An Empirical Study on the Generation of Anaphora in Chinese

Ching-Long Yeh[*]
Tatung Institute of Technology

Chris Mellish[†]
University of Edinburgh

The goal of this work is to study how to generate various kinds of anaphora in Chinese, including zero, pronominal, and nominal anaphora, from the syntactic and semantic representation of multisentential text. In this research we confine ourselves to descriptive texts. We examine the occurrence of anaphora in human-generated text and those generated by a hypothetical computer equipped with anaphor generation rules, assuming that the computer can generate the same texts as the human except that anaphora are generated by the rules. A sequence of rules using independently motivated linguistic constraints is developed until the results obtained are close to those in the real texts. The best rule obtained for the choice of anaphor type makes use of the following conditions: locality between anaphor and antecedent, syntactic constraints on zero anaphora, discourse segment structures, salience of objects and animacy of objects. We further establish a rule for choosing descriptions if a nominal anaphor is decided on. We have implemented the above rules in a Chinese natural language generation system that is able to generate descriptive texts. 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 anaphora. The results of the comparison show that the rules are fairly effective in dealing with the generation of anaphora in Chinese.

1. Introduction

The field of natural language generation has made a great deal of progress in the generation of multisentential text in recent years (McKeown 1985; Maybury 1990; Dale 1992; Hovy 1993). Most of the well-known systems first select and organize the message contents to be generated and then map the organized results into a sequence of surface sentences. When mapping into the surface form, the selection of appropriate forms for anaphora is very important to make the generated text a cohesive unit (McDonald 1980; Dale 1992). In this paper, our goal is the computer generation of anaphora in Chinese.

In Chinese, anaphora can be classified as zero, pronominal, and nominal forms, as exemplified in (1) by ϕ_1^i, *ta*^{i} 'he' and *nage ren*^{i} 'that person', respectively (Chen 1987).[1] Zero anaphora are generally noun phrases that are understood from the context and do not need to be specified. In contrast, in this paper, we use the term nonzero

[*] Department of Computer Science and Engineering, 40 Chungshan North Road, Section 3, Taipei, 104, Taiwan. Email: chingyeh@cse.ttit.edu.tw

[†] Department of Artificial Intelligence, 80 South Bridge, Edinburgh EH1 1HN, Scotland. Email: c.mellish@ed.ac.uk

[1] We use a ϕ_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 ϕ 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.

anaphora to denote those that are specified in discourse, namely, pronominal and nominal anaphora.

(1) a. *Zhangsan^i jinghuang de wang wai pao,*
Zhangsan frightened NOM towards outside run
'Zhangsan was frightened and ran outside.'
b. *ϕ^i_1 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. *ϕ^i_2 renchu na ren^j shi shui.*
(he) recognize that person is who
'(He) recognized who that person is.'

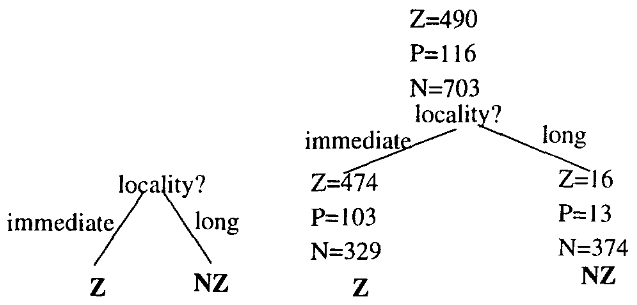
This research starts with establishing possible rules for the generation of anaphora in Chinese. Previous work suggests obtaining these rules from consulting the results of linguistic study, including general principles, such as the Gricean maxims (Grice 1975) used in (Dale and Haddock 1991; Reiter and Dale 1992; Dale 1992) and focus theory, as used in (Dale 1992). A shortcoming of previous work is that it is unclear to what extent the resulting rules are effective in dealing with the generation of anaphora. In an attempt to overcome this, we adopt an empirical approach to obtaining rules based on observations of real texts.

The basic methodology used is to start with a set of human-generated Chinese texts and the simplest possible anaphor generation rule (a rule that only considers the locality of anaphora). We then progressively add extra tests to the rule, based on independently motivated but simple linguistic principles. At each stage, we conduct experiments that compare the anaphora occurring in the human-generated text with those in the texts that would be generated by a computer taking the same syntactic and semantic content as the human texts and generating Chinese anaphora according to the rule being tested (this has to be simulated by hand). This process continues until a rule with promising performance on the data is obtained. The objective is thus to answer the question of how complex a rule must be to account for the complexity of anaphor generation exhibited by the test data.

This paper presents one sequence of rules developed using the above methodology and evaluates the effectiveness of the new linguistic principles taken into account at each point. At present, we have chosen only one intuitively plausible way to generate increasingly complex rules, with refinements introduced as they occurred to us (though not motivated by the data). Clearly the work could and should be extended to consider all possible combinations of the principles in all possible orders.

Except where noted below, the preselected Chinese data serves as an independent test of the effectiveness of the different rules, which are based on principles that have been independently suggested in the literature. However, the fact that the chosen data determine the termination condition for the development means that the rules could be overfitting the chosen data. Therefore a selection of the rules have been implemented in a Chinese natural language generation system and their results are further evaluated by means of an experiment using native speakers.

This paper concentrates on the use of zero, pronominal, and nominal anaphora in Chinese generated text. We are not concerned with lexical anaphora (Tutin and Kittredge 1992) where the anaphor and its antecedent share meaning components,



Matched	Overgenerated	Undergenerated	Total
861 (66%)	432 (33%)	16 (1%)	1,309

Figure 1
Decision tree, classification tree, and result for Rule 1.

while the anaphor belongs to an open lexical class. For example, *flower* can be used as a lexical anaphor for *rose* (Tutin and Kittredge 1992).

In Sections 2 to 3.3, we establish the rules for the generation of anaphora in Chinese. We consider the case of zero anaphora (Section 2) first, followed by nonzero anaphora (Section 3), which divides into pronouns (Section 3.1) and nominal anaphora (Sections 3.2 and 3.3). Next, in Section 4, we describe the implementation of the generation rules in our Chinese generation system and show the result of evaluating the anaphora in the text generated by systems employing different rules. Finally, Section 5 presents the conclusions.

2. Zero Anaphora

Initially we consider simply the decision of whether a generated anaphor should be a zero pronoun (Z) or some nonzero phrase (NZ).

2.1 Rule 1: Locality

Although there are no clear rules delineated in previous linguistic work, we, nevertheless, can summarize a very simple rule, Rule 1 as shown below and in an associated decision tree in Figure 1, for the generation of zero anaphora.

Rule 1
If an entity, *e*, in the current clause was referred to in the immediately preceding clause, then a zero anaphor is used for *e*; otherwise, a nonzero anaphor is used.

This is clearly a very simple rule, but it is interesting to see how well it performs. We now describe an experiment comparing the anaphora generated by a hypothetical computer employing this rule and those occurring in real text to see how well it works. The same basic format is used for subsequent experiments on more refined rules.

