A Corpus-based Case Study on the POS Tagging of Self-referential Lexemes in the Contemporary Chinese Dictionary

Jun Zhang
Zhaoqing University, Zhaoqing, China
Heng Zhang
Nanchang Normal University, Nanchang, China

Abstract—The POS tagging in the 5th edition of the CCD has been revised in the 6th and the 7th editions. The noun POS of most sports and science lexemes are deleted, and their senses of noun (self-referential senses) are included into verbs. However, most of these lexemes can be used as nouns intuitively, and their noun POS and senses should exist. Based on the grammatical functions of words (Xv & Tang, 2006) and the two-level word class categorization theory (Wang, 2014), this study conducts a corpus-based case study of a science lexeme “guina”. The result shows that “guina” not only has self-referential usage, but has high token frequency, with 133 occurrences accounting for 42.8% of the total usages, and rich type frequency widely distributed in “guina + (的) + NP”, “NP + (的) + guina”, “VP + guina”, “guina + VP” and “PP + guina”, which conforms to the criterion of conventionalization. Therefore, it is necessary to tag the noun POS and to set up the self-referential sense for “guina”. This research has an implication for solving the POS tagging problem of self-referential lexemes in the CCD.

Index Terms—part-of-speech, the Contemporary Chinese Dictionary, corpus, self-referential lexemes

I. INTRODUCTION

The Contemporary Chinese Dictionary (CCD) is an original normative dictionary of modern Chinese, which is compiled by the Institute of Linguistics of the Chinese Academy of Social Sciences and published by the Commercial Press. It is one of the most important reference works for learning Chinese and an important blueprint dictionary for compiling Chinese learner’s dictionaries and Chinese-Foreign Language bilingual dictionaries (Zhang, 2010a, 2010b; Hu, 2013, 2014). Its authoritativeness and scientificity are second to none, and it has the reputation of milestone in the history of Chinese dictionary making (Cao & Wu, 2002). Since the 1st edition was officially published in 1978, the CCD has been published to the 7th edition in 2016.

Part-of-speech (POS) tagging is of great significance in language teaching, bilingual dictionary making and natural language processing. The 1st edition of the CCD only showed the POS information for most functional words, common pronouns and quantifiers in definitions. The 3rd edition began to tag POS for disyllabic and polysyllabic Chinese characters. A comprehensive POS tagging was not achieved until the 5th edition (Xv & Tan, 2006; Jiang, 2013). However, due to the complexity and flexibility of the sentence structures of modern Chinese and the lack of corresponding morphological changes or marks when the same word appears in different syntactic positions, the problem of POS tagging in modern Chinese has not been solved well and has been perplexing modern Chinese grammarians and dictionary compilers for several decades. Although the 5th edition of the CCD has achieved comprehensive POS tagging, the accuracy of the POS tagging has been constantly questioned (e.g. Wang, 2009, 2010, 2013; Hou, 2017; Yang, 2019).

In view of this, the editorial board of the CCD has made two revisions on the basis of the 5th edition, and published the 6th edition and the 7th edition respectively. In order to ensure the consistency of the POS tagging within the dictionary as much as possible, the noun POS of most sports and science lexemes such as “kualan” (跨栏), “huabing” (滑冰), “sheji” (射击), “yanyi” (演绎), “guina” (归纳) and “shijian” (实践) are systematically deleted and the senses of which are included into verbs. Here we take “guina” as an example:

【归纳】

归纳并使有条理（多用于抽象事物）：大家提出的意见，～起来主要就是这三点。

一种推理方法，由一系列具体的事实概括出一般原理（跟“演绎”相反）。（5th edition of the CCD）
The grammatical function criterion of POS classification has been clearly established and adopted by both modern Chinese grammarians and the CCD compilers, but due to the incomplete understanding of the nature of POS and the failure to clarify the relationship between individual words (word tokens) at the parole level and vocabulary words (word type) at the language level, there is no consensus on at which level words should be tagged with POS in a dictionary. According to the two levels that words exist and the linguistic view of the complex adaptive system, Wang (2014) puts forward the Two-level Word Class Categorization Theory, which has been perfected and applied in a series of studies (e.g. Wang & Huang, 2017; Wang & Yang, 2017; Wang, Huo & Deng, 2019).

