

Since as described in Chap. 7 we did not implement the constraint of "beginning of breaks" in Fig. 8.2, we ignore it in the following test. We chose three rules, termed TR*1* TR*2* and TR*3*, with different complexities among the possible candidates as the targets of the test. The rules are shown in Fig. 8.3. The first one uses locality, syntactic constraints and animacy. The second and the third rules have one additional constraint, namely, discourse segment boundaries and salience, respectively, added to their predecessors. In the following, we use the above rule names to represent the systems.

### 8.3.2 The test task

The task can be divided into an annotation and a comparison stage. Each of twelve native speakers of Chinese was given a number of test sheets to finish. On each sheet is a text generated by our generation system. Each anaphor position in a generated text was left empty and all candidate forms of the anaphor, including zero, pronominal, and

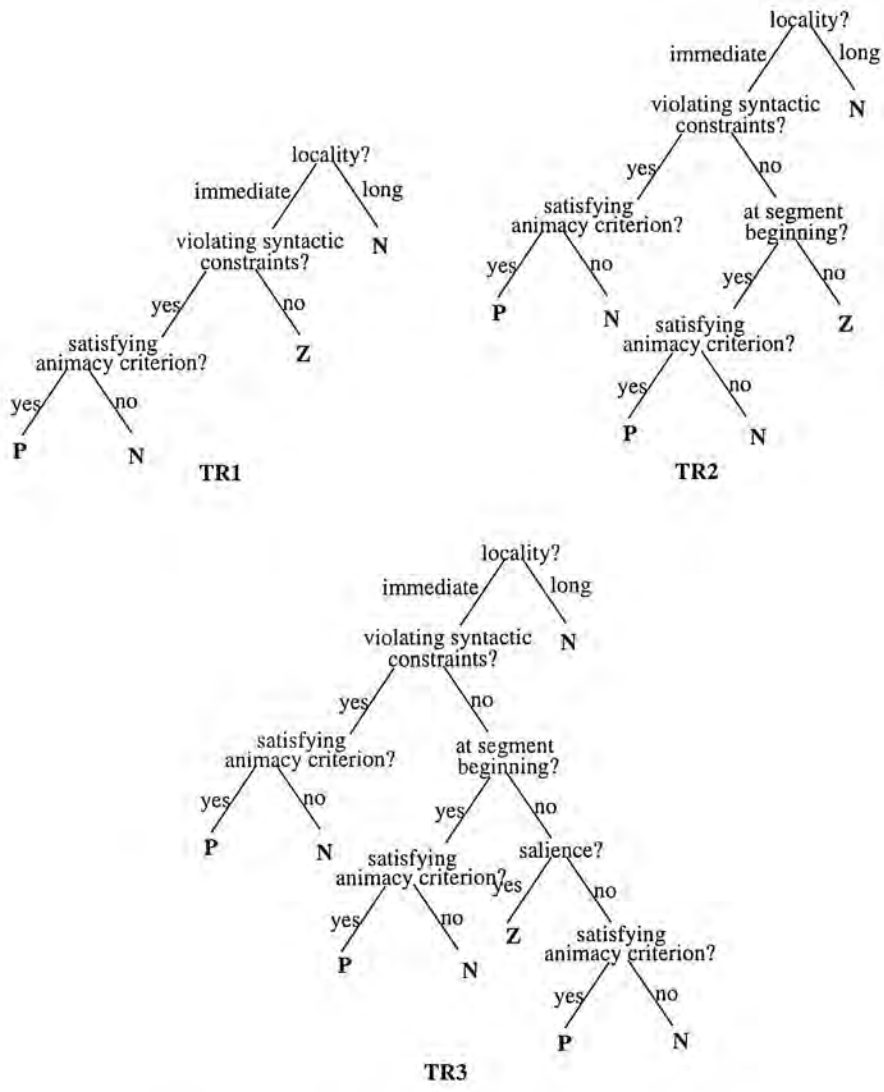



Figure 8.3: Rules used in the comparison systems.

full, or reduced description were put under the empty space. The task for a speaker to perform was to annotate which form he or she preferred for each anaphor position on the sheets. We selected five texts generated by our system for the test. The numbers of clauses in the texts are 5, 12, 12, 21 and 34; the numbers of anaphors in the texts are 4, 11, 11, 20 and 34. For example, the test text, Text 4, corresponding to the plan tree shown in Fig. 7.9 is shown in Fig. 8.4. See Appendix D for the other four test texts. For convenience, we summarise the occurrence of anaphors in the test texts in a graphical form as shown in Fig. 8.5. In the figure, each box represents a clause and at the right end is the accompanying punctuation mark. Each box is divided into three parts which represent the topic, the subject and the direct object positions of the clause. The numbers in a box, except for the first occurrences in the text, are the indices of anaphors in the corresponding clauses. Initial references are indicated by bold italics. For example, in Text 2, the numbers 1, 2, 3 and 4 occurring in the first, 5th, 8th and 10th clauses, respectively, are initial references; others are anaphors.

After the annotations were collected, we carried out comparisons between the speakers' results and the generated texts to investigate the performance of the test rules. In each comparison, we noted down the number of matches between the computer generated text and the human result. In the following, we use $C_{ij}$ to denote the text indexed $j$ generated by the system equipped with Rule TR$i$, where $i$ is 1 to 3 and $j$ is 1 to 5; and $H_{kl}$ to denote the resulting text indexed $l$ of speaker $k$, where $k$ is 1 to 12 and $l$ is 1 to 5. The comparison work is summarised procedurally as below.

```
for each rule TRi
    for each speaker j
        for each text k
            compare C_ik with H_jk and
            note down the number of matches of anaphors between them
```

## 8.4 Results

In this section, we investigate the result of the comparisons made in the last section. Before presenting the comparison results, we give a summary of the occurrence of

台灣水韭1 屬於 水韭科植物，____1 是 台灣特有種。____1 ____莖2
(Z,它,台灣水韭) (Z,它,台灣水韭)

____2 外觀 三裂瓣狀，____2 寬 約 2 公分。____1 ____的葉子3
(Z,它,短球莖,球莖) (Z,它,短球莖,球莖) (Z,它,台灣水韭)

____3 外觀 細長，____3 ____4 匙狀， ____3 切面 半圓形，
(Z,它,翠綠的葉子,葉子) (Z,它,翠綠的葉子,葉子) (Z,它,翠綠的葉子,葉子)

____3 內側 平， ____3 度 約 20公分，____3 側 弧凸，
(Z,它,翠綠的葉子,葉子) (Z,它,翠綠的葉子,葉子) (Z,它,翠綠的葉子,葉子)

____3 長 在 ____莖2 ____1 ____圓形的孢子囊5
(Z,它,翠綠的葉子,葉子) (Z,它,台灣水 )

____5 長度 約 0.8公分， ____5 在 ____基部4 。
(Z,它,長橢圓形的孢子囊,孢子囊) (Z,它,長橢圓形的孢子囊,孢子囊)

____1 ____的根11 ____11 有 根毛組織，____11 造 簡單
(Z,它,台灣水韭) (Z,它,典型的根,根) (Z,它,典型的根,根)

____11 有 維管束單條，____11 有 通氣道。
(Z,它,典型的根,根) (Z,它,典型的根,根)

Figure 8.4: An example of a test text for evaluation.

