https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2181 (Article)

Word-Class Labeling in the *New Chinese–Vietnamese Dictionary:* A Data-Based Approach

Ni Gong, *College of Foreign Languages, Lanzhou University of Finance and Economics, Lanzhou, China*
(854369354@qq.com)
(https://orcid.org/0009-0007-3607-9012)
and
Yongqiang Ma, *College of Foreign Languages, Lanzhou University of Finance and Economics, Lanzhou, China*
(yongqiangma2008@foxmail.com)
(https://orcid.org/0000-0002-3776-8721)

Abstract: Word-class labeling, which significantly affects the quality and practical value of dictionaries, has remained a persistent challenge in Chinese–foreign language lexicography. Grounded in quantum thinking and transdisciplinary methodology, this study applies the Two-Level Lexical Categorization Theory to systematically examine the word-class labeling in the *New Chinese–Vietnamese Dictionary*. Drawing on a self-built database and corpus-based usage survey, the study evaluates the current labeling practices, explores the underlying reasons for the problems, and proposes strategies for improvement. The study reveals that while the typical class membership of the entries is accurately labeled, there are still notable defects: (1) mistakenly recognizing non-word morphemes as lexical entries; (2) artificially minimizing the number of heterosemous entries and failing to capture the true usage of Modern Chinese entries contained in the dictionary. These problems may partially stem from certain biased conceptions of Modern Chinese word classes, and the misapplication of the "Principle of Parsimony/Simplicity" under the constraints of scientism/Newtonian thinking. To address these issues, this paper argues that the *New Chinese–Vietnamese Dictionary* and other Chinese–foreign language dictionaries should abandon the "Principle of Parsimony/Simplicity" in word-class labeling, adopt corpus-based evidence to reflect actual Modern Chinese usage, and thereby enhance dictionary reliability and usability for learners and translators.

Keywords: CHINESE–FOREIGN LANGUAGE DICTIONARY, *NEW CHINESE–VIETNAMESE DICTIONARY*, WORD-CLASS LABELING, HETEROSEMOUS ENTRIES, CORPUS-BASED, QUANTUM THINKING

Opsomming: Woordklasetikettering in die *New Chinese–Vietnamese Dictionary:* 'n Datagebaseerde benadering. Woordklasetikettering, wat 'n beduidende invloed op die kwaliteit en praktiese waarde van woordeboeke het, bly 'n voortdurende uitdaging in die Chinees–vreemdetaalleksikografie. Gegrond op kwantumdenke en transdissiplinêre metodologie, pas hierdie studie die teorie van tweevlak- leksikale kategorisering toe om die woordklasetikettering in die *New Chinese–Vietnamese Dictionary* sistematies te ondersoek. Deur gebruik te maak van

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2181 (Article)

'n selfsaamgestelde databasis en korpusgebaseerde gebruiksopname, evalueer die studie die huidige etiketteringspraktyke, ondersoek die onderliggende redes vir die probleme, en stel strategieë vir verbetering voor. Die studie toon dat, hoewel die tipiese woordklaskategorisering van die inskrywings akkuraat geëtiketteer is, daar steeds noemenswaardige gebreke is: (1) die verkeerdelike herkenning van nie-woord-morfeme as leksikale inskrywings; (2) die kunsmatige vermindering van die aantal heterosemiese inskrywings en die versuim om die ware gebruik van Moderne Chinese inskrywings wat in die woordeboek opgeneem is, vas te lê. Hierdie probleme kan gedeeltelik spruit uit sekere bevooroordeelde opvattings oor Moderne Chinese woordklasse, en die verkeerde toepassing van die 'Beginsel van Spaarsamigheid/Eenvoud' onder die beperkings van sciëntisme/Newtoniaanse denke. Om hierdie kwessies te hanteer, stel hierdie artikel voor dat die *New Chinese–Vietnamese Dictionary* en ander Chinees–vreemdetaalwoordeboeke die 'Beginsel van Spaarsamigheid/Eenvoud' in woordklasetikettering laat vaar, en korpusgebaseerde bewyse volledig insluit om die gekonvensionaliseerde taalgebruik in Moderne Chinees akkuraat weer te gee, om sodoende die woordeboek se betroubaarheid en bruikbaarheid vir beide leerders en vertalers te verbeter.

Sleutelwoorde: CHINEES–VREEMDETAALWOORDEBOEK, *NEW CHINESE–VIETNAMESE DICTIONARY*, WOORDKLASETIKETTERING, HETEROSEMIESE INSKRYWINGS, KORPUSGEBASEERD, KWANTUMDENKE

1. Introduction

As a fundamental task in lexicography, word-class labeling directly affects the quality and practical value of dictionaries. The categorization of word classes remains a cutting-edge research focus across theoretical linguistics, neurolinguistics, and linguistic typology (Spike 2020). In analytic languages such as Modern Chinese (Mandarin) and Modern English, grammatical multifunctionality has been a persistent point of debate across different theoretical approaches to word-class categorization (Vapnarsky and Veneziano 2017; Van Lier 2023). Due to its typological characteristics, including the absence of morphological inflections, heavy reliance on syntactic criteria, and a fluid boundary between morphemes and words, word-class identification in Modern Chinese is a particularly challenging issue. This renders word-class labeling in the compilation of Chinese and Chinese–foreign language dictionaries quite difficult, especially in distinguishing lexical entries from non-lexical entries and handling heterosemous/multi-category entries (Lu 2024). Moreover, authoritative Modern Chinese dictionaries, which serve as significant references for the compilation of Chinese–foreign language dictionaries, also exhibit inconsistencies and flaws in their treatment of heterosemous entries (Yang and Wang 2018). Therefore, a systematic examination of word-class labeling accuracy in Chinese–foreign language dictionaries is warranted.

Nevertheless, previous research on Chinese lexical categorization has largely centered on word tokens in syntax, paying limited attention to word types or lexemes in the lexicon, especially in the context of lexicography. Besides, existing studies demonstrate several limitations, including inadequate systematic analysis

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2181 (Article)

of labeling practices, insufficient empirical validation, and disproportionate emphasis on Chinese–English dictionaries while neglecting Chinese and non-major language pairs (e.g., Chinese–Vietnamese dictionaries). Given the important impact of accurate and consistent word-class labeling on L2 learners' vocabulary acquisition, the research gaps need to be addressed.

Furthermore, with the deepening of economic and cultural exchanges between China and Vietnam, the demand for high-quality Chinese–Vietnamese learning dictionaries has been greatly increasing. In response, *the New Chinese–Vietnamese Dictionary* (hereinafter referred to as "NCVD") was published in 2013 under the National Publication Foundation Project. Since its publication, the NCVD has garnered widespread acclaim and has become an essential reference book for Chinese language learners in Vietnam, marking a milestone in the history of reference book publishing. Yet no systematic investigation has been conducted to evaluate the word classes, leaving both its accuracy and its reflection of actual Chinese usage unverified. Given that NCVD is compiled primarily for Vietnamese learners of Chinese, accurate word-class labeling is particularly crucial, as Vietnamese, like Chinese, is an analytic language with limited morphological marking. Unlike previous studies on Chinese–English dictionaries, NCVD serves Vietnamese learners — an under-researched user group — making this the first systematic investigation of word-class labeling in a Chinese–Vietnamese dictionary.