The theory holds that the categorization of word classes occurs at two levels: 1) the categorization of individual words at the parole level; 2) the categorization of vocabulary words at the language level. The former refers to a speaker’s propositional speech behavior (reference, statement and modification), while the latter is reflected as an unconscious self-organizing process of a speech community, the core of which is conventionalization or qualitative change.

Regarding how to determine whether a certain usage has been conventionalized, Wang & Chen (2014) propose four criteria: 1) token frequency; 2) type frequency; 3) diachronic distribution; 4) register distribution. Diachronic distribution and register distribution have important reference value, but token frequency and type frequency are decisive for judging whether a usage has been conventionalized. The decisive role of token frequency lies in the promotion of the fixation or conventionalization of individual words, while that of type frequency lies in the promotion of the fixation or conventionalization of more abstract schemas (Evans & Green, 2006, p. 188), which is closely related to the productivity of language structures (Bybee, 2010, p. 95). Therefore, this study intends to conduct a
comprehensive survey of the token frequency and type frequency of “guina” with the aid of the Modern Chinese Corpus of the National Language Commission.

III. METHODS

A. Research Tool

What a general dictionary describes is language facts or the actual uses of language, thus the compilation of a general dictionary is naturally inseparable from the support of natural language data. As collections of natural language texts or discourses, corpora play an important role in dictionary making and provide the most authentic and effective contextual support for the selection and establishment of lemmas, senses division, definition writing, POS tagging, examples selection and writing, grammatical and pragmatic information annotation, etc., which all require statistical operations and abstract analysis of a large number of related language data (Zhang & Yong, 2007, p. 105-106).

The Modern Chinese Corpus of National Language Commission is a large-scale balanced corpus, which contains 9487 language samples (texts) with a total of 100 million Chinese characters. Among them, the annotated corpus (a subset of the modern Chinese general balanced corpus) has about 50 million Chinese characters. The initial data (when the corpus was first built) show that the language data in the corpus are from humanities and social sciences accounting for 60%; natural sciences (including agriculture, medicine, engineering and technology) accounting for 6%; newspapers and comprehensive publications accounting for 26%; practical writing, such as various government documents, notices, letters, brochures, advertisements, etc. accounting for 8%. The language data in the corpus are collected from 1919 to 2002, and most of them are for the last 20 years. The language data provided for online search have been divided and tagged on the basis of word unit and can be searched by “word” and “word class”. As a general corpus, the National Language Commission Modern Chinese Corpus can represent the whole picture of modern Chinese in terms of characters, vocabularies, grammars and semantics (http://corpus.zhonghuayuwen.org/).

B. Research Process and Data Collection

We first type “guina” into the searching column of the Modern Chinese Corpus of the National Language Commission, and then choose the searching conditions of “whole word matching”, “labeled data” and “data source” in the condition column. Finally, we retrieve 291 language examples (sentences and passages) with a total of 318 occurrences (sometimes 2, 3 or even 4 occurrences appear in one language example). After manual identification, all of the 291 language examples with 318 occurrences are valid language data. In view of the relatively small number of occurrences (sometimes 2, 3 or even 4 occurrences appear in one language example). After manual identification, all of the 291 language examples with 318 occurrences are valid language data. In view of the relatively small number of occurrences (sometimes 2, 3 or even 4 occurrences appear in one language example).