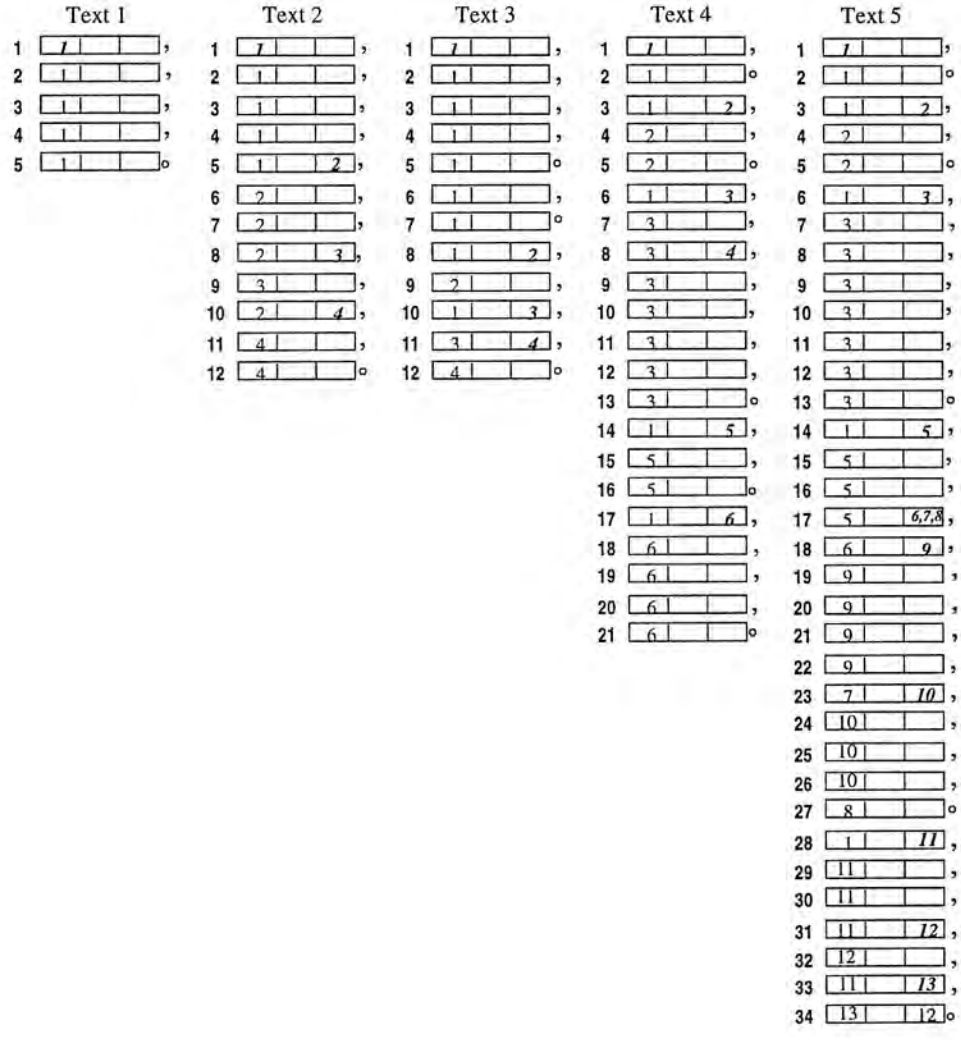



Figure 8.5: Occurrence of anaphors in the test texts.

Table 8.1: Occurrence of anaphors in the generated texts.

| | Text 1 | | | Text 2 | | | Text 3 | | | Text 4 | | | Text 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| System | Z | P | N | Z | P | N | Z | P | N | Z | P | N | Z | P | N |
| TR*1* | 4 | 0 | 0 | 10 | 0 | 1 | 9 | 0 | 2 | 17 | 0 | 3 | 26 | 0 | 8 |
| TR*2* | 4 | 0 | 0 | 10 | 0 | 1 | 8 | 0 | 3 | 16 | 1 | 3 | 25 | 1 | 8 |
| TR*3* | 4 | 0 | 0 | 7 | 0 | 4 | 5 | 0 | 6 | 12 | 1 | 7 | 17 | 1 | 16 |

Table 8.2: Occurrence of anaphors in the speakers' results.

| | Text 1 | | | Text 2 | | | Text 3 | | | Text 4 | | | Text 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Speaker | Z | P | N | Z | P | N | Z | P | N | Z | P | N | Z | P | N |
| 1 | 4 | 0 | 0 | 10 | 1 | 0 | 10 | 1 | 0 | 15 | 3 | 2 | 24 | 2 | 8 |
| 2 | 4 | 0 | 0 | 7 | 0 | 4 | 5 | 2 | 4 | 14 | 2 | 4 | 20 | 3 | 11 |
| 3 | 4 | 0 | 0 | 8 | 0 | 3 | 5 | 3 | 3 | 12 | 1 | 7 | 18 | 0 | 16 |
| 4 | 4 | 0 | 0 | 7 | 1 | 3 | 6 | 2 | 3 | 10 | 1 | 9 | 15 | 1 | 18 |
| 5 | 3 | 1 | 0 | 6 | 2 | 3 | 7 | 1 | 3 | 11 | 2 | 7 | 16 | 3 | 15 |
| 6 | 4 | 0 | 0 | 7 | 0 | 4 | 8 | 2 | 1 | 15 | 4 | 1 | 23 | 3 | 8 |
| 7 | 4 | 0 | 0 | 8 | 1 | 2 | 4 | 1 | 6 | 15 | 3 | 2 | 21 | 2 | 11 |
| 8 | 4 | 0 | 0 | 11 | 0 | 0 | 9 | 1 | 1 | 14 | 1 | 5 | 26 | 1 | 7 |
| 9 | 2 | 1 | 1 | 5 | 1 | 5 | 5 | 2 | 4 | 6 | 2 | 12 | 8 | 4 | 22 |
| 10 | 4 | 0 | 0 | 8 | 1 | 2 | 8 | 1 | 2 | 11 | 3 | 6 | 16 | 3 | 15 |
| 11 | 2 | 1 | 1 | 4 | 2 | 5 | 7 | 2 | 2 | 7 | 3 | 10 | 13 | 5 | 16 |
| 12 | 4 | 0 | 0 | 8 | 0 | 3 | 6 | 1 | 4 | 10 | 1 | 9 | 16 | 1 | 17 |
| Average | 4 | 0 | 0 | 7 | 1 | 3 | 7 | 2 | 3 | 12 | 2 | 6 | 18 | 2 | 14 |

anaphors used in both sides in Tables 8.1 and 2.

In the following, we first show the comparison results with respect to zero vs non-zero anaphors. More precisely, anaphor positions in both sides are matched if they are either both zero or both non-zero anaphors. Second, we show the comparison results taking into account all kinds of anaphors, i.e., zero, pronominal and nominal anaphors. Third, we show the comparison result in detail, i.e., further distinguishing full and reduced descriptions of nominal anaphors. The results are shown in Tables 8.3, 4 and 5.

As shown in Fig 8.5, the anaphors in Text 1 form a "topic chain" within a single "sentence". [1] These anaphors are all zeroed according to the conditions of locality

[1] See Sec. 2.5 for the detail about "topic chain."

Table 8.3: Matches of zeroes and non-zeroes.

| System | Speaker | Text 1 | Text 2 | Text 3 | Text 4 | Text 5 |
|---|---|---|---|---|---|---|
| TR*1* | 1 | 4 | 11 | 10 | 18 | 30 |
| | 2 | 4 | 8 | 7 | 17 | 28 |
| | 3 | 4 | 7 | 7 | 15 | 26 |
| | 4 | 4 | 8 | 6 | 13 | 26 |
| | 5 | 3 | 7 | 9 | 14 | 22 |
| | 6 | 4 | 8 | 8 | 18 | 25 |
| | 7 | 4 | 7 | 6 | 18 | 22 |
| | 8 | 4 | 10 | 9 | 17 | 29 |
| | 9 | 2 | 6 | 7 | 9 | 15 |
| | 10 | 4 | 9 | 10 | 14 | 22 |
| | 11 | 2 | 5 | 7 | 10 | 17 |
| | 12 | 4 | 9 | 6 | 13 | 22 |
| | Average | 4 | 8 | 8 | 15 | 26 |
| % | | 100% | 73% | 73% | 75% | 76% |
| TR*2* | 1 | 4 | 11 | 9 | 19 | 29 |
| | 2 | 4 | 8 | 8 | 18 | 29 |
| | 3 | 4 | 7 | 8 | 16 | 25 |
| | 4 | 4 | 8 | 7 | 14 | 24 |
| | 5 | 3 | 7 | 10 | 15 | 25 |
| | 6 | 4 | 8 | 9 | 19 | 32 |
| | 7 | 4 | 7 | 7 | 19 | 28 |
| | 8 | 4 | 10 | 10 | 18 | 33 |
| | 9 | 2 | 6 | 8 | 10 | 17 |
| | 10 | 4 | 9 | 11 | 15 | 25 |
| | 11 | 2 | 5 | 8 | 11 | 20 |
| | 12 | 4 | 9 | 7 | 14 | 25 |
| | Average | 4 | 8 | 8 | 16 | 26 |
| % | | 100% | 73% | 73% | 80% | 76% |
| TR*3* | 1 | 4 | 8 | 6 | 15 | 21 |
| | 2 | 4 | 11 | 11 | 18 | 31 |
| | 3 | 4 | 10 | 11 | 20 | 33 |
| | 4 | 4 | 9 | 6 | 14 | 24 |
| | 5 | 3 | 10 | 10 | 17 | 29 |
| | 6 | 4 | 11 | 8 | 17 | 28 |
| | 7 | 4 | 8 | 10 | 15 | 26 |
| | 8 | 4 | 7 | 7 | 14 | 25 |
| | 9 | 2 | 7 | 7 | 14 | 25 |
| | 10 | 4 | 10 | 8 | 17 | 25 |
| | 11 | 2 | 6 | 7 | 13 | 24 |
| | 12 | 4 | 10 | 10 | 16 | 31 |
| | Average | 4 | 9 | 8 | 16 | 27 |
| % | | 100% | 82% | 73% | 80% | 79% |