In this paper, the selected texts are restricted to the exposition type, which explain an idea or discuss a problem. Three sets of articles consisting of scientific questions and answers written by multiple authors, and an introduction to Chinese grammar, are selected as the test data (more details can be found in Yeh [1995]). In this data, there are 490 zero pronouns, 116 pronouns, and 703 nominal anaphora, making a total of 1,309 anaphora. The experiment is executed in three steps:

1. Zero and nonzero anaphora within the selected texts are identified.[2]
2. Each anaphor is given values according to the conditions in the current rule. For example, for Rule 1, an anaphor is determined to be immediate if its antecedent occurs in the immediately preceding clause; otherwise it is long-distance. We can then classify the anaphora corresponding to the decision tree of the rule, as in Figure 1. In the figure, Z and NZ denote zero and nonzero anaphora, respectively. Later we will use P and N to distinguish between pronouns and nominal anaphora.[3]
3. We assume that a hypothetical computer employing the current rule can generate the same text as the test data except for the anaphora, which are determined by the rule to be tested. We simulate this computer by hand and note down the difference between the anaphora generated by the computer and those in the test data.

In step 3, we categorize the differences between the results as: **matched, over-generated** and **under-generated** types. If a reference created by the simulated computer is the same as the one in the real text, then it belongs to the matched type. If a zero anaphor is created by the hypothetical computer, while the corresponding position in the real text is a nonzero anaphor, then it belongs to the overgenerated type. Conversely, if a zero anaphor is found in some position in the real text, while a nonzero anaphor is created by the computer, then it belongs to the undergenerated type.

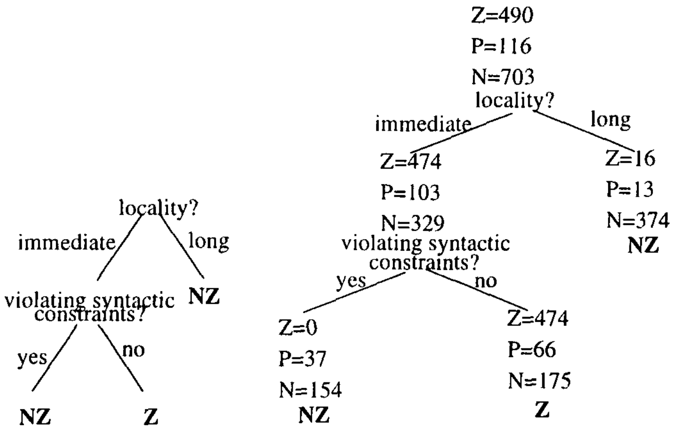
From the classification tree, the number of the matched type is the total number of zero and nonzero anaphora associated with zero and nonzero leaf nodes in the classification tree. The over- and under-generated types are counted as the numbers of nonzero and zero anaphora associated with zero and nonzero leaf nodes in the tree. The result of using Rule 1 on the test data is shown in Figure 1. In the table, the matched rate of the test data is 66%, which obviously shows an unpromising performance of the computer employing Rule 1. Apparently, what we need to do is to find more constraints to enhance Rule 1. As shown in the classification trees of the test data, the numbers of nonzeros are far greater than their counterparts, zeros, in the long-distance cases of anaphora. Thus, in the following, we will not make any refinement to the long-distance cases because little progress would be obtained.

2.2 Rule 2: Adding Syntactic Constraints

Li and Thompson (1979, 1981) formulated a negative rule stating that zero anaphora are not allowed in certain syntactic positions regardless of discourse factors: the NP right after a coverb, and the pivotal NP in a serial verb construction. Therefore, we enhanced Rule 1 by adding the above syntactic constraints on zero anaphora, which becomes Rule 2, as shown in Figure 2.

[2] This is not necessarily a trivial task, as of course there is no physical evidence for zero anaphora in text. Indeed, there is some question as to whether the notion of zero pronoun is the best way of accounting for the syntactic facts about languages such as Chinese. Since we are looking at things from a generation perspective, we have considered a zero pronoun to occur when an important semantic element is not overtly specified in the text. In practice, this criterion probably produces similar results to approaches considering verb subcategorization (Walker, Iida, and Cote 1994).

[3] Note that we only deal with third person pronouns in Chinese; thus, in the table, and the following, pronominal anaphora, or pronouns, refer to third person cases. In this paper, we treat the first and second person pronouns as nominal anaphora.



Matched	Overgenerated	Undergenerated	Total
1,052 (80%)	241 (18%)	16 (1%)	1,309

Figure 2
Decision tree, classification tree, and result for Rule 2.

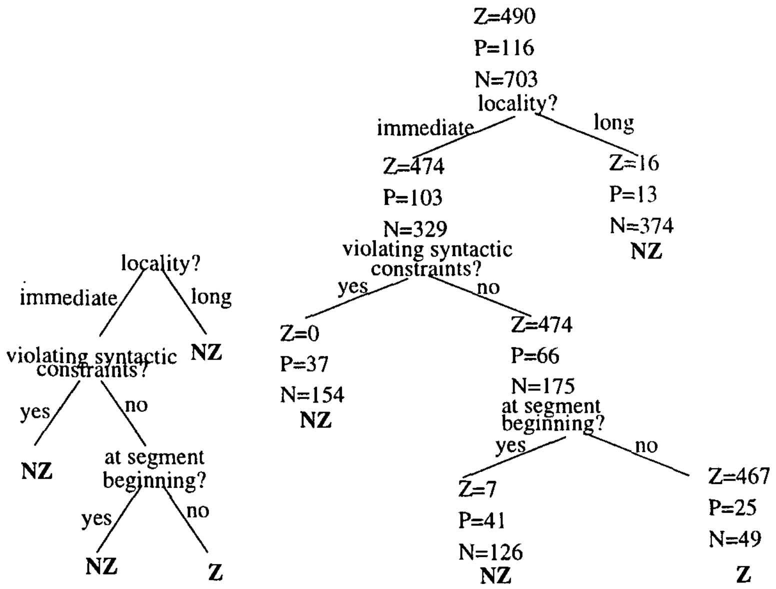
Rule 2
If an entity, *e*, in the current clause was referred to in the immediately preceding clause and does not violate any syntactic constraint on zero anaphora, then a zero anaphor is used for *e*; otherwise, a nonzero anaphor is used.

We then established for each anaphor in the test data whether a zero anaphor in this position would violate these syntactic constraints or not and obtained a new classification tree, as shown in Figure 2. The matched rate of Rule 2 is 80%, as shown in the same figure. Though Rule 2 improves its predecessor's performance, the result still discourages us from using it for the generation of zero anaphora in Chinese. As shown in Li and Thompson (1979) and Frosz and Sidner (1986), the structure of discourse is a significant factor affecting the use of anaphoric forms. Thus, we employed the notion of discourse structure as the basis for enhancing the rule.

