Criteria for the Classification of Lexical Categories in a Syntax-Oriented Parsing System

Yu-Ling Una Shiu* and Keh-Yih Su**

* **Institute of Linguistics**
National Tsing Hua University
Hsinchu, Taiwan, R.O.C.

** **Department of Electrical Engineering**
National Tsing Hua University
Hsinchu, Taiwan, R.O.C.

ABSTRACT

In Natural Language Processing systems, different classification of lexical categories will lead to different set of rules, and thus different kinds of analyses, therefore the choice of a good category system is very important to the efficiency and to the memory load of the overall parsing system. Unfortunately, although every parsing system has a set of lexical categories, the issues as to whether these category systems are properly chosen, and the factors for evaluating the adequacy of the classification of lexical categories has generally been ignored. Especially, things go worse in research areas, such as Mandarin NLP field, where many fundamental issues are just beginning to be explored, the lack of a good category system apts to obstruct an in-depth research.

In this paper, we propose eight criteria for the classification of lexical categories in a syntax-oriented parsing system. These criteria are **syntax dominance, descriptive power, simplicity, explicitness, mutual exclusion, collective exhaustiveness, applicational efficiency**, and **conventionality**. Each of them is clearly defined and illustrated by Mandarin examples. Furthermore, the tradeoffs among these criteria are also taken into consideration. These criteria and the discussions of tradeoffs will be helpful in serving as a guide for designing and evaluating a category system.

I. Introduction

In Natural Language Processing (NLP) systems, the classification of lexical categories is the fundamental work. NLP first takes a lexical analysis which reads and converts the input into a stream of tokens, which are then analyzed by the parser. So the choice of the inventory of tokens and the assignment of tokens to words, i.e., the classification of lexical categories, are the very jobs of lexical analysis. Since the quality of such classification will usually affect the performance of its following stages (such as syntactic analysis, semantic analysis), a good category system is indispensible to NLP.

Although there are some existing Mandarin lexical category systems, most of them are not proposed for computer parsing (such as [Chao 68], [Lyu 80]). Their adequacy in NLP is quite doubtful. At the mean time, though a variety of parsing algorithms have been proposed, the syntax-oriented parsing algorithm is still a very popular one within the NLP community ([Su 87], [Char 1986]).[1] Because syntax-orientded parsing systems and non-syntax-oriented ones may have different requirements on their lexical categories, even if some category systems are constructed mainly for parsing (such as [CKIP 86, 88]), further examination is still needed to ensure their suitability for syntax-oriented parsing systems. However, it is quite surprising that there are still no objective, explicit, and rigorous criteria available in literatures to judge the quality of a category system. The lack of a set of clearly-defined criteria makes the comparison and examination of different lexical category systems a very hard task, if possible at all.

In response to such demand, we propose eight criteria in this paper to evaluate different category systems within the syntax-oriented parsing framework. Each criterion is clearly defined and illustrated by Mandarin examples. Furthermore, tradeoffs among these criteria are also discussed. These criteria and the discussions of tradeoffs can serve as a helpful guide for analysis and evaluation of a category system.

II. Criteria for the Classification of Lexical Categories

In order to facilitate the analysis and comparison of different category systems within the syntax-oriented parsing framework, eight criteria are set up as follows:

1. Syntax Dominance : *considering only the phenomena of syntactic distribution*

In a syntax-oriented parsing system, words should be classified only according to their syntactic distribution. To make this criterion more clear, we can regard each lexical entry as a set of attribute-value pairs [Gazd 85]. Each pair encodes a piece of linguistically significant information, such as its character(s), its phonological representation, its possible position(s) in the sentence, its semantic meaning(s), its pragmatic function(s), etc.. Then, this criteria in fact says that, the only attributes we should take into consideration at the stage of lexical category classification are those capable of informing us what positions in the sentence a given entry can occupy. The reason is that if we intend to encode every pieces of information by the classification of lexical categories, the resultant consequence will be that we have to assign every word an independent lexical category. Such classification is certainly meaningless and useless. Thus, our consideration must be selective. While, in a syntax-oriented parsing

system, the main purpose of its analysis is to render correct syntactic structures for input sentences. Naturally, under this approach, any non-syntactic information is irrelevant.