Table 8.4: Matches of zeroes, pronouns and nominals.

| System | Speaker | Text 1 | Text 2 | Text 3 | Text 4 | Text 5 |
|---|---|---|---|---|---|---|
| TR*1* | 1 | 4 | 10 | 9 | 16 | 28 |
| | 2 | 4 | 8 | 6 | 16 | 26 |
| | 3 | 4 | 7 | 5 | 15 | 26 |
| | 4 | 4 | 8 | 5 | 13 | 23 |
| | 5 | 3 | 7 | 8 | 14 | 24 |
| | 6 | 4 | 8 | 7 | 15 | 29 |
| | 7 | 4 | 7 | 6 | 16 | 26 |
| | 8 | 4 | 10 | 9 | 17 | 32 |
| | 9 | 2 | 6 | 7 | 9 | 16 |
| | 10 | 4 | 8 | 9 | 14 | 24 |
| | 11 | 2 | 5 | 6 | 10 | 21 |
| | 12 | 4 | 9 | 5 | 13 | 24 |
| | Average | 4 | 8 | 7 | 14 | 25 |
| % | | 100% | 73% | 64% | 70% | 74% |
| TR*2* | 1 | 4 | 10 | 8 | 17 | 27 |
| | 2 | 4 | 8 | 6 | 17 | 27 |
| | 3 | 4 | 7 | 5 | 16 | 25 |
| | 4 | 4 | 8 | 5 | 14 | 24 |
| | 5 | 3 | 7 | 9 | 15 | 25 |
| | 6 | 4 | 8 | 7 | 16 | 30 |
| | 7 | 4 | 7 | 7 | 17 | 27 |
| | 8 | 4 | 10 | 9 | 18 | 33 |
| | 9 | 2 | 6 | 8 | 9 | 16 |
| | 10 | 4 | 8 | 10 | 15 | 25 |
| | 11 | 2 | 5 | 7 | 11 | 20 |
| | 12 | 4 | 9 | 6 | 14 | 25 |
| | Average | 4 | 8 | 7 | 15 | 25 |
| % | | 100% | 73% | 64% | 75% | 74% |
| TR*3* | 1 | 4 | 7 | 5 | 13 | 19 |
| | 2 | 4 | 11 | 9 | 17 | 29 |
| | 3 | 4 | 10 | 8 | 20 | 33 |
| | 4 | 4 | 9 | 4 | 14 | 24 |
| | 5 | 3 | 9 | 9 | 16 | 27 |
| | 6 | 4 | 11 | 6 | 14 | 26 |
| | 7 | 4 | 8 | 10 | 13 | 25 |
| | 8 | 4 | 7 | 6 | 14 | 25 |
| | 9 | 2 | 7 | 7 | 13 | 24 |
| | 10 | 4 | 9 | 7 | 16 | 24 |
| | 11 | 2 | 6 | 6 | 13 | 24 |
| | 12 | 4 | 10 | 9 | 16 | 31 |
| | Average | 4 | 9 | 7 | 15 | 26 |
| % | | 100% | 82% | 64% | 75% | 76% |

Table 8.5: Matches of zeroes, pronouns and full and reduced descriptions.

| System | Speaker | Text 1 | Text 2 | Text 3 | Text 4 | Text 5 |
|---|---|---|---|---|---|---|
| TR*1* | 1 | 4 | 10 | 9 | 16 | 27 |
| | 2 | 4 | 8 | 6 | 16 | 24 |
| | 3 | 4 | 7 | 5 | 15 | 24 |
| | 4 | 4 | 8 | 5 | 13 | 23 |
| | 5 | 3 | 7 | 8 | 14 | 23 |
| | 6 | 4 | 8 | 7 | 15 | 28 |
| | 7 | 4 | 7 | 6 | 16 | 25 |
| | 8 | 4 | 10 | 9 | 17 | 32 |
| | 9 | 2 | 6 | 7 | 9 | 14 |
| | 10 | 4 | 9 | 9 | 14 | 23 |
| | 11 | 2 | 5 | 6 | 10 | 20 |
| | 12 | 4 | 8 | 5 | 13 | 23 |
| | Average | 4 | 8 | 7 | 14 | 24 |
| | | 100% | 73% | 64% | 70% | 71% |
| TR*2* | 1 | 4 | 10 | 8 | 17 | 26 |
| | 2 | 4 | 8 | 6 | 17 | 25 |
| | 3 | 4 | 7 | 5 | 16 | 23 |
| | 4 | 4 | 8 | 5 | 14 | 24 |
| | 5 | 3 | 7 | 9 | 15 | 24 |
| | 6 | 4 | 8 | 7 | 16 | 29 |
| | 7 | 4 | 7 | 7 | 17 | 26 |
| | 8 | 4 | 10 | 9 | 18 | 33 |
| | 9 | 2 | 6 | 8 | 9 | 14 |
| | 10 | 4 | 8 | 10 | 15 | 24 |
| | 11 | 2 | 5 | 7 | 11 | 19 |
| | 12 | 4 | 9 | 6 | 14 | 24 |
| | Average | 4 | 8 | 7 | 15 | 24 |
| | | 100% | 73% | 64% | 75% | 71% |
| TR*3* | 1 | 4 | 7 | 5 | 13 | 18 |
| | 2 | 4 | 11 | 9 | 16 | 25 |
| | 3 | 4 | 10 | 8 | 19 | 29 |
| | 4 | 4 | 9 | 4 | 14 | 24 |
| | 5 | 3 | 9 | 9 | 16 | 26 |
| | 6 | 4 | 11 | 6 | 14 | 24 |
| | 7 | 4 | 8 | 10 | 13 | 23 |
| | 8 | 4 | 7 | 6 | 14 | 25 |
| | 9 | 2 | 7 | 7 | 12 | 20 |
| | 10 | 4 | 9 | 7 | 16 | 23 |
| | 11 | 2 | 6 | 6 | 12 | 21 |
| | 12 | 4 | 10 | 9 | 16 | 30 |
| | Average | 4 | 9 | 7 | 15 | 24 |
| | | 100% | 82% | 64% | 75% | 71% |

and syntactic constraints in the three test rules. All three systems produce the same result for Text 1 and, hence, unsurprisingly all three systems have the same matching rate as shown in Tables 8.3, 4 and 5. This shows that the above two conditions are highly reliable. Let's take a closer look at the individual improvement of the test rules over their predecessors. For convenience, in the following, we use $M_{i,j,k}$ to denote the number of matches between the $k$th texts of speaker $i$ and system TR$j$. A system TR$j$+*1* is deemed to have an improvement over its predecessor TR$j$ on the $k$th text if $M_{i,j+1,k}$ is greater than $M_{i,j,k}$. We summarise in Tables 8.6, 7 and 8 the improvements of the object systems on the test texts. Note that the rows of TR*1* are left empty because it has no predecessor with which to compare. Since the three object systems produce the same output for Text 1, both systems TR*2* and TR*3* trivially produce zero improvement on Text 1.

