Mandarin Chinese Words and Parts of Speech

A Corpus-based Study

Chu-Ren Huang, Shu-Kai Hsieh and Keh-Jiann Chen

First published 2017

ISBN: 978-1-138-94944-7 (hbk)
ISBN: 978-0-367-59837-2 (pbk)
ISBN: 978-1-315-66901-4 (ebk)

DOI: 10.4324/9781315669014-9

6 Introduction to CKIP Parts of Speech System

Funder: The Hong Kong Polytechnic University

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

6 Introduction to CKIP Parts of Speech System

This part is an extension of CKIP (Chinese Knowledge Information Processing) group's POS Analysis of Contemporary Chinese ("CKIP_POS") originally published in 1986. Following three years of extensive study and analysis of over 40,000 lemmas in the Mandarin Daily Dictionary (*GuoYuRibao*) a revised version of CKIP_POS was published in 1989. The next step was the development and introduction of the Information-based Case Grammar by CKIP for natural language parsing (ICG; Chen and Huang 1990). Based on ICG, we built a Chinese electronic lexicon that consisted of approximately 80,000 lemmas, each with their lexical classes, phonetic annotations, frequencies, semantic classification, etc.

This publication addresses the need to provide users and the public with a full account of our POS classification framework and its criteria, a keenly felt need since the release of the full electronic lexicon to the public in 1992. However, unlike CKIP_POS or its revision, the book features not only traditional explanations, definitions and examples for each headword, but also additional criteria for classification in each lexical entry, especially those for borderline cases. Moreover, the example sentences in this book are extracted primarily from the Sinica Corpus (Chen et al. 1996) and represent authentic language use.

Prior to a detailed discussion of PoS classifications, Section 6.1 will give an introductory account of the lexical entries as well as the tagset in our lexicon; Section 1.2, introduces the syntactic features.

6.1 Word and its POS Tag in the CKIP Lexicon

The entries in the CKIP lexicon include not only words,[1] but also sub-lexical units smaller than words, as well as phrases and idioms. However, compounds that are highly productive or that can be derived based on grammatical rules are not included in the lexicon. For instance, the *Determiner-Classifier compounds* and *Replicative words* can be treated compositionally with morphological rules, hence they are not included, POS information is assigned to each lexical entry.[2] For instance, tag A is attached to non-predicative adjectives (see Chapter 8); tag I is attached to interjections (see Chapter 14). General tagging principles will be discussed after Chapter 6. In the following section, only the criteria for specific tag assignments are introduced.

DOI: 10.4324/9781315669014-9

6.1.1 Annotation Guidelines for Bound Morphemes

In addition to free lexical entries, there are numerous bound components (morphemes) that have to co-occur with other independent components. Currently, we mark these components with a *b* feature. E.g. 木 'wood' in 樹木 'tree' and 木 材 'wood'; 述 'to state' in 描述 'describe' and 述 職 'duty report'; 式 'style' in 程式 'program' and 式 子 'formulae'. Some di-syllabic bound components can also be found, such as 無度 'endless' in 揮霍無 度 'endless squander' and 需索無 度 'endless demand'; or 不力 'without conviction' 工作不 力 'work without conviction' or 執行不 力 'execute without conviction'. It is worth noting that the same word form in a lexical entry can function both as a word with a POS tag and an bound morpheme tagged with a feature. For example:

聲 'sound'

1 (Nfi) Classifier that modifies the action verbs. For example, 叫一聲 'called once', 喊一聲 'yelled once'.
2 (b) For example, 鋼琴聲 'sound of piano', 鈴聲 'sound of bell', 鼓聲 'sound of drum', 撞擊聲 'sound of impact'.

聲 is tagged as Classifier (Nfi) in 1, and as productive suffix in 2. Without tagging with *b*, the automatic POS tagger will yield incorrect results.

6.1.2 Annotation Guidelines for Sentences

In addition to words, there exist 12 'sentences' in the CKIP Lexicon. Most of these sentences are fixed expressions. Syntactically, they are well-structured with Subject, Verb and Object; semantically, they are independent units requiring no argument participation (e.g. 家醜不可外揚 'Don't air your dirty laundry in public', 家書抵萬金 'A letter from home is as good as gold', 家家有本難唸的經 'Each family has its own difficulties'). Currently, the tag S is attached to them, indicating their sentential attribute and distinguishing them from other POS taggers. However, not every phrase or idiom is tagged with S. If these phrases and idioms can be used as verb, then priority will be given to stative intransitive verb (VH11), such as 米珠薪桂 'price is skyrocketing; literally "rice priced like pearl, and firewood priced like cinnamon"', 林木參天 'tall woods reaching the sky', as seen in the following sentences.

