ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 10, No. 8, pp. 879-885, August 2020
DOI: http://dx.doi.org/10.17507/tpls.1008.05

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 + (of) + NP “,” NP + (of) + guina” and “VP + 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 learners’ dictionaries and Chinese-Foreign Language bilingual dictionaries (Zhang, 2010a, 2010b; Hu, 2013, 2014). Its authoritativeness and scientificity are second to none[1], 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[2], 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) were systematically deleted and the senses of which are included into verbs. Here we take “guina” as an example:

【归纳】<**动**> 归拢并使有条理（多用于抽象事物）：大家提出的意见，～起来主要就是这三点。<**名**> 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”相对）。(《现汉》第 5 版）

[1] http://www.china.com.cn/guoqing/2012-07/16/content_25917832.htm

[2] In the 5th edition, the POS tagging for the same type of words is not consistent or even contradictory with each other, for example, all the lexemes of sports do have the sense “one of the sports events”, but some are labeled as mono-category words of verb, while others are labeled as bi-category words of noun and verb.

【归纳】<动> ① 归拢并使有条理（多用于抽象事物）：大家提出的意见，～起来主要就是这三点。② 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”相对）。(《现汉》第 6、7 版）

In the 5th edition, “guina” is labeled as both verb and noun, but it is only labeled as verb in the 6th and 7th editions. The noun POS is not only deleted, but the sense “一种推理方法” (a method of reasoning) is included into the verb. Based on the principle of grammatical functions of words, which is adopted by the compilers of the *CCD* since 5th edition (Xv & Tang, 2006), and the two-level word class categorization theory (Wang, 2014), this study conducts a corpus-based case study of the usage patterns of “guina”, so as to answer the following questions:

1) Is there a noun usage of “guina”?

2) If so, does the noun usage need to be labeled independently?

II. Theoretical Basis

A. *The Criterion of POS Classification*

In an article explaining the POS tagging in the 5th edition, Xv & Tan (2006, p. 26), the compilers of the *CCD*, points out: “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 function should be considered together, but the actual operation of tagging is mainly based on the grammatical functions of words.” The grammatical functions of words mainly include two aspects: 1) the ability to perform a certain syntactic function and the size of this ability, for example, whether a word can be used as a subject, predicate, attributive or complement; 2) the ability to collocate with other words, for example, whether a word can be modified by an adverb, quantifier or followed by “le” (了), “zhe” (着) and “guo” (过). At the same time, they also elaborate the specific judging criteria for 12 major word classes or POS such as noun, verb, adjective, and adverb. Since this study mainly deals with the problem of the nominalization of verbs, we will cite the two criteria adopted in the *CCD* for judging nouns and verbs (Xv & Tan, 2006, p. 26).

Nouns: Can be used as a subject and an object (猫捉老鼠); can be used as an attributive (木头桌子，邻居的孩子); can be modified by quantifiers (一盏灯，三辆汽车); generally cannot be modified by adverbs (不青年，很桌子). The grammatical meaning of noun is the name of people and concrete or abstract things; the definition in a dictionary is manifested as nounness.

Verb: Can be used as a predicate (他知道); can be used with “le”, “zhe” and “guo” ; can be negated by “bu” (不) and “mei” (没) (不看，没回来); most have objects (吃苹果) or complements (洗干净); generally cannot be modified by degree adverbs such as “hen” (很) and “tai” (太). If it can be modified by degree adverbs and have objects, it is still classified as a verb (很喜欢他). The grammatical meaning of a verb is to express the action or behavior of a person and the change or existence of things; the definition in dictionaries is manifested as verbness (买：拿钱换东西).

It should be noted that the POS tagging criterion is consistent with the currently mainstream view of modern Chinese scholars on POS, that is, the classification of POS should be based on the grammatical functions of words, and the meaning is only for reference (e.g. Chen, 1978, p. 38-57; Zhu, 1982, p. 37; Lv, 1979, p. 33; Lu, 1994; Guo, 1999; Shen, 2009; Fan, 2016).

B. *The Two-level Word Class Categorization Theory*

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, namely the categorization of individual words at the parole level and 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 the 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 for judging whether a certain usage has been conventionalized, but token frequency and type frequency are decisive for judging whether a usage has been conventionalized. The former is to promote the fixation or conventionalization of individual words, while the latter is to promote 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 language examples in the corpus come 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 from the past 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 language data, an exhaustive analysis and counting of the POS of the retrieved data are performed. The result shows that the verb usage of “guina” accounts for 97.5%, a total of 310 occurrences, and noun usage accounts for only 2.5%, a total of 8 occurrences (see Table 1).