Text 2 similarly contains a single "sentence" as the previous text but has topic shifts in addition to "topic chains" within the "sentence" as shown in Fig. 8.5. Since no discourse segment boundaries occur within the "sentence", the discourse segment boundary constraint in TR*2* has no effect on this test text, which means both TR*1* and TR*2* produce the same output. However, there are three topic shifts within the "sentence", namely, clauses 5 and 6, 8 and 9, and 10 and 11, as shown in Fig. 8.5. The shifts would make the rule containing the salience constraint, TR*3*, obtain different output from those without this constraint, TR*1* and TR*2*. The figures in Tables 8.3, 4 and 5 show that the average matching rates of TR*3* increase from 73% to 82%. Furthermore, as in Tables 8.6, 7 and 8, TR*3* receive 10 improvements in this case. Again, the above shows the effectiveness of the salience constraint.

We then examine another middle-sized test text, Text 3, which is broken into three "sentences," as shown in Fig. 8.5. Recall that the beginning of a "sentence" is the beginning of a discourse segment in our implementation. [2] Furthermore, there are three topic shifts occurring in Text 3, i.e., clauses 8 and 9, 10 and 11, and 11 and 12. The constraint of discourse segment beginnings in TR*2* and TR*3* and the salience constraint in TR*3* would therefore have some effects on the output texts. In Tables 8.3, 4 and 5, all three systems gain the same average matching rates on Text 3. As for the

[2] See Sec. 7.4.2 for details.

Table 8.6: Improvement over the predecessor on zero and non-zero anaphors.

| System | Text 1 | Text 2 | Text 3 | Text 4 | Text 5 |
|---|---|---|---|---|---|
| TR*1* | | | | | |
| % | | | | | |
| TR*2* | 0 | 0 | 10 | 12 | 9 |
| % | 0% | 0% | 83% | 100% | 75% |
| TR*3* | 0 | 10 | 4 | 6 | 6 |
| % | 0% | 83% | 33% | 50% | 50% |

Table 8.7: Improvement over the predecessor on zero, pronominal and nominal anaphors.

| System | Text 1 | Text 2 | Text 3 | Text 4 | Text 5 |
|---|---|---|---|---|---|
| TR*1* | | | | | |
| % | | | | | |
| TR*2* | 0 | 0 | 6 | 11 | 8 |
| % | 0% | 0% | 50% | 92% | 67% |
| TR*3* | 0 | 10 | 4 | 6 | 6 |
| % | 0% | 83% | 33% | 50% | 50% |

Table 8.8: Improvement over the predecessor on zero, pronominal anaphors and nominal descriptions.

| System | Text 1 | Text 2 | Text 3 | Text 4 | Text 5 |
|---|---|---|---|---|---|
| TR*1* | | | | | |
| % | | | | | |
| TR*2* | 0 | 0 | 6 | 11 | 9 |
| % | 0% | 0% | 50% | 92% | 75% |
| TR*3* | 0 | 10 | 4 | 6 | 5 |
| % | 0% | 83% | 33% | 50% | 42% |

individual improvement,TR*2* has 10 improvements, while TR*3* has 4 improvements on Text 3 in case of zeros and non-zeros as shown in Table 8.6. However, as the comparison distinguishes pronominal and nominal anaphors for non-zeros, TR*2*'s improvement decreases from 10 to 6, while TR*3* remains unchanged. This shows the effect of adding the constraint of discourse segment beginning into the anaphor generation rule. Also the salience constraint is effective to a certain extent on Text 3.

Then we examine the more complicated texts, Texts 4 and 5. As shown in Tables 8.3, 4 and 5, TR*2* receives better average matching rates than TR*1* on Text 4, while having the same rates as TR*1* on Text 5. From the point of view of individual improvement, TR*2* receives 12 improvements over TR*1* on Text 4 and 9 on Text 5 as shown in Table 8.6, and the improvement slightly decreases as non-zeroes are further distinguished as shown in Tables 8.7 and 8. As for TR*3*, the average matching rates on Texts 4 and 5 are no less than TR*1* and TR*2*. The improvements of TR*3* on Texts 4 and 5, though not so significant as TR*2*, are about 50% of the total improvement. In brief, according to the average matching rates and individual improvements of the object systems as shown in the tables, the more sophisticated constraints a rule contains, the better it performs. In other words, TR*2* is better than TR*1* and TR*3* is better than TR*2*.

In brief, the system TR*3* produces the best matching rates for all individual test texts, as shown in Tables 8.3, 4 and 5. On average, the matching rates of TR*3* for three comparisons are 83, 79 and 78%, compared with those of the other systems, TR*1*'s 79, 76, and 76%, and TR*2*'s 80, 77 and 77%. Although TR*3* produces the best results, still some native speakers, Speakers 1 and 8, for example, disagreed with it. As shown in Tables 8.3, 4 and 5, TR*3*'s match rates with Speakers 1 and 8 are the lowest of the three systems for all individual test texts. Observing the results of TR*3* and Speakers 1 and 8, we found that almost all the mis-matches occur at topic shifts, where TR*3* generated non-zero forms, while the speakers preferred zeroes. This shows the difference of concepts of salience used between the two speakers and TR*3*. Although they disagreed with TR*3* on the concept of salience, most of other speakers agreed with TR*3*. Thus TR*3* is reliable.

## 8.5 Summary

In this chapter, we evaluated the quality of anaphors in the texts generated by using various rules. As shown in the results of comparisons between the anaphors created by computers and native speakers of Chinese, the individual constraints we collected in Chaps. 4 to 6 are effective to a large extent in the generation of anaphors in Chinese. Also they can be implemented successfully. The comparison results reveal the fact that a Chinese natural language generation system employing the combination of these constraints would produce more effective anaphors than using individual constraints.

# Chapter 9

# Summary and Future Directions

## 9.1 Summary

In this thesis, we first carried out empirical work on the generation of anaphors in Chinese. The work consisted of experiments comparing anaphors in human-generated test data with those generated by using anaphor generation rules, assuming the same postulated semantic structure as the test data. The test data was a corpus derived from two scientific question-and-answer books for children and a Chinese grammar book,and this was used to evaluate hypotheses concerning anaphor distribution. The experimental results showed that a rule using locality constraints, syntactic constraints on zero anaphora, discourse segment structure, salience of objects and animacy of objects can effectively deal with the generation of zero, pronominal and nominal anaphors. We further repeated the same experiment using rules for appropriate descriptions if a nominal anaphor is decided on by the previous rule. The experimental results show that a rule considering locality between the anaphor and the antecedent within the discourse segment structure can effectively generate appropriate nominal descriptions.

Secondly we implemented the above rules in a Chinese natural language generation system. The system was built up by adopting concepts from conventional natural language generation systems which are basically composed of a text planner and a linguistic realisation component. The task of choosing anaphoric forms occurs immediately after the message contents are well organised by the text planner. A discourse model is maintained to provide the information for the decision. After this is finished, the results are then passed to the linguistic component to be realised as surface

sentences. The output texts show that the rules can work successfully in a real system.

We finally compared the anaphors in some typical texts generated by our system using anaphor generation rules with different complexities to those created by a group of native speakers of Chinese, using the same content for the generated text but with the anaphor positions left empty. The comparison results show that the anaphors in the text generated by using the rules consisting of all constraints collected from the empirical study are closer to those in human texts than those previously tested. On average, the matching rates of the best results are 83, 79 and 78% with respect to zero vs non-zero, all kinds of anaphors, and further considering descriptions for nominal anaphors, respectively. In brief, the contributions of this thesis can be summarised as below.

- Effective rules for the generation of anaphors in Chinese, including zero, pronominal and nominal anaphors, and appropriate descriptions for nominal anaphors have been developed;
- A successful implementation of these rules in a Chinese natural language generation system has been demonstrated; and
- An evaluation of anaphors generated by the system using various rules has been presented.

## 9.2 Future Directions

We suggest several issues related to this work which require further investigation. These include the extension of empirical work to deal with anaphors in other types of texts, the use of connectives in generated text to create cohesive discourse, an improvement on the constraints for pronominal anaphora, the extension of text planning mechanism and discourse segmentation method, the extension of current knowledge representation and linguistic mechanisms to deal with more complicated sentences.