The Two-level Lexical Categorization Theory (Wang 2014a, 2023) addresses word-class categorization at both *parole* and *langue* levels (a distinction originating with De Saussure 1916/1959), offering a useful analytical framework for long-standing issues in analytic languages such as Modern Chinese and English. Guided by this theoretical framework, the present study systematically examines the identification of lexical versus non-lexical entries and the representation of heterosemous entries, drawing on both the self-built "Word-Class Labeling Database of NCVD" and corpus-based evidence. The present study focuses specifically on "X变(X biàn)" two-character entries (e.g., "改变(gǎibiàn)" [change]) as a representative case for examining the accuracy of heterosemous entry representation in lexicography. Through this analysis, the study critically evaluates the current labeling practices, identifies the underlying causes of the existing problems, and proposes corresponding strategies for enhancing the quality of NCVD and other Chinese–foreign language dictionaries. These refinements aim to optimize their utility for language learners and educators in both dictionary consultation and pedagogical contexts.

2. Literature review

2.1 Current status of Modern Chinese word-class research

Word-class classification has long been a debated topic in Chinese linguistics (Lu 2024: 33). Research on Chinese word classes focuses on two core issues: the

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2181 (Article)

relationship between lexical categories and syntactic functions, and the criteria for word-class determination, particularly regarding the treatment of grammatical flexibility/multifunctionality (Vapnarsky and Veneziano 2017; Lu 2024). Zhu (1985: 4-5), a prominent Chinese linguist, argued that Chinese nouns, verbs and adjectives are intrinsically multifunctional (e.g., Verbs and adjectives can all serve as subjects, predicates, objects, or modifiers without requiring inflectional markers or morphological changes, etc.), indicating that there is no one-to-one correspondence between Chinese word classes and syntactic functions/grammatical roles as seen in Indo-European languages. This viewpoint has influenced Modern Chinese and Chinese–foreign language dictionaries, as well as pedagogical materials for both native Chinese instruction and teaching Chinese as a foreign language. In lexicography, Zhu's proposal has been further formalized as the "Principle of Parsimony/Simplicity" (Shen 2016; Guo 2018), which advocates minimizing the number of heterosemous entries. However, empirical evidence indicates that this framework fails to account for the authentic characteristics of Modern Chinese and creates theoretical inconsistencies (Wang 2020).

Moreover, there are theoretical divergences in word-class analysis, as scholars tend to conflate the objects of word-class categorization in Modern Chinese. Specifically, the structuralist view (Lehmann 2013; Guo 2018) treats Chinese word classes as classifications of word types, which, however, inadequately addresses part-of-speech tagging challenges in natural language processing. Conversely, the generative and construction grammar approaches (Chomsky 1970; Bisang 2013) treat word classes as attributes of word tokens only, creating intractable problems for the representation of heterosemous entries in lexicography. Departing from these unidimensional approaches to word-class categorization, Wang (2023) proposed the quantum linguistic framework — the Two-level Lexical Categorization Theory. The theory holds that lexical items exhibit wave-particle duality analogous to quanta in quantum physics; word tokens manifest particle-like states with single-category properties in concrete linguistic contexts, while word types display wave-like behavior through their categorical wave function, possessing multifunctional potential (representing a superposition of categorical states). Briefly, this approach accounts for both the categorization of word tokens and word types while considering the stability of word classes in dictionaries and their flexibility in actual usage, providing a comprehensive and reasonable perspective for analyzing word-class issues and offers important guidance for improving part-of-speech tagging quality in natural language processing and word-class labeling in lexicography.

In conclusion, due to the typological characteristics of Modern Chinese, although the Two-level Lexical Categorization Theory offers a useful analytical framework, different perspectives on word-class categorization persist. This theoretical predicament has resulted in discrepancies in word-class representation practices in Modern Chinese dictionaries and Chinese–foreign language bilingual dictionaries.

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2181 (Article)

2.2 The previous studies on word-class labeling in Chinese–foreign language dictionaries

Research on word-class labeling in Chinese–foreign language dictionaries has provided valuable theoretical frameworks and practical guidance for dictionary compilation, thereby contributing to the development of bilingual lexicography. The significance of word-class labeling in dictionary compilation has been well documented. Word-class labeling influences the precision of definitions (Tan 2024), sense division, translation, illustrative examples and the accurate and efficient lexical acquisition for learners of Chinese as a foreign language (Wang and Yin 2023).

Studies on word-class labeling in Chinese–foreign language dictionaries have primarily focused on Chinese–English dictionaries, and have addressed issues such as lexical and non-lexical entry distinction, translation, and heterosemous entry representation.

Notable achievements have been made in the study of word-class labeling in Chinese–English dictionaries, though significant divergences persist in scholarly assessments. For instance, Wu and Wang (2022) found that five authoritative Chinese–English dictionaries uniformly classify "temporal words" (e.g., "半夜(bànyè)" [midnight] and "过去(guòqù)" [past]) as single-category nouns, disregarding their other grammatical functions. Similarly, building upon the framework of the Two-level Lexical Categorization Theory (Wang 2014a, 2023), Wang and Yin (2023) compared the representation of "identical medical terms" (e.g., "瘫痪(tānhuàn)" [paralysis; palsy] and "治疗(zhìliáo)" [treat; treatment]) across four major Chinese–English dictionaries, revealing both inconsistencies in word-class labeling for these entries and discrepancies in labeling for semantically symmetrical term pairs within the same dictionary. However, existing research demonstrates marked divergences in assessing word-class labeling practices across Chinese–English dictionaries. Specifically, regarding the *Chinese–English Dictionary* (third edition), Li (2013) positively evaluated its treatment of heterosemous entries through selective examples, but Wang (2020) documented frequent omissions or mislabeling of Chinese content words, as well as the inconsistent handling of antonymous, synonymous, or semantically related lexical items both within and across dictionaries.

Chinese–English lexicography has achieved relative maturity, with a growing body of research, whereas studies on other Chinese–foreign language pairs (especially non-major languages like Chinese–Vietnamese) remain scarce. Wang (2014b, 2015) investigated the word-class labeling in *the Concise Chinese–German Dictionary* and *Chinese–Italian/Italian–Chinese Bilingual Dictionary*, identifying important deficiencies in the two dictionaries' standards and procedures for word-class determination, translation methods, and the distinction between lexical and non-lexical entries. His studies have not fully examined the representation of heterosemous entries, the underlying causes for the problems, and potential solutions for dictionary improvement, but they provided key insights and guidance for Chinese–foreign language dictionary studies and compilation. Notably, beyond

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2181 (Article)

Wang's pioneering work, our literature review shows that word-class labeling issues in other Chinese–non-major language dictionaries, including Chinese–Vietnamese ones, have to date received comparatively little scholarly attention.

2.3 Limitations in previous studies on word-class labeling in Chinese–foreign language dictionaries

Existing research on word-class labeling in Chinese–foreign language dictionaries has offered useful insights into enhancing dictionary quality, but several limitations remain to be addressed.

First, holistic and systematic research is relatively scarce. Most studies have discussed the representations of heterosemous entries in Chinese–foreign language dictionaries through one or several examples of the entries, but they fail to conduct a full-sample analysis of current labeling practices. Besides, their exploration of the root causes of the problems and potential solutions is insufficient. Second, empirical research remains underdeveloped. Current studies mainly rely on traditional theoretical speculation, limited sampling analyses, or isolated case studies, failing to incorporate systematic investigations using large-scale corpora, which restricts the depth and generalizability of findings. Third, research coverage is uneven. The previous studies focus disproportionately on Chinese–English dictionaries, leaving word-class labeling in Chinese dictionaries of non-major languages (including Chinese–Vietnamese) understudied. While foundational research on Vietnamese word classes exists (e.g., Honey 1956), such work has rarely been incorporated into Chinese–Vietnamese lexicography.