1 台北的物價米珠薪桂 'Taipei's product prices are skyrocketing.'
2 三貂嶺一帶林木參天 'The area around San Tiago is verdant with tall woods reaching the sky.'

6.1.3 Annotation Guidelines for Determiner–Measure Compounds

Since Determiner and Classifier/Measure can productively form a DM compound in a compositional manner, which renders exhaustive enumeration impossible, we have proposed a set of morphological rules to deal with DM compounds. As a result, DM compounds (e.g. 一本, 整輛) are not included in the lexicon, except when they also carry other semantic functions, as shown in the following:

這樣 - Dh (方式副詞), Nfc (定量式) 'this way, such' [3]
三重 - Nca (地方名詞), Nfd (定量式) 'SanChung (a town in New Taipei City)'
千萬 - Dbab (評價副詞), Nfzz (定量式) 'At any rate'.

6.1.4 Annotation Guidelines for Reduplicated Words

Reduplicated words (e.g. 快快樂樂 (from 快樂 'happy') , 打打 from 打 'hit'), like DM compounds, can be derived using morphological rules; hence, they are not included in the lexicon. Two exceptions are as follows:

- Words with reduplicated form yet without corresponding morphological rules of derivation. E.g. Nouns with replicated forms 風風雨雨 'winds and rains; turbulent times',[4] 事事物物 'things and objects, all things big and small'; Adverb with replicated forms 常常 'often'; Words in reduplicated form both without corresponding un-reduplicated roots: 鬧鬧攘攘 'busy and bustling', 漸漸 'gradually', 家家戶戶 'each and every family'.[5]
- Words with reduplicated forms yet their syntactic/semantic behavior cannot be predicted of morphological rules. These words will be listed separately in the lexicon and assigned with multiple POS tags. E.g. 乖乖 'guai guai'

1 Naa (Material nouns, FOOD, a brand of rice snacks)
2 VH11 (Stative intransitive verbs, replicative form of 乖)

6.1.5 Annotation Guidelines for Verb–Complement Compounds

The definition of Verb-Complement (VC) Compounds in Chinese remains a topic of ongoing research. Basically, a VC compound consists of at least two predicative morphemes. The verb component usually refers to action, while the complement component refers to results or directions (e.g. 打開 'to open',, 跑過來 'to run over here'). For all of its straightforward structure, a VC's syntactic and semantic properties are complicated. Syntactically, some VC compounds are composed of two intransitive verbs, but the two intransitives combine into a transitive when compounded (e.g. 「走」'walk' +「破」'break' → 走破 'wear to broken by walking' (他走破一雙鞋) 'He walked through a pair of shoes'). Other VC compounds comprise one transitive and one intransitive, but the two verbs combine into an intransitive when compounded e.g. 「灌」'pour' +「醉」

'drunk' -> 灌醉 'to cause (someone) to be drunk by forcing him/her to drink' (張三灌醉了李四) 'Zhangsan made Lisi drunk by making him drink excessively'. Although the meaning of most VC compounds can usually be inferred from their composing parts, such as 吃飯 'eat+rice, to eat a meal', 喝醉 'drink+drunk, to get drunk'; some cannot, such as (116) and (117), where in (116) 看 'look'+ 來 'come' the meaning derivation is not transparent; and 容 'contain'+ 下 'down' meaning derivation is not transparent in (117).

(116) 他 看起來 很好
tā kànqïlai hénhao
'S/He looks great.'

(117) 車子 可 容下 二人
chēzi kě róngxià èrrén
'The car can accommodate two people.'

Given the productivity and the non-compositionality of the VC compounds, their sub-classes cannot be exhaustively listed in the lexicon. Currently, we do not assign them to any specific subclasses. They are assigned with the VR feature instead. We also allow a VC compound to carry the semantic features of its two components: verbs and complements, from which the syntactic and semantic features of the VC can be inferred. For instance, we found that verbs carrying movement features can be combined with complements carrying direction features. This is cognitively sound in that the motion implies the moving direction. Therefore, we have a morphological rule V[+movement]+R[+direction]→VR, which correctly predicts that action verbs such as 「跑」[+movement] can combine with directional complements 「上」[+upward],「下」[+downward], 「進」[+inward], 「出」[+outward] to form grammatical VR compounds such as 「跑上」「跑下」「跑進」「跑出」, whereby the compositional meaning can be inferred as well. We plan to study how to predict a VC's syntactic behavior based on these features in the future.