*Other text types.* As we pointed out at the beginning of this thesis, we concentrate on descriptive texts in this thesis. However, there are other types of text that need to be studied, such as narration, exposition and argument, as summarised in [Maybury 90].

The empirical study described in this thesis provides a clear framework to carry out research work upon other types of text. One simple change is to replace the test data with the target type of text and then carry out experiments as before. The rule we used in this thesis can be employed as the starting basis of the new study.

Another possible direction is to consider anaphors occurring in texts in a dialogue environment, which we did not mention at all. In particular, in the situation where short texts are generated frequently by a system as response to the user, the discourse structure and focus space tend to be different from those we are concerned with in this thesis. This kind of study may need to pay more attention to the discourse structure and focus space mechanism, and hence, their implementation in a Chinese generation system.

*Connectives in discourse.* In this thesis, we did not address the issue of connectives in discourse. In fact, as shown in previous studies, connectives, such as *for example* and *therefore*, are important linguistic devices, in addition to anaphora, the creation of cohesive discourse. Connectives are closely related to the meaning of rhetorical relations implicit in discourse [Mann & Thompson 87, Knott & Dale 92]. In our system, text planning operators play the role of rhetorical relations in discourse. Thus, in the future, operators should be extended to be able to provide information for the decision about what connectives to use.

*Constraints for pronominal anaphora.* As described in Chap. 5, the animacy condition plays an important role towards the success of generating pronouns. The experimental results show that the general concept of animacy cannot account for some pronouns in the test data. This seems to involve complicated factors, such as world knowledge, the speaker's style, etc. We have proposed to improve this by extending the animacy condition and considering different types of discourse segments. [1] Our experience of implementation suggests that we need more empirical study to render these ideas effective. Since the occurrence of pronouns in our test data is unable to support further empirical study, we need to select more texts to investigate.

*Discourse segmentation.* Discourse structure is an important constraint in our anaphor

[1] See Sec. 5.3 for details.

generation rules. In this work, we adopt Grosz and Sidner's discourse structure theory [Grosz & Sidner 86] as the basis to carry out empirical study and implementation. Based on the concept provided by the theory, we found that the discourse segment boundaries largely match the occurrence of "sentential marks" in Chinese written texts. [2] The experiments in Secs. 4.3.3 and 6.4 were based on the above idea. In the implementation, the problem is to divide the text planning trees into discourse segments. The anaphor generation rules are implemented on the basis of this discourse segment structure. As described in Sec. 7.4.2, the segmentation heavily relies on the idea of text planning. The current implementation adopts the idea of text planning directly from the TEXPLAN system [Maybury 90]. The discourse segmentation is realised in a straightforward way by considering the level of nodes in the text planning trees. To improve the approach, further study is required to investigate how to integrate the idea of discourse segmentation into text planning operators. Another possible direction is to investigate discourse segmentation with other kinds of text planner, like the RST planner [Hovy 93].

*Knowledge representations and linguistics mechanism.* At the moment, the linguistic component of our Chinese generation system only deals with simple sentences that do not contain embedded clauses. In the future, we will extend the linguistic component to encompass the capacity to process nominalisation clauses, serial verb constructions, prepositional phrases, etc. To do so, the semantic and syntactic representations of message units need to be extended to include new sentence structures. Furthermore, in the future, it will be necessary to develop a more sophisticated representation for the domain knowledge base so that the content selection component can take advantage of the representation to get the message contents.

[2] See Sec. 4.3.3 for details.

# Bibliography


[Block & Horacek 90] Russell Block and Helmut Horacek. Generating referring expressions using multiple knowledge sources. In *Proceedings of the 13th International Conference on Computational Linguistics*, 1990.

[Chafe 79] W. Chafe. The flow of thoughts and the flow of language. In T. Givon, editor, *Syntax and Semantics, Vol. 12: Discourse and Syntax*. Academic Press, New York, 1979.

[Chao 68] Y. R. Chao. *A Grammar of Spoken Chinese*. University of California Press, Berkeley, CA, 1968.

[Chen 84] P. Chen. A discourse analysis of third person zero anaphora in chinese. Technical report, Indiana University Linguistics Club, Bloomington, Indiana, 1984.

[Chen 86] P. Chen. *Referent Introducing and Tracking in Chinese Narratives*. Unpublished PhD thesis, University of California, Los Angeles, CA, 1986.

[Chen 87] P. Chen. Hanyu lingxin huizhi de huayu fenxi (a discourse approach to zero anaphora in chinese) (in chinese). *Zhongguo Yuwen (Chinese Linguistics)*, pages 363–378, 1987.

[Clocksin & Mellish 94] W. F. Clocksin and C. S. Mellish. *Programming in Prolog*. Springer-Verlag, 4th edition, 1994.

[Dale & Haddock 91] R. Dale and N. Haddock. Content determination in the generation of referring expressions. *Computational Intelligence*, pages 252–265, 1991.

[Dale 86] R. Dale. The pronominalization decision in language generation. Technical Report DAI Research Paper No. 276, Department of Artificial Intelligence, University of Edinburgh, Edinburgh, Scotland, 1986.

[Dale 88] R. Dale. The generation of subsequent referring expressions in structured discourses. In M. Zock and G. Sabah, editors, *Advances in Natural Language Generation, An Interdisciplinary Perspective*, volume 2, pages 58–75. Pinter Publishers, London, 1988.

[Dale 92] R. Dale. *Generating Referring Expressions: Constructing Descriptions in a Domain of Objects and Processes.* The MIT Press, Cambridge, Massachusetts, 1992.

[Gazdar & Mellish 89] Gerald Gazdar and Chris Mellish. *Natural Language Processing in Prolog: an Introduction to Computational Linguistics.* Addison-Wesley, 1989.

[Grice 75] P. H. Grice. Logic and conversation. In P. Cole and J. L. Morgan, editors, *Syntax and Semantics 3: Speech Acts.* Academic Press, New York, 1975.

[Grosz & Sidner 86] B. J. Grosz and C. L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

[Grosz 77] B. J. Grosz. The representation and use of focus in dialogue. Technical Report 151, SRI International, Menlo Park, CA, 1977.

[Grosz *et al.* 83] B. J. Grosz, A. K. Joshi, and S. Weinstein. Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21th Annual Meeting of the Association for Computational Linguistics*, pages 44–50, Boston, 1983.

[Horacek 95] Helmut Horacek. More on generating referring expressions. In *Proceedings of the Fifth European Workshop on Natural Language Generation*, Leiden, The Netherlands, 1995.

[Hovy 90] E. Hovy. Approaches to the planning of coherent text. In C. L. Paris, W. R. Swartout, and W. C. Mann, editors, *Natural Language in Artificial Intelligence and Computational Linguistics.* Kluwer Academic Publishers, 1990.

[Hovy 93] E. Hovy. Automated discourse generation using discourse structure relations. *Artificial Intelligence*, 63(1-2):341–385, 1993.

[Huang 91] Y. Huang. A neo-gricean pragmatic theory of anaphora. *Journal of Linguistics*, 27:301–335, 1991.

[Huang 94] Y. Huang. *The Syntax and Pragmatics of Anaphors: A Study with Special Reference to Chinese.* Cambridge University Press, 1994.

[Knott & Dale 92] A. Knott and R. Dale. Using linguistic phenomena to motivate a set of rhetorical relations. Technical Report HCRC/RP-39, Human Communication Research Centre, University of Edinburgh, Edinburgh, Scotland, 1992.

[Li & Thompson 78] C. N. Li and S. A. Thompson. An exploration of Mandarin Chinese. In W. Lehmann, editor, *Syntactic Typology: Study in the Phenomenology of Language.* Harvester Press, 1978.

[Li & Thompson 79] C. N. Li and S. A. Thompson. Third-person pronouns and zero-anaphora in Chinese discourse. In T. Givon, editor, *Syntax and Semantics: Discourse and Syntax*, volume 12, pages 311–335. Academic Press, 1979.

[Li & Thompson 81] C. N. Li and S. A. Thompson. *Mandarin Chinese: a Functional Reference Grammar.* University of California Press, Berkeley, CA, 1981.