3. Research design

3.1 Research questions

The study seeks to address the following questions:

RQ 1: How accurate is the word-class labeling in NCVD, particularly in distinguishing lexical from non-lexical entries, and representing heterosemous entries?
RQ 2: What are the main causes of inaccuracies in word-class labeling in NCVD?
RQ 3: What strategies can be employed to improve word-class labeling in NCVD?

3.2 Theoretical basis

This study adopts the Two-level Lexical Categorization Theory (Wang 2014a, 2023), originally proposed at the 36th Annual Conference of the German Linguistic Society. The theory integrates quantum thinking, the four axioms of trans-

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2181 (Article)

disciplinary methodology, and the complex adaptive systems (CAS) perspective (Beckner et al. 2009; Bybee 2010) to address word-class categorization in analytic languages, particularly Modern Chinese (Mandarin). The essence of the theory is that lexical items are linguistic quanta (Aerts and Beltran 2022; Busemeyer and Bruza 2025: 313) that exhibit wave-particle duality analogous to quanta in quantum physics. The theory's essential tenets can be summarized as follows:

(1) A word token is a first-order material entity bound to specific spatiotemporal contexts (localized), exhibiting the particle state of word class (i.e., a single-category word). In contrast, a word type is a second-order abstract entity transcending spacetime (non-localized), manifesting as the wave function of word class, possessing multifunctional potential, including exploitation or even multi-categorization (i.e., a superposition state of word-class information) (see also Zhu 2010: 32-35).

(2) The word-class categorization of word tokens reflects the agency of speech individuals in specific contexts, while that of word types reflects the collective agency of the speech community (i.e., inter-subjectivity or subject-subject interaction). Both, however, involve subject-object interaction.

(3) The grammatical multifunctionality of word types constitutes word-class superposition, which can be divided into conventionalized superposition (i.e., multi-categorization, where a word type has established multiple word classes in the communal lexicon) and non-conventionalized superposition (i.e., exploitation, where a word type exhibits novel grammatical functions in specific contexts without having acquired conventionalized status; see Hanks 2013).

(4) The word-class determination of word tokens should follow classical logic, integrating their syntactic functions (marked morphosyntactically), pragmatic functions, and conceptual representations (semantic events) in given contexts. The outcome is a definite single category, analogous to the eigenstate of a collapsed wave function.

(5) The word-class determination of word types should follow quantum logic, employing corpus-based usage pattern analysis. Based on four criteria — token frequency, type frequency, temporal span, and register distribution — a conventionalized judgment is made, yielding either a single-category word or a heterosemous/multi-category word (i.e., a grammatical property superposition state).

The theory provides a coherent framework of word-class categorization by incorporating both word tokens and word types. For the present study, this framework offers concrete guidance in distinguishing lexical from non-lexical entries, identifying heterosemous entries, and applying corpus-based criteria. The theory's applicability has been demonstrated in studies on temporal words (Wu and Wang 2022), transdisciplinary methodology (Wang 2023), medical entries (Wang and Yin 2023), and "X击" compounds (Wang and Zhao 2025). These applications support the theory's relevance to the current investigation of NCVD.

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2181 (Article)

3.3 Data collection and analysis

This study employs an empirical approach, combining dictionary database and corpus analysis with case study methods, to comprehensively investigate word-class labeling in NCVD. The study examines all word-class labeled entries in NCVD (full sample). The following steps were taken:

(1) Creating the database: the word-class labeling for each entry in NCVD was coded into an Excel table to create the "Word-Class Labeling Database of NCVD", which serves as the foundation for subsequent analyses.

(2) Performing statistical analyses: this is performed to examine the identification between lexical and non-lexical entries, as well as the representation strategies adopted for heterosemous entries.

(3) Selecting case studies: this study focuses on "X变(X biàn)" two-character entries (i.e., the disyllabic entries ending with "变(biàn)", such as "改变(gǎibiàn)" [change] and "转变(zhuǎnbiàn)" [shift]) in NCVD as a representative test case. These entries were variably selected due to their functional diversity in authentic language use, which presents both a challenge for compilers to determine their word classes and an opportunity for researchers to assess the accuracy of heterosemous entries representation in lexicography (Wang 2020: 72). Crucially, a simplistic ontological approach might classify all such entries uniformly as verbs based solely on their shared meaning of "change", but the actual usage patterns reveal broader syntactic functionality that demands more nuanced representation. Through analyzing these cases, the study intends to further reveal the limitations in NCVD's current word-class labeling practices.

(4) Conducting comparative analyses with MCD6: the study conducted comparative analysis with *Modern Chinese Dictionary* (sixth edition) (hereinafter referred to as "MCD6"), which is the primary reference resource in the compilation of NCVD. Since NCVD was compiled based on MCD6, the seventh edition (published later) is not applicable as a baseline for comparison. The corresponding "X变(X biàn)" entries in NCVD were identified and compared with those in MCD6.

(5) Surveying corpus-derived usage patterns:
(i) The selection of corpora
This study utilizes the Chinese Corpus from the Peking University Modern Chinese Corpus (hereafter "CCL Corpus") as the primary data source. The 2024 CCL edition contains 4.75 billion characters of naturally occurring Modern Chinese data, spanning diverse registers including newspapers, literature, general texts, classical Chinese, dialogues, and other domains, thereby comprehensively capturing authentic syntactic usage of Chinese words (word tokens) at the *parole* level.

(ii) Corpus-based analysis
Guided by the methodological framework of the Two-level Lexical Cate-

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2181 (Article)

gorization Theory, the study retrieved and analyzed the data from the CCL Corpus to examine the class membership of the lexical item "改变(gǎibiàn)". This lexical item is the most frequently used among "X变(X biàn)" two-character entries in natural language contexts. Specifically, at the *parole* level, the study analyzed contextual usage patterns of "改变(gǎibiàn)" in concordance lines to determine its word class, examining syntactic, pragmatic and cognitive correlations. At the *langue* level, the study quantified token frequency, type frequency distribution, temporal span, and register variation to identify conventionalized propositional speech acts, thereby establishing its word class as a word type within the communal language system — lexicographically represented word classes in theory.

4. Research results

4.1 The identification of lexical entry and non-lexical entry in NCVD

Accurate identification of lexical entries and non-lexical entries is crucial for foreign language learners in their use of vocabulary and sentence construction. The survey reveals that NCVD assigns word-class labels to the entries with fewer than three characters (i.e., single-character and two-character entries), whereas for three-character entries, only nouns and quantifiers are labeled. Additionally, NCVD has identified combinations of numerals and quantifiers as "numeral-quantifier entries". There are a total of 25 "numeral-quantifier entries[1]" in NCVD, including "百倍(bǎibèi) (hundredfold), 两次(liǎngcì) (twice)" (See Endnote 1 for all "numeral-quantifier entries"). Being not a major word class in NCVD's labeling system, numeral-quantifier entries are excluded from the scope of this investigation for the sake of cross-dictionary comparison. The word formation status of the other entries in NCVD is shown in Tables 1 and 2.