Take Mandarin sentential particles as an example.[2] This set of items, such as 'incoative' *le*, 'presuppositional' *de*, 'friendly reminding' *o*, etc., always occur in the sentence-final position. If we recognize Mandarin sentential particles as an independent lexical category, say PART, their characteristic syntactic distribution can be nicely captured by a rule, as shown in (1):

(1) S′ —> S (PART)

However, among these particles, three of them carry interrogative information, namely, 'Yes-No question' *ma*, 'expecting addressee's participation' *ne*, and 'conjectual' *ba*. Although these interrogative particles obviously have a special semantic meaning (or pragmatic function) which can turn declaratives into questions, they do not differ with other sentential particles in their syntactic distribution. Thus, according to the criterion of **Syntax Dominance**, interrogative particles should not form a separate lexical category.

However, one point worth noting here. Systems using a syntax-oriented parsing algorithm are not unable to manage semantic information at all. But they always handle semantic messages by checking semantic attributes in lexical entries rather than by the classification of lexical categories.

2. Descriptive Power : *adequate descriptions of linguistic phenomena provided by a category system*

Generally speaking, a category system which can adequately account for more syntactic phenomena is better than one which can do less. Consider Mandarin sentential particles again. Two of their detailed distributional phenomena are observed. First, in a simple sentence, two sentential particles may co-occur in succession, with *le* and *de* alternating in penultimate position and other sentential particles absolute sentence-final position. Second, in a complex sentence, only *le* and *de* can be attached to an embedded clause, others must have a scope over the whole matrix sentence ([Shiu 88]). Following the above observation, category systems which contain only one sentential particle category like (2) will fail to capture these empirical facts. The best solution they could offer is shown in (3):

(2) PART : *le, de, ma, ne, ba, o,* etc.

(3) a. S″ —> S ′ (PART)

b. S′ —> S (PART)

c. VP —> V (S′)

However, this analysis allows too many ungrammatical sentences, such as below:

*(4) Da-shiung shiantzai bu du-buo **ba le** ?
Da-shiung now not gamble BA LE

*(5) Yi-jing yi-ding huei sheng-chi **o de** !
Yi-jing surely will get angry O DE!

A more powerful Mandarin category system should have two separate categories for sentential particles, as shown below:

(6) a. PART′ : *le, de*

b. PART″ : *ma, ne, ba, o,* etc.

With the two separate lexical categories in (6), we can came up with a more adequate analysis, presented in (7), which can correctly rule out ungrammatical sentences , like (4), (5), and sanction grammatical sentences, such as (8), and (9).

(7) a. S″ —> S′ (PART2)

b. S′ —> S (PART1)

c. VP —> V (S′)

(8) Da-shiung shiantzai bu du-buo **le ba** ?
Da-shiung now not gamble LE BA
' (I suppose) Now, Da-shiung won't go gambling (any more), will he ? '

(9) Yi-jing yi-ding huei sheng-chi **de o** !
Yi-jing surely will get angry DE O !
' (Let me tell you) Surely, Yi-jing will get angry ! '

However, we have to mention that the criterion of **Descriptive Power** must balance with the next criterion **Simplicity**. Their tradeoffs will be discussed in Section III.

3. Simplicity *: using as few categories as possible; avoiding any redundant classification*

This criterion is supported by both computational and linguistic considerations. Computationally, the set of lexical categories is relevant to the set of grammar rules. Usually, an increase in the number of lexical categories implies the simultaneous need of a larger set of grammar rules. In a syntax-oriented parsing system, the parsing table is constructed by expanding the grammar rules, therefore, the growth of rules will certainly lead to the enlargement of the parsing table, and thus the increase of memory load. The number of lexical categories is therefore preferred to be as few as possible. Linguistically, simplicity usually correlates with maximal degree of generalization. Any redundancy will certainly destroy the elegance of an approach. Thus, the most economical solution is usually the best choice.

