CLASSIFICATIONS AND COOCCURRENCE RESTRICTIONS IN CHINESE SIMPLE NOUN PHRASES

Li-Li Chang* Juei-Chu Huang* Li-Ping Chang* Wen-Chen Wei*
Ya-hsia Cheng* Keh-jiann Chen** Shih-shyeng Tseng* Ching-chun Hsieh**

* Computing Center, Academia Sinica
** Institute of Information Science, Academia Sinica

ABSTRACT

In this paper, we try to give an answer to the problems about what kinds of information should be given to every component of a Chinese simple noun phrase (NP) in order to parse it. We provide a new subcategorization criteria for the set of nouns , measures and determinatives. The cooccurrence restrictions between subcategories are also specified.

1. INTRODUCTION

Generally speaking, in Mandarin, the word order of the syntactic categories is specified by grammar rules which can be viewed as a sequence of basic categories. In a noun phrase, the basic categories are the determinatives (D), measures (M), and nouns (N). They always follow the same order, i.e. D-M-N. The grammar rule in which only the basic syntactic categories are distinguished does not provide sufficient information of determining a legal noun phrase , since there are cooccurrence restrictions between the constituents in a phrase. Therefore, for each word, we like to establish a limited domain of words which are possible to cooccure with it. Besides, such information will become necessary if we want to parse and generate sentences.

The simple NP in our definition refers to a serial construction.

(1)	DD	ND	M	N
	SD	QD		
eg:	這	一	塊	蛋糕。
	那	兩	間	房子。

where DD denotes demonstrative determinatives
SD denotes specifying determinatives
ND denotes numerical determinatives
QD denotes quantitative determinatives

In order to capture the cooccurrence restriction between each connected pair in a simple NP, we encounter the following two problems.

It is always the case that the classification is not complete and cannot provide sufficient information for the purpose of parsing. Therefore, the first type of problem is how much information can be included in the category and how to patch the remained information. For example, in Mandarin, there are some types of measures which cooccur with the nouns arbitrarily. Such information cannot be captured from any other syntactic or semantic aspects. Hence, under each noun its specific measures must be attached.

The second type of problem is the difficulties in giving a clear-cut classification for the lexicon. The nouns cause the most serious problem in classification because there are many nouns which may be classified into two or two more categories. Such ambiguous classification will cause problems in parsing. We will adapt the concept of 'feature' to distinguish the nouns in order to simplify the complexity in the classification. Under the principle of feature, every noun will be classified into only one category.

We propose our strategies of handling the subclassification of nouns , cooccurrence relations between nouns and measures, and will not discuss them in this paper for Chao[1] has given a clear classification , which is represented in Appendix A.

2. SUBCLASSIFICATION OF NOUNS

In the Indo-European language, such as English, the four basic types of nouns can be distinguished morphosyntactically. As regards to Mandarin, an isolated language, the differences among the four types of nouns are not reflected on their inflections, but partly on their cooccurrence restriction with the measures. The measures in Mandarin are versatile and plenty. Different types of nouns cooccur with different types of measures as shown by Chao [1] in the following:

1. Individual nouns are associated with their specific individual measures (個體量詞).
2. Collective nouns do not take individual measures, but can take temporary measures (暫時量詞) or partitive measures (部份量詞).
3. Abstract nouns are nouns which can only

take certain group measures (群體量詞), measures for verbs (動作動詞的量詞) , and partitive measures, but not individual measures or standard measures (標準量詞).

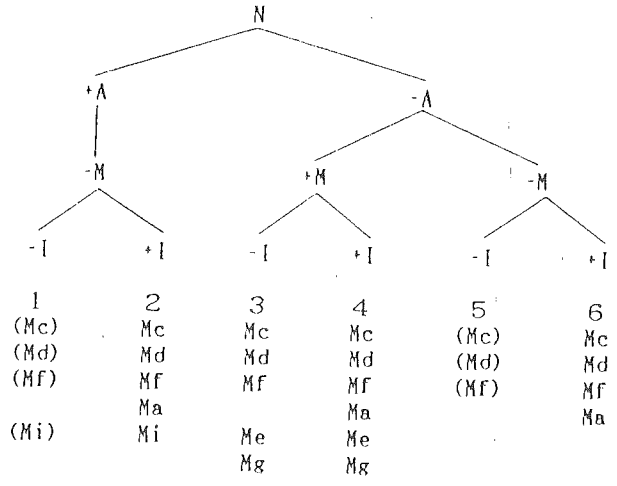
4. Mass nouns do not have specific individual measures, but can be modified by D-M compounds in which M is

a) a standard measure
b) a container measure (容器量詞) or a temporary measure
c) a partitive measure
d) a shape in which the mass can be gathered, i.e. a group measure.

