Determinative-Measure Compounds in Mandarin Chinese

Formation Rules and Parser Implementation

*Ruo-ping Jean Mo**, *Yao-Jung Yang**, *Keh-Jiann Chen**, *Chu-Ren Huang***
**The Institute of Information Science, Academia Sinica*
***The Institute of History and Philology, Academia Sinica*
Nankang, Taipei, Taiwan,
Republic of China

Abstract

We deal with the identification of the determinative-measure compounds (DMs) in parsing Mandarin Chinese in this paper. The number of possible DMs is infinite, and cannot be listed exhaustively in a lexicon. However, the set of DMs can be described by regular expressions, and can be recognized by a finite automaton. We propose to identify DMs by regular expression before parsing.

After investigating large linguistic data, we find that DMs are formed compositionally and hierarchically from the simpler constituents. Based upon this fact, some grammar rules are constructed to combine determinatives and measures. Moreover, a parser is also formed to implement these rules. By doing so, almost all of the unlisted DMs are recognized. However, if only the DM recognition procedure is fired, many ambiguous results appear, too. Yet with our word segmentation process, these amiguities are greatly reduced.

I. Introduction

A determinative-measure compound (DM) in Mandarin Chinese is composed of one or more determinatives, together with an optional measure.(1) It is used to determine the reference or the quantity of the noun phrase that co-occurs with it. It may sometimes function as a noun phrase by itself.(2) However, despite the fact that the categories of determinatives and measures are both closed, the combinations of them are not.

(1) 這 三 本
D D M
this three CL
"these three books"

(2) 他 喜 歡 這 三 個
he like this three CL
"He likes these three."

(3) 三 百 二 十 一
three hundred two ten one
"three hundred and twenty one"

五 萬 四 千 三 百 二 十 一
five ten-thousand four thousand three hundred two ten one
"fifty four thousand three hundred and twenty one"

九 億 零 五 萬 四 千 三 百 二 十 一
nine hundred-million zero five ten-thousand four thousand three hundred two ten one
"nine hundred million fifty four thousand three hundred and twenty one"

(4) 九 點 三十 分
nine o'clock thirty minute
"half past nine"

三月 八 日 星期五 上午 九 點 三十 分
March eight day Friday morning nine o'clock thirty minute
"nine-thirty a.m., Friday, March eighth"

民國八十年三月 八 日 星期五 上午 九 點 三十 分
1991 March eight day Friday morning nine o'clock thirty minute
"nine-thirty a.m., Friday, March eighth, 1991"

(5) 二 十 五 件
two ten five CL
"twenty five items"

這 二 十 五 件
this two ten five CL
"these twenty five items"

其餘 二 十 五 件
other two ten five CL
"the other twenty five items"

Take example (3) as an illustration: it is obvious that the total numbers of possible combinations of numerals are enumerable but infinite. Because of the productivity of these DMs, listing them directly in the lexicon becomes almost impossible. Consequently, the process of finding proper word breaks for Chinese sentences is incomplete without DMs in the lexicon. Therefore our design of a word segmentation system utilizes both the words listed in the lexicon and those generated by DM rules. We have the following reasons to support our strategy. First, from the processing point of view, it is better to recognize compound words as early as possible and DMs can be considered as compounds. Since the structure of DMs seems to be exocentric, they are not similar to other endocentric phrase structures and can not be analyzed by head driven parsing strategies. Second, the set of DM forms is a regular language which can be expressed by regular expressions and recognized by finite automata. It is well known that the grammar of Mandarin contains central embedding and must be expressed by context-free grammars. [7] This also suggests that the processing of DMs should be separated from the processing of other phrases. Third, the set of determinatives and measures usually serve only a single grammatical function which are comparatively simpler than other categories which play multiple grammatical functions due to the lack of inflections in Chinese. We believe that DMs can be identified at the level of lexical analysis and this fact has been proven by our experiments. We design a regular grammar interpreter with a chart parser to identify the DMs for input sentences. The flexible design of this interpreter allow us to modify the grammar rules generating DMs without changing the interpreter.

In the next section, the structures of DMs and their representations are given. The third section states the design of the interpreter and its application to improve the DM rules. The fourth section shows the experimental results and discussions. The last section concludes with remarks on other applications of the DM identification system.