2.3 Rule 3: Adding Discourse Structure

Grosz and Sidner (1986) 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. An important idea in the theory is the effect of the linguistic expressions in utterances constituting the discourse and the discourse segment structure on each other. On the one hand, linguistic expressions can be used to convey information about the discourse segment structure. On the other hand, the discourse segment structure constrains the interpretation of linguistic expressions. What concerns us here is the interrelationship between the forms of referring expressions and the discourse segment structures.

Li and Thompson (1979) propose the idea that the use of nonzero anaphora has to do with the segment boundaries in a discourse. A zero anaphor used to refer to some entity in the previous clause might be expected to indicate the continuation of a discourse segment, while a nonzero anaphor occurring in the same situation



Matched	Overgenerated	Undergenerated	Total
1,212 (93%)	74 (6%)	23 (2%)	1,309

Figure 3
Decision tree, classification tree, and result for Rule 3.

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 utterance 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 anaphora, to make Rule 3, as shown in Figure 3.

Rule 3

If an entity, *e*, in the current clause was referred to in the immediately preceding clause, does not violate any syntactic constraint on zero anaphora, and is not at the beginning of a discourse segment, then a zero anaphor is used for *e*; otherwise, a nonzero anaphor is used.

To determine the applicability of the new constraint to each anaphor, we had 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, by hand, according to perceived discourse segment intentions. Since our annotations were based on intuition, we tested them by comparing them 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 (see the Appendix) about the idea of discourse structure and the task to be done, namely, annotate the discourse segment boundaries according to the intentions of the discourse segments. The speakers reached a good level of agreement among themselves (obtaining a value of 0.76 for the kappa statistic [Siegel and Castellan 1988]) and adding our own annotations to the pool resulted in a similar level of agreement (kappa = 0.764). On average, 89% of our annotation markers match those of the speakers. From the above comparison, we judged that the annotations we made were highly reliable for the purpose of the experiment. The result also shows that the sentential marks in the test data closely correlate to the boundaries between discourse segments. 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; on the other hand, commas are inserted between clauses within a "sentence" as separators (Liu 1984).[4] 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.

The classification trees and results of the experiment are shown in Figure 3. By taking into account the effect of discourse segment structure, we obtained 93% matches in the test data. The result shows that Rule 3 is helpful for the decision as to whether to use a zero anaphor.

2.4 Rule 4: Adding Topic Continuity

Although the zero anaphora generated using Rule 3 look considerably similar to those in the test data, there are, nevertheless, still a number of overgenerations for the test data. Tai (1978), Li and Thompson (1979), and Chen (1984, 1986), have noticed that zero anaphora frequently occur in **topic chains** where a referent is referred to in the first clause, and then several more clauses follow talking about the same referent (the topic), but with it omitted; (1b) in Section 1 is an example. Here, we use the feature of topic-prominence in Chinese (Li and Thompson 1981) to further refine the previous rule.

In Chinese, the topic of a sentence is what the sentence is about and always comes first in the sentence; the rest of the sentence is comment upon the topic (Li and Thompson 1981). The topic is always either **definite** (refers to something that the reader already knows about), or **generic** (refers to a class of entities). The subject of a sentence, on the other hand, is the NP that has a "doing" or "being" relationship with the verb in the sentence. By distinguishing between topics and subjects in sentences, we have the following types of sentences: sentences with both subject and topic, sentences in which the subject and topic are identical, sentences with no subjects, and sentences with no topic (Li and Thompson 1981). A sentence without a topic is used to introduce a new entity into the discourse. In the remaining types of sentences, the topic can be found at the beginning of the sentence.

The basic idea here is to investigate the positions of the antecedent and the anaphor in their respective clauses. 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 anaphora in their respective utterances into topic and nontopic cases.

For each anaphor, its antecedent's position is classified as either topic or direct object. Thus we have the types of antecedent-anaphor pairs shown in Figure 4. Since in the new rule the condition of topic continuity in clause will be considered to refine the zero leaf node in the decision tree of Rule 3, we focus on investigating the corresponding anaphora in the classification trees. The numbers of the various types of

[4] The sentential mark also has two auxiliaries, question and exclamation marks, which are used to express "sentences" with certain tones.

Types of antecedent-anaphor pair

	Position of antecedent	Position of anaphor
Type A:	topic	topic
Type B:	object-1	topic
Type C:	topic	nontopic
Type D:	object-1	nontopic
Type E:	others	topic
Type F:	others	nontopic

Occurrence of antecedent-anaphor pairs

Anaphor	A	B	C	D	E	F	Total
Z	403	47	5	3	0	9	467
P	10	3	11	0	1	0	25
N	10	11	2	1	25	0	49

Figure 4
Types and occurrence of antecedent-anaphor pairs in the subset of test data corresponding to zero leaf of Rule 3.

antecedent-anaphor pairs in the test data, according to this classification, are shown in Figure 4.

Obviously, for columns A and B in the table, nonzero cases, namely the sums of pronouns and nominals, are in the minority of the test data. Chen (1987) found a higher percentage of zero anaphora occurring in the topic position with their antecedent most frequently in the topic or object positions of the immediately previous clause, which strongly supports the idea of letting anaphora of Types A and B be zero. Zero anaphora of Types A and B are generally understood because they are salient (Li and Thompson 1981). Anaphora of Types C to F are not as salient as Types A and B; thus we group Types C to F as nonsalient. The total number of zero cases for the nonsalient type is 17(4%) in the test data; the total number of nonzeros for the same type is 40(63%). Thus we let anaphora of the nonsalient type be nonzero. By letting Types A and B be zero, and others be nonzero, we obtained a new rule, Rule 4.

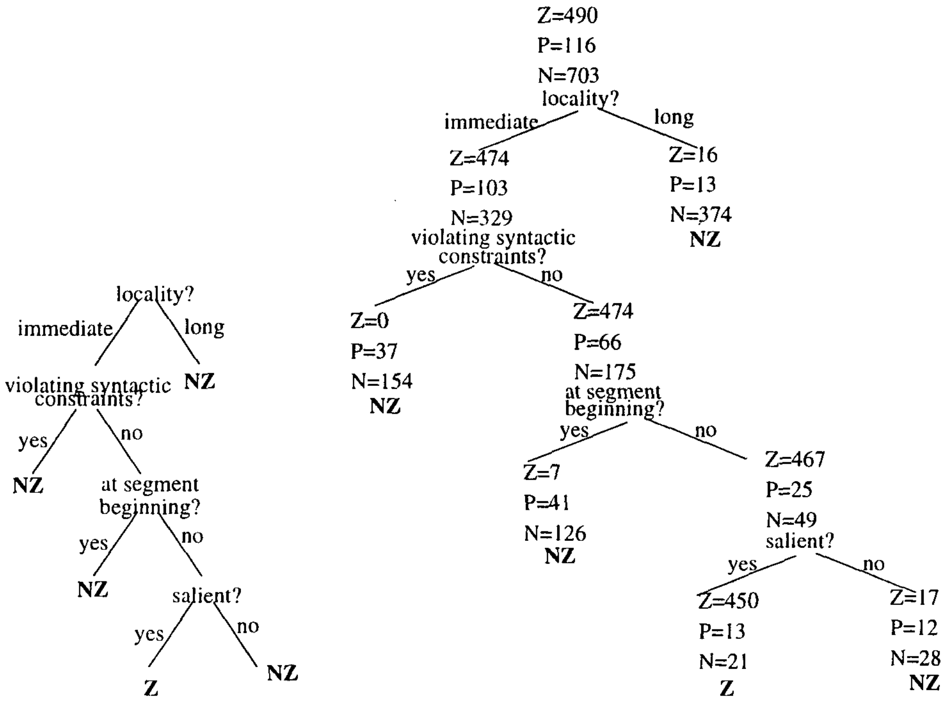
Rule 4
If an entity, *e*, in the current clause was referred to in the immediately preceding clause, does not violate any syntactic constraint on zero anaphora, is not at the beginning of a discourse segment, and is salient, then a zero anaphor is used for *e*; otherwise, a nonzero anaphor is used.

