# Categorial Fluidity in Chinese and its Implications for Part-of-speech Tagging

**Oi Yee Kwong** **Benjamin K. Tsou**
Language Information Sciences Research Centre
City University of Hong Kong, Kowloon, Hong Kong
{rlolivia, rlbtsou}@cityu.edu.hk

## Abstract

This paper discusses the theoretical and practical concerns in part-of-speech (POS) tagging for Chinese. Unlike other languages such as English, Chinese lacks morphological marking in association with categorial alternations. We consider such categorial fluidity a continuum, and any categorial shift a transition, with special focus on the verb-noun shift. Preliminary observations are reported on this phenomenon from empirical data, and we suggest that POS tagging should not only be theoretically valid but also sufficiently capture the extent of categorial fluidity as reflected by the data.

## 1 Introduction

There are currently a number of POS-tagged Chinese corpora available, based on different tagsets and theoretical frameworks. Some are more semantic-oriented in determining the syntactic category of a word. For instance, CKIP (1993) had a very fine-grained classification of Chinese verbs based on thematic structures. Others are mainly based on syntactic distribution (e.g. Yu *et al.*, 1998; Xia, 2000) where POS tags are assigned mainly depending on the syntactic properties of the target words.

Given the many-to-many relation between grammatical function and lexical category in Chinese, it is often not straightforward as to how certain words should be tagged in particular sentences. For example, should the verb 懷疑 *huai2yi2* (to suspect) be given the same tag in all cases in (1)[1]?

(1) a. 我懷疑他是賊 (I suspect he is a thief)
*wo3 huai2yi2 ta1 shi4 zei2*
b. 他滿臉懷疑表情 (He wears a suspicious look)
*ta1 man3lian3 huai2yi2 biao3qing2*
c. 這只是我的懷疑 (This is only my suspicion)
*zhe4 zhi3shi4 wo3 de5 huai2yi2*

Most Chinese grammarians would suggest that Chinese words have predefined lexical categories mainly based on their syntactic properties, which should not be contingent upon the grammatical function the words play in particular sentences (e.g. Zhu, 2001). In this study, however, we will show that categorial fluidity (i.e. the relative flexibility of a word being used for different grammatical functions) should be captured in the tagging of large corpora to provide an important resource for the study of this special linguistic phenomenon, as well as for lexicography and natural language processing.

In Section 2, we first discuss POS ambiguity and categorial fluidity in Chinese and the difficulty thus posed on tagging. Then in Section 3, we report on a preliminary empirical study of the categorial fluidity and shift between Chinese verbs and nouns. The implications for POS tagging are explored in Section 4, followed by a conclusion with future work in Section 5.

## 2 POS Ambiguity in Chinese

### 2.1 The Difficulty

The major problem in POS tagging, for all languages alike, is probably that of ambiguity. Where a word has multiple potential POS tags, the ambiguity has to be resolved in context.

The problem of POS ambiguity is especially salient for Chinese, mainly for two reasons. First, categorial change in Chinese words is not often associated with morphological marking.

[1] The digits following the Hanyu pinyin indicate the tone.

Thus the same word form can have more than one syntactic categories, and this difference is not marked by any derivational affixes. For example, the word 領導 *ling3dao3*, can be a verb (to lead) or a noun (leader), as in (2).

(2) a. 領導國家前進 (to lead the country forward)
*ling3dao3 guo2jia1 qian2jin4*
b. 他是我們的領導 (He is our leader)
*ta1 shi4 wo3men2 de5 ling3dao3*

Second, the same Chinese word can have different grammatical functions in individual sentences. There is no one-to-one relationship between grammatical function and syntactic category. As in (3), the word 唱歌 (sing) is the predicate in (a) but is the modifier in (b).

(3)a. 他唱歌 (he sings)
*ta1 chang4ge1*
b. 唱歌技巧 (singing skill)
*chang4ge1 ji4qiao3*

## 2.2 Three Levels of Ambiguity

POS ambiguity is a serious problem for Chinese in a small portion of the Chinese lexicon, but especially among the most frequently used lexical items (Liu, 2000). Moreover, we may think of categorial fluidity as a continuum, from genuine ambiguities to specific cases of "word play". Any categorial shift undergoes a transition, moving from the "word play" end toward the genuine ambiguity end. Thus we distinguish three levels of ambiguity, namely:

- *Regular ambiguity*: a word has multiple POSs which are well accepted and described in any existing lexicon, as in (4).
- *Transitional ambiguity*: a word undergoes a process of categorial shift, where it originally belongs to a particular syntactic category and gradually assumes usage of another category as well. Many prepositions in Chinese are evolved from verbs, as in (5). Categorial shift from verb to noun is also common, as discussed in the next section.
- *Innovative ambiguity*: sometimes words are deliberately used in peculiar ways to create a special effect, as in (6). Such individual cases cannot be regarded as genuine ambiguity, until the special use becomes common enough.