II. The Structures and Representations of DMs

Earlier studies of DMs concentrate mainly on 1. listing members of determinative and measure sets, 2. proposing classifications, and 3. describing agreements between measures and their nominal heads. Chao[8], for instance, divides determinatives into four subclasses:

(6) (i) demonstrative determinatives: 這, 那, 哪
(ii) specifying determinatives: 每, 各, 別, 旁, 本, 某, 上, 下, 前, 後, 今, 昨, 明, 去
(iii) numeral determinatives: 二, 百分之三, 四百五 十 etc.
(iv) quantitative determinatives: 一, 滿, 全, 整, 半, 幾, 多, 多少,許多, 好些, 好多, 好幾, 很多

Measures, on the other hand, are divided into nine classes by Chao[8] 1. classifiers, e.g. 本 "a (book)", 2. classifiers associated with V-O constructions, e.g. 手 "hand", 3. group measures, e.g. 對 "pair", 4. partitive measures, e.g. 些 "some", 5. container measures, e.g. 盒 "box", 6. temporary measures, e.g. 身 "body", 7. standard measures, e.g. 公 尺 "meter", 8. quasi-measures, e.g. 國 "country", and 9. measures with verbs, e.g. 次 "number of times". However, earlier studies do not analyze the internal structure of DMs, which is crucial to their recognition and formation.

In what follows, we will first adjust the various determinative sets based on their productivity and co-occurrence restrictions, and then discuss the internal structure of DMs,

as well as the rules to construct them. As for the measures, although they also play a role in forming DMs, the choice of them largely depends on the nature of the entity denoted by the nominal heads. Since this paper focuses on the DM itself, the problem of agreement between measures and nominal heads will not be pursued here.

2.1. The Determinative Sets

In general, determinatives are classified in terms of their meanings. However, if we take typical grammatical properties such as productivity and co-occurrence restrictions into account, we find that some of the classifications based upon meanings are questionable.

Instead we propose three criteria to classify various determinatives. They are 1. productivity, 2. syntactic similarities, and 3. semantic meanings. The determinatives are quite different in terms of productivity. For instance, 今 "today", 明 "tomorrow", 去 "last", 昨 "yesterday" precede no other determinatives. In fact, they can only co-occur with a few measures such as 日 "day", 天 "day", and 年 "year". Since their usage is fixed, we will put all the possible combinations of those determinatives and the measures, such as 今天，明天，今年，明年，今日，明日，昨天，昨日，去年， in the lexicon. On the other hand, the determinatives with high productivity will be classified according to their syntactic and semantic similarity.

Although Mandarin Chinese allows two or more determinatives to be juxtaposed, not every determinative can co-occur with the others. 别 "other", and 旁 "side", for example, are incompatible with other determinatives. But 這 "this" is relatively free: it can be adjoined to either a numeral or a quantitative determinative :

(7) *别 三 家
other three home

*旁 半 天
side half day

這 三 名
this three CL
"these three persons"

這 半 年
this half year
"this half year"

Therefore, co-occurrence relations will be the major syntactic criteria employed to subclassify determinatives.

The primary function of a determinative is to restrict or quantify the references of the following noun phrases. From the data collected, a variety of other words also have much the same function and distribution as those well-discussed determinatives. 近 "near" and 將 近 "near" are two such words. Like 這 "this" in (8), 近 "near" in (9) also modifies the following noun phrase and determines which period the event expressed by the verb phrase occurs[1] Actually, these two words can be substituted with each other in this context. Based upon this principle, those with similar function and distribution as determinatives will also be included in the determinative set.

---

[1] Another reason for treating 近 "near" and 將 近 "near" as determinatives comes from the grammatical theory we adopt. According to one assumption of the Lexical Mapping Theory, every verb must have a subject.[3][7] However, this condition will not be held if we analyze 近 "near" and 將 近 "near" as verbs whenever they appear:

(i) 我 家 離 台 北 很 近
I home from Taipei very near
"My home is near Taipei."

端 午 節 將 近
Dragon-Boat-Festival near
"It's almost Dragon Boat Festival."

(ii) 近 十 時 三 十 分 我 回 到 家
near ten o'clock thirty minute I back home
"I got home at about ten-thirty"

In sentence (ii), 近 "near" cannot take 張 三 "Jangsan" as its subject. In fact, no subject may occur before it. In order not to violate the more accepted condition, we will classify 近 "near" and 將 近 "near" as determinatives besides verbs.

(8) 這 二 十 年 來 他 每 天 晨 泳
this twenty year come he every day morning swim
"He's gone for swim every morning for the past twenty years."

(9) 近 二 十 年 來 他 每 天 晨 泳
near twenty year come he every day morning swim
"He's gone for swim every morning for about twenty years."

Due to space limitations, we will simply list our revised determinative sets in Appendix I and related DM rules in Appendix II without further discussion.