Table 1: Distribution of lexical and non-lexical entries across character-length categories

Character-length		Single-character	Two-character	Three-character	Four-character	Total
Lexical entries[2]	Number	8484	36805	5873	0	51162
	Percentage	16.58%	71.94%	11.48%	0	100%
Non-lexical entries[3]	Number	392	3	1335	8377	10107
Subtotal		8876	36808	7208	8377	61269

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2181 (Article)

Table 2: Distribution of lexical entries across character-length categories

Character-length	Single-character	Two-character	Three-character	Four-character
Number of lexical entries	8484	36805	5873	0
Total number of entries	8876	36808	7208	8377
Percentage	95.58%	99.99%	81.48%	0

As shown in Tables 1 and 2, in NCVD, two-character entries form the largest category, demonstrating a 99.99% lexicality rate and comprising 71.94% of all lexical entries; Three-character entries show an 81.48% lexicality rate while accounting for 11.48% of the dictionary's lexical entries; Single-character entries exhibit a 95.58% lexicality rate and represent 16.58% of total lexical entries. According to Wang's (2011: 74) investigation on *Modern Chinese Dictionary*, two-character entries achieve a 99.98% lexicality rate and constitute 82.91% of the lexical entries, while three-character entries display a 90.35% lexicality rate and make up 10.54% of the entries. However, the two dictionaries diverge dramatically in their treatment of the single-character entries. Specifically, the NCVD shows a 95.58% lexicality rate for the single-character entries, which is 3.45 times higher than the 27.71% rate documented in the *Modern Chinese Dictionary*, where the single-character entries account for only 5.84% of the total. To identify the underlying causes of this striking discrepancy in lexicality rates, the study further analyzed the identification of the single-character entries in NCVD.

4.2 Analysis of lexicality in single-character entries

The survey reveals that NCVD's overestimated lexicality rates for the single-character entries result from treating two types of entries as independent lexical entries, while MCD6 treats them as non-lexical: (a) individual senses of polysemous characters, and (b) monosemous bound morphemes. For example, NCVD has assigned word-class labels to six meanings (excluding its use as a surname) of the character "金(jīn)" (gold). In Modern Chinese, however, these meanings, including "metal", "money", "ancient percussion instrument", "metaphorical preciousness", and "golden color", represent non-word morphemes, as they primarily function as word-formation components rather than independent lexical items.

On the other hand, NCVD classifies certain non-free single-character entries as independent entries inconsistently with MCD6. For instance, "苞(bāo)" (husk) is labeled as an adjective despite exclusively appearing in two-character lexical items like "苞米(bāomǐ)" (corn) and "苞谷(bāogǔ)" (maize) in Modern Standard Mandarin; "烬(jìn)" (ash) is treated as a noun though it only occurs in two-character lexical items, such as "灰烬(huījìn)" (ashes) and "余烬(yújìn)" (embers);

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2181 (Article)

"蕾(lěi/léi)" (bud/lace) is categorized as a noun when it serves solely as a word-formation component in terms like "花蕾(huālěi)" (flower bud) and "蕾丝(léisī)" (lace).

Similarly, single-character lexical items such as "攫(jué)" (to grasp), "娲(wā)" (Nuwa), "袜(wà)" (sock), "椭(tuǒ)" (oval), "鸵(tuó)" (ostrich), "涯(yá)" (edge), "睚(yá)" (corner of the eye), "荧(yíng)" (glow), "诊(zhěn)" (diagnose), and "缜(zhěn)" (meticulous) are labeled as nouns in NCVD where MCD6 treats them as bound morphemes. As treated in MCD6 and MCD7, these characters are rarely used independently in Modern Standard Mandarin; instead, they primarily function exclusively as bound morphemes, appearing only in two-character entries like "攫取(juéqǔ)" (to seize), "女娲(nǚwā)" (Nuwa), "袜子(wàzi)" (socks), "椭圆(tuǒyuán)" (ellipse), "鸵鸟(tuóniǎo)" (ostrich), "天涯(tiānyá)" (coastline), "睚眦(yázì)" (corners of the eyes), "荧光(yíngguāng)" (fluorescence), "诊断(zhěnduàn)" (diagnosis), and "缜密(zhěnmì)" (meticulous). Thus, NCVD's treatment of these non-word morphemes as free lexical items differs from that of MCD6 and artificially increases its proportion of single-character entries, which deviates from the actual word-formation patterns of Modern Mandarin Chinese.

A chi-square test was conducted to compare the proportion of single-character entries classified as lexical entries in NCVD versus the Modern Chinese Dictionary (based on data from Wang 2011: 74). NCVD exhibits a lexicality rate of 95.58% (8,484 out of 8,876), whereas the Modern Chinese Dictionary shows a rate of only 27.71% (3,005 out of 10,845). The chi-square test revealed a statistically significant difference between the two dictionaries ($\chi^2 = 8660.4$, $df = 1$, $p < 0.001$), indicating that NCVD adopts a substantially broader criterion for treating single-character entries as independent lexical items.

4.3 The representation strategy of heterosemous entries in NCVD

The treatment of heterosemous entries in a dictionary serves as a critical benchmark for assessing its lexicographical quality. To systematically evaluate the word-class labeling accuracy for heterosemous entries in NCVD, this study conducts a comprehensive examination of both the dictionary's general treatment of heterosemous entries and in-depth case studies of "X变(X biàn)" entries and "改变(gǎibiàn)" as representative examples. MCD6 defines "verbs" as "words that denote actions, existence, or changes in people or things". Under a purely ontological-semantic framework, this definition would lead to the misclassification of all "X变(X biàn)" entries as verbs. Moreover, to thoroughly assess NCVD's representation strategies, the analysis incorporates a synchronic comparison with MCD6.

4.3.1 The overall representation of heterosemous entries in NCVD

This study provides a detailed analysis of the distribution of heterosemous entries in NCVD. The overall representation of these entries is presented in Table 3.

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2181 (Article)

Table 3: Distribution of heterosemy in NCVD

Category	Single-category entries	Heterosemous entries	Total
Number	47659	3503	51162
Percentage	93.15%	6.85%	100%

According to Table 3, among the 51,162 entries labeled with word classes in NCVD, 47,659 (93.15%) are single-category entries, while only 3,503 (6.85%) are heterosemous ones. This distribution closely aligns with the data from MCD6, which reports 52,579 single-category entries (93.35%) and 3,746 heterosemous ones (6.65%) (Yang and Wang 2018). In other words, the recognition of word classes for Modern Chinese entries in NCVD is largely consistent with that in MCD6.

4.3.2 The word-class labeling for "X变(X biàn)" entries

(1) Statistical results of word-class labeling for "X变(X biàn)" entries in NCVD
The survey results indicate that NCVD contains 41 "X变(X biàn)" entries, of which 40 are single-category entries, while only one ("惨变(cǎnbiàn)") exhibits heterosemy. Manifestly, verbs account for 68.29% of all "X变(X biàn)" entries (see Table 4 for details).

Table 4: Word-class labeling of 41 "X变(X biàn)" entries in NCVD

Word class	V.	N.	Adj.	N.+V.	Total
Number	28	11	1	1	41
Percentage	68.29%	26.83%	2.44%	2.44%	100%
Entry	癌变，兵变，病变，多变，改变，畸变，渐变，剧变，裂变，流变，霉变，叛变，权变，善变，衰变，顺变，突变，蜕变，胁变，衍变，演变，窑变，诱变，折变，应变[1]，转变，哗变，政变	婚变，急变，巨变，量变，情变，世变，事变，形变，灾变，应变[2]，质变	可变	惨变	

(2) Comparison of word-class labeling for "X变(X biàn)" entries in NCVD and MCD6
Comparative analysis reveals that MCD6 contains 44 "X变(X biàn)" entries, with 36 overlapping with NCVD. The two dictionaries show strong agreement in their

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2181 (Article)

classification, both identifying 35 entries as single-category entries and recognizing only one ("惨变(cănbiàn)") as heterosemous. See Table 5 for detailed comparisons.