[Liu 77] Y. C. Liu. *Biaodianfuhao de yongfa (Usage of Punctuation Marks) (in Chinese).* Guoyurebao Chubanshe, Taipei, Taiwan, 1977.

[Liu 81] Feng-Hsi Liu. Zero-anaphora in Mandarin Chinese. In R. A. Hendrick, C. S. Masek, and M. F. Miller, editors, *Papers from the Seventeenth Regional Meeting of the Chicago Linguistic Society*, pages 197–204, Chicago, Illinois, 1981.

[Liu 84] Y. C. Liu. *Zuowen de fangfa (Approaches to Composition) (in Chinese).* Xuesheng Chubanshe, Taipei, Taiwan, 1984.

[Mann & Thompson 87] W. C. Mann and S. A. Thompson. Rhetorical structure theory: a theory of text organisation. In L. Polanyi, editor, *The Structure of Discourse.* Ablex, 1987.

[Maybury 90] M. T Maybury. *Planning Multisentential English Text Using Communicative Acts.* Unpublished PhD thesis, Cambridge University, 1990.

[McDonald 80] D. D McDonald. *Natural Language Generation as a Process of Decision Making under Constraints.* Unpublished PhD thesis, MIT, 1980.

[McKeown 85] K. R. McKeown. *Text Generation.* Cambridge University Press, 1985.

[Meteer & McDonald 91] M. Meteer and D. McDonald. Evaluation for generation. In J. G Neal and S. M. Wlater, editors, *Natural Language Processing Systems Evaluation Workshop*, pages 127–131, NY, 1991. Rome Laboratory.

[Moore & Paris 94] J. D. Moore and C. L. Paris. Planning text for advisory dialogues: capturing intentional and rhetorical information. *Computational Linguistics*, 19(4):651–694, 1994.

[Reiter & Dale 92] E Reiter and R. Dale. A fast algorithm for the generation of referring expressions. In *Proc. of the 14th International Conference on Computational Linguistics*, pages 232–238, 1992.

[Reiter & Dale 95] E Reiter and R. Dale. Computational interpretation of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 1995.

[Reiter 90] E. Reiter. Generating descriptions that exploit a user's domain knowledge. In R. Dale, C. Mellish, and M. Zock, editors, *Current Research in Natural Language Generation.* Academic Press, 1990.

[Reiter 91] E. Reiter. A new model of lexical choice for nouns. *Computational Intelligence*, 7(4):240–251, 1991.

[Reiter 94] E. Reiter. Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible? In *Proc. of the 1994 International Natural Language Generation Workshop*, 1994.

[Shieber 86] S. M. Shieber. An introduction to unification-based approach to grammar. Technical Report Lecture Notes, No. 4, CSLI, Stanford University, 1986.

[Sidner 83] C. Sidner. Focusing in the comprehension of definite anaphora. In M. Brady and R. C. Berwick, editors, *Computational Models of Discourse*, pages 267–330. MIT Press, 1983.

[Tai 78] James H. Y. Tai. Anaphoric constraints in Mandarin Chinese narrative discourse. In J. Hinds, editor, *Anaphora in Discourse.* Linguistic Research, Edmonton, Alberta, 1978.

[Tate 85] A. Tate. A review of knowledge-based planning techniques. *Knowledge Base Review*, 2:4–16, 1985.

[Teng 75] S. H Teng. *A Semantic Study of the Transitivity Relations in Chinese.* University of California Press, Berkeley, CA, 1975.

[Thompson 77] H. Thompson. Strategy and tactics: a model for language production. In *Papers from the 13th Regional Meeting.* Chicago Linguistic Society, 1977.

[Tutin & Kittredge 92] A. Tutin and R. Kittredge. Lexical choice in context: generating procedural texts. In *Proc. of the 14th International Conference on Computational Linguistics (COLING-92)*, pages 763–769, Nantes, France, 1992.

[Yu 55] Z. C. Yu. *Biaodianfuhao de yongfa bijiao (Comparisons of Use of Punctuation Marks) (in Chinese).* Jiangsu Renmin Chubanshe, Nanjing, Jiangsu, China, 1955.

# Appendix A

# Test for Discourse Segmentation

The instructions for discourse segmentation, given in Chinese, are as follows.

> **Description:** There are five articles to be examined in this investigation. Each article is accompanied by a question-style topic. The content of an article is to answer the question accompanying it. Therefore the purpose (or intention) of the whole article is obviously for answering its own question. Reading carefully, you will find that an article can be divided into a string of segments according to their respective purposes (or intentions). Let's call each of them a sub-purpose (or sub-intention). Therefore the purpose (or intention) of an article is obviously composed of a string of sub-purpose (or sub-intention). In other words, every sub-purpose (sub-intention) serves as a part of the whole intention of an article. Furthermore, in an article, a sub-purpose (or sub-intention) can be a subsidiary of other sub-purposes (or sub-intentions), just like sub-purposes (sub-intention) are subsidiaries of the whole intention. That is, a sub-purpose can subsume others. Therefore, we have a hierarchical intentional structure for an article, for example, Fig. 4.6 in Sec. 4.3.3.
>
> **Tasks:** After thoroughly understanding the above description, for each article, complete the following tasks.
>
> 1. Mark the boundaries of segments; and
> 2. Draw the hierarchical intentional structure.

In the test, four native speakers of Chinese were asked to annotate the boundaries between discourse segments and draw the hierarchical discourse structures for five articles selected from the test data. For each article, we list the number of matches between our boundary markers and each of the speakers'; moreover, we note down the number of markers in our results but not in the counterparts, and vice versa. In Table A.1, the rows of *match*, *over* and *under* correspond with the results of the above types of comparisons, respectively. In the table, the total number of markers in our results are listed under each article. For each article, the percentage of match for speaker $i$ is

defined as the number of matches with $i$ over the total number of our marks. Similarly, the percentages of over and under are defined as their respective numbers over the number of our marks. By averaging the match, over and under percentages, they are 84, 16 and 8, respectively. These figures indicate that on average 84% of our marks match those of the speakers. As for mismatches between both sides, 16% of our marks do not appear in the speakers results; the marks in the speakers' results only occupy 8% of our marks. From the above comparisons, the annotations of both sides, to a large extent, are similar to each other. Thus, the annotations we made are highly reliable for the purpose of applying the algorithm of Rule 3.

From both the authors' and speakers' results, we found that the discourse segment boundaries were considerably related to the sentential and question marks in the text [1]. We therefore classified the annotations of discourse segment boundaries as coincident, over- and under-generated with respect to the sentential and question marks, corresponding to $=$, $>$ and $<$ columns in Table A.2. There are 6, 4, 7, 6 and 3 such punctuation marks in the selected articles, respectively. In the table, there is 100, 88, 96 and 100% coincidence between the boundaries and the marks for speakers 1 to 4, respectively, and 100% for the author's. For the over-matches between the boundaries and the marks, i.e., the $>$ columns in the table, there are 10, 15, 7 and 10% in the speakers' result, which are 11% in average, and 16% in the author's result. There are less clear similarities for the situations of the under-matches. Among them, one situation is discourse segment boundaries accompanied by discourse-level adverbial phrases. In our annotations, there were 3 such cases; in the speakers' result, there were 1, 3, 1 and 0 such cases, respectively. Another is discourse segment boundaries occurring with non-zero topic anaphors within a sequence of sentences having the same topic. There was 1 in our annotations and 2 in speaker 4's result. From the above comparisons, the annotations of both sides, whether coming from people who understand the theory of discourse structure or not, to a large extent are similar to each other. Thus the annotations we made are reliable for applying the algorithm of Rule 3.

[1] As described in Sec. 2.4, among the officially used punctuation marks, sentential mark (.) and comma are frequently used as the separator of sentences. The former is used to express the completeness of intention for a sequence of sentences. Question marks always occur at the end of question sentences.