2.2. The Internal Structures and Formation Rules of DMs

As was mentioned at the very beginning of this paper, a DM can contain one or more determinatives together with an optional measure. Closer investigation shows that the composition of the determinative can be complicated: it may consist of only one kind of iterating determinative, like numerals, or have several determinatives belonging to different subsets. For example, in 其他數百名 "hundreds of the other," three different kinds of determinatives are concatenated. In addition, these adjoining determinatives are not freely ordered. They have to conform with some linear precedence restrictions.[2]

(10) a. 這 二 十 人
this twenty person
"these twenty persons"

a' *二 十 這 人
twenty this person

b. 其 餘 近 一 百 五 十 名 團 員
other near one hundred fifty CL member
"the other almost one hundred and fifty members"

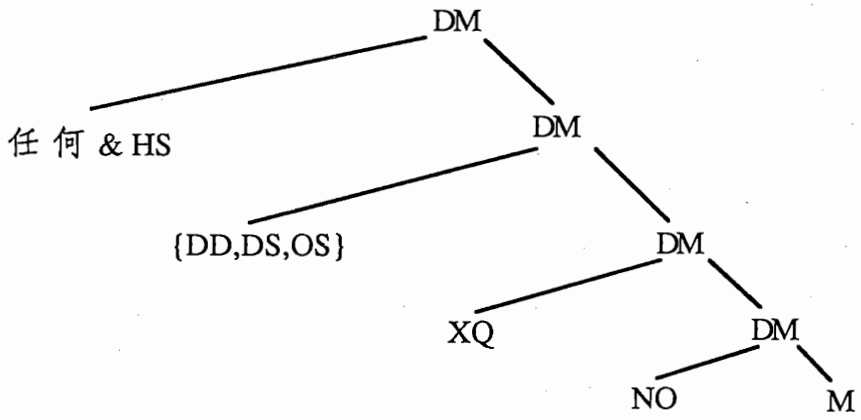
b' *一 百 五 十 其 餘 近 名 團 員
one hundred fifty other near CL member

[2]Similar restrictions also appear among numeral compounds. However, such restrictions depend on mathematic knowledge, not linguistic knowledge. In this paper, we do not handle these restrictions.

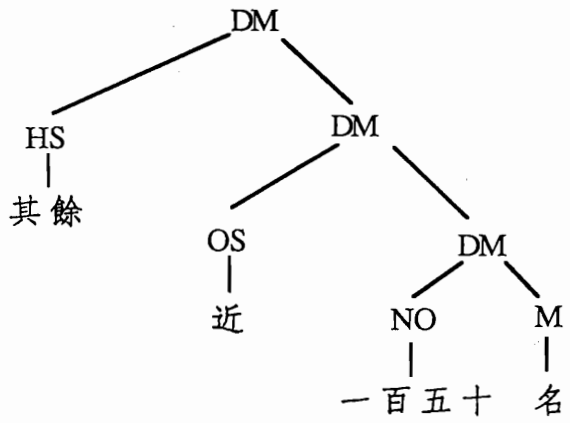
In unmarked cases, a numeral compound occurs at the leftmost position of a compound made up of determinatives. Similar precedence relations also exist among other determinative sets.

In order to account for the precedence order, we propose a tree structure for the general construction of DMs. This tree structure represents two facts: first, that a DM compound is formed compositionally and hierarchically from the simpler constituents such as numerals and measures. Second, that two determinatives belonging to the same level generally do not co-occur.

(11) The Tree Structure of DMs[3]



(12) 其餘近一百五十名



[3]Here M=measure, NO=numerals, XQ are various quantitative determinatives such as interrogative quantitative determinatives. As for DD, DS, and OS, they are demonstrative determinatives, definite specific determinatives, and ordinal specific determinatives respectively. Finally, HS refers to those specific determinatives which have the meaning of "the other".

Based on this tree structure, our DM formation rules begin with the combinations of recurring numerals, numeral compounds, post-nominal modifiers (PNMs), and measures.[4] New DMs can be formed by attaching other determinatives to these basic numeral compounds. Co-occurrence restrictions developed by other linguists as well as by ourselves will be taken into consideration at this stage. For instance, demonstrative determinatives can not co-occur with interrogative determinatives, or with those listed in the DS set. Some example rules can be seen in (13). Please refer to Appendix II for a complete list of rules.

(13)
IN1--> NO* ;
IN2--> NO* {多 ,餘 ,來 } ({萬 ,億 ,兆 }) ;
DN--> (IN1) {點 } IN1 ;
FN--> IN1 {分 之 } IN1 ;
FN--> IN1 {又 } FN ;
NOP1-->IN1 (DESC) ({半 }) (LM) ;
NOP2-->DESC ({半 }) LM ;
NOP3-->IN1 {平 方 ,立 方 } Nfga ({的 }) ;
NOP4-->IN1 (M) PNM ({的 }) ;
NOP5-->M (PNM) ({的 }) ;
NOP6-->{IN2,DN,FN} (LM) ;

Three remarks can be made about the above rules. First, as observed in Lu [12], some adjectives such as大 "big",小 "small",整 "whole" and 長 "long" may be inserted into a DM.[5] However, the measures that can follow 長 "long" are more restricted;

[4] In general, a determinative precedes a measure. But those listed in the PNM set, such as 半 "half", 整 "whole", the situation is quite to the contrary in that most determinatives have to occur after PNM measures.