(4)a. 一張枱布 (a piece ("spread") [of] tablecloth)
*yi1 zhang1 tai3bu4*
b. 飯來張口 (rice ready, open mouth)
*fan4 lai2 zhang1 kou3*
c. 張先生 (Mr. Zhang)
*zhang1 xian1sheng1*

(5)a. 陽光透過窗戶 (sunlight passes through window)
*yang2guang1 tou4guo4 chuang1hu4*
b. 透過討論找答案 (through discussion find answer)
*tou4guo4 tao3lun4 zhao3 da2an4*

(6) 他很小丑 (He [is] very clown[ish])
*ta1 hen3 xiao3chou3*

As said, when innovative uses become frequent enough, at some point they might be considered, or at least treated as, genuine ambiguities. Nevertheless, it is the process of transition that is of linguistic interest, and the usage frequencies from corpus data gathered over time would give good evidence for it. As a result, when tagging a corpus, the different uses of the same word must be sufficiently represented to enable this kind of longitudinal empirical study to be carried out. In the following section, we will first discuss a preliminary study along this line, and then in Section 4, we will further discuss the implications on the requirements of POS tagging.

# 3 Categorial Fluidity and Shift

Very often when a verb is nominalised, it loses part of a verb's syntactic features. For instance, a nominalised transitive verb cannot take an object in the normal VO order any more, as in (7).

(7)a. 產生影響 (produce influence)
*chan3sheng1 ying3xiang3*
*b. 產生影響別人 (produce influence others)
*chan3sheng1 ying3xiang3 bie2ren2*

At the same time, it gains some syntactic features of nouns, such as modification by adjectives or numbers, as in (8).

(8)a. 一項宣佈 (one announcement)
*yi1xiang4 xuan1bu4*
b. 一項重要的宣佈 (one important announcement)
*yi1xiang4 zhong4yao4 de5 xuan1bu4*

The asymmetry between nominalisation and verbalisation in Chinese was observed in Tai (1997).

In other words, verbs are more freely deverbalised than nouns denominalised. This fluidity between verbal and nominal status of verbs can in theory be generalised to many, if not all, verbs. Nevertheless, does such categorial shift happen to all verbs? What words undergo such shift and how fast? At what point is the shift significant enough to become genuine ambiguity? These questions should be best answered by empirical data.

### 3.1 A Preliminary Corpus-Based Study

As a preliminary study, we tagged a small subset of the LIVAC corpus[2] (Tsou *et al.*, 2000) with an earlier version of the tagset to be described in Section 4.2. This subcorpus consists of newspaper texts from Hong Kong, with about 520K word tokens and about 32K word types (excluding numbers). There are around 6.6K verbs, excluding auxiliary and copula verbs. Out of these, 672 were found also tagged nouns or nominalised verbs. In other words, about 10% of all verbs in the subcorpus were playing a role somewhere in the verb-noun categorial shift.

A simple ratio (Eq.1) was computed for all the 672 verbs to give a picture of the categorial shift as reflected by the preliminarily tagged data.

$$r = log\ (verb\ uses\ /\ noun\ uses) \qquad \text{(Eq.1)}$$

The *log* ratio was used to give a linear scale. If verb usage outnumbers noun usage to a certain extent, i.e. when $r >> 0$, it suggests that the word is originally a verb and has just started to shift. If verb usage and noun usage are more or less equal, i.e. when $r \approx 0$, then either the shift is mature enough or there is genuine ambiguity. Meanwhile, if noun usage outnumbers verb usage by a lot, it would mean that either the verb has over shifted or the word is originally a noun and is occasionally denominalised (i.e. beginning to shift).

Preliminary analysis of the data shows that about 58% of the words have $r \geq 0.3$, that is, verb usage at least doubles noun usage for these words. About 23% have $r$ between –0.3 and 0.3. These are the words of which verb usage and noun usage are quite balanced. The third group, with $r \leq -0.3$, occupies about 17%. These words are more used as nouns than verbs, at least twice as much. A sample for each group with examples of verb and noun usages is shown in (9), (10), and (11) respectively. The above figures again reflect the asymmetry between deverbalisation of verbs and denominalisation of nouns.

(9) $r \geq 0.3$ e.g. 發現 (to discover / discovery)
a. 被途人發現 (discovered by passers-by)
*bei1 tu2ren2 fa1xian4*
b. 獲得重大發現 (have great discovery)
*huo4de2 zhong4da4 fa1xian4*

(10) $-0.3 \leq r \leq 0.3$ e.g. 服務 (to serve / service)
a. 服務公眾 (to serve the public)
*fu2wu4 gong1zhong4*
b. 電話輔導服務 (telephone counselling service)
*dian4hua4 fu3dao3 fu2wu4*

(11) $r \leq -0.3$ e.g. 關係 (to relate to / relation)
a. 一些關係中國政經發展大局的問題
(problems which relate to the political and economic development of China)
*yi1xie1 guan1xi4 zhong1guo2 zheng4jing1 fa1zhan3 da4ju2 de5 wen4ti2*
b. 友好合作關係 (friendly cooperative relation)
*you3hao3 he2zuo4 guan1xi4*

## 4 Implications for POS Tagging

Chinese POS tagging can so far be grouped into two approaches. One holds that words have predefined POSs independent of sentential contexts. So as long as the form and the meaning do not change, the POS does not change. Another approach is of the view that grammatical function within a particular sentence determines POSs. So a word is sometimes tagged as verb and sometimes as noun depending on its syntactic distribution in any given sentence. The debate between the two approaches has never been settled, and in this section we will discuss if there can be a compromised approach by considering the practical and theoretical concerns of POS tagging.