Chao has mentioned that there exists the phenomena of class overlapping on two groups of nouns.

"In a minority of cases, a word may be an individual noun or a mass noun by way of class overlap. For example, 麵包 'bread' is an individual noun in 一個麵包 'a (loaf of) bread', but a mass noun in 一塊或一片麵包 'a piece of bread'。"

The second class overlapping Chao has noted is that "nouns for many abstraction" can also be "grammatically individual nouns". He has given two examples: 兩個學說 'two theories' and 一個夢 'a dream'。



A: abstract
M: mass
I: individual
Mg: standard measures
Mi: measures of action
Ma: individual measures
Mc: group measures
Md: partitive measures
Me: container measures
Mf: temporary measures

fig. 1 a hierarchical classification of the nouns

For the convenience of parsing, it is better to classify a word into only one syntactic category. We decide to further classify the mass nouns and the abstract nouns into two subsets. The first subset contains the ordinary mass or abstract nouns. The second subset contains the nouns of class overlapping. Thus, the second subset of the mass and abstract nouns are also individual nouns, as shown on figure 1. On figure 1, the hierarchical structure of our classification also are represented. Besides, the types of measures which may cooccur with each subtype of nouns also are listed under each subtype.

In our system, the concept of 'feature' is adapted. Only three features are available in the the system, i.e. [+/-abstract], [+/-mass] and [+/-individual]. We may notice that in our system, there are no combination of two features [+A] and [+M], because a mass noun is used to refer to the reality, never be an abstraction. Of course, there are also no combination of [+A], [+M] and [+I].

3. COOCCURRENCE RESTRICTIONS BETWEEN MEASURES AND NOUNS

The cooccurrence restrictions between the measures and nouns are complex. There are at least two types of restrictions which are caused by measures. The first type of restriction could be clearly understood from the semantic aspects. The Standard Measures and the Container Measures belong to this type. Because the liquids have weight and can occupy a certain space, they can cooccure with the Standard Measures of liquids, weight, or capacity, and the container measures which can hold liquids. Since the cooccurrence restriction is determined by the meaning, we will construct a conceptual structure to indicate their relations[6]. For every category of nouns, there will be a corresponding group of measures which may cooccur with it.

The second type of association of measures and nouns is primarily arbitrary and only roughly in terms of meaning. The individual measures, group measures and partitive measures belong to this type. For example, the individual measures for the animals are restricted to 隻、條、頭、尾 and 匹, but for a particular animal, such as 魚 'fish', only 尾 and 條 could be used. Another example, the group measures for the people are 群、批、幫 and 房. For 學生 'student', we can only say 一群/一批學生; for the relatives 親戚, we can only say 一群/一房親戚; and for 流氓 'rascal', we can only say 一群/一批/一幫流氓.

Every noun is cooccured with their specific measures, so that in dictionaries the specific measures must be cited under each individual noun. In our system, the specific measures have to be explicitly stated in each word. However, we don't know how many measures to be cited are enough. For the word 學生 'student', we can have individual measures 個 'an individual', 位 'polite form for ge', 名 'an individual', group measures 種 'species', 類 'category', 標 'sort', 班 'squad', 群 'crowd', 批 'batch', 排 'row', 列 'series', 行 'column', and partitive measures 些 'some', 部份 'part', 半 'half', 堆 'pile', 點 'a few'. Some are used with 學生 frequently, some are used in special occasions, and some are seldom, maybe never, but permitted in meaning. Should we cite all the possible measures of a

noun but waste much time in finding measures that are seldom used, or should we just cite measures often being used but lose some possible assoication? We decide to choose the latter one and use a strategy to make up the lost. Just like the way we handle standard measures, we need a conceptual structure of nouns. After a conceptual structure of nouns has constructed, we built a table collecting measures which are possible to associate with each category of nouns, as shown in Appendix B. Thus we can reduce the domain of measures associated with a noun while parsing. But if we want to generate a new sentence, only the cited measures can easily be applied.