Table 5: Word-class labeling of 36 shared "X变" entries in NCVD and MCD6

Category	Single-category entry			Heterosemous entry	Total
Word class	Verb	Noun	Subtotal	Noun+Verb	
Number	24	11	35	1	36
Percentage	66.67%	30.56%	97.23%	2.77%	100%

Tables 4 and 5 reveal several significant findings regarding "X变(X biàn)" entries in NCVD and MCD6: (i) The data demonstrate that verbs constitute the dominant word class for "X变(X biàn)" entries in NCVD, accounting for 68.29% of all entries; (ii) There exists a striking imbalance in category distribution, with single-category entries representing 97.23% of the entries compared to just 2.77% for heterosemous entries; (iii) Only one entry ("惨变(cănbiàn)") displays dual noun-verb functionality, representing a singular type of heterosemy; (iv) The two dictionaries exhibit perfect agreement in their word-class labeling for all 36 shared "X变(X biàn)" entries.

(3) Comparison of word classes for "改变(gǎibiàn)" in NCVD and corpus-based surveys

To comprehensively examine NCVD's word-class labeling issues and demonstrate the proposed strategies for enhancing the accuracy of word-class labeling, this study selected the lexical entry "改变(gǎibiàn)" as a case study. Both NCVD and MCD6 uniformly classify "改变(gǎibiàn)" as a verb. The consistency between the two dictionaries, however, reflects shared adherence to the 'Principle of Parsimony/Simplicity' rather than empirical evidence. A corpus-based investigation is therefore required to verify whether this classification aligns with actual usage patterns in Modern Chinese. Guided by the Two-level Lexical Categorization Theory, this study analyzed usage patterns of "改变(gǎibiàn)" in the CCL corpus to determine its conventionalized word classes at the *langue* level.

(i) The token frequency and type frequency distribution of "改变(gǎibiàn)"

According to the CCL system documentation, users can set a maximum download of 20,000 concordance lines per query. 20,000 lines for "改变" were retrieved. To ensure a random sample, the 20,000 lines were imported into Excel, where each line was assigned a random number using the RAND() function. The

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2181 (Article)

dataset was then sorted by the random numbers, and the first 2,000 lines were selected for analysis. After excluding invalid and repetitive entries, 1,971 valid lines were retained. A random validation sample of 500 lines was analyzed. The two coders annotated 496 and 495 valid lines, respectively. The first coder identified 424 verbal uses (85.48%) and 72 nominal uses (14.52%); the second identified 425 verbal uses (85.86%) and 70 nominal uses (14.14%). Inter-coder agreement was 94.9% ($\kappa = 0.90$), demonstrating high reliability. The distribution in the validation sample (approximately 85.7% verbal vs. 14.3% nominal) was consistent with that of the main sample (87.52% vs. 12.48%).

Table 6: Distribution of "改变" usages in the CCL corpus

Word class	Frequency	Percentage
Verb	1725	87.52%
Noun	246	12.48%
Total	1971	**100%**

As shown in Table 6, "改变(gǎibiàn)" operates as a verb in 87.52% of the sample. This verbal usage is mainly found in two syntactic patterns: (1) The structure "NP + 改变(gǎibiàn)" (e.g., "今天情况改变了" [The situation has changed today.]), which performs a constative speech act expressing the speaker's recognition of unalterable situations; (2) The transitive construction " (NP) + 改变(gǎibiàn) + NP" (e.g., "这就严重地改变了整个国际局势" [This has seriously changed the entire international situation.]). This pattern realizes a directive speech act implying the potential for intentional modification.

In contrast, "改变(gǎibiàn)" manifests nominal properties in 12.48% of the sample, functioning mainly as either a subject or object representing abstract concepts. When used nominally, its propositional acts shift toward: (1) The "(NP) + VP + 改变(gǎibiàn)" pattern (e.g., "我们迄今尚未看到有什么改变" [We have not seen any changes so far.]), which serves as the goal of volitional acts; (2) The modified noun phrase structure "NP + (的)(de) + 改变(gǎibiàn)" (e.g., "部分改变是因为地心的热力作用" [Some changes are caused by the thermal action of the Earth's core.]), which encodes stative propositions about sociocultural transformations.

(ii) Temporal span and register variation of "改变(gǎibiàn)"

Regarding temporal span, the survey based on the CCL corpus reveals that the verbal usage of "改变(gǎibiàn)" first emerged in Eastern Han dynasty texts (25-220 CE), exemplified in Kong Congzi·Lian Cong I: "顾惟世移，名制改变，文体义类" [Considering that times have shifted, names and systems have changed, and literary styles and meanings have evolved.], already displaying intransitive verb syntax in Chinese. Its nominalization developed during the Six Dynasties (222-589 CE), as shown in Pei Songzhi's Three Kingdoms commentary: "绍惧进之改变" [Shao feared Jin's change.], where the "NP + 之(zhi) + 改变(gǎibiàn)" structure demonstrates nominal properties.

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2181 (Article)

In terms of register distribution, the CCL corpus-based investigation reveals that the different word-class usages of "改变(gǎibiàn)" span multiple domains including literature, dialogues, newspapers, and general texts. Both its verbal and nominal usages first emerged in literary works such as novels, before subsequently appearing in journalistic texts, translated works, spoken conversations, practical writings, TV/movie scripts, and web-based corpora. Notably, the nominal pattern "NP + (的)(de) + 改变(gǎibiàn)" displays robust productivity across registers, particularly in literary works such as fiction and in social conversational speech.

To cross-validate these findings, we consulted the authoritative reference work *Ciyuan* (third edition). The entry confirms that the nominal usage of "改变"(e.g., "发生了改变"[... has undergone changes]) is recognized, supporting our corpus-based evidence for its verb-noun heterosemy.

The analysis above demonstrates that both the verbal and nominal usages of "改变(gǎibiàn)" have achieved conventionalized status within the community language system. These findings provide compelling empirical evidence for recognizing "改变(gǎibiàn)" as a verb-noun heterosemous entry, thereby necessitating a revision of NCVD's current verb-only classification.

5. Discussions

5.1 Current status and existing problems of word-class labeling in NCVD

Chinese–Vietnamese lexicography remains underdeveloped, with limited dictionary production and ongoing challenges in both theoretical frameworks and practical compilation methods. Despite these challenges, NCVD has made remarkable achievements in word-class labeling. For instance, NCVD demonstrates innovation in lexical selection by incorporating both Modern Chinese expressions (e.g., "善变(shànbiàn)" [capricious], "多变(duōbiàn)" [changeable]) and specialized terminology from disciplines like physics (e.g., "可变(kěbiàn)" [variable], "裂变(lièbiàn)" [fission]) — entries that are not included in MCD6 and MCD7. Furthermore, compilers have achieved generally accurate labeling of primary word-class categories for entries.

Nevertheless, NCVD displays notable inconsistencies and limitations in its word-class labeling that require critical attention.