TABLE 1
THE POS OF “GUINA” IN THE ANNOTATED CORPUS

Word class	number	proportion
Verb	310	97.5%
Noun	8	2.5%
Total	318	100.0%

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**/v 假说/n ，/w 力图/v 排斥/v 玄/a 思/v 妙/a 想/v ，/w 以/p 达到/v 客观/a 规律/n 。/w

2. 他/r 特别/d 提出/v **归纳**/v 不同/a 于/p 综合/a ：/w 综合/a 是/vl 从/p 同一/a 命题/v 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**/v 则/c 是/vl 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” (假说) (hypothesis), the two together are modified by the quantifier “gezhong” (各种) (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 2
THE ACTUAL POS OF “GUINA”

WORD CLASS	NUMBER	PROPORTION
VERB	185	58.2%
NOUN	133	42.8%
TOTAL	318	100%

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)[3] + 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 3
THE TYPE FREQUENCY OF THE NOUN USAGES OF “GUINA”

Structure	number	proportion	Example
guina + (的)+ NP	80	60.2%	归纳（处理）系统、归纳（方）法、归纳的范围、归纳原则、归纳逻辑、归纳的作用、归纳主义、归纳（推理）的人、
NP + (的) + guina	17	12.8%	科学归纳、音位系统的归纳、音位归纳、经验归纳、实验的归纳、同类事物的归纳
VP + guina	14	10.5%	进行归纳、视为归纳、是归纳、用归纳
guina + VP	9	6.8%	归纳所要求（的）、归纳（既可以）是、归纳发挥（作用）、归纳（则）是、归纳得到
PP + guina	9	6.8%	从归纳、通过归纳、以归纳（为主）、对（……）归纳
guina + PP	3	2.3%	归纳不同于、归纳在……中（的运用）
Adj + guina	1	0.8%	这样的归纳
合计	133	100%	

(Notes: To ensure the accuracy of the results, the usages of verb and noun 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”, and “VP + 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*, points 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 POS classification, while the grammatical function is the external performance of POS, the two are closely related. In POS tagging, the grammatical meaning and the grammatical function of a word should be considered together, but the specific operation is mainly based on the grammatical function.” 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 judging 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”(翻译) (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”, “tuili”, 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 of a certain class can be used in different syntactic positions and the words of the same class can be used in the same way as it, this kind of usage is included in the functions of this word, and not regarded as a multi-category word. For example, “laodong” (劳动) can appear in four grammatical positions: subject

[3] “的” in Chinese equals to “of” in English

(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 judging a word is a multi-category word or not, the principles of analogy, quantity, and meaning should be referred to. If words of the same class of a certain word can be used in the same way as it and the number of the words is large, and at the same time there is no obvious change in meaning, other usages of such words only can be regarded as the inherent functions of them. 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 word class 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 not the classification for other purposes but the need for syntactic analysis (Hu, 1995). It is also an essential instrument for grammatical analysis (Shen, 2009). Xv & Tan (2006) also argue in the article explaining the POS tagging in the *CCD* that POS is the grammatical classification 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 function 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 use of some words is so special that it is difficult to classify; 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 impossible, 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 derived from excerpted cards. However, there were still 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).

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 includes 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. 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. This way of handling the POS tagging 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 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 scholar 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 are emerging 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), should emerge 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 due to the concern that words of the same category of a certain word can be used in the same way as this word in theory. To determine whether these words need to be tagged with a certain POS, only the corpus-based usage pattern survey can give an answer. If the 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. As for new meanings or usages of words, compilers must not 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 in 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 are the POS of headwords.

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". The criterion for classifying POS is determined by the purpose of classification (Hu, 1995), hence no matter it is tagging the POS for mono-category words or for multi-category words, this criterion should always be adhered. To ensure the objectivity and accuracy of POS tagging, the corpus-based investigation on the usage patterns of lexemes also should be made.

This article not only aims to provide significance for the POS tagging of self-referential lexemes in the *CCD*, but also aims to introduce the theoretical and practical dilemmas of Chinese grammar studies, especially the POS problem, to the international grammar and lexicography community. Modern Chinese is a heterogeneous system which takes vernacular Chinese as its basis and mixes with some classical Chinese vocabularies and grammatical rules, that is, a mixture of different historical levels of grammar and vocabulary (Guo, 1999). And, 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

[1] Bybee, J. (2007). Frequency of use and the organization of language. Oxford: Oxford University Press.

[2] Bybee, J. (2010). Language, usage, and cognition. Cambridge: Cambridge University Press.

[3] Bybee, J., & Hopper, P. (2001). Frequency and the emergence of linguistic structure. Amsterdam: John Benjamins.

[4] Cao, X, Z., & Chao, J, Z. (2002). The historical position of the Contemporary Chinese Dictionary. Chinese Lexicography (Eds). Beijing: The Commercial Press.

[5] Chen, C, X., & Huang, J, H. (1994). Micro-structure of bilingual dictionary. *Lexicographical Studies* (05): 145-153.

[6] Chen, W, D. (1978). A brief introduction to grammar. Shanghai: Shanghai Education Press.

[7] Du, X. (2019). On the compilation of the Contemporary Chinese Dictionary in modern times. *Sinogram Culture* (05): 5-7.

[8] Evans, V., & M. Green. (2006). Cognitive linguistics: An introduction. Edinburgh: Edinburgh University Press.

[9] Fan, X. (2016). Reflections on Chinese empty words. *Journal of Shanghai Normal University (Philosophy & Social Sciences Edition)* (06): 105-115.

[10] Guo, R. (1999). The part-of-speech tagging problem of language dictionary. *Studies of Chinese Language* (02): 150-158.