[5] Lu [12] lists seven such adjectives: 大 "big", 小 "small",長 "long", 寬 "thick", 薄 "thin",滿 "full" and整 "whole". However, owing to dialect variations, 厚 "thick" and薄 "thin" never appear between determinatives and measures in Taiwan Mandarin. As for 滿 "full", we follow Chao [4] as well as CKIP [10] and regard it as a determinative of the WQ subcategory which denotes the concept of wholeness.

actually, only six measures can co-occur with the word 長 "long". Since its productivity is rather low, we will list all the combinations of 長 "long" and the compatible measures directly in the measure set. Second, in Mandarin Chinese a DM may be followed by a clitic 的 "DE" to indicate that it serves as a modifier (Huang [6]), like 三磅的肉 "three pounds of meat" or 兩桌的客人 "two tables of guests." But not every measure can co-occur with the 的 "DE": most classifiers, for example, are incompatible with 的 "DE". For processing efficiency, we list the various combinations of 的 "DE" and the immediately preceding measures in the measure set, too. [6] Third, for the convenience of language analysis, we also consider complex time expressions (14) and reduplicated DMs (15) as a single unit and express them by our DM rules.

(14) 中華民國 八十 年 九月 十日 二時 十分
R.O.C. eighty year September ten day two o'clock ten minute
"ten after two, September tenth, 1991"

(15) 一 個 個
one CL CL
"one by one"

一 瓶 瓶 的
one CL CL DE
"bottle by bottle"

From the above discussion, it is shown that the structures of DMs are quite complicated.

---

[6]However, this does not imply that when 的 "DE" follows a DM, it will be always correct to combine them together. In certain cases, this 的 "DE" should be attached to larger phrase of which the DM is only one of its constituents. For example, in the following two sentences, the 的 "DE" adjacent to the DMs is actually a relativizer relativizing the whole verb phrases.

(i) 坐 前 兩 排 的 學 生
sit front two row DE student
"the students who sit in the first two rows"

(ii) 已 經 喝 完 一 杯 咖 啡 的 人 請 站 起 來
already drink finish one cup coffee DE person please stand up
"Those who finished the first cup of coffee please stand up."

In order to test and modify our determinative sets and formation rules, we construct a rule interpreter and a chart parser.

III. An Interpreter for Regular Grammar and Its Application to Improve DM Rules

We have already shown that DMs in Mandarin Chinese can be expressed by a set of "Regular Expressions". We construct a regular expression interpreter and a chart parser in order to recognize DMs in input sentences. By testing the real input data from corpus, we can iteratively improve our classification and rule sets.

Our system is not the first DM parser. Chuang [11], based on the classifications developed in CKIP [10], creates a set of grammar rules and a program to implement the rules. However, the coverage of his grammar rules is incomplete and his program is procedure oriented which means it has to be modified once the grammar rules have changed.

Our system is divided into two parts: transformation-interpretation, and parsing.[7] At the transformation-interpretation stage the system transfers the grammar rules into a simpler format. The rules are originally in the form of regular expressions.[8] They are transformed into the format known as Chomsky Normal Form (Aho & Ullman [1]). The reason why we did not write it in Chomsky Normal Form originally is because it is easier to write the

---

[7]This original interpreter of the grammar and DM parser is designed and developed by Charles Lee of Stanford University and Yao-Jung Yang cooperatively. All other programs mentioned in this paper are written by Yao-Jung Yang.

[8]In this paper, all the Determinatives and Measure words are defined in symbol sets placed together with the grammar rules. By doing so, it is very convenient to modify the rules as well as the sets when we are running tests. However, this strategy will not be adopted in actual implementation because it will cause data redundancy. The lexical information of Determinatives and measure words must be attached to the words after dictionary lookup.

rules first in the form of regular expressions. On the other hand, it is much easier for computer to interpret the Chomsky Normal Form. The benefit of the interpreting approach is that we can modify the rules over and over again without changing the program. The following is a fragment of our grammar rules before and after the transformation-interpretation stage:

(16)

```
NO      = { ○,一,二,兩,三,四,五,六,七,八,九,十,
            廿,卅,百,千,萬,億,兆,零,幾};

IN1     ->      NO*;

IN2     ->      NO*   { 多,餘,來 }  ( { 萬,億,兆 } );
```

```
NO = {○,一,二,兩,三,四,五,六,七,八,九,十,廿,卅,
      百,千,萬,億,兆,零,幾};
IN1 -> NO IN1;
IN1 -> NO;
_0 = {多,餘,來};
_1 = {萬,億,兆};
_2 -> NO _2;
_2 -> NO;
_3 -> _0 _1;
_3 -> _0;
IN2 -> _2 _3;
_4 = {點};
_5 -> _4 IN1;
```

The parsing part of the program is built according to the concept of "Chart Parsing". The reason why we choose chart parsing as our basic strategy is because the chart data structure can hold all information about words which can then be used in the latter stage of the Information-Based Case Grammar (ICG) (Chen & Huang [5]) parsing process.