For illustration, consider Mandarin common nouns and proper names. Many category systems make distinction between them for the reason that only common nouns can be preceded by determiner-measure compounds (D-M compounds). However, this observation is not entirely right. Consider the following counterexamples:

(10) **Na wei Li shiaujie** tzou guo lai le.
That M Li Miss walk along LE
' That woman called Miss Li are walking along. '

(11) Women ban shang you **liang ge Wang-shiau-wu**.
We class LOC have two M Wang-shiau-wu

' There are two persons named Wang-shiau-wu in our class. '

Sentence (10) and (11) show that there is no problem for proper names to co-occur with D-M compounds. Removing this distinction, proper names and common nouns in fact have the same possible syntactic positions. Thus, they should be combined into one single category according to the criterion of **simplicity**.

4. Explicitness : *defining precise scope for each category by using rigorous definitions*

The definition of each lexical category must be rigorous enough to yield desirable classification. Since NLP system is usually a teamwork, the use of vague or ambiguous definitions for lexical categories will result in a lot of mis-labeling which will destroy the consistency of the overall category system.

For example, if nouns are vaguelly defined as 'names of entities', it will be hard to judge whether 'linguistics'*yu yian shiue* or 'dragon'*lung* should be considered as nouns or not, because *yu yian shiue* is not a concrete object, and *lung* does not exist at all. But they indeed share similar syntactic distribution with other nouns. Thus, a more explicit classification of lexical categories should define Mandarin nouns as 'what can be modified by D-M compounds or possessive expressions'. Such definition can then correctly assign the category noun to *lung* and *yu yian shiue* without confusions.

5. Mutual Exclusion : *complementary classification; avoiding overlapping*

If the domains of different categories overlap within a category system, it implies that words in the overlapping areas will be assigned with more than one category labels. Such multi-categoried words are the main source of lexical ambiguities. Except for some homographs which will be inevitable in raising such parsing difficulty, a good lexical category system should avoid this type of ambiguity with mutual exclusive classification.

For instance, some Mandarin category systems follow the traditional classification of English and recognize verbs and adjectives as two separate categories. But, in fact the so-called adjectives can be further divided into two groups (Figure 1), namely, predicative adjectives and non-predicative adjectives. Predicative adjectives, like *piauliang, gau, gaushing,* etc., are very similar to verbs in their syntactic behaviors in that they can form A-NOT-A constructions, be modified by adverbs, act as the main predicates of sentences, etc. (Figure 2). This is exemplified as follows:

(12) a. Yi-jing **lai-bu-lai** ?
Yi-jing come-not-come
' Does Yi-jing come or not come ? '

b. Yi-jing **piauliang-bu-piauliang** ?
Yi-jing pretty-not-pretty
' Is Yi-jing pretty or not pretty ? '

(13) a. Da-shiung **hen shihuan** Yi-jing.
Da-shiung very like Yi-jing
' Da-shiung likes Yi-jing very much. '