4. SUBCLASSIFICATION ON THE DETERMINATIVES

Most determinatives are bound and monosyllabic morphemes. They can form D-M compounds with measures to modify nouns. Some of them are used to refer a noun, others are used to quantify a noun.

Chao [1] has classified determinatives into four types: demonstrative determinatives (DD), specifying determinatives (SD), numerical determinatives (ND) and quantitative determinatives (QD). DD contains 這, 那 and 哪. SD includes 每、各、別、另、旁、本、某、上、下、前、後、今、明、昨、去. We have added other SD, i.e. 諸、歷、列、本、此、該、當、茲、適、敝、令、賢、什麼、啥、何、頭、次、. ND includes numerals and 幾. However, in our system, 幾 is shifted to QD. We regard 天干 'heaven's stems' and 地支 'earth's branches' as ND. QD are determinatives which do not give exact numbers, but express relative quantities, or in the case of interrogatives, unknown quantities. They are 一、滿、全、整、半、幾、多、多少、許多、好些、好多、好幾、很多. We add 數、若干、幾多 to this group. In Chao[1], all determinatives can be followed by measures. But in these addition, we have changed the scope of determinatives and allowed them to occur before nouns without the association of measures. The reason is that we have found some morphemes which are similar to determinatives semantically but hard to be classified into any other syntactic categories. So we treat them as determinatives.

We find that DD and SD are linearly preceding the ND and QD. But it is not necessary that each determinative in DD and SD may precede each one in ND and QD. We further classify DD, SD, ND and QD into subsets according to their meaning, and indicate the cooccurrence restrictions among them as shown in table 1.

In parsing, for each word, it would be better to establish a limited domain of words which are possible to follow it. This might make the parsing more efficient. Table 1 is used to provide such information about the determinatives. For example, for the first subset of DD, the other determinatives which would follow it are only four subtypes: two in QD and two in ND. For the second subset of SD, it will be immediately followed by the measures, no other determinatives may be inserted.

			指示定詞(DD)		修飾定詞(SD)									
			一般	疑問	列舉		他指		順序		指屬		疑問	
									一般	時間	遠屬	近屬		
			這那	哪	每各任何	諸歷列	另某	別旁	上下前後頭末	今明昨去在	本此該當	適敝今賢	什麼啥何	
數詞(一QD)	全部	一全滿整	✓											
	部分	半數若干幾	✓	✓	✓		✓		✓					
	時間	多少若干幾幾多												
	程度	許多很多好多好幾好些少許												
數詞定詞(一ND)	正整數			✓	✓	✓		✓		✓				
	分數	小數		✓	✓	✓		✓		✓				
	序數													

Table 1: the cooccurrence restrictions between DD, SD and ND, QD. The symble '✓' indicates that they may cooccur together.

5. CONCLUSION

There is a problem which is inevitable to the subcategorization. As we have mentioned that it is necessary to subclassify the syntactic categories according to the cooccurrence restrictions, which are partly due to syntactic restrictions, but mostly due to the semantic restrictions. That means, the criteria of subclassification would be based on the meaning. Because such criteria can't be completely explicit, there would be some words which may cause a few ambiguities. For example, there are some types of nouns which are hard to be classified by the criteria of "abstract", such as

the ceremonies : 典禮、儀示、畫展、暴動、文字獄

the principles : 信條、憲章、規則、校規、條約
the signs : 符數、小數、數目、記號、文字
the plays : 戲劇、布袋戲、京戲

This paper is part of the result of the project of Chinese Word Knowledge Base which is being conducted at Computer Center of Academia Sinica. In this project, the word set is based on 國語日報辭典[2]. Before classify the Ds, Ms and Ns, we have given a detailed study on their cooccurrence restriction and classification structure[3][4][5]. 19,260 '體詞 'substantives' have been managed, and only few cases, about 2%, cause controversy during classification. The data are built not only for parsing, but also for generating. It would be more efficient if a conceptual structure is built.