However, in the statistical process, we find that the POS tagging of “guina” in the corpus is not accurate, for example:

1. 有/v 了/u 这些/r 事例/n 和/c 比较/d , /w 再/d 由此/d 提出/v 各种/r 归纳/n 假说/n , /w 力图/v 排斥/v 玄/a 神/v 妙/a 想/v , /w 以/p 达到/v 客观/a 规律/a 。 /w

2. 他/r 特别/d 提出/v 归纳/n 不同/a 于/p 综合/a : /w 综合/a 是/v1 从/p 同一/a 命题/n 的/u 细节/n 提炼/n 出/vd 完整/a 的/u 概念/n 或/c 理论/n , /w 而/c 归纳/n 则是/v1 从/p 已知/v 论/k 及/c 未知/v 。 /w

In the first example, “guina” is used as an attributive of “jiashuo” (hypothetical) (hypothesis), the two together are modified by the quantifier “gezhong” (various) and served as the object of the predicate “tichu” (提出) (propose); In the second example, the first “guina” and the following “zonghe” (综合) (synthesis) are two co-ordinate components, they together serve as the object of the predicate “tichu”, the second “guina” serves as the subject of the entire clause.

Therefore, the two authors re-analyzed the retrieved language data on the basis of the grammatical functions of words (Xv & Tan, 2006) and the two-level word class categorization theory (Wang, 2014). The result shows that the verb usages of “guina” are 185 occurrences accounting for 58.2%; the noun usages are 133 occurrences accounting for 42.8% (see Table 2).

<table>
<thead>
<tr>
<th>TABLE 1</th>
<th>THE POS OF “GUINA” IN THE ANNOTATED CORPUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word class</td>
<td>number</td>
</tr>
<tr>
<td>Verb</td>
<td>310</td>
</tr>
<tr>
<td>Noun</td>
<td>8</td>
</tr>
<tr>
<td>Total</td>
<td>318</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE 2</th>
<th>THE ACTUAL POS OF “GUINA”</th>
</tr>
</thead>
<tbody>
<tr>
<td>WORD CLASS</td>
<td>NUMBER</td>
</tr>
<tr>
<td>VERB</td>
<td>185</td>
</tr>
<tr>
<td>NOUN</td>
<td>133</td>
</tr>
<tr>
<td>TOTAL</td>
<td>318</td>
</tr>
</tbody>
</table>
After counting the token frequency of the verb and noun usages of “guina”, we analyze and count the type frequency of all the usages of noun. The result shows that “guina” is distributed in the structures “guina + (的) + NP” (60.2%), “NP + (的) + guina” (12.8%), “VP + guina” (10.5%), “guina + VP” (6.8%), “PP + guina” (6.8%), “guina + PP” (2.3%) and “Adj + guina” (0.8%) (see Table 3).

<table>
<thead>
<tr>
<th>Structure</th>
<th>Type Frequency</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>guina + (的) + NP</td>
<td>80 60.2%</td>
<td>归纳（处理）系统，归纳（方法）、归纳的范围、归纳原则、归纳理论、归纳的作用、归纳主义、归纳（推理）的人。</td>
</tr>
<tr>
<td>NP + (的) + guina</td>
<td>17 12.8%</td>
<td>科学归纳、专著系统的归纳、专著归纳、经验归纳、实验的归纳、同类事物的归纳</td>
</tr>
<tr>
<td>VP + guina</td>
<td>14 10.5%</td>
<td>进行归纳、视为归纳、是归纳、用归纳的。</td>
</tr>
<tr>
<td>guina + VP</td>
<td>9 6.8%</td>
<td>归纳所要求（的）、归纳（可以）是、归纳发挥（作用）、归纳（原则）是、归纳得到的。</td>
</tr>
<tr>
<td>PP + guina</td>
<td>9 6.8%</td>
<td>从归纳、通过归纳、以归纳（为主）、对（……）归纳</td>
</tr>
<tr>
<td>Adj + guina</td>
<td>3 2.3%</td>
<td>归纳不同于、归纳在……中（的运用）</td>
</tr>
<tr>
<td>合计</td>
<td>133 100%</td>
<td>这样的归纳</td>
</tr>
</tbody>
</table>

(Note: To ensure the accuracy of the results, the token frequency of verb and noun usages and the type frequency of noun usages are analyzed and counted separately by the two authors. After that, the two authors compared their statistical results with each other, any inconsistency is discussed fully before the final decision is made.)