(1) Mistakenly recognizing non-word morphemes as lexical entries

NCVD incorrectly labels non-word morphemes as lexical entries, resulting in the proportion of single-character lexical entries beyond authentic Modern Chinese usage patterns. The primary criterion for determining whether a single-character entry qualifies as a lexical entry is its capacity for independent use in language. In Modern Chinese, characters used in proper nouns, idioms, and classical Chinese, such as nouns, verbs, and adjectives, are typically not labeled with word classes. Nevertheless, NCVD fails to clearly distinguish between lexical and non-lexical usages of polysemous single-character entries, as seen in

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2181 (Article)

the Modern Chinese Dictionary. This issue is markedly evident in entries that can be used independently in specialized fields but not in everyday language. As previously illustrated, characters such as "金(jīn)" (gold), "苞(bāo)" (husk), "烬(jìn)" (ash), "蕾(lěi/léi)" (bud/lace) are mainly used as morphemes for word formation in Modern Chinese, yet NCVD erroneously classifies them as independent lexical entries. This inaccurate identification results in a substantially higher proportion of single-character lexical entries than actually exist in Modern Chinese.

(2) Artificially minimizing the number of heterosemous entries and failing to fully reflect the true usage of Modern Chinese entries contained in the dictionary
Heterosemous entries are both prevalent and typologically diverse in Modern Chinese (Yang and Wang 2018: 10-11), yet NCVD includes only 3,503 such entries (6.85% of the total). This limitation is particularly evident in "X变" entries, where only 1 out of 41 items is recognized as having both noun and verb functions. The treatment of "改变" further shows that compilers' systematic omission of heterosemous word-class attributes has led to an insufficient number, low ratio, and restricted typological range of heterosemous entries in the dictionary, ultimately failing to reflect actual language usage patterns in Modern Chinese.

5.2 Causes and countermeasures for word-class labeling problems in NCVD

5.2.1 Causes of word-class labeling problems in NCVD

(1) Holding biased conceptions of Modern Chinese word classes
The distinction between lexical and non-lexical entries must account for both register (spoken vs. written language) and domain (specialized terminology vs. everyday vocabulary). Modern Chinese lexical categories manifest marked prototype effects, where morphemes, words, and phrases form a continuum rather than discrete categories. Compilers can reliably identify prototypical members of word classes based on intuition alone, but their judgments tend to be less accurate for non-prototypical cases (Wang 2011: 76). The survey reveals that NCVD compilers failed to uniformly apply established criteria, adequately consider register/domain variations, or move beyond classical categorization theory. These problems may partially stem from biased conceptions of Modern Chinese word classes.

(2) Excessive reliance on authoritative Modern Chinese dictionaries for word-class labeling
Compilers of Chinese–foreign language bilingual dictionaries typically rely on authoritative Modern Chinese dictionaries as their primary reference sources. NCVD, for example, extensively consulted MCD6 when handling 36 "X变(X biàn)" two-character entries, including the entry "改变(gǎibiàn)". Nevertheless, the word-class labeling in MCD6 has not conducted comprehensive big-data-

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2181 (Article)

based investigations, particularly lacking systematic examination of each individual lexical entry (Yang 2025: 153). Moreover, the representation of heterosemous entries in MCD6 displays theoretical inconsistency and practical contradictions, failing to provide systematic or unified treatment (Yang and Wang 2018: 11). Corpus-based findings in this study further confirm that NCVD's over-dependence on MCD6 has led to inaccuracies in lexical categorization.

(3) Misapplying the "Principle of Parsimony/Simplicity" under the constraints of scientism (classical Newtonian mindset)

The "Principle of Parsimony/Simplicity" (Shen 2016; Guo 2018) advocates a single-layered, static ontological perspective on word classes, embodying scientistic (Newtonian) paradigms in linguistic theory (Wang 2020: 78). The study of Modern Chinese continues to grapple with several unresolved issues, including the language's distinctive typological features, operational criteria for word-class identification, the ontological status of heterosemy, the actual quantity of heterosemous entries, and their optimal lexicographic representation. Among these, heterosemy has proven to be exceptionally contentious (Lu 2024), which is partly because there is no consensus in Chinese linguistics on the object of word categorization — whether it should be based on individual word tokens or abstract word types. For over a century, research on word classes has made significant progress, but it remains deeply constrained by Newtonian thinking, with persistent deficiencies in ontology, logic, epistemology, and axiology (Wang 2022: 10-13). The word-class labeling in NCVD similarly exhibits these limitations.

From an ontological perspective, the word-class determination of "X变(X biàn)" entries tends to adhere to a scientism-driven, single-layered and static conception of lexical categories. This approach not only obscures the object of study in lexical classification, but also disregards the constructive role of language users in category formation, consequently failing to capture both the diachronic evolution and synchronic layering of their categorical properties (Wang 2022: 10). Specifically, when lexicographers determine the class membership of "X变(X biàn)" two-character entries like "改变(gǎibiàn)", they make no clear ontological distinction between word tokens in syntactic contexts and word types in lexicon in a communal language, nor do they adequately account for the potential subjectivity and subject-object interaction involved in the word-class categorization process. Even when noting innovative usages, they lack proper criteria to determine their conventionalization status. This has led to the omissions of certain word classes (e.g., nouns) and senses in dictionary entries, demonstrating that lexicographers remain confined by this single-layered, static ontological perspective and overlooked the cognitive construal of language users in word-class categorization.

From a logical perspective, word-class determination in NCVD remains constrained by classical logic, where admitting multiple class membership or heterosemy is perceived to violate the three axiomatic laws of first-order logic: the law of identity, the law of non-contradiction, and the law of excluded middle.

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2181 (Article)

For decades, Chinese linguists have acknowledged the existence of grammatical multifunctionality in Chinese words (without explicitly differentiating between word tokens and word types) while consistently adhering to Zhu's (1985: 4-5) "Principle of Parsimony/Simplicity", which requires the fewest possible heterosemous entries. More specifically, the principle suggests treating words with dual grammatical functions either as homographs belonging to different word classes or as manifestations of multi-functional word classes, thereby avoiding the recognition of heterosemous entries to achieve "simplicity" in grammatical analysis (Shen 2016; Guo 2018). Many authoritative Chinese language textbooks, textbooks for teaching Chinese as a foreign language, and Chinese/Chinese–foreign language dictionaries have been compiled following the "Principle of Parsimony/Simplicity" (Shen 2016; Guo 2018). However, this over-reliance on the principle has caused extremely negative impacts on word-class labeling in Modern Chinese/Chinese–foreign language dictionaries, leading to apparent flaws in their representation of heterosemous entries (Wang 2020: 78). This study contends that NCVD's recognition of only 2.77% of "X变(X biàn)" entries as heterosemous, while omitting the noun category for "改变(gǎibiàn)", derives exactly from its adherence to the classical logic embedded in mainstream word-class theory. Furthermore, this study argues that classical logic's law of identity should apply only to semantic equivalence between word tokens in syntactic contexts. In contrast, a word type embodies a multifunctional potential comprising multiple semantic components (corresponding to multiple senses), and its word-class determination should adhere to quantum logic principles (Wang 2023: 9).

From an epistemological standpoint, lexicographers adhere to reductionism, determinism, and linear causality, employing abstract semantics as the sole criterion for word-class determination. This approach neglects the difference in the criteria for judging the word classes of word tokens and word types (Wang 2022: 12). Thus, even when lexical items in subject or object positions convey meanings of "action, activity, or change", they are still categorically recognized as verbs. This study reveals that in NCVD, heterosemous entries account for merely 6.85% of the total entries. Among the 41 "X变(X biàn)" entries examined, 97.23% are treated as monosemous and mono-categorial, with 68.29% classified exclusively as verbs. These results clearly reflect a scientism-driven word-class ideology. The present survey based on Two-level Lexical Categorization Theory demonstrates that both verbal and nominal usages of the general word "改变(gǎibiàn)" meet conventionalization criteria. Thus, "改变(gǎibiàn)" should be recognized as a verb-noun heterosemous entry in the lexicon of Modern Mandarin speakers, rather than misclassified as a verb-only category.