Table A.1: Comparison of author's and speakers' results.

| | Article 1 markers=6 | | | | Article 2 markers=4 | | | | Article 3 markers=9 | | | | Article 4 markers=7 | | | | Article 5 markers=5 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| speaker | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| match | 6 | 4 | 6 | 6 | 4 | 4 | 4 | 4 | 7 | 6 | 6 | 8 | 6 | 6 | 6 | 6 | 3 | 4 | 3 | 3 |
| over | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 3 | 3 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 2 |
| under | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 2 | 2 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| average % of match = 84 | | | | | | | | | | | | | | | | | | | | |
| average % of over = 16 | | | | | | | | | | | | | | | | | | | | |
| average % of under = 8 | | | | | | | | | | | | | | | | | | | | |

Table A.2: Coincidence of discourse segment boundaries and punctuation marks in the text.

| | Speaker 1 | | | Speaker 2 | | | Speaker 3 | | | Speaker 4 | | | Authors | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | = | > | < | = | > | < | = | > | < | = | > | < | = | > | < |
| Article 1 | 6 | 0 | 0 | 4 | 0 | 2 | 6 | 0 | 0 | 6 | 0 | 0 | 6 | 0 | 0 |
| Article 2 | 4 | 1 | 0 | 4 | 0 | 0 | 4 | 1 | 0 | 4 | 2 | 0 | 4 | 0 | 0 |
| Article 3 | 7 | 2 | 0 | 6 | 0 | 1 | 6 | 1 | 1 | 7 | 1 | 0 | 7 | 2 | 0 |
| Article 4 | 6 | 0 | 0 | 6 | 2 | 0 | 6 | 0 | 0 | 6 | 0 | 0 | 6 | 1 | 0 |
| Article 5 | 3 | 0 | 0 | 3 | 2 | 0 | 3 | 0 | 0 | 3 | 0 | 0 | 3 | 1 | 0 |

# Appendix B

# Test of Speaker's Preference on Test Data

The experiment in Section 4.4 showed that, using Rule 4, 10 and 16 zero and non-zero anaphors in the Set 1 test data satisfies the non-salience condition. Rule 4 always produces non-zeros in this case. We asked five native speakers of Chinese, according to their intuitions, their preference on these 26 cases. For each case, they were asked to choose one of zero, non-zero and both equally good. We noted down the number of preferences for each case, as shown in Table B.1. In the table, indices from 1 to 10 are the 10 zero anaphors in the test data and the rest are the 16 non-zero anaphors. The opinions of equal were interpreted as agreement with Rule 4, namely, preferring a non-zero form for each anaphor in the test. Thus, in the table, the agreement number of each anaphor is the preference number of equal plus non-zero. The last column in the table indicates the percentage of speakers' agreement with Rule 4 for each anaphor. By averaging the percentages in the last column, we obtained 72% of speakers' agreement with Rule 4.

Table B.1: Result of the test about preference of zero and non-zero anaphors.

| Index | Anaphor in text | Rule 4's prediction | Number of preference | | | Agreeing number | |
|---|---|---|---|---|---|---|---|
| | | | Zero | N-zero | equal | number | % |
| 1 | Z | NZ | 3 | 2 | 0 | 2 | 40 |
| 2 | Z | NZ | 2 | 0 | 3 | 3 | 60 |
| 3 | Z | NZ | 2 | 1 | 2 | 3 | 60 |
| 4 | Z | NZ | 3 | 1 | 1 | 3 | 60 |
| 5 | Z | NZ | 2 | 1 | 2 | 3 | 60 |
| 6 | Z | NZ | 4 | 0 | 1 | 1 | 20 |
| 7 | Z | NZ | 3 | 1 | 1 | 2 | 40 |
| 8 | Z | NZ | 3 | 1 | 1 | 2 | 40 |
| 9 | Z | NZ | 3 | 0 | 2 | 2 | 40 |
| 10 | Z | NZ | 2 | 1 | 2 | 3 | 60 |
| 11 | NZ | NZ | 2 | 1 | 2 | 3 | 60 |
| 12 | NZ | NZ | 0 | 5 | 0 | 5 | 100 |
| 13 | NZ | NZ | 1 | 4 | 0 | 4 | 80 |
| 14 | NZ | NZ | 0 | 5 | 0 | 5 | 100 |
| 15 | NZ | NZ | 1 | 4 | 0 | 4 | 80 |
| 16 | NZ | NZ | 1 | 3 | 1 | 4 | 80 |
| 17 | NZ | NZ | 1 | 4 | 0 | 4 | 80 |
| 18 | NZ | NZ | 0 | 4 | 1 | 5 | 100 |
| 19 | NZ | NZ | 0 | 4 | 1 | 5 | 100 |
| 20 | NZ | NZ | 1 | 3 | 1 | 4 | 80 |
| 21 | NZ | NZ | 1 | 4 | 0 | 4 | 80 |
| 22 | NZ | NZ | 1 | 2 | 2 | 4 | 80 |
| 23 | NZ | NZ | 0 | 5 | 0 | 5 | 100 |
| 24 | NZ | NZ | 0 | 5 | 0 | 5 | 100 |
| 25 | NZ | NZ | 1 | 3 | 1 | 4 | 80 |
| 26 | NZ | NZ | 0 | 5 | 0 | 5 | 100 |

# Appendix C

# Rhetorical Predicates and Semantic Mapping Rules

As described in Secs. 7.3.2 and 7.3.3, when the process of text planning reaches a surface act node, the text planner consults rhetorical predicates to extract associated message contents from the domain knowledge base for the surface act node and converts them into semantic structures through a set of mapping rules. In the following, we show part of the rhetorical predicates and semantic mapping rules [1] used in our current implementation.

*Rhetorical predicates.* A rhetorical predicate, for example, *logical definition*, is a frame abstracting information from the domain knowledge base. The rhetorical predicate of *logical definition* consists of the following fields:

**Name:** `logical-definition`
**Entity:** *an entity index*
**Superclass:** ...
**Differentia:** ...

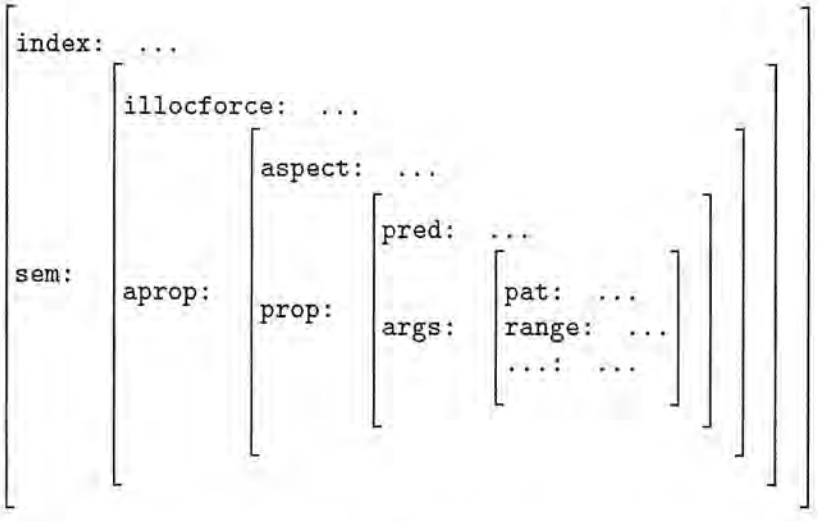
When a surface act is encountered, `assert(logical_definition(p1))`, for example, the text planner consults the corresponding rhetorical predicates, as shown above. The entity field is instantiated with the entity in the surface act,here `p1`. Then it gets the values of `superclass` and `differentia` of `p1` from the domain knowledge base, which are `p2` and `taiwan_unique_species`, respectively, as shown in Fig. 7.3. Other rhetorical predicates used in our system are shown in Fig. C.1.

*Semantic mapping rules.* After the associated message contents are extracted from the domain knowledge base, the text planner converts them into semantic structures, taking the extracted message contents and the surface act as the argument. Each rhetorical predicate is associated with a semantic mapping rule. Through this rule, the text planner constructs the semantic structures, repeated below, for the extracted message contents.

[1] The concepts described here basically come from Maybury's TEXPLAN system. A more detailed description can be found in [Maybury 90].

| | | | |
|---|---|---|---|
| **Name:** | has_subpart | **Name** | constituency |
| **Entity:** | ... | **Entity:** | ... |
| **Subpart:** | ... | **Comps:** | ... |
| **Name:** | classification | **Name** | attribution |
| **Entity:** | ... | **Entity:** | ... |
| **Subtypes:** | ... | **Attrs:** | ... |

Figure C.1: Rhetorical predicates used in our system.