C. Research Results

Through the reanalysis and statistics of a total of 318 occurrences in the 291 language examples, it is found that “guina” not only has noun (self-referential) usages, but has high token frequency with 133 occurrences accounting for 42.8% of the total usage, and varied type frequency widely distributed in the structures “guina + (的) + NP”, “NP + (的) + guina”, “VP + guina”, “guina + VP” and “PP + guina”. Judging from the total number, proportion and the distribution of the type frequency of the noun usages of “guina” (mainly served as an attributive, object and subject), it is concluded that the noun usage of “guina” has reached the criterion of conventionalization, and should be labeled with the noun POS and set up self-referential sense independently.

IV. Discussion

A. The Criterion of the POS Classification of Mono-category and Multi-category Words

As for the POS tagging in the 5th edition of the CCD, Xv & Tan (2006, p. 26), the editors of the CCD, point out: “POS is the grammatical classification of words, which can explain the usages and functions of words,” “The grammatical meaning is the internal basis of the classification of POS, while the grammatical function is the external performance of POS, the two are closely related. In POS tagging, the grammatical meaning and the grammatical functions of words should be considered together, but the actual operation is mainly based on the grammatical functions.” Through the investigation of the POS tagging in the CCD, it is found that the editors followed the above principle while judging the mono-category words and the metonymic use of words, but when it comes to the self-referential usage of lexemes, the principle “if the meaning is unchanged, the POS should be unchanged as well” is followed (see Wang, 2009; Jiang, 2013; Hou, 2017). For example, the POS and the sense of the metonymic use of the verb “fanyi” (翻泽) (action for the doer metonymy, which refers to the people who carry on the job of translation or interpretation) are established in the 5th, 6th and 7th editions of the CCD, but its self-referential usage, both the POS and its sense, are not.

The principle that “if the meaning is unchanged, the POS should be unchanged as well” can be traced back to the Modern Chinese Grammar by Wang in 1943, the Grammatical Rhetoric Speech and the About the principal issues on the POS of Chinese by Lv & Zhu in 1951 and in 1954 respectively. Lv & Zhu argue in the Grammatical Rhetoric Speech that when the meaning of a word is unchanged, the class to which it belongs should be unchanged as well (Lv & Zhu, 2013, p. 10). The reason why this principle is so popular among Chinese scholars and lexicographers is that they believe that the number of the words like “guina” and “fanyi” is very large. If their POS and self-referential senses are tagged and established separately, then the number of multi-category words will become very large (Lu, 1994; Tan, 2001).

Lu (1994) argues that if a word can be used in different syntactic positions and the words of the same type can be used in the same way as it, this kind of usage is included in the functions of this word, but not regarded as a multi-category word. For example, “laodong” (劳动) can appear in four grammatical positions: subject (劳动光荣), predicate (他不劳动), object (他爱劳动) and attributive (要关心劳动人民), but since there are a great deal of Chinese characters like “laodong”, we cannot take it as a multi-category word, otherwise, the proportion of multi-category words will be too large. Therefore, “laodong” is only regarded as a verb, not as a bi-category word of verb and noun. Tan
(2001, p. 294-295) also believes that whether to tag a word as a multi-category word or not, the principles of analogy, quantity, and meaning should be referred to, if a word can be used in different syntactic positions and other words of the same type can be used in the same way as the word, meanwhile the number of the words is large and there is no obvious change in meaning, other usages of such words can only be regarded as the inherent functions of them. However, Lu’s view has changed in recent years. He believes that the verb and adjective which appear in the position of subject and object cannot be simply considered to be nominalized, or just the inherent functions or usages of the verb and adjective themselves, but should distinguish between “nominalization” and “omission” (Lu, 2015).

As illustrated before, this view of POS classification does not only affect the POS tagging of Chinese dictionaries, but also affect the POS tagging of Chinese corpora. However, which criterion should be adopted to POS classification and to deal with multi-class membership is mainly determined by the purpose of POS tagging. POS is the classification of grammatical functions of words, it is an essential instrument for syntactic and grammatical analysis (Hu, 1995; Shen, 2009). In the article explaining the POS tagging in the CCD, Xv & Tan (2006) also argue that POS is the classification of grammatical functions of words and can be used to explain the usages and functions of words. In addition, in terms of the purpose of POS tagging in a dictionary, lexicographers all hold that it is mainly used to present the grammatical information of lexemes (e.g. Chen & Huang, 1994; Zhang & Yong, 2007, p. 122; Svensén, 2009, p. 136). In this respect, POS also should be viewed as the grammatical functions that a word serves.