The decision tree and classification tree are shown in Figure 5.

The result in the same figure shows that the matched rate increased from 93% to 94%. Note that, although the new material in Rule 4 was motivated by the prior work of Chen and others, the exact form of the new constraint was formulated after considering the distribution of anaphora in the data, which means that an improvement (on this data) was almost inevitable.

3. Overt Noun Phrases

We now consider how to distinguish between pronouns and nominal anaphora.



Matched	Overgenerated	Undergenerated	Total
1,235 (94%)	34 (3%)	40 (3%)	1,309

Figure 5
Decision tree, classification tree, and result for Rule 4.

3.1 Animacy and Overt Pronominals

As shown in the classification trees of Rule 4 in Figure 5, pronouns are in the minority of the nonzeros in the test data, and indeed this is clearly the case in the language in general. A simple way to refine the previous anaphor generation rule is to let the nonzero parts in the rule be nominal. The decision tree and classification tree can then be obtained from Figure 5 by changing all nonzeroes (NZs) into nominals (Ns).

To demonstrate the result of using the new decision tree, we extended the definition of matched, overgenerated and undergenerated types used previously for zero and nonzero anaphora to zero, pronominal, and nominal anaphora. The number of matched cases for zero, pronoun, and nominal in the test data can be obtained by summing up anaphora of the correct type associated with the leaf nodes labeled Z, P, and N in the classification trees, respectively. The overgenerated cases of zero anaphora, for instance, are the sum of nonzero anaphora associated with the leaf nodes labeled Z in the classification trees. Conversely, the undergenerated cases of zero anaphora, for instance, are the sum of zero anaphora associated with the leaf nodes labeled with nonzeros. The overgenerated and undergenerated cases of pronouns and nominals can be obtained in a similar way. The result from using full NPs for nominal anaphora is shown in Table 1. Hereafter, we use **overall matched** to refer to the total number of matched anaphora, across all the classes. The number of overall matched cases is thus 1,132 (450 + 682), out of 1,309 anaphora in total. In general, we can convert this to a percentage by dividing by the total number of anaphora. Thus the percentage of

Table 1
Result of choosing full NP NZ in Rule 4.

Matched			Overgeneration			Undergeneration		
Z	P	N	Z	P	N	Z	P	N
450 (34%)	0 (0%)	682 (52%)	34 (3%)	0 (0%)	143 (11%)	40 (3%)	116 (9%)	21 (2%)

overall matched cases is 86%. This rate looks quite promising; however, it does not truly reflect the use of different nominal forms.

Li and Thompson (1979), and Chen (1986) showed that pronouns are frequently used when the anaphora 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 condition of animacy noticed by Chen can be determined according to the features of the referent and hence is easily implementable. In an examination of inanimate anaphora, Chen (1986) found that there were only a few instances of pronouns; in other words, most pronominal anaphora are animate. On the other hand, the percentage of inanimate anaphora being encoded in nominal forms is higher than that of pronouns. Thus we employ the animacy of the referent as a constraint to refine Rule 4 and obtain a new rule, Rule 5, as shown in Figure 6.

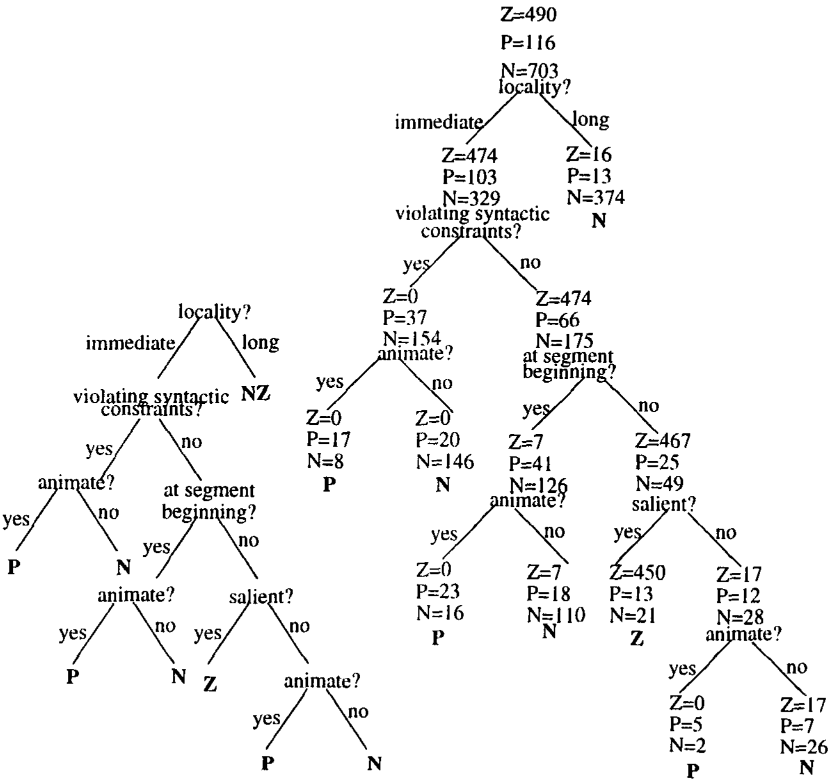
Rule 5
If an entity, *e*, in the current clause was referred to in the immediately preceding clause, does not violate any syntactic constraint on zero anaphora, is not at the beginning of a discourse segment, and is salient, then a zero anaphor is used for *e*; otherwise, a nonzero anaphor is used. If a nonzero anaphor is animate, then it is pronominalized; otherwise, it is nominalized.

In general, animate objects characterize living things, especially animal life. We adopted this concept to determine the animacy of anaphora. The result of using Rule 5 is shown in the table of Figure 6. Although the increase in the overall matched rate was not significant, 39% (45/116) of the pronouns in the test data, however, were matched by using the new rule.