As described in Sec. 7.3.3, the values of `index`, `illocforce` and `aspect` in the semantic structure are obtained from an automatic counter and the corresponding surface act. A semantic rule is concerned with choosing a predicate and mapping the arguments from the extracted message contents. For example, considering the semantic mapping rule of the *logical definition* predicate, the `pred` is assigned with *be*; the `pat` role is obtained from the **Entity** in the extracted message content and the `range` roles of the two semantic structures are obtained from the **Superclass** and **Differentia** in the content. The semantic rules corresponding to the rhetorical predicates in Fig. C.1 are shown in Fig. C.2. Note that we only show one case of the *attribution* rule because of its complexity. The *attribute* and *value* correspond to an attribute-value pair of the **Entity**. In this case, the value of the pair is taken as the predicate.

| | | | |
|---|---|---|---|
| **Predicate:** | has_subpart | **Predicate:** | constituency |
| **Mapping:** | pred←have | **Mapping:** | pred←have |
| | pat←**Entity** | | pat←**Entity** |
| | range←**Subpart** | | range←**Comps** |
| **Predicate:** | classification | **Predicate:** | attribution |
| **Mapping:** | pred←divide_into | | pred← *value* |
| | pat←**Entity** | | exp←**Entity** |
| | range←**Subtype** | | patient←*attribute* |

Figure C.2: Semantic mapping rules used in our system.

# Appendix D

# Test Texts Used in the Evaluation

The five test texts used in the evaluation chapter are shown in Figs. D.1, D.2 and D.3.

## 文件編號：Text 1

面天山$_1$ 屬於 錐狀火山，___$_1$ 高度 977公尺，___$_1$ 山形 渾圓整齊，
(Z,它,面天山) (Z,它,面天山)

___$_1$ 頂部 有 矢竹草坡，___$_1$ 頂部西側 有 向天池。
(Z,它,面天山) (Z,它,面天山)

## 文件編號：Text 2

面天山$_1$ 屬於 錐狀火山，___$_1$ 高度 977公尺，___$_1$ 山形 渾圓整齊，
(Z,它,面天山) (Z,它,面天山)

___$_1$ 頂部 有 矢竹草坡，___$_1$ 頂部西側 有 向天池$_2$，___$_2$ 深 約 45公尺，
(Z,它,面天山) (Z,它,面天山) (Z,它,向天池)

___$_2$ 直徑 約 150公尺，___$_2$ 長 沼澤植物$_3$，___$_3$ 有 四角藺 和 燈心草，
(Z,它,向天池) (Z,它,向天池) (Z,它,沼澤植物)

___$_2$ 北側 有 天山$_4$ ___$_4$ 山形 圓扁，___$_4$ 海拔 880公尺。
(Z,它,向天池) (Z,它,向天山) (Z,它,向天山)

## 文件編號：Text 3

幻湖$_1$ 屬於 火口湖，___$_1$ 面積 約 2800平方公尺，___$_1$ 深 約 2 公尺，
(Z,它,夢幻湖) (Z,它,夢幻湖)

___$_1$ 位於 七星山東麓，___$_1$ 海拔 約 860公尺。___$_1$ 是 溼地形社會，
(Z,它,夢幻湖) (Z,它,夢幻湖) (Z,它,夢幻湖)

___$_1$ 是 保護區。___$_1$ 有 沒區$_2$ ___$_2$ 長有 沉水植物 和 挺水植物，
(Z,它,夢幻湖) (Z,它,夢幻湖) (Z,它,淹沒區)

___$_1$ 有 淹沒區$_5$ ___$_5$ 長有 生植物$_6$ ___$_6$ 有 五節芒 和 類地毯草。
(Z,它,夢幻湖) (Z,它,非淹沒區) (Z,它,陸生植物)

文件編號：Text 4

台灣水韭[1] 屬於 水韭科植物，___[1] 是 台灣特有種。___[1] ___莖[2]
(Z,它,台灣水韭) (Z,它,台灣水韭)

___[2] 外觀 三裂瓣狀，___[2] 寬 約 2 公分。___[1] ___的葉子[3]
(Z,它,短球莖,球莖) (Z,它,短球莖,球莖) (Z,它,台灣水韭)

___[3] 外觀 細長，___[3] ___[4] 匙狀， ___[3] 切面 半圓形，
(Z,它,翠綠的葉子,葉子) (Z,它,翠綠的葉子,葉子) (Z,它,翠綠的葉子,葉子)

___[3] 內側 平， ___[3] 度 約 20公分，___[3] 側 弧凸，
(Z,它,翠綠的葉子,葉子) (Z,它,翠綠的葉子,葉子) (Z,它,翠綠的葉子,葉子)

___[3] 長 在 ___莖[2] ___[1] ___圓形的孢子囊[5]
(Z,它,翠綠的葉子,葉子) (Z,它,台灣水 )

___[5] 長度 約 0.8公分， ___[5] 在 ___基部[4] 。
(Z,它,長橢圓形的孢子囊,孢子囊) (Z,它,長橢圓形的孢子囊,孢子囊)

___[1] ___的根[11] ___[11] 有 根毛組織，___[11] 造 簡單
(Z,它,台灣水韭) (Z,它,典型的根,根) (Z,它,典型的根,根)

___[11] 有 維管束單條，___[11] 有 通氣道。
(Z,它,典型的根,根) (Z,它,典型的根,根)

## 文件編號：Text 5

台灣水韭[1] 屬於 水韭科植物，___[1] 是 台灣特有種。___[1] ___莖[2]
(Z,它,台灣水韭) (Z,它,台灣水韭)

___[2] 外觀 三裂瓣狀，___[2] 寬 約 2 公分。___[1] ___的葉子[3]
(Z,它,短球莖,球莖) (Z,它,短球莖,球莖) (Z,它,台灣水韭)

___[3] 外觀 細長，___[3] ___[4] 匙狀， ___[3] 切面 半圓形，
(Z,它,翠綠的葉子,葉子) (Z,它,翠綠的葉子,葉子) (Z,它,翠綠的葉子,葉子)

___[3] 內側 平， ___[3] 度 約 20公分，___[3] 側 弧凸，
(Z,它,翠綠的葉子,葉子) (Z,它,翠綠的葉子,葉子) (Z,它,翠綠的葉子,葉子)

___[3] 長 在 ___莖[2] ___[1] ___圓形的孢子囊[5]
(Z,它,翠綠的葉子,葉子) (Z,它,台灣水 )

___[5] 長度 約 0.8公分， ___[5] 在 ___基部[4] ，
(Z,它,長橢圓形的孢子囊,孢子囊) (Z,它,長橢圓形的孢子囊,孢子囊)

___[5] 成 ___子囊[6]，___子囊[7] ___孢子囊[8] ___[6] 有 ___子[9]
(Z,它,長橢圓形的孢子囊,孢子囊) (Z,它,大孢子囊)

___[9] 外觀 四面體形，___[9] 徑 約 310 到 390微米，
(Z,它,大孢子) (Z,它,大孢子)

___[9] 數量 約 300 到 500個，___[9] 有 疣狀突起，
(Z,它,大孢子) (Z,它,大孢 )

___[7] 有 ___子[10] ___[10] 觀 二面體形，
(Z,它,小孢子囊) (Z,它,小孢子)

___[10] 徑 約 15 到 25微米，___[10] 有 尖刺狀突起，
(Z,它,小孢子) (Z,它,小孢 )

___[8] 有 大孢子 和 小孢子。___[1] ___的根[11] ___[11] 造 簡單，
(Z,它,混生孢子囊) (Z,它,台灣水韭) (Z,它,典型的根, )

___[11] 有 根毛組織，___[11] 有 ___道[12] ___[12] 有 隔膜組織，
(Z,它,典型的根,根) (Z,它,典型的根,根) (Z,它,通氣道,氣道)

___[11] 有 ___束單條[13] ___[13] 在 ___[12] 。
(Z,它,典型的根,根) (Z,它,維管束 條) (Z,它,通氣道,氣道)