Appendix A:

Chao has classified measures into nine types, as shown in the following.

a. Individual measures
b. Individual measures associated with V-O
c. Group measures
d. Partitive measures
e. Container measures
f. Temporary measures
g. Standard measures
h. Quasi-measures
i. Measures for verbs of Action

Appendix B:

The coocurrence restrictions between nouns and the individual, group and partitive measures.

0. common measures for every type of Ns:
個、種、類、堆、些、部分、半、點、樣
1. celestial body:
顆、條、道、片、朵、塊、層、重、滴、粒、輪、枚、彎、牙
2. topography:
a. plane: 塊、片、帶、處、畦、襲
b. mountain: 座、重、塊、片、處、帶
c. river, road: 條、道、段、帶、處、彎、股
3. time: 段
4. physical phenomena: 片、層、股、道、束
5. mineral: 塊、粒、顆、層、片、批、排、列
6. animal: 隻、條、對、群、批、窩、胎、頭、尾、匹
7. plant:
common measures for this type:
棵、株、片、批、排、列
a. woody: 枝、條
b. herb: 根、叢、枝
c. flower: 朵、簇、束、把、枝、瓣、襲、畦、棚、抱、叢、捧
d. vegetable: 顆、粒、塊、根、把、畦、襲、札、棚、朵、節、道
e. fruit: 顆、粒、根、片、串、塊
8. food:
頓、餐、客、塊、份、根、條、只、顆、粒、枚、滴、方、味、丸、桌、套、道、泡、批、捲、口
9. costume and accessories:
件、套、條、襲、腰、塊、面、方、匹、疋、根、頂、雙、隻、只、粒、顆、排、列、對、副、串、組、枚、批
10. architecture:
戶、所、棟、幢、間、座、家、片、層、處、帶、排、列、批、片、進、道、扇、塊、口、堵、垛、根、溜
11. furniture:
套、件、堂、張、把、隻、只、組、批
12. living article:
件、條、頂、領、捲、床、塊、面、把、根、支、只、口、雙、枝、襲、張、本、軸、座、批、炷、組、柄
13. stationery:
a. pen and ink stick: 根、枝、管、桿、對、組、錠、塊、方、條、批
b. character, paper, book, picture:
疊、綑、批、排、列、本、冊、部、套、頁、行、卷、期、張、篇、份、幅、封、件、張、捲、團、幅、軸、手、羅、幀
14. machinery:
台、架、部、把、根、支、批、柄、管、組、面、件、張、條、副、具
15. person:
位、名、群、批、幫、起、伙、票、對、組、員、介、窩、排、列、隊、支、任、屆、具、班、路、房、干
16. organ:
條、根、道、雙、隻、副、片、張、對、塊、層、枚、顆、排、把、綹、莖、撮
17. physiology:
把、泡、滴、顆、粒、串、條、團、口
18. game: 場、項、局、盤、屆
19. medical:
場、項、劑、帖、粒、顆、根、支、管、服、貼、滴、片、錠、張、批、味
20. weapon:
部、隻、條、輛、架、艘、枝、桿、根、座、顆、尊、挺、發、列、排、粒、枚、件、把、張、批、門
21. traffic equipment:
部、輛、條、艘、隻、架、只、節、列、乘、頂
22. art activity:
部、場、齣、幕、折、串、輪、系列
23. speech, poem and song:
句、段、席、串、首、支、闋、條、曲
24. abstract:
個、種、類、樣、派、堆、些、部分、點、絲、半、羅、場、項、股、串、層、重、團、樁、件、檔子、宗、門、則、副

References:

1. Chao. Yuen-ren,"A Grammar of Spoken Chinese," University of California Press, Berkeley and Los Angeles, 1968.
2. He. Zung ed. "Guoyu Zrbau Tzdian" Taipei: Guoyu Zrbau She.
3. Liu. Shu-shiang "Shiandai Hanyu Babaitsz" Peking: Shang-wu Yin-shu-guan, 1980.
4. Liu. Shu-shiang,"Hanyu Yufa Fenshi Wenti" Peking: Shang wu yin shu guan, 1979.
5. Tang. Ting-chr "Guoyu Bianshing Yufa Yianjiou" Taipei: Student Book Co., Ltd., 1976.
6. Sowa. J. F. "Conceptual Structure: Information Processing in Mind and Machine" California: Addison-Wesley Publishing Company, 1984.