However, for actual testing, the program still has to be equipped with a preprocessor and a postprocessor: the former breaks the input article into sentences. Based upon the fact that in general no DMs can cross a punctuation marker, this article can be broken into substrings with punctuations as delimiters before being fed to the parser.[9] The latter reads

---

[9]But, in certain cases, a comma or punctuation mark " " is inserted into a numeral phrase. For example, we may have 五,六 月 間 "during May and June", 第 三、四、五 層 "the third,fourth, and fifth floors", 三、四 百 人

the output chart files produced by the core parser and does a filtering process to eliminate redundant or intermediate results. These two processors can be executed separately from the core parser so that the core parser is kept more adaptive to the other usages.

All the programs mentioned above are developed in the C language on the Borland Turbo C System. Some fragments of the input data and their output forms will be presented and analyzed in the following section.

IV. Discussion of Results

During test runs, postprocessed output is evaluated based upon two factors: first, are all of the DMs in the input article recognized, and second, how many are overgenerated? The former is concerned with the completeness of the rules; the latter is concerned with their accuracy. In the following we define some statistical values for the purpose of analysis.

N_{act} = the number of DMs in the testing article.
N_{ove} = the number of substrings which are recognized but are not DMs.
N_{mis} = the number of DMs in the testing article which are not recognized.
N_{rec} = the number of DMs which are recognized by our system.

"three or four hundred people", 十五，六歲 "fifteen or sixteen years old", etc. These marks either indicate a list (cf. the first two examples), or present an omission resulting from repetition (cf. the last two examples):

三、四百人 = 三百 or 四百人
十五，六歲 = 十五 or 十六歲

At this moment in time, these phrases cannot be correctly recognized for our rules do not take punctuation marks into consideration. The problem should be solved with an appropriate preprocessor.

After testing over 16 articles picked from a corpus,[10] we have:

The recognition rate = (N_{act} - N_{mis}) / N_{act} = 100%
The missing rate = N_{mis} / N_{act} = 0%
The overgeneration rate = N_{ove} / N_{rec} = 39.57%

Article#	N_{act}	N_{rec}	N_{ove}	N_{mis}
1	16	22	6	0
2	71	86	15	0
3	22	40	18	0
4	12	25	13	0
5	13	29	16	0
6	4	22	18	0
7	22	42	20	0
8	20	28	8	0
9	20	38	18	0
10	9	14	5	0
11	22	28	6	0
12	22	33	11	0
13	28	50	22	0
14	26	46	20	0
15	36	59	23	0
16	19	37	18	0
Total	**362**	**599**	**237**	**0**

From the missing rate, it shows that the completeness of the system is perfect. As for the soundness, the overgeneration rate seems to be quite high. However, after carefully studying the test result, we find that the overgenerations are mainly caused by ambiguous word segmentation. Thus these ambiguities can be avoided if we incorperate the DM recognition and word segmentation processes in parallel.

The ambiguities can be further classified into the following different cases:

[10]The corpus is supported regularly by two daily news associations: the Liberty Times and the United Daily News. The amount of data supported per month is about 4 M bytes.

1. Ambiguities resulting from lexical ambiguity. (i.e. polysemy of lexical items)

(17) a. 張 三 一 天 只 要 上 一 節 課
Jangsan one day only want up one CL class
"Jangsan has only one class a day."

b. 草皮 上 三 棵 老 橡 樹
lawn up three CL old oak
"There are three oaks on the front lawn."

c. 上 一 節 課 張 三 沒 來
up one CL class Jangsan not come
"Jangsan was not here for the first part of the class."

This kind of overgeneration is caused by the multi-categorization of individual lexical items. For example, 上 "up" may function as a verb, a localizer, or a determinative. In (17a), it is a verb; in (17b), it is a localizer; only in (17c) does it function as a determinative. We devise the following resolution principle to solve this kind of overgeneration.

> Resolution Principle 1: If the first character of the longest matched DMs is a lexical entry with multi-categories, such as 上 , 下 , 大 , 前 and 後, then both the longest matched DM and the DM without the first character are kept and the ambiguity will be resolved in the parsing stage.

2. Ambiguities resulting from improper word-breaks involving lexical items.

(18) 應 該 把 欠 我 的 錢 還 給 我 了 吧
should BA owe I DE money return I ASP
"(You) shoud return the money you owe me."