3.2 Full NP Descriptions

The surface structure of a Chinese nominal anaphor is a noun phrase that consists of a head noun optionally preceded by associative phrase, articles, relative clauses, and adjectives (Li and Thompson 1981). In Chinese, whether one chooses articles for nominal descriptions depends on complicated factors (Teng 1975; Li and Thompson 1981). Observing the test data, we found that nominal anaphora are not commonly marked with articles.[5] Thus, we chose not to use articles for descriptions of nominal anaphora in our system. The nominal descriptions investigated in the remainder of this section are thought of as noun phrases of the above scheme without articles. Nominal anaphora do not have unique forms as their zero and pronominal counterparts do. The description can be the same as the initial reference, parts of the information in the

[5] See Yeh (1995) for detailed descriptions.



Matched			Overgeneration			Undergeneration			
Z	P	N	Z	P	N	Z	P	N	Total anaphora
450	45	656	34	26	98	40	71	47	1,309
34%	3%	50%	3%	2%	7%	3%	5%	4%	

Figure 6
Decision tree, classification tree, and result for Rule 5.

Table 2
Occurrence of various types of nominal anaphora in the test data.

Bare	Full	Reduced	New	Other	Total
471 (69%)	85 (12%)	72 (11%)	31 (5%)	23 (3%)	682

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 paper, we focus on the first two cases. 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.

We can classify nominal descriptions into the types shown in Figure 7. The breakdown of the matched nominal anaphora in the test data, in terms of the above classification, is shown in Table 2. Note that first and second person pronouns in the test data are classified as Type Bare in the table.

Types of nominal anaphora.

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

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

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

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

Other Otherwise.

Examples of nominal anaphora.

	Initial references	**Nominal anaphora**
Bare	*zuqiu* 'football'	*zuqiu* 'football'
Full	*tie-tong* 'iron barrel'	*tie-tong* 'iron barrel'
Reduced	*tie-tong* 'iron barrel'	*tong* 'barrel'
New	*shui* 'water'	*yuan-wan-zhong de shui* 'water in the round bowl'
Other	*qian* 'money'	*neixie chaopiao* 'those notes'

Figure 7
Types and examples of nominal anaphora.

The figures in Table 2 show that full descriptions, namely, Types Bare and Full, are frequently used for nominal anaphora. Thus we first choose full descriptions for all N's. As shown in Table 2, there are 556 (471 + 85) full descriptions used among 682 matched nominal anaphora. Thus the overall matched rate becomes 77%, if we take different descriptions of nominal anaphora into account. Obviously this shows that the choice of full NP for nonzeros is not promising. In the next subsection, we improve this by considering the use of reduced and full descriptions.

3.3 Reduced Descriptions within Segments

Previous work on the generation of referring expressions focused on producing minimal distinguishing descriptions (Dale and Haddock 1991; Dale 1992; Reiter and Dale 1992) or descriptions customized for different levels of hearers (Reiter 1990). Since we are not concerned with the generation of descriptions for different levels of users, we look only at the former group of work, which 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. Minimal distinguishing descriptions pursue efficiency in producing an adequate description that can identity the intended referent unambiguously with a given context set. Dale (1992) used the global focus space (Grosz and Sidner 1986), 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 anaphora 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 anaphora if the *context sets* at both places contain the same elements. On the other hand, in general, a description with more distinguishing information is used for the second anaphor if **distractors** have entered into the context set. 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*. The entity, *the big cat*, is not a distractor to *the black dog* because it is of different category, *cat*.

Grosz and Sidner (1986) claim that discourse segmentation is an important factor, though obviously not the only one, governing the use of referring expressions. If the idea of context set were restricted to local focus space (Grosz and Sidner 1986), then the resulting descriptions would be to some extent sensitive to local aspects of discourse structure. Although the algorithms would be refined due to the introduction of more discourse structure, they would essentially still serve the purpose of distinguishing potential referents.

The beginnings of discourse segments, in a sense, indicate shifts of intention in a discourse (Grosz and Sidner 1986). In this situation, it may be preferred that subsequent references be full descriptions rather than reduced ones or pronouns, to emphasize the beginning of discourse segments, even if the referents have just been mentioned in the immediately previous utterance. See Grosz and Sidner (1986) and Dale (1992) for some examples that illustrate this idea. Figure 8 indicates that a similar situation may happen in Chinese discourse.

Among the groups of initial and subsequent references, 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 Figure 9. At the beginning of the second "sentence," it appears in a full description and then in four reduced descriptions in the rest of the "sentence."[6] It is not mentioned in the third "sentence." When it is reintroduced into the fourth "sentence," it appears in another full 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 (g), can be interpreted as emphasizing 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 at the beginning of a "sentence" because it is the first mention in the "sentence." Thus, we would generalize 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 (2d) and *fangwan* 'the square bowl' in (2e).[7]

(2)
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,*

[6] See Section 2.3 for an explanation of "sentence."
[7] This is also obtained from the test data.

a. *fengzheng^i ϕ fangdao gaokong shangqu yihou,*
b. *la fengzheng^i de xian^j zhenme ye la bu zhi,*
c. *ϕ^j zongshi xiang xia wan,*
d. *zhe shi weishenme ne?*
e. *yuanlai, buguan fang fengzheng^i de xian^j you duome xi,*
f. *ϕ^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. *ϕ^k ba xian^j xiangxi zhuai,*
j. *xian^j ϕ jiu la bu zhi le.*
k. *qishi, fengzheng^i ye you zhongliang,*
l. *yinwei feng^m chui zhe fengzheng^i,*
m. *ϕ^m shi fengzheng^i xiang shang sheng,*
n. *suoyi fengzheng^i bingbu xiang xia chen.*
o. *zheyang, ϕ 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. *ϕ^j yue la bu zhi.*

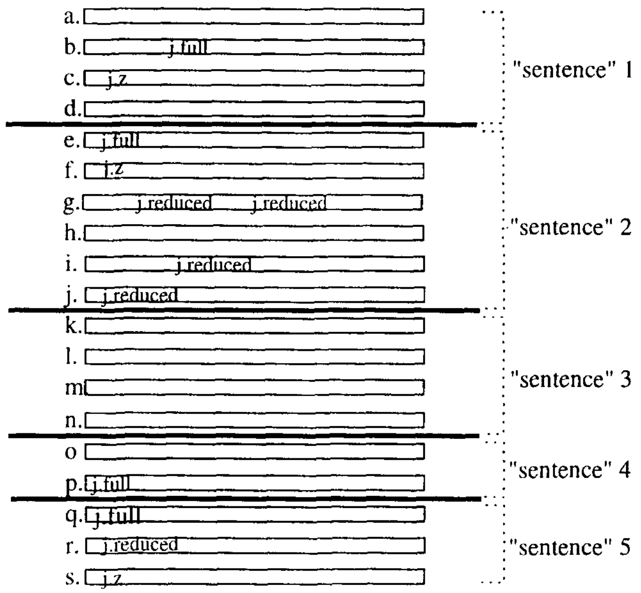
Translation:

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 cannot 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 8
A sample Chinese written text.