### 4.1 Practical Concerns

There are at least two practical concerns in Chinese POS tagging. First, we want to obtain information on the distribution of word uses. It may defeat the purpose of corpus tagging if every word is tagged independent of sentential contexts. Second, we need to distinguish between different syntactic structures, as in (12).

[2] http://www.rcl.cityu.edu.hk/livac

(12) 紅燒排骨
*hong2shao1 pai2gu3*
a. V-O: to roast spare ribs
b. Mod-Head: roasted spare ribs

If (12) is always tagged as a verb followed by a noun, the two different readings cannot be distinguished. Hence apparently it would be preferable to assign different tags to a word when it is used in different syntactic distribution. However, this introduces a crisis for the whole classification of Chinese syntactic categories because any given word cannot be prescribed a category and any category cannot be adequately described. Moreover, it upsets the compactness and the organisation of the lexicon.

## 4.2 The Tagset

Hence even if the ambiguities are still in the transitional stage and should not be kept in the lexicon, it is still preferable to have them captured and reflected in a tagged corpus.

Xia's (2000) approach does not accommodate such transitional usages, but assigns a verb tag if the word is used as a verb and a noun tag if it is used as a noun. CKIP (1993) assigns the same tag to the same word, but uses features to encode specific grammatical functions, such as using [+nom] to indicate a nominalised usage of a verb.

For the current study, we reckon that the categorial fluidity information should be captured as early as the tagging begins. So we attempted to resolve the conflict between theoretical and practical concerns via the design of the tagset. In that case the final lexicon would be theoretically valid on the one hand, and the tagged corpus would be of practical use for NLP tasks (e.g. parsing) on the other. In our approach, each tag consists of a letter code for the general classification (i.e. noun, verb, etc.) of the word, and another for the sub-classification according to the particular context. For example, when a verb is used as the subject on its own (i.e. no modification, not in a phrase, etc.), we tag it as Vx. Thus on the one hand we still recognise it as a verb, but on the other hand we can distinguish it from its normal form of usage. Moreover, theoretically we do not treat it as nominalised but if for any practical reason someone might want it this way, he or she can easily extract all the Vx-tagged words and mark them nouns.

# 5 Future Work and Conclusion

At present our tagset covers 14 general lexical categories and altogether 43 small categories (sub-classification). Both the tagset and the operational guidelines have resulted from continuous revision based on our experience of actually tagging the corpus and observation of the categorial fluidity phenomenon.

The tagging task is ongoing with the latest revised tagset and guidelines to produce a clean and accurately tagged training corpus to be used for the automatic tagging of the remaining corpus. The long-term goal is to produce a very large tagged corpus for use in lexicography and other natural language processing tasks. Focus will also be on a more detailed synchronous and longitudinal study of the verb-noun and other categorial shifts, with data from different regions over time.

This paper has thus discussed POS ambiguity on Chinese, with a focus on the categorial shift from verbs to nouns, and the implications this phenomenon might have for POS tagging.

# References


Chinese Knowledge Information Processing Group (CKIP). 1993. 中文詞類分析（三版） Technical Report no.93-05, Academia Sinica, Taiwan.

Liu, K. 2000. *Zhongwen Wenben Zidong Fenci He Biaozhu.* Beijing: Commercial Press.

Tai, J.H-Y. 1997. Category Shifts and Word-Formation Redundancy Rules in Chinese. 《中國境內語言暨語言學》, 3: 435-468.

Tsou, B.K., Tsoi, W.F., Lai, T.B.Y., Hu, J. and Chan, S.W.K. 2000. LIVAC, A Chinese Synchronous Corpus, and Some Applications. In *Proceedings of the ICCLC International Conference on Chinese Language Computing,* Chicago, pages 233-238.

Xia, F. 2000. *The Part-Of-Speech Tagging Guidelines for the Penn Chinese Treebank (3.0).* IRCS Technical Report IRCS-00-07. Institute for Research in Cognitive Science, University of Pennsylvania.

Yu, S., Zhu, X. and Li, F. 1998. Representation of Grammatical Knowledge in the Chinese Lexicon. In B.K. T'sou, T.B.Y. Lai, S.W.K. Chan and W.S-Y. Wang (Eds.), *Quantitative and Computational Studies on the Chinese Language*, pages 353-372.

Zhu, D. 2001. *Xiandai Hanyu Yufa Yanjiu.* Beijing: Commercial Press.