要 統 一 分 發 這些 信
want unify distribute these letter
"distribute these letters at the same time"

訂 了 好些 週 刊
order ASP many weekly magazine
"ordered many weekly magazines"

73% of the ambiguities in our test results belong to this type. We also found out that if an ambiguous word break occurs between a lexical word and a DM, the lexical word has the priority, as exemplified in (18). Therefore, we have the following resolution principle:

> Resolution Principle 2: If ambiguous word breaks occur between the words in the lexicon and the DMs, the words in the lexicon should have higher priority to get the shared characters.

3. Ambiguities resulting from improper word breaks involving proper names.

(19) a. 同 時 促成 了 宮本 對 死亡 的 認識
same-time cause ASP miyahon to death DE know
"At the same time Miyahon was forced to realize the meaning of death."

b. 警一分局 忙得不可開交
the-first-precint extremely busy
"The first precint was extremely busy."

The number of proper names is unlimited and therefore can not be exhaustively listed in the lexicon. Thus we are not able to apply resolution principle 2 if the ambiguous word breaks appear between proper names and DMs. So far, we do not have any good solution principles to solve this problem. Fortunately, only 6.33% of ambiguities are of this type.

As was mentioned in the previous paragraph, most of the ambiguities can be disambiguated by word segmentation. This does in fact happen after word segmentation is tried. For instance, in example (18) 應該 "should", 統一 "unify", 分發 "distribute", 週刊 "magazine" are words in lexicon, and thus get the priority in becoming units. Since the characters 該 "should", 一 "one", 分 "distribute", 週 "week" are part of words, no overgenerated DMs in these sentences exist any longer. However, to those overgenerations resulting from lexical ambiguity, the ambiguous word segmentations will still be kept. The following is our new test result derived from combining DM parsing and word break procedure. The recognition rate is $(N_{act}-N_{mis})/N_{act}$=99.17%, and the

ambiguity rate is N_{amb}/N_{act}=12.71%. By ambiguities (N_{amb}) we mean those caused by ambiguous word segmentations and those resulting from proper names.

Article#	N_{act}	N_{amb}	N_{mis}
1	16	1	0
2	71	0	2
3	22	3	0
4	12	2	0
5	13	1	0
6	4	2	0
7	22	3	0
8	20	1	0
9	20	1	0
10	9	1	0
11	22	3	0
12	22	2	1
13	28	14	0
14	26	5	0
15	36	6	0
16	19	1	0
Total	**362**	**46**	**3**

Another type of ambiguity which we do not consider as overgeneration is that some DMs are intrinsically ambiguous with multiple structures, as in (20), or multiple functions as in (21).

(20) 這 下 三 天 也 做 不 完 了
this down three day also do not finish ASP
"Even in three days, we can not finish it."

1) 這 下 三 天
2) 這 下 三 天

(21) a. 簡 直 像 作 夢 一 樣
almost like dream the same
"just like a dream"

b. 她 用 手 一 指
she use hand one point
"She pointed with her finger."

c. 不 知 該 帶 孩子 去 那 兒 玩
not know should bring child go where play
"(I) don't know where to take the children to play."

For such cases, the ambiguity remains to be resolved by parsing.

V. Applications and Concluding Remarks

We pointed out at the beginning of this paper that the combinations of DMs are infinite, and thus can not be exhaustively listed in the dictionary. Moreover, they occur quite frequently in the text. In order to solve this unavoidable problem in parsing, we build a DM parser to be a supplement of the lexicon.

The motivation for us to build this DM parser is to support the word segmentation module of the project developed in the Institute of Information Science, Academia Sinica, whose final goal is to establish a knowledge representation model of Mandarin Chinese. However, the word segmentation module depends heavily on a dictionary, which does not hold a complete list of DMs. With this parser, all those previously unrecognized DMs can be recognized.

Another application of our DM parser involves improving the efficiency of the phonetic input of the Mandarin Chinese. The most common idea to improve the efficiency of the phonetic input method is to utilize a lexicon with a phonetic code of every Chinese word as a key index because the more syllables a word has, the fewer homophones it possesses. With this parser, we can recognize the DMs by their phonetic spelling and greatly reduce the homophonic ambiguity.

In this paper, a DM parser together with some test results are presented. After scrutinizing a large amount of linguistic data, we form some grammar rules to combine determinatives and measures whenever they appear, and a parser to implement these rules. By doing so, all unlisted DMs are recognized. As for the test result, the recognition rate is quite satisfactory, although many pseudo DMs are overgenerated. Nonetheless, these

overgenerated "DMs" are disambiguated by incorperating word segmentation and the DM recognition processes in parallel.