6.2 POS Annotation

There are eight major POS classes in the CKIP Lexicon—verbs, non-predicative adjectives, nouns, adverbs, prepositions, connectives, particles, and interjections. Other than non-predicative adjectives and interjections, all the POS classes are further divided into sub-classes based on their semantic and syntactic behaviors (see Appendix). For instance, nouns are classified into material nouns, individual nouns, individual abstract nouns, abstract nouns, and collective nouns. Verbs are first classified into action and stative verbs, then further into subclasses such as intransitive verbs, quasi-transitive verbs, di-transitive verbs, sentential object verbs, verb-phrase object verbs, etc. More details will be addressed in the following sections.

Due to the lack of morphological markers, the difficulty we have often encountered when analyzing the data is that the same lemma can play different syntactic roles. For example, a verb can often serve as the main verb but also the noun-modifier, as in 評估報告 'to evaluate a report' or 'an evaluation report', 漂亮的女孩 'beautiful girl'; many can occur in the position of nouns (so-called nominalized verbs) as in (118). Similarly, nouns can function as a modifier, such as 蘋果臉 'apple-face', and also act as a predicative , as in (119).

(118) 他 的 調查 顯示出 不同 的 結果
tā de diàochá xiǎnshìchū bùtóng de jiéguǒ
he DE investigate reveal out different DE result
'His investigation showed a different result.'

(119) 她 很 寶貝 她 的 頭髮
tā hěn bǎobèi tā de tóufǎ
she very treasure/baby her DE hair
'She treasures/pampers her hair.'

The distinction is also made between a word with polyfunctionality and a word where multiple syntactic categories are assigned. These two are treated differently in our framework based on observations of their actual use in large corpora. Syntactic features are attached to the former, which will be illustrated in detail in section 6.2.1; section 6.2.2 will address the conditions under which a word will be analyzed in terms of multiple syntactic categories.

6.2.1 Polyfunctionality of Words

Some of the syntactic categories in Chinese serve polyfunctionally in various contexts, but are consistent in syntactic behaviors; therefore, certain syntactic features are given instead of different POS taggers, in the hope of facilitating the parsing task in natural language processing. Four constructions are discussed as follows:

Firstly, most nominals and verbs in Chinese can serve as modifiers; however, we do not assign them a multiple POS function, but specify the syntactic information within the representation model of nominals in Information-based ICG Grammar (ICG, Chen and Huang 1990), as shown in (X):

Secondly, a large number of simple verbs and verbs followed by DE (的) or DI (地), stative verbs in particular, can serve as the manner-adverb of the main verb in a sentence. For instance, 「賣力」'dedicate' is the main verb in 「他很賣力」 'S/He is dedicated' ; while in 「他很賣力的工作」'S/He works dedicately', it serves as the modifier of the main verb 「工作」'work'; another example is 「感動地」'emotionally' in 「他感動地掉下眼淚」'S/He cried emotionally'.

In the above cases, we do not assign different POS tags to the word, but annotate it with features such as `+way` or `+de`, etc (Wei 1991). One can use these features in the parser when deciding between 'main verb' or 'modifier',

or during the analysis of Non-Predicative adjective (A) or determiner-measure (DM) compounds, as shown in (X) - (X).

(120) 他們 非法 入境 (A)+de
tāmen fēifǎ rùjing
'They entered the country illegally.'

(121) 他 一個一個 數 (DC)+way
tā yīgèyīgè shǔ
'He counts one by one.'

Thirdly, the time nouns in Chinese often serve as temporal modifiers (Chang, 1988), such as in (X). In many English dictionaries, words like 'tomorrow' have two syntactic classes: noun and adverb. In our framework, only the nominal tag is assigned to '明天 (tomorrow)'. Although nouns and adverbs differ considerably in occurring positions and syntactic functions, time nouns in Chinese often form a larger temporal unit with temporal noun phrases to modify the whole sentence. The information of nouns carrying temporal features will be submitted to the parser so as to identify the role of modification without needing to assign multiple POS tags.

(122) 他 明天 不 來
tā míngtiān bù lái
'He won't come tomorrow.'