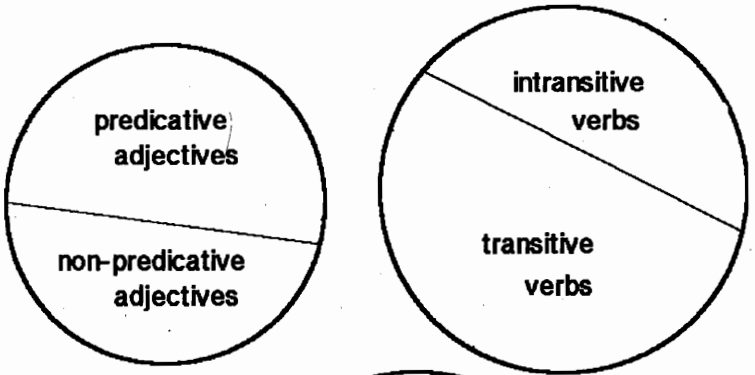


Figure 1

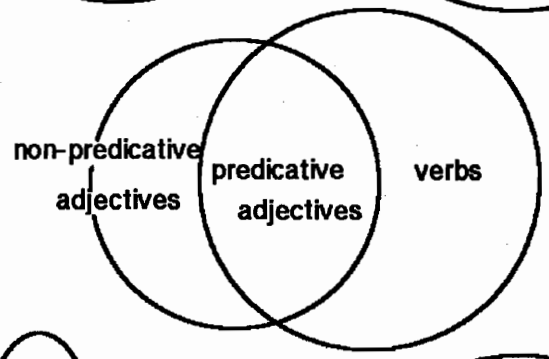


Figure 2

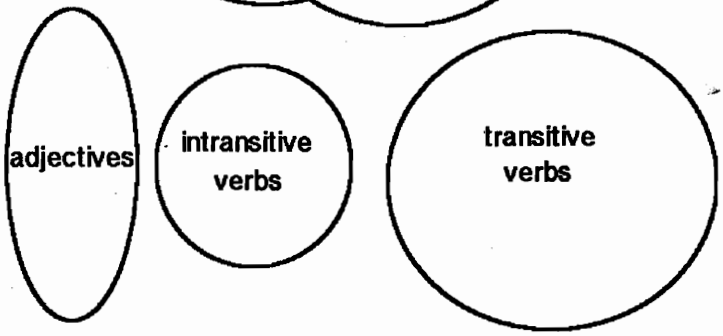


Figure 3

b. Da-shiung **hen gau** .
Da-shiung very tall
' Da-shiung is very tall. '

(14) a. Da-shiung bu **shuohua**
Da-shiung not talk
' Da-shiung does not talk. '

b. Da-shiung bu **gaushing** .
Da-shiung not happy
' Da-shiung is not happy. '

If the above syntactic positions are specified both in the definitions of verbs and adjectives, all predicative adjectives will unfortunately bear two category lables. Such analysis greatly increases the load of a syntax-oriented parsing system.

A better solution is to classify non-predicative adjectives, like 'the most important' *shouyau*, 'chronic' *manshing*, etc. as a single category. Further, since non-predicative

adjectives are usually intransitive, they can combine with intransitive verbs to form another lexical category. Thus, according to the criterion of **Mutural Exclusion**, the plausible classification of verbs and adjectives may be like as shown in Figure 3.

Careful readers may notice that this criterion may also contradict with the criterion of **Simplicity**. Their tradeoffs will also be discussed later.

6. Collective Exhaustiveness *: thorough classification; the union of every classes should be equal to the whole set of data*

This criterion requires that each word must be classified into at least one class. Words without category labels will fail to be accessed by the parser. Thus, even a small set of words, like Mandarin particles, which do not occur in English at all, should not be neglected in a Mandarin category system.

7. Applicational Efficiency *: yielding the most desirable and economical processing with regard to its special application domain*

A sophisticated category system should also take its application domain into consideration. Different application fields may have different requirements for their category systems. For example, in a Machine Translation System (MTS), the categories of a source language and those of a target language may had better carry some kind of correspondance. The MTS with this kind of lexical category classification can deal with its transfer rules better. Under this consideration, a Chinese-to-English MTS may prefer Mandarin predicative adjectives, separated from intransitive verbs, also form an independent category. Since adjectives and verbs really distinct in English, such separation in Mandarin will make the interlanguage corresponse easier to be captured and is helpful in render correct translations.

8. Conventionality *: following established conventions*

If the above criteria are equally satisfied, a category system with more conventional notations, and more standard definitions will be more transparent to the linguists and have better mnemonic quality. For instance, the notations of lexical categories such as noun, verb, adjective ... etc. have been widely-adopted. And most people in linguistics or NLP community have a general idea about their scope of classification. Thus, a catgory system following this way of classification will be conceptually easier and make more ready-made analyses available. This criteria is therefore proposed as the last point.