Since POS is the classification of words in terms of grammatical functions and mainly reflects the grammatical information of lexemes, in the process of tagging mono-category words and dealing with multi-class membership, this criterion should always be adhered to ensure the consistency of POS tagging criterion and to avoid the systematic problems of POS tagging in a dictionary, which is also the fundamental requirement of the systematic principle of dictionary making (Zhang & Yong, 2007, p. 206). Admittedly, even if the principle of grammatical functions of words is adhered from the beginning to the end in the course of POS tagging, it does not mean that all problems in POS tagging can be solved or no new problems appear. On the one hand, there are many difficulties in the POS tagging in modern Chinese: 1) The usages of some words are so special that they are difficult to be classified; 2) The syntactic components of some words in real use are not easy to determine; 3) The usage of some words is unclear, especially the classical Chinese words and some technical terms; 4) The identity of some words is not easy to determine (Guo, 1999). On the other hand, POS is not a clear-cut concept but a continuum, so it is impossible to avoid the gray areas. This dilemma is also experienced in the process of judging the word-class membership of “guina”. Therefore, the idea to classify all words clearly according to a certain principle and through several classification procedures is almost impossible (Lu, 1994, 2015), but to ensure the consistency of the criterion of POS tagging as far as possible undoubtedly plays an important role in solving the systematic problem of the POS tagging in the CCD.

B. The Procedure of POS Tagging

Before the advent of corpus, dictionary making also had a certain empirical basis, that is, the materials of dictionary making (e.g. senses division, definitions writing, and examples selection or writing) were mostly from excerpted cards. In this tradition of dictionary making, there are many subjective factors involved, which lead to the fact that a dictionary does not reflect the language as itself, but reflect the language as editors imagine (Svensén, 2009, p. 40-45).

As a normative dictionary, the CCD has an important role in the promotion of mandarin and the standardization of modern Chinese (Pan, 2000; Jiang, 2019; Du, 2019). However, the standardization of modern Chinese mainly lies in the phonetic transcription, the writing of Chinese characters, and the grammar of modern Chinese. As a general dictionary, meaning, usage, and other information related to language facts are still descriptive. As Wang argues in the introduction of a special column of lexicographical studies that what a general language dictionary include are the standardized language units, meaning, and usage, which represent the language knowledge of lexicon at communal language system level (see Wang & Huang, 2017). Thus the POS tagging and the sense establishment should be descriptive and objective as well.

Through the investigation of the literature on the making of the CCD since 5th edition (e.g. Jiang, 2013; Hou, 2017), it is found that the making of the CCD has been making use of corpora in certain aspects, but as for POS tagging, it is unknown whether the compilers have conducted a comprehensive investigation on the usage patterns of all lexemes on the basis of corpora. But judging from the number and types of all the POS problems in the CCD, we have reason to believe that the POS tagging is largely subjective and does not make use of corpora. The way of handling the POS tagging issue is consistent with the dominant view of the Chinese grammar community on POS, namely a word can only belong to a certain class and multi-category words must be a minority (Zhu, 1982; Lu, 1994; Guo, 1999; Zhou, 2015).

Zhu (1982, p. 39) argues that when we separate the two classes of words A and B, some words can be allowed to belong to both classes, but if most of the A-class words belong to the B-class, or most of the B-class words belong to the A-class, the division of A and B classes is of little meaning. Guo (1999) holds that the bi-category words of verb and noun like “yanjiu” (研究) (research) and “jiancha” (检查) (check) are very large. If a homogeneous strategy is adopted, the number of bi-category or multi-category words will become too large and destroy the simplicity principle of POS tagging. Zhou (2015) also argues that in the tagging of multi-category words, the most commonly used criterion is the
quantitative principle, namely the multi-category words must be a minority, otherwise the classification of word class is invalid.