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.



Key: j.z: referent j in zero form.
j.full: referent j in full noun phrase.
j.reduced: referent j in reduced noun phrase
▬ : "sentence" boundary.

Figure 9
Occurrence of referent *j* in the discourse in Figure 8.

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 anaphora in Chinese.

Preference Rule
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.

We examined the nominal anaphora matched by using Rule 5 with the ones generated by the preference rule. The result is shown in Table 3. As shown in the table, by using the preference rule, in addition to the fact that the majority of the nominal anaphora using full descriptions are matched, a considerable number of reduced descriptions are matched as well, giving an overall match of 88%. If we only consider

Table 3
Result of using the preference rule on the test data.

Matched	Bare	Full	Reduced	New	Other	Total	%
yes	459	67	53	0	0	579	88%
no	0	13	13	31	20	77	12%

Types Bare, Full, and Reduced, namely, full and reduced descriptions in the test data, the match rates become 96% (579/605). Both figures show that the preference rule is promising in the choice of full or reduced descriptions for nominal anaphora.

4. Implementation and Evaluation Result

In this section, we briefly describe the implementation of the rules in our Chinese natural language generation system. We then present an evaluation of the anaphora in some texts generated by our system.

4.1 Implementation

The rules obtained in the previous sections have been implemented in the referring expression component of our Chinese natural language generation system (Yeh 1995) that generates paragraph-sized texts for describing the plants, animals, etc., in a national park. Basically, the main goal of our work is to generate coherent texts by taking advantage of various forms of anaphora in Chinese. The system, like conventional ones (McKeown 1985; Maybury 1990; Dale 1992; Hovy 1993), is divided into strategic and tactical components. Since we do not aim at inventing new concepts in content planning, we borrow the idea of text planning in Maybury's TEXPLAN system (Maybury 1990) as the basis of the strategic component. As for the tactical component, we have constructed a simple Chinese grammar in the PATR formalism (Shieber 1986), which is sufficient for our purpose at the current stage.

On accepting an input goal from the user, the system invokes the text planner according to the operators in the plan library to build a hierarchical discourse structure that satisfies the input goal. After the text planning is finished, the decision of anaphoric forms and descriptions is then carried out by traversing the plan tree. Within the traversal, when a reference is met, if it is a subsequent one, then the program consults Rule 5 to obtain a form: zero, pronominal, or nominal. If the nominal form is chosen, then the preference rule is consulted to get a description.

In the domain knowledge base, each entity, in addition to the information for the head noun in the surface form, is accompanied by a property list that will be realized in the modification part of the surface noun phrase for the initial reference. We build up the semantic structure of an initial reference by taking all the elements in the property list, along with the substance of the entity, corresponding to the head noun in the surface noun phrase. To simplify the work, for the moment, only one element is stored in the property list. When a full description is chosen for a subsequent reference, its semantic structure contains the same property and substance information as the initial reference. On the other hand, if a reduced description is decided on, only the substance is taken into the semantic structure. In the future, we will extend the property list by allowing multiple elements in the list.

The tests of locality, syntactic constraints, and salience are straightforward to implement because the system has complete knowledge of the discourse to be generated and its syntactic structure. Only the tests of discourse structure and animacy are difficult, and for these we have had to approximate what a more sophisticated system might be able to do. Currently, we examine the decomposition field of a planning operator by hand to determine "sentence" boundaries and fix this for all applications of the operator. Thus we assume that there is a distinguished level of structure in a discourse plan that is relevant for this purpose (this may be expressible in terms of Maybury's distinction between rhetorical acts and speech acts). For the animacy constraint, we have had to determine by hand whether each individual object in our domain is likely to be treated as animate or not.

4.2 Evaluation

The linguistic principles embodied in our rules were all independently proposed, so in some respects the previous data served as both training and test data in the development of the rules. Furthermore, the assumed contextual information, for example, discourse structures, may be difficult to access in a real implementation. Thus, the performance of a *real* anaphor generation algorithm based on the rules proposed here may be different from the experimental results we obtained. In this section, we attempt a post-evaluation by asking some native speakers of Chinese to judge the quality of the anaphora generated by a real system based on the rules.

Evaluation is becoming an increasingly important issue for natural language generation systems (Meteer and McDonald 1991), though, unfortunately, there are still no generally accepted methods. In this work, we were particularly concerned to find a method of evaluation that reflected directly on the anaphor generation of the system (unlike "black box" evaluation of the kind we had done before [Levine and Mellish 1995]). We were also wary of asking human subjects to estimate the "readability" or "coherence" of texts (though this seemed to work well for Acker and Porter [1994]). In this evaluation, we chose three Chinese natural language generation systems to compare. Each system is assumed to have the same system components, as described in Section 4.1, except that the referring expression component of each system is equipped with a different anaphor generation rule. Given an input to a test system, anaphora in the resulting texts will be determined by the rule used in the referring expression component of the system. The rules, TR*i*, $i = 1, \ldots, 3$, used in the test systems are shown in Figure 10. TR*1* corresponds to our Rule 2, together with an animacy test to distinguish between pronouns and nominal anaphora. TR2 adds the constraint on discourse structure and TR3 adds to this the salience constraint (and is the same as Rule 5). The intention was to test a range of rules and hence get an indication of how much better (if at all) the more sophisticated rules are than the simpler ones.

The evaluation task can be divided into an annotation stage and a comparison stage. In the annotation stage, each of 12 native speakers of Chinese is given five test sheets corresponding to five texts generated by our generation system. The numbers of clauses in the texts are 5, 12, 12, 21, and 34; the numbers of anaphora in the texts are 4, 11, 11, 20, and 34.

Each anaphor position in a generated text was left empty and all candidate forms of the anaphor, including zero, pronominal, and full and reduced descriptions were put under the empty space. The speaker was asked to annotate which form he or she preferred for each anaphor position on the test sheets. After the annotations were collected, we compared the speakers' results with 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. This approach is the same as that used in Knight and Chander (1994) for the problem of article generation, except that in our case we had to use generated, rather than naturally