[11] Hou, R, F. (2017). Revisions of and reflections on the part-of-speech in the Contemporary Chinese Dictionary. *Lexicographical*

Studies (04): 43-54+94.

[12] Hu, M, Y. (1995). Investigation on the part-of-speech of modern Chinese. *Studies of Chinese Language* (05): 381-389.

[13] Hu, W, F. (2013). Users' need and the construction of micro-structure in Chinese-English learners' dictionary: An empirical study based on Chinese EFL learners. *Foreign Language and Literature* (02): 72-78.

[14] Hu, W, F. (2014). On the micro-structure of Chinese-English dictionaries: From the perspective of autonomy and dependence model. *Foreign Language and Literature* (03): 63-68.

[15] Jiang, L, S. (2013). An overview of the Contemporary Chinese Dictionary. *Lexicographical Studies* (02): 1-19+93.

[16] Jiang, L, S. (2019). Speech on the 40th anniversary of the official publication of the Contemporary Chinese Dictionary. *Sinogram Culture* (05): 3-4.

[17] Kretzschmar, W. Jr. (2015). Language and complex systems. Cambridge: Cambridge University Press.

[18] Lu, J, M. (1994). On the heterosemy of lexemes. *Studies of Chinese Language* (01): 28-34.

[19] Lu, J, M. (2015). On the characteristics of Chinese word classes. *Chinese Linguistics* (04): 2-7+95.

[20] Lv, S, X. (1979). The problems of Chinese grammar analysis. Beijing: The Commercial Press.

[21] Lv, S, X., & Zhu, D, X. (2013). Grammatical rhetoric speech. Beijing: The Commercial Press.

[22] Pan. X, L. (2000). Century review of Chinese normative dictionaries. *Lexicographical Studies* (04): 31-44.

[23] Shen, J, X. (2009). My view of word classes in Chinese. *Linguistic Sciences* (01): 1-12.

[24] Svensén, B. (2009). A handbook of lexicography: The theory and practice of dictionary-making. Cambridge: Cambridge University Press.

[25] Tan, J, C. (2001). Provisional classifiers: The shift of parts of speech and their marking in dictionaries. *Studies of Chinese Language* (04): 291-301+383.

[26] Wang, R, Q. (2009). Grammatical metaphor and the establishment of self-designation senses in Chinese dictionaries: A corpus-based study. *Foreign Language and Literature* (01): 100-108.

[27] Wang, R, Q. (2010). A validity study of the word class system in modern Chinese as seen from *the Contemporary Chinese Dictionary (5th ed.)*. *Foreign Language Teaching and Research* (05): 380-386+401.

[28] Wang, R, Q. (2013). A study of multiple class membership in modern Chinese with a comment on the significance of the linguistic theories of Ferdinand de Saussure. *Foreign Language and Literature* (01): 12-20.

[29] Wang, R, Q., & Chen, H, M. (2014). A corpus-based study of the relationship between verbs and constructions: The conventionalization of transitive *sneeze*. *Foreign Language Teaching and Research* (01): 19-31.

[30] Wang, R, Q., & Huang, C, N. (2017). Heterosemy of self-reference lexemes in modern Chinese from the perspective of the two-level word class categorization theory. *Foreign Language and Literature* (01): 87-96.

[31] Wang, R, Q., & Yang, X. (2017). The word class problem of "chuban" and debates over endocentric constructions: A study from the perspective of the two-level word class categorization theory. *Chinese Linguistics* (04): 26-35+95- 96.

[32] Wang, R, Q., Huo, Z, Z., & Deng, J. (2019). A study of the representation strategies of heterosemy in the New Century Chinese-English Dictionary (2ed ed.). *Foreign Language and Literature* (02): 11-22.

[33] Wang, R. Q. (2014). Two-level word class categorization in analytic languages // *Proceedings of the 36th annual conference of the German Linguistic Society*. University of Marburg, Germany: 345-347.

[34] Xu, S., & Tan, J, C. (2006). Remarks on the part-of-speech tagging in the Contemporary Chinese Dictionary (5th ed.). *Lexicographical Studies* (01): 25-35.

[35] Yang, X. (2019). On the revision of labeling the word class of multi-category words in *the Contemporary Chinese Dictionary*. *Lexicographical Studies* (01): 20-28+119.

[36] Zhang, Y, H., & Yong, H, M. (2007). Contemporary lexicography. Beijing: The Commercial Press.

[37] Zhang, Y, H. (2010a). Theoretical and definitional problems in CDIL: A contrastive study of native-oriented ordinary dictionaries and foreign-oriented learner's dictionaries. *Journal of Guangdong University of Foreign Language Studies* (05): 5-9.

[38] Zhang, Y, H. (2010b). A contrastive study between foreign-oriented Chinese dictionary and general Chinese dictionary. *Academic Research* (09): 151-160.

[39] Zhou, R. (2015). Reflections on "multi-category" in mandarin's word classes. *Linguistic Sciences* (05): 504-516.

[40] Zhu, D, X. (1982). Lecture on grammar. Beijing: The Commercial Press.

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.