From an axiological perspective, there is a striking disconnect between theoretical word-class research and lexicographical practice. This disconnection mainly manifests in three aspects: theoretical constructs fail to incorporate lexicographic evidence, word-class determination criteria in the dictionary disregard actual usage patterns, and word-class labeling systems deviate from the cognitive essence of language. A typical example is NCVD's artificial reduction

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2181 (Article)

of heterosemous entries, which fails to represent linguistic reality and consequently impairs learners' and users' comprehension and application of Modern Chinese vocabulary. Lexicography is essential for achieving a complete grammatical analysis of a language, particularly when every lexical entry requires word-class labeling (Munro 2005: 307). Only through systematic investigation of dictionary data can researchers truly capture the ontological essence of word-class phenomena and effectively address word-class labeling issues. Ultimately, the explanatory adequacy of word-class theory needs to be empirically verified through its implementation in lexical annotation practices in the dictionary.

5.2.2 Countermeasures for word-class labeling problems in NCVD

(1) Breaking free from the constraints of the "Principle of Parsimony/Simplicity" and aligning with the actual usage patterns of the Chinese language
Given the persistent problems in word-class labeling for Modern Chinese, a critical re-examination of the "Principle of Parsimony/Simplicity" (Shen 2016; Guo 2018) is imperative. The lexicographic research shows that this principle may work for word tokens (the first-order material entities), but it inadequately accounts for the categorical evolution of word types (the second-order entities) (Wang 2022: 11). In other words, this principle becomes problematic when representing heterosemy across word types at the level of the community language system. On the other hand, as an essential vehicle for language standardization, cognitive and learning support, as well as cross-cultural communication, dictionary compilation must be grounded in the objective reality of Chinese language use, faithfully reflecting its authentic features. This fact-based lexicographical approach not only helps establish normative language standards but also provides language learners with accurate and reliable resources, while facilitating effective dissemination and exchange of linguistic culture. Briefly, lexicographic practice must transcend the limitations of the "Principle of Parsimony/Simplicity" to properly represent authentic Chinese language usage.

(2) Drawing on the guidance of the Two-level Lexical Categorization Theory in word-class labeling
The Two-level Lexical Categorization Theory combines classical logic and fuzzy logic, taking into account the categorization of both word type and word token, providing an empirically grounded explanation for the word-class labeling of the entries/lexemes in Chinese/Chinese–foreign language dictionaries and the part-of-speech tagging in Chinese/Chinese–foreign language corpora, as well as clear and reasonable criteria and procedures for judging word-class/part-of-speech. Besides, several empirical studies based on corpora have confirmed that this theory can effectively address the issue of identifying and representing heterosemous entries in Chinese dictionary compilation (Wang 2020). Therefore, lexicographers may benefit from adopting quantum thinking, re-examine the distinctive features of Modern Chinese, prioritize research on lexical categorization, and implement a unified framework for word-class classification.

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2181 (Article)

(3) Fully utilizing corpus-based usage pattern analysis
Medium and large balanced corpora reliably and comprehensively reflect the syntactic behavior of individual words in actual language use. These corpora provide essential support throughout the dictionary compilation process, including lemma selection, entry formulation, and word-class labeling, thereby enhancing a dictionary's accuracy, efficiency, completeness, and consistency (see Dalpanagioti 2019; Jackson 2022: 196; Van Lier 2023; Wang 2024). Corpus-based usage analysis has thus become an indispensable methodology in modern lexicography. Regrettably, some lexicographical representations are not fully based on corpus data due to both the corpus's shortcomings (e.g., insufficient coverage of certain linguistic phenomena) and practical constraints. Crucially, even when corpus evidence is available, some compilers may skip thorough analysis and instead opt for traditional methods or smaller datasets to expedite the compilation process. Furthermore, Chinese word-class research still suffers from theoretical inadequacies, and even authoritative dictionaries frequently contain inconsistent word-class labeling. Corpus-based usage analysis effectively addresses these long-standing challenges by comprehensively and objectively documenting actual lexical usage patterns, scientifically distinguishing between words and non-words, and accurately handling cross-categorical phenomena. These combined advantages significantly enhance the overall quality of dictionary compilation (see also Atkins and Rundell 2008; Ptasznik 2020).

6. Conclusion

Based on the self-constructed "Word-Class Labeling Database of NCVD" and the CCL corpus, this study systematically investigates the distinction between words and non-words, as well as the representation strategies for heterosemous entries in NCVD. Building on previous work, this study applies the Two-level Lexical Categorization Theory to a Chinese–Vietnamese dictionary for the first time, providing full-sample evidence and concrete revision strategies. The results reveal that while NCVD has achieved remarkable progress in word-class labeling, it still shows several noteworthy limitations, including treating non-lexical entries as lexical (inconsistently with MCD6) and artificially minimizing the number of heterosemous entries. To further enhance the quality of the dictionary, this study proposes that Chinese–Vietnamese lexicography should avoid over-reliance on authoritative Modern Chinese dictionaries and instead adopt corpus-based evidence to accurately represent authentic Modern Chinese usage patterns, potentially enhancing the dictionary's practical value for learners and translators.

Nevertheless, several limitations of this study should be acknowledged. First, the case study focuses on "X变" entries; extending this analysis to a broader range of heterosemous types will further strengthen the generalizability of the findings. Second, we adopt a binary lexical/non-lexical distinction to align with NCVD's system, yet a three-way categorization (lexical, non-lexical, mixed; see Wang 2011) would better capture polysemy — a direction for future research.

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2181 (Article)

Third, while the CCL corpus provides rich data, it may contain unverified social media content; cross-validation with print sources such as *Cihai* or *Ciyuan* is therefore recommended. Finally, although the present study focuses on Chinese word-class labeling, its findings have implications for Vietnamese learners of Chinese. Future research should investigate how this learner group uses word-class information in dictionary consultation.

Endnotes

1. The 25 numeral-measure entries are "百倍(bǎibèi) (hundredfold), 半点儿(bàndiǎnr) (a little bit), 半截儿(bànjiér) (half section), 半晌(bànshǎng) (half morning), 半天(bàntiān) (half day), 两次(liǎngcì) (twice), 俩(liǎ) (two), 千斤(qiānjīn) (thousand catties), 仨(sā) (three), 首次(shǒucì) (first time), 首届(shǒujiè) (first session), 万贯(wànguàn) (ten thousand strings of cash), 万年(wànnián) (ten thousand years), 万顷(wànqǐng) (ten thousand acres), 万丈(wànzhàng) (ten thousand feet), 一帮(yìbāng) (a gang), 一点儿(yìdiǎnr) (a little), 一度(yídù) (once), 一刻(yíkè) (a moment), 一溜儿(yíliùr) (a streak), 一丝(yìsī) (a trace), 一下(yíxià) (a moment), 一线(yīxiàn) (a line), 一些(yìxiē) (some), 一阵(yízhèn) (a burst)".
2. This paper refers to the entries and their meanings labeled with word classes as "lexical entries".
3. This paper refers to the entries and their meanings not labeled with word classes as "non-lexical entries", including morphemes and phrases.

Acknowledgements

We are grateful to Prof. Elsabé Taljard and two anonymous reviewers for their insightful comments and valuable suggestions. All remaining errors are our own. Correspondence should be addressed to Ni Gong.

This study is supported by the National Social Science Foundation of China (Grant No. 22XYY021) for the project "Quantum Linguistics: Theory and Practice".