Wang & Huang (2017) clearly state that the scholars who advocate the above principle do not distinguish the categorization processes of vocabulary words and individual words, and take (communal) language, which is a product of cultural heritage, as a natural product that has nothing to do with use. In POS tagging, they rely more on introspection and ignore empirical investigations on actual use of language. Language is essentially a complex adaptive system, language structures emerge from the use of language, the frequency of use is very important for the cognitive representation and conventionalization of language structures, the so-called language knowledge is the description or generalization of the actual use of language (Bybee & Hopper, 2001, p. 1; Bybee, 2007, p. 5, 2010, p. 1-2; Kretzschmar, 2015, p. 19, etc.).

If language structures or language knowledge emerge from language use, POS, as a kind of language knowledge (grammatical units), emerges from language use as well. POS is not an unchangeable object, the initial usage of a word may belong to any categories, but with the emergence and conventionalization of other usages, the word may evolve into a two-category, three-category or even multi-category word. Taking the word “back” as an example, in the 9th edition of the Oxford Advanced English Dictionary, “back” is a multi-category word of noun, verb, adjective and adverb. However, according to the Online Etymology Dictionary, the earliest usage of “back” is a noun, after which the usage of adverb, adjective and verb are derived in turn.

The POS is derived from the use of language refutes the view that the number of multi-category words should be minimized or multi-category words must be a minority (see Lu, 1994; Tan, 2001). To determine whether certain words need to be tagged with a certain POS or not, the corpus-based usage pattern survey can give an answer. If words of the same category of a certain word are indeed used in the way as this word and this kind of usage is conventionalized, the corresponding POS should be tagged so as to describe the actual usage of language objectively and accurately. The realistic principle of dictionary making also requires dictionary compilers to face up to the facts of language use and to describe language phenomena objectively (Zhang & Yong, 2007, p. 213-214). As for new meanings or usages of words, compilers cannot depend on introspection and ignore the actual use of language, but conduct a corpus-based survey according to certain principles and methods, and then to determine whether they can be included into a dictionary or not (Zhang & Yong, 2007, p. 213-214). As every sense in an entry is generalized from the typical environment of language use as well as specific language materials (Chen & Huang, 1994), so is the POS.

V. CONCLUSION

Based on the criterion of grammatical functions of words and the two-level categorization theory, this paper conducted a corpus-based case study on the POS tagging of a science lexeme “guina” in the 6th and 7th editions of the CCD. The result shows that “guina” not only has self-referential usage, but also has high token frequency and rich type frequency, which conforms to the standard of conventionalization. Therefore, it is necessary to tag the noun POS and to set up the self-referential sense for “guina”. Since the criterion for classifying POS is determined by the purpose of classification (Hu, 1995), whether it is tagging the POS for mono-category words or for multi-category words, the criterion of grammatical functions of words should always be adhered. In order to ensure the objectivity and accuracy of POS tagging, the corpus-based investigation on the usage patterns of lexemes also should be made. Modern Chinese is a heterogeneous system which takes vernacular Chinese as its basis and mixes with some classical Chinese vocabularies and grammatical rules, namely, a mixture of different historical levels of grammar and vocabulary (Guo, 1999). The differences between classical Chinese and vernacular Chinese in the use of words and grammatical rules determine that the modern Chinese grammatical system and POS tagging lack the clarity that most Indo-European languages have.

REFERENCES


© 2020 ACADEMY PUBLICATION


---

**Jun Zhang** a teacher at Zhaoqing University, Guangdong, China. He obtained his M. A. degree in Applied Linguistics. His research interest covers Lexicography and Cognitive Linguistics.

**Heng Zhang** a PHD candidate in English Linguistics and a lecturer at Nanchang Normal University, Jiangxi, China. He is also an English counselor providing professional development to EFL teachers around the local areas. His research interest covers sociolinguistics, semantics, and pragmatics.