III. Tradeoffs among the Criteria

So far, we have explained and illustrated the proposed criteria. Ideally, a good category system should closely obey all of them. However, the complicated linguistic phenomena of natural languages and the inherent limitations of computational devices preclude them from coexisting optimally within most category system. Thus, we will briefly discuss the main tradeoffs among them. The considerations are separately represented as follows:

1. Descriptive Power & Simplicity

With the same number of lexical categories, a category system having more descriptive power is superior; with the same descriptive power, a category system having a smaller set of lexical categories is more desirable. But if there are contradictions between these two criteria, the tradeoffs between them should be carefully considered.

For example, there are dependences between Mandarin measures and nouns. Different classes of nouns require different classes of measures, such as plants can be preceded by *ke, ju, tsung*, etc., animals by *jr, chiun, wo,* etc., and furnitures by *tau, tsu, jian,* etc.. If we classify nouns and measures into various categories according to their co-occurrence, the descriptive power of the category system will be increased. But such classification is trivial and will create numerous additional lexical categories which rersults in a plenty of additional rules. For the sake of memory load, this analysis is not welcome. The preferred alternative is to encode such co-occurrence restrictions by using attributes. By virtues of checking attributes, we can still correctly constrain the co-occurrence between measures and nouns, and reach a desirable consequence without using a clumsy set of grammar rules.

Thus, based on the consideration of descriptive power and simplicity, we suggest that only general syntactic phenomena should be handled by the classification of lexical categories.

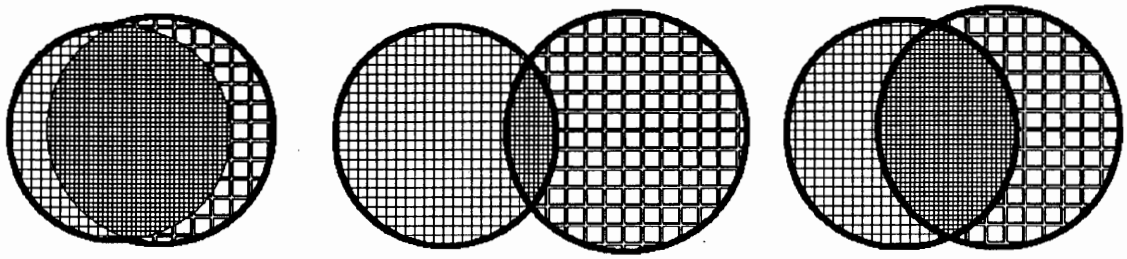
2. Simplicity & Mutual Exclusion

If two classes of words share some possible syntactic positions, we suggest the following solutions :

A If the syntactic distribution of two classes overlaps a lot, these two classes should be combined into one lexical category, and using condition check in grammar rules to specify their distinction (Figure 4. a). For example, some pronouns really cannot be preceded by D-M compounds, such as '(politely) your father' *lin-tzuen*, '(modestly) my son' *shiau-chiuan.* But their other position are just similar to those of nouns. Thus, more efficiently, we should regard them as a single lexical category, and using condition check to prevent pronouns co-occurring with measures.

B If the distribution of two classes of words overlaps just in a few cases, they should be classified into two separate lexical categories, and with the few words appeared in the overlapping positions labeled two category lables (Figure 4. b).

C If the overlapping section of the distribution of two classes of words is approximately equal in size to the two distinct non-intersecting sections, then the most efficient way is to classify all three of them into independent categories (Figure 4. c). This is the method suggesting for handling traditional adjectives and verbs.



a. b. c.

Figure 4 Overlapping between the distribution of two classes of words

3. Simplicity & Applicational Efficiency

If some important applicational advantages can be captured by further classification of lexical categories, the criterion of **Simplicity** has to give way to such special demand. For example, as we have mentioned above, though the adjective category in Mandarin is not theoretically needed and will add the number of lexical categories, however, in a Chinese-to-English MTS, this category is still preferred for the efficiency of overall processing.

4. Descriptive Power & Applicational Efficiency

Generally speaking, a category system with more descriptive power is favored. But if concerning applicational efficiency, some reservation about descriptive power may be made. Some classification of lexical categories may really add to the descriptive power, however, if such information is of little use or even irrelevant to its applicational domain, such classification is just useless and not worthy of implementation.