However, at this moment, no semantic features have been taken into consideration. They are not only important to the interpretation of DMs, but also useful for the reduction of ambiguous readings. This is because if co-occurrence restrictions between determinatives and measures can be found, many pseudo DMs will no longer appear. But these restrictions largely depend on the semantic compatibility existing between determinatives and measures. We hope in the near future that these semantic features may be added to our rules to reduce overgenerations, and thus reduce ambiguous readings.

After undergoing large amounts of testing, the rules and sets are proved to be quite complete. The next step is to revise the DM parser program to a finite-automata version instead of an interpreter version in order to improve the performance and reduce the program size. By doing so, the DM parser can be more easily embedded into the whole parsing project.

References

[1] Aho, Alfred V. and Jefferey D. Ullman. 1972. *The Theory of Parsing, Translation, and Compiling*. London: Prentice Hall International.

[2] Allen, James. 1987. *Natural Language Understanding*. Menlo Park, California: The Benjamin/Cummings Publishing Company.

[3] Bresnan, Joan and Jonni M. Kanerva. 1989. Locative Inversion in Chichewa: A Case Study of Factorization in Grammar. *Linguistic Inquiry* 20: 1-50.

[4] Chao, Yuen-Ren. 1968. *A Grammar of Spoken Chinese*. Berkeley: University of California Press.

[5] Chen, Keh-Jiann and Chu-Ren Huang. 1990. Information-based Case Grammar, Proceedings of Coling 90: 54-59.

[6] Huang, Chu-Ren. 1987. Mandarin Chinese NP de: A Comparative Study of Current Grammatical Theories, Ph.D. dissertation Cornell University.

[7] .....1991. Mandarin Chinese and The Lexical Mapping Theory--A Study of the Interaction of Morphology and Argument Changing. *Bulletin of the Institute of History and Philology* 62.2.

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

[9] Lyons, John. 1977. *Semantics II*. New York: Cambridge University Press.

[10] 詞庫小組 (CKIP), 1989, 國語的詞分類修訂版, 南港: 中央研究院計算中心

[11] 莊德明 (Chuang, De-Ming),1986,一套中文定--量詞處理系統的研究與設計, 國立清華大學碩士論文

[12] 陸儉明 (Lu, Jianming),1987,"數量詞中間插入形容詞情況考察",語言教學與研究 1987年第四期,pp 53-72

[13] 黃居仁 (Huang, Chu-Ren), 1989,試論漢語的數學規範性質, 中央研究院歷史語言研究所集刊 60.1:47-71.

Appendix I

NO = {〇,一,二,兩,三,四,五,六,七,八,九,十,廿,卅,百,千,萬,
億,兆,零,幾};

ON = {甲,乙,丙,丁,戊};

DESC = {大,小};

PNM = {多,餘,半,出頭,好幾,開外,整,正,許,足,之多};

Ndabe = {清晨,凌晨,早晨,早上,晚上,上午,中午,下午,晨間,午間,
晚間,半夜,午夜,晨,午,晚,傍晚,深夜,曷午,子時,丑時,
寅時,卯時,辰時,巳時,午時,未時,申時,酉時,戌時,亥時};

Ndaac = {光緒,乾隆,廣德,昭和, etc.};

Ndaad = {民國,中華民國,西元,公元, etc.};

Ndabb = {春天,春季,夏天,夏季,秋天,秋季,冬天,冬季};

Ndabd1 = {星期一,星期二,星期三,星期四,星期五,星期六,星期日,
星期天,禮拜一,禮拜二,禮拜三,禮拜四,禮拜五,禮拜六,
禮拜日,禮拜天,週一,週二,週三,週四,週五,週六,週日};

Ndabf = {上旬,中旬,下旬,暑假,寒假,春假, etc.};

TPNM = {半,多,許,整,正};

WQ = {一,全,滿,整,成,一切,一整};

QQ = {多少,若干,幾多};

DQ = {多,許多,很多,好多,好些,少許,多數,少數,大多數};

PQ = {半,若干,有的};

DD = {這,那,哪};

OS = {上,下,前,後,頭,末,次,另,某,近,將近};

DS = {本,貴,敝,什麼,啥,何,別,旁,他};