References

Aerts, D. and L. Beltran. 2022. Are Words the Quanta of Human Language? Extending the Domain of Quantum Cognition. *Entropy* 24(1): 6. https://doi.org/10.3390/e24010006

Atkins, B.T.S. and M. Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Beckner, C., R.A. Blythe, J. Bybee, M.H. Christiansen, W. Croft, N.C. Ellis, J. Holland, J. Ke, D. Larsen-Freeman and T. Schoenemann. 2009. Language Is a Complex Adaptive System: Position Paper. *Language Learning* 59(s1): 1-26. https://doi.org/10.1111/j.1467-9922.2009.00533.x

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2181 (Article)

Bisang, W. 2013. Word-Class Systems between Flexibility and Rigidity: An Integrative Approach. Rijkhoff, J. and E. van Lier (Eds.). 2013. *Flexible Word Classes: Typological Studies of Underspecified Parts of Speech:* 275-303. Oxford: Oxford University Press.

Busemeyer, J.R. and P.D. Bruza. 2025. *Quantum Models of Cognition and Decision*. Second edition. Cambridge: Cambridge University Press.

Bybee, J. 2010. *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

Chomsky, N. 1970. Remarks on Nominalization. Jacobs, R.A. and P.S. Rosenbaum (Eds.). 1970. *Readings in English Transformational Grammar:* 184-221. Washington, DC: Georgetown University Press.

Dalpanagioti, T. 2019. From Corpus Usages to Cognitively Informed Dictionary Senses: Reconstructing an Mld Entry for the Verb Float. *Lexicography* 6(2): 75-104. https://doi.org/10.1007/s40607-019-00059-5

De Saussure, F. 1916/1959. *Course in General Linguistics.* (Translated by W. Baskin.) New York: Philosophical Library.

Guo, Rui. 2018. *Modern Chinese Parts of Speech: Systems Research.* London/New York: Routledge.

Hanks, P. 2013. *Lexical Analysis: Norms and Exploitations*. Cambridge, MA: MIT Press.

Honey, P.J. 1956. Word Classes in Vietnamese. *Bulletin of the School of Oriental and African Studies* 18(3): 534-544. https://doi.org/10.1017/S0041977X00088029

Jackson, H. 2022. *The Bloomsbury Handbook of Lexicography*. London: Bloomsbury Academic.

Lehmann, C. 2013. The Nature of Parts of Speech. *STUF — Language Typology and Universals* 66(2): 141-177. https://doi.org/10.1524/stuf.2013.0008

Li, Xiang. 2013. Analysis of Word-Class Labeling in the *Chinese–English Dictionary.* Third edition. *Lexicographical Research* 5: 53-59. https://doi.org/10.16134/j.cnki.cn31-1997/g2.2013.05.016

Lu, Jianming. 2024. *On the Classification of Words in Chinese*. Beijing: The Commercial Press.

Munro, P. 2005. From Parts of Speech to the Grammar. *Studies in Language* 30(2): 307-349. https://doi.org/10.1075/sl.30.2.07mun

Ptasznik, B. 2020. Which Defining Model Contributes to More Successful Extraction of Syntactic Class Information and Translation Accuracy? *Lexikos* 30: 363-385. https://doi.org/10.5788/30-1-1545

Qi, Guangmou. 2013. *New Chinese–Vietnamese Dictionary*. Nanning: Guangxi Education Press.

Shen, Jiaxuan. 2016. *Nouns and Verbs.* Beijing: The Commercial Press.

Spike, M. 2020. Fifty Shades of Grue: Indeterminate Categories and Induction in and out of the Language Sciences. *Linguistic Typology* 24(3): 465-488. https://doi.org/10.1515/lingty-2020-2061

Tan, Jingchun. 2024. Promoting Effects of Part-of-Speech Tagging on Dictionary Definition. *Chinese Linguistics* 4: 80-91.

Van Lier, E. 2023. *The Oxford Handbook of Word Classes.* Oxford: Oxford University Press.

Vapnarsky, V. and E. Veneziano. 2017. *Lexical Polycategoriality: Cross-linguistic, Cross-theoretical and Language Acquisition Approaches.* Amsterdam/Philadelphia: John Benjamins.

Wang, Renqiang. 2011. A Cognitive Study of Wordwood in Modern Chinese Based on *the Contemporary Chinese Dictionary.* Fifth edition. *Foreign Language and Literature* 1: 71-77. Available from: https://kns.cnki.net/kcms2/article/abstract?v=SCWY201101015

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2181 (Article)

Wang, Renqiang. 2014a. The Two-Level Word Class Categorization in Analytic Languages. *Proceedings of the 36th Annual Conference of the German Linguistic Society，5–7 March 2014, Philipps University Marburg, Germany:* 345-347. Marburg: University of Marburg. Available from: https://kns.cnki.net/kcms2/article/abstract?v=WGYJ201502003

Wang, Renqiang. 2014b. A Study of Word Class Labeling in *Concise Chinese–German Dictionary*. *Contemporary Foreign Language Studies* 11: 37-41. https://doi.org/10.3969/j.issn.1674-8921.2014.11.007

Wang, Renqiang. 2015. Research on Word Class Labeling in Chinese–Italian Bilingual Learning Dictionary: A Case Study of *Chinese–Italian Bilingual Dictionary*. *Foreign Language and Literature Research* 4: 20-40.

Wang, Renqiang. 2020. A Big Data Perspective on the Representation Strategy of Heterosemy in *A Chinese–English Dictionary.* Third edition. *Technology Enhanced Foreign Language Education* 6: 71-79.

Wang, Renqiang. 2022. The Methodological Dilemma in the Study of Word Classes in Scientism: The Second Paper in the Series on the Quantum Turn in Linguistics. *Foreign Language Education* 1: 9-16. https://doi.org/10.16362/j.cnki.cn61-1023/h.2022.01.002

Wang, Renqiang. 2023. The Transdisciplinary Methodology of the Two-level Lexical Categorization Theory. *Foreign Language Education* (1): 8-16. https://doi.org/10.16362/j.cnki.cn61-1023/h.2023.01.004

Wang, Renqiang and Rui Yin. 2023. The Word Class Labeling of Medical Entries in Chinese–English Dictionaries: A Transdisciplinary Perspective. *English Studies* 2: 164-177.

Wang, Renqiang and Shiyu Zhao. 2025. Wave-Particle Duality of Lexical Categories in Chinese "X击(X jī)" Parallel Verb Compounds: Quantum Thinking and Corpus Evidence. *Foreign Languages Research* 4: 41-50+113. https://doi.org/10.13978/j.cnki.wyyj.2025.04.013

Wu, Ming and Renqiang Wang. 2022. Research on the Word Class Labeling of Time Entries in Modern Chinese. *Foreign Languages Research* 3: 45-52. https://doi.org/10.13978/j.cnki.wyyj.2022.03.005

Yang, Xu. 2025. Word Class Labeling of "X化(X huà)"and Related Issues. *Essays on Linguistics* 1: 146-154.

Yang, Xu and Renqiang Wang. 2018. Investigating the Representation Strategies of Multi-Category Lexemes in the *Contemporary Chinese Dictionary.* Sixth edition. *Journal of Guangdong University of Foreign Studies* 4: 5-13. https://doi.org/10.3969/j.issn.1672-0962.2018.04.001

Zhu, Dexi. 1985. *Grammatical Questions and Answers*. Beijing: The Commercial Press.

Zhu, Dexi. 2010. *Lectures on Grammatical Analysis*. Beijing: The Commercial Press.