Fourthly, verb are often nominalized in Chinese. Chinese verbs frequently serve as nominals, sharing their syntactic properties when modified by DM compounds (Tang 1989) (e.g. 他主張完成那二項研究 'S/He argues that the two research (projects) should be completed'). Though the verb is nominalized in this case (Yeh, et al. 1992), we annotate it with syntactic features rather than with different PoS, both for reducing the complexity during automatic PoS assignment and for a deeper grasp of the intriguing interactions between verbs and nominals. It is important to note that in nominalization, although the syntactic behavior has changed, the argument structure is preserved. For instance, nominalized verbs still inherit the original argument structure, such as「認同」'to identify with' has two arguments (THEME and GOAL), and these two arguments are retained in nominalization, and differ only in their realization forms, as illustrated by a and b (Yeh et al. 1992).[6]

Figure 6.1 Shared argument mapping of deverbal nouns

6.2.2 Multiple Syntactic Classification of Words

The following guidelines are proposed as the conditions under which multiple syntactic classes will be assigned to words.

1 Homonyms or homographs
Multiple assignments will be applied to homonyms, that is, words coincidentally sharing the same form while having different senses.

For instance, the word form「重」has distinct senses that fit into different syntactic classes (‘again’:adverb/‘heavy’:stative verb); other examples include「會」(‘meeting’ noun; ‘will’ verb),「只要」(‘only if’ conjunction; ‘only’ adverb), etc. Please note again that polysemous words will not be assigned with multiple POS tags.

2 Common nouns lose their referring function, then acquire their verbal characteristics.

For example,「油」(‘oil’ vs. ‘greasy’, (123)),「火」(‘fire’ vs. ‘mad’, (124)) and「寶貝」(‘baby’ vs. ‘pamper’, (125)) are assigned with stative intransitive and stative transitive tags, as can be seen in 6.2.2.

(123) 這 種 菜 很 油
zhè zhǒng cài hěn yóu
‘This type of dish is quite greasy.’

(124) 王 老師 很 火
wáng lǎoshī hěn huǒ
‘Teacher Wang was very angry.’

(125) 陳 小姐 很 寶貝 她 的 頭髮
chén xiǎojie hěn bǎobèi tā de tóufǎ
‘Ms. Chen pampers her hair.’

3 Word forms with clear distinctions in both syntactic function and semantic content.

For example,「結果」in (126a) and (126b) are tagged with a sentential adverb ‘in consequence’ and a common noun ‘result’, respectively;「不過」in (127a) and (127b) have different meanings while functioning as the adverb ‘just’ and conjunction ‘although’. Hence, these words will be assigned with multiple POS tags, due mainly to their specific behaviors.

(126) (a) 結果 他 什麼 也 不 說
jiēguǒ tā shénme yě bù shuō
‘In consequence, he said nothing.’

(b) 他 知道 結果 了
tā zhīdào jiēguǒ le
‘He knew the result.’

(127) (a) 他 不過 吃 你 一口 蘋果
tā búguò chī nǐ yīkǒu píngguǒ
'He just had a bite of your apple.'

(b) 不過, 他 還沒 滿 二十歲
búguò, tā huánméi mǎn èrshísuì
'Although, he is not yet 20.'

Following this guideline, a word form is assigned with up to four PoS tags in our lexicon. For instance, the wordform「點」is assigned with four tags: abstract nouns (一個點) 'one dot'; action verbs with single object (點 菜) 'to order dishes'; proximate classifiers (一點 意見) 'some opinion'; temporal classifier (下午三點) 'three pm', each with distinctive syntactic behaviors.

Notes

1 Regarding the definition of Chinese wordhood, please refer to Tang (1989: 9).
2 For ease of reference, a concise description of a Part-of-Speech tag set can be found in Appendix I.
3 For DM compounds, only the heads (i.e. classifiers) are POS-tagged. Nfc, Nfd and Nfzz thus stand for different types of classifiers. Please refer to section 9.1.6.
4 Unlike verbs and adjectives, there are no regular noun reduplications in Chinese and reduplicated forms are lexically determined.
5 Note that there are no corresponding free word forms (i.e.「鬧攘」,「漸」「家戶」in Chinese lexicon).
6 Currently, we use the [+argument] feature to mark the verbs that take the arguments after being nominalized (Yeh et al. 1992). This practice is under review. This feature has been discarded in favor of the [+NV] feature.