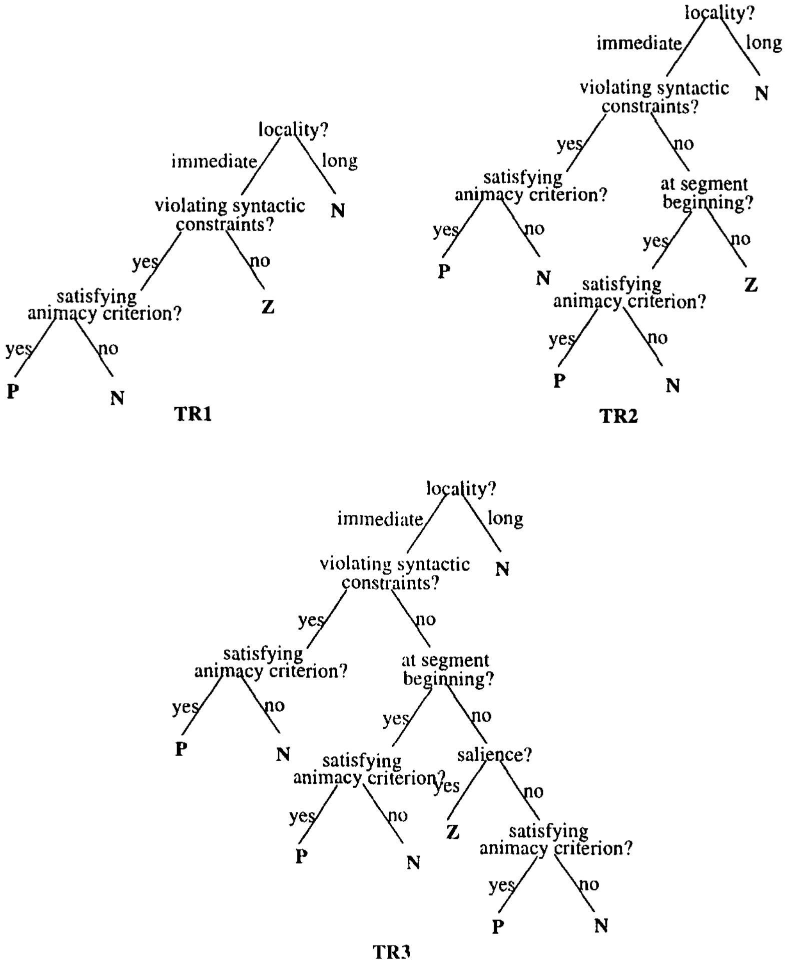

locality?
immediate
long
violating syntactic constraints?
N
yes
no
satisfying animacy criterion?
Z
yes
no
P
N
TR1
locality?
immediate
long
violating syntactic constraints?
N
yes
no
satisfying animacy criterion?
at segment beginning?
yes
no
P
N
yes
no
satisfying animacy criterion?
Z
yes
no
P
N
TR2
locality?
immediate
long
violating syntactic constraints?
N
yes
no
satisfying animacy criterion?
at segment beginning?
yes
no
P
N
yes
no
satisfying animacy criterion?
salience?
yes
no
yes
no
P
N
Z
satisfying animacy criterion?
yes
no
P
N
TR3

Figure 10
Rules used in the compared systems.

Table 4
Average match rates between the results of test systems and native speakers.

System	Text 1	Text 2	Text 3	Text 4	Text 5
TR1	3.6	7.8	6.8	14	23.8
	90%	70%	62%	70%	70%
TR2	3.6	7.8	7.3	14.9	24.3
	90%	70%	66%	75%	71%
TR3	3.6	8.7	7.1	14.6	24
	90%	79%	65%	73%	71%

occurring, texts, because otherwise our system would not have had access to the appropriate syntactic and semantic information. The average matching rates of the texts generated by the test systems with native speakers' results are shown in Table 4. On average, the matching rate of TR3 is 76%, compared with the other systems, the matching rate of TR1 is 72% and of TR2 is 74%.

Table 5
Agreement of annotations among speakers.

Speaker	Text 1	Text 2	Text 3	Text 4	Text 5
1	3.9	8	7.5	14.3	24
2	3.9	9.5	7.8	16.1	26.5
3	3.9	9.1	7.8	15.8	26.3
4	3.9	8.9	6.6	15.4	23.9
5	3.3	8.5	8.3	15.2	25.4
6	3.9	9.5	8.3	14.1	26.5
7	3.9	8.3	7.1	15	26.2
8	3.9	8.1	7.9	15.8	26.4
9	2.4	6.8	7	12.1	20.5
10	3.9	8.6	8.1	14.5	25
11	2.3	5.7	7	12.7	21.2
12	3.9	9.4	7.8	15.3	26.3
Average	3.6	8.4	7.6	14.7	24.9
	90%	76%	69%	73%	73%

This average matching rate, however, is lower than the matching rates we obtained in the empirical studies described previously. The problem is partly because the test texts used in the former comparison are human-created, while the test texts used here are computer-generated. The grammatical structures of the computer-generated texts are simplified; they are not as sophisticated as human texts. When asked to decide their preferences for anaphora in the computer-generated texts, speakers may find the information shown in the test texts less complete than what they are used to in creating their own texts and hence it may be difficult for them to make decisions. In the empirical study, the human-created texts perhaps provided enough information for the hypothetical computer to decide on an appropriate anaphoric form.

A more important reason why the matching rates are lower with speakers than with the hypothetical computer may be that in some circumstances, more than one solution may be acceptable and the speakers may not always choose the same one as the computer. This hypothesis can be investigated by looking at the extent to which the speakers agree among themselves.

To see how the speakers agree among themselves, we compared speakers' annotations. The comparison result is shown in Table 5. For each speaker, the number for each test text is the average of matches with the other eleven speakers. At the end of the table are the average numbers for the speakers' agreement among themselves. The figures in the table show that the speakers do not achieve agreement among themselves for the use of anaphora in this test. These figures are further supported by the use of the kappa statistic. The overall kappa value for all speakers is about 0.41, which represents only "moderate" agreement. The measure of agreement gets worse if only the zero/ pronoun/ nominal distinction is considered or if zero and nonzero pronouns are lumped together. Only two speakers agree with one another with a kappa value of more than 0.7 (none with a value of greater than 0.8). The speakers as a whole agreed with kappa greater than 0.7 on only 30 out of the 80 anaphora, with complete agreement only 14 times. To get an overall agreement of greater than 0.8 would require reducing the set of speakers from 12 to a carefully selected 3.

Since all systems produce the same result on Text 1, unsurprisingly they all have the same matching rate, as shown in Table 4. Text 2 contains three topic shifts that would make the rule containing the salience constraint, TR3, obtain different output from those without this constraint. TR1 and TR2 produce the same output and hence they obtain the same matching rate, 70%. TR3 obtains higher matching rates than the other two, 79%, which shows the effectiveness of the salience constraint in it.

Another middle-sized test text, Text 3, is broken into three "sentences" and contains three topic shifts. The constraints on discourse segment beginnings in TR2 and TR3 and the salience constraint in TR3 would therefore have some effects on the output texts. The matching rate, as shown in Table 4, increases from 62% to 66% for TR2, which shows that the constraint on discourse segment beginnings in TR2 is effective. TR3 obtains a 65% matching rate, on average, which is 1% lower than its predecessor TR2. However, this decrease in average matching rate does not negate the effectiveness of the salience constraint in TR3. TR2's text differs from TR3's in the three topic shifts: TR2 generates zero anaphora for these shifts, while TR3 generates full descriptions. The speakers varied greatly in choosing anaphoric forms for these topic shifts: among 12 speakers, 4 chose all full descriptions, 3 used all zero anaphora, and the other 5 chose zero, pronominal, and nominal anaphora. Thus, 4 of the 12 speakers completely agree with TR3, while 3 agree with TR2. This shows that the salience constraint in TR3 is still effective.