IV. Conclusion

Before a satisfactory category system can be determined, a good set of criteria for evaluating them should be set up first. This paper proposes eight criteria for the classification of lexical categories in a syntax-oriented parsing system. Each of them is examplified by Mandarin data. Since there are no standard lexical category system in Mandarin NLP as yet, this set of criteria can helpfully serve as a guide for designing a good category system and as a reference for examining existing category systems. Further, some tradeoffs among these criteria are indicated. Since there are still too many controversies in Mandarin syntax, it is beyond our ability to provide a detailed quantification of these tradeoffs. Nevertheless, the general direction is pointed out.

NOTES

1 The distinction between syntax and semantics is not a clear-cut matter, thus we use the term **Syntax-oriented**.

2 The Romanization system adopted in this paper is Mandarin Phonetic Symbols II (MPS II), which is formally announced by the Ministry of Education in 1986.

ACKNOWLEDGEMENTS

We are indebted to Prof. Keh-Jiann Chen for making valuable advices on an earlier draft of this paper, to Prof. Ting-Chi Tang for helpful comments on Mandarin examples, and to Prof. Chu-Ren Huang and Prof. Samuel Wang for useful suggestions. We are also grateful to all the members in BTC R&D Center and CKIP, for their discussions and goodwill. Special thanks are due to Behavior Tech. Computer Corp. (BTC), for her full financial support. Responsibility of errors is, of course, ours.

REFERENCES

[Chao 68] Chao, Yuen-Ren. A Grammar of Spoken Chinese. Berkeley : University of California Press (1968).

[Char 86] Charniak, Eugene & Drew Mcdermott. Introduction to Artifical Intelligence. Addison-Wesley Publishing Company, Inc. (1986).

[CKIP 86] 中文詞知識庫小組, 國語的詞類分析, 技術報告T002, 中研院計算中心：台北南港 (1986).

[CKIP 87] Chang, Li-li, Huang, juei-chu, Chang, Li-ping, Wei, Wen-chen, Cheng, Ya-hsia, Chen, Keh-jiann, Tseng, Shih-shyeng, Hsieh, Ching-chun. "Classification and Co-occurrence Restrictions in Chinese Simple Noun Phrases". The Chinese Language Society. (1987)

[CKIP 88] 中文詞知識庫小組, 國語的詞類分析（修訂版）, 中研院計算中心：台北南港 (1988).

[Gazd 85] Gazdar, Gerald, Ewan Klein, Geoffrey Pullum, an& Ivan Sag. Generalized Phrase Structure Grammar. Oxford : Blackwell. (1985)

[Lyu 80] 呂叔湘, 現代漢語八百詞, 商務印書館香港分館 (1980).

[Lyu 81] 呂叔湘, 試論非謂形容詞, 中國語文, 1981, no. 2, (1981).

[Nire 87] Nireburg, Sergei. ed. Machine Translation : Theoretical and Methodological Issues. Cambridge : Cambridge University Press. (1987)

[Shiu 88] Shiu, Yu-Ling & Chu-Ren Huang. "Unification-based Analysis and Parsing Strategy of Mandarin Particle Questions". To appear in 1988 Proc. of International Computer Symposium, Taipei, Taiwan, Dec. 15–17, 1988. (1988)

[Su 87] Su, Keh-Yih, Jing-Shin Chang, & Hsue-Hueh Hsu. "A Powerful Language Processing System for English-Chinese Machine Translation." 1987 Proc. of Int. Conf. on Chinese and Oriental Language Computing, pp. 260–264, Chicago, June 15–17, 1987. (1987)

[Tang 77] 湯廷池, 動詞與形容詞之間, 華文世界, 1977, 9 : pp. 30-40, also in 國語語法研究論集,

1979, pp. 161-167. 台北：學生書局. (1977)

[Wilk 75] Wilks, Yorick. "An Intelligent Analyzer and Understander of English", Communications of the ACM, Vol 18, no 5, pp. 264–274, May 1975. (1975)