Appendix II

```
IN1     ->      NO*;
IN2     ->      NO* {多,餘,來} ({萬,億,兆});
DN      ->      (IN1) {點} IN1;
FN1     ->      (IN1 {又}) IN1 {分之} {IN1, DN} ({強,弱}) ({的});
FN2     ->      (IN1 {又}) IN1 {分之} {IN1, DN};
ONP     ->      ON (LM);
NOP1    ->      IN1 (DESC) ({半}) (LM);
NOP2    ->      DESC ({半})  LM ;
NOP4    ->      IN1 (M) PNM ({的});
NOP3    ->      IN1 {平方, 立方} Nfga ({的});
NOP5    ->      M (PNM) ({的});
NOP6    ->      {IN2,DN,FN2} ( LM );
NOP     ->      {FN1,NOP1,NOP3,NOP4,NOP5,NOP6} ;
WQP     ->      WQ (LM);
WQP     ->      WQ (Nff ({的}));
QQP     ->      QQ (NOP5);
QQP     ->      QQ {的} ;
DQP1    ->      {好幾} ({NOP1,NOP2,NOP3,NOP5}) ;
DQP2    ->      {DQ1, DQ2} (LM);
DQP3    ->      {最多, 最少} (DQ3) {NOP1, NOP3, NOP4, NOP6} ;
DQP4    ->      DQ3 {NOP1, NOP3, NOP4, NOP6} ;
DQP5    ->      {DQ1, DQ2} {的} ;
DQP     ->      {DQP1,DQP2,DQP3,DQP4,DQP5} ;
```

```
PQP1    ->      {數} ({NOP1,NOP2,NOP3,NOP5}) ;
PQP2    ->      PQ (NOP5);
PQP     ->      {PQP1, PQP2} ;
XQP     ->      {WQP,QQP,DQP,PQP} ;
CNP     ->      IN1 {年} {IN1,ON} {班} ;
DSP1    ->      DS (LM) ;
DSP2    ->      {該} ({NOP, PQP}) ;
DSP     ->      {DSP1, DSP2} ;
OSP1    ->      {第} NOP1;
OSP2    ->      {每,各} {XQP,NOP,DSP2} ;
OSP3    ->      OS {PQP,NOP1,NOP3,NOP6} ;
OSP3    ->      {前, 後} DESC {半} LM ;
DDP1    ->      DD ({ WQP, DQP, PQP, NOP, NOP2 });
DDP2    ->      {此} ({OSP1, NOP});
OHSP    ->      ({其它, 其他, 其餘} ({的})) {任何} ({NOP1, DSP});
OHSP    ->      ({其它, 其他, 其餘} ({的})) {任何} ({的});
OSP     ->      {OSP1,OSP2,OSP3} ;
HOSP    ->      ({任何}){其它, 其他, 其餘} ({XQP,DDP1,OSP,NOP,ONP});
HOSP    ->      ({任何}){其它, 其他, 其餘} ({的});
STDM    ->      IN1 {分} (IN1 {秒} (IN1));
STDM    ->      IN1 {秒} (IN1);
TDM1    ->      IN1 {時,點,小時} (STDM) (TPNM);
TDM1    ->      IN1 {時,點,小時}  IN1 {刻} (TPNM);
TDM2    ->      ({Ndaac,Ndaad)}{元}{年}({元}{月}(IN1 ({日,號})));
TDM2    ->      ({Ndaac,Ndaad)}IN1 {年}({元}{月}(IN1 ({日,號})));
```

```
TDM2    ->      ({Ndaac,Ndaad}){元}{年}(IN1 {月}(IN1 ({日,號})));
TDM2    ->      ({Ndaac,Ndaad})IN1 {年}(IN1 {月}(IN1 ({日,號})));
TDM3    ->      ({Ndaac,Ndaad}){元}{年}{元}{月份};
TDM3    ->      ({Ndaac,Ndaad})IN1 {年}{元}{月份};
TDM3    ->      ({Ndaac,Ndaad })){元}{年}IN1{月份};
TDM3    ->      ({Ndaac,Ndaad})IN1 {年}IN1{月份};
TDM4    ->      IN1 {月}(IN1 ({日,號}));
TDM4    ->      {元,正,上,下,每,本}{月}(IN1 ({日,號}));
TDM5    ->      IN1 {日,號};
TDM6    ->      {TDM2,TDM4,TDM5}({Ndabb,Ndabd1,Ndabf})(Ndabe)(TDM1);
TDM7    ->      Ndabd1  (Ndabe)  TDM1 ;
TDM8    ->      Ndabd1  (Ndabe)  (TDM1) ;
TDM9    ->      Ndabe  TDM1 ;
TDM10   ->      {每, 上, 下, 本} ({個}) TDM8 ;
LLP     ->      IN1 {度} (IN1 {分} (IN1 {秒}));
ADP     ->      (IN1 {段})(IN1 {巷})(IN1 {弄})IN1({之} IN1)
                {號}(IN1 {樓}) ;
TMP     ->      {攝氏,華氏} ({零下}) {IN1,DN} {度} ;
DM      ->      {FN1,ONP,NOP1,NOP2,NOP3,NOP4,NOP6,XQP,CNP,DSP,OSP,
                OHSP,DDP1,DDP2,HOSP,STDM,TDM1,TDM2,TDM3,TDM4,TDM5,
                TDM6,TDM7,TDM9,TDM10,LLP,ADP,TMP} ;
```