Next, we examine the more complicated texts, Texts 4 and 5. As shown in Table 4, the increases in matching rates show the effectiveness of the constraint on discourse segments beginning in TR2. Again, the average matching rates of TR3 are sightly lower than TR2 for these two texts. However, similar to the situation in Text 3, the speakers have varied agreement on the choice of anaphora for the topic shiftings in these two texts. For Text 4, 3 speakers completely agree with TR2 and 1 speaker agrees with TR3. As for Text 5, 2 speakers completely agree with TR2, while the others partly agree with TR2 and TR3.

The discussions above show that the salience constraint in TR3 is sometimes effective in getting small improvements in the output texts. In brief, the more sophisticated constraints a rule contains, the better it performs. Both TR2 and TR3 perform better than TR1. TR3 performs better than TR2 for texts with simple discourse segment structure. For the texts having complicated discourse segment structures, TR2 is slightly better than TR3 on average matching rates. Adding the results of the rules to those of the speakers leads to a slight decrease in kappa for TR1 but progressively better (though only from 0.41 to 0.43) values for kappa for TR2 and TR3. This indicates that the better rules seem to disagree with the speakers no more than the speakers disagree among themselves. There are nine anaphora where the kappa score including TR3 is less than that for the speakers alone (in many other cases, the results are better). These seem to involve places where the speakers were more willing to use a zero pronoun (where the system used a reduced nominal anaphor) and where the speakers reduced nominal anaphora less than the system did.

5. Conclusion

In this paper, we present empirical work on the generation of anaphora in Chinese. The initial set of results suggests that most anaphora, including zero anaphora, and full and reduced descriptions for nominal anaphora, can be effectively generated by a rule using simple syntactic, semantic, and discourse constraints. The results obtained from an implementation of this rule, however, correlated less well with human performance. It is hard to determine the reason for this, though the problems of reliably implementing all the constraints, presenting the anaphora within natural-looking texts and, above all, coping with the disagreements between native speakers, all probably make a contribution.

The factors affecting the use of pronouns are very complicated; thus it is difficult to get computable rules. Introducing the constraint of animacy of objects in the rule can resolve part of the problem. We do not handle the generation of long-distance pronouns, which were rare in our texts. A possible solution would be to employ the concept of stacked focus space in Grosz and Sidner's discourse structure theory (Grosz and Sidner 1986; Dale 1992).

In the final rule, the implementation of the test of the beginning of a discourse segment is not quite as straightforward as the other constraints. In our current implementation, we rely on the hierarchical structure of the message content to be generated as the basis for dividing the message into segments, which is effective in improving the texts generated by our Chinese natural language generation system. The evaluation result also shows that the rule using all constraints collected from the empirical study performs better than one with simpler constraints.

In the future, this work needs to be further developed to deal with anaphora in other types of texts and the use of connectives in generated text to create cohesive discourse. In addition the constraints for pronominal anaphora could be improved, and the implementation extended to satisfy other types of applications.

6. Acknowledgements

We would like to thank the three anonymous reviewers of this paper, whose detailed comments have been extremely valuable.

Appendix: Instructions 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 to answer 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 subpurpose (or subintention). Therefore the purpose (or intention) of an article is obviously composed of a string of subpurposes (or subintentions). In other words, every subpurpose (subintention) serves as a part of the whole intention of an article. Furthermore, in an article, a subpurpose (or subintention) can be a subsidiary of other subpurposes (or subintentions), just like subpurposes (subintentions) are subsidiaries of the whole intention. That is, a subpurpose can subsume others. Therefore, we have a hierarchical intentional structure for an article,
>
> **Task:** 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.

References

Acker, Liane and Bruce Porter. 1994. Extracting viewpoints from knowledge bases. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pages 547–552. AAAI Press/MIT Press.

Chen, Ping. 1984. A discourse analysis of third person zero anaphora in Chinese. Technical Report, Indiana University Linguistics Club, Bloomington, IN.

Chen, Ping. 1986. *Referent Introducing and Tracking in Chinese Narratives*. Ph.D. thesis, University of California, Los Angeles, CA.

Chen, Ping. 1987. Hanyu lingxin huizhi de huayu fenxi [A discourse approach to zero anaphora in Chinese]. *Zhongguo Yuwen [Chinese Linguistics]*, pages 363–378.

Dale, Robert. 1992. *Generating Referring Expressions: Constructing Descriptions in a Domain of Objects and Processes*. The MIT Press, Cambridge, MA.

Dale, Robert and Nicholas Haddock. 1991. Content determination in the generation of referring expressions. *Computational Intelligence*, pages 252–265.

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

Grosz, Barbara J. and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

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

Knight, Kevin and Ishwar Chander. 1994. Automated postediting of documents. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pages 779–784. AAAI Press/MIT Press.

Levine, John and Chris Mellish. 1995. The idas user trials: Quantitative evaluation of an applied natural language generation system. In *Proceedings of the Fifth European Workshop on Natural Language Generation*, pages 75–94, Leiden, The Netherlands.

Li, Charles N. and Sandra A. Thompson. 1979. Third-person pronouns and zero-anaphora in Chinese discourse. In T. Givón, editor, *Syntax and Semantics: Discourse and Syntax*, volume 12. Academic Press, pages 311–335.

Li, Charles N. and Sandra A. Thompson. 1981. *Mandarin Chinese: A Functional Reference Grammar*. University of California Press, Berkeley, CA.

Liu, Y. C. 1984. *Zuowen de fangfa [Approaches to Composition]*. Xuesheng Chubanshe, Taipei, Taiwan.

Maybury, Mark T. 1990. *Planning Multisentential English Text Using Communicative Acts*. Ph.D. thesis, Cambridge University.

McDonald, David D. 1980. *Natural Language Generation as a Process of Decision Making under Constraints*. Ph.D. thesis, MIT.

McKeown, Kathleen R. 1985. *Text Generation*. Cambridge University Press.

Meteer, Marie and David McDonald. 1991. Evaluation for generation. In J. G. Neal and S. M. Walter, editors, *Natural Language Processing Systems Evaluation Workshop*, pages 127–131, NY. Rome Laboratory.

Reiter, Ehud. 1990. 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.

Reiter, Ehud and Robert Dale. 1992. A fast algorithm for the generation of referring expressions. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 232–238.

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

Siegel, Sidney and N. J. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill.

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

Teng, Shou-Hsin. 1975. *A Semantic Study of the Transitivity Relations in Chinese*. University of California Press, Berkeley, CA.

Tutin, Agnès and Richard Kittredge. 1992. Lexical choice in context: Generating procedural texts. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, pages 763–769, Nantes, France.

Walker, Marilyn, Masayo Iida, and Sharon Cote. 1994. Japanese discourse and the process of centering. *Computational Linguistics*, 20(2):193–232.

Yeh, Ching-Long. 1995. *Generation of Anaphors in Chinese*. Ph.D. thesis, University of Edinburgh, Edinburgh, Scotland.