Pertanika J. Soc. Sci. & Hum. 34 (2): 995 - 1012 (2026)

PERTANIKA JOURNALS

SOCIAL SCIENCES & HUMANITIES

Journal homepage: http://www.pertanika.upm.edu.my/

An Analysis of the Vietnamese Dictionary from a Computational Linguistics Perspective

Trang Thi My Phan[1,2]*, Hai Van Ba Phan[3], Tri Quoc Do[3], Dien Dinh[3], and Phuong Thi Minh Tran[4]

[1]Faculty of Literature and Linguistics, University of Social Sciences and Humanities (USSH-VNUHCM) - HCMC National University, 10-12 Đinh Tiên Hoàng street, Sài Gòn ward, 70000 Hồ Chí Minh city, Vietnam
[2]Faculty of Fundamental Sciences, Saigon Technology University, 180 Cao Lỗ street, Chánh Hưng ward, 70000 Hồ Chí Minh city, Vietnam
[3]Computational Linguistics Center (CLC), University of Science (US-VNUHCM) - HCMC National University, 227 Nguyễn Văn Cừ street, Chợ Quán ward, 70000 Hồ Chí Minh city, Vietnam
[4]Faculty of English Linguistics and Literature, University of Social Sciences and Humanities (USSH-VNUHCM) - HCMC National University, 10-12 Đinh Tiên Hoàng street, Sài Gòn ward, 70000 Hồ Chí Minh city, Vietnam

ABSTRACT

Dictionaries are essential resources for exploring a language's lexicon, providing insights into word formation, usage, and linguistic relationships. With the advancement of computational linguistics, applying statistical methods to dictionary data enables researchers to discover the lexical characteristics of a language. This study explored the *Vietnamese Dictionary* from a computational linguistics perspective, applying statistical techniques like frequency analysis, part-of-speech (POS) distribution analysis, multi-POS (words that can function as more than one part of speech) coefficient analysis, and polysemy coefficient analysis to investigate letter distribution, POS characteristics, and polysemy levels. The findings indicate that the most frequently occurring letters are *n, h, a, i, t, g, c*, and *u*, while letters like *q, x, d, v, e, s, ă, k,* and *r* occur less frequently. Letters like *t, c, n, đ, b, l,* and *h* occur most often in initial positions. Nouns account for the largest proportion of lexical entries (44.7%), followed by verbs (31.58%) and adjectives (21.22%). The multi-POS coefficient analysis shows that 90.11% of words have one part of speech, 8.84% can function in two, and fewer than 1% span three or more, highlighting the low syntactic flexibility of the Vietnamese lexicon in terms of POS variation. Polysemy coefficient analysis indicates that particles, pronouns, and verbs exhibit the highest degrees of polysemy. These findings reveal the distributional characteristics

ARTICLE INFO

Article history:
Received: 27 May 2025
Accepted: 10 April 2026
Published: 30 April 2026

DOI: https://doi.org/10.47836/pjssh.34.2.20

E-mail addresses:
trang.phanthimy@stu.edu.vn (Trang Thi My Phan)
phanvanbahai@gmail.com (Hai Van Ba Phan)
doquoctri2003@gmail.com (Tri Quoc Do)
ddien@fit.hcmus.edu.vn (Dien Dinh)
minhphuongtrn@hcmussh.edu.vn (Phuong Thi Minh Tran)
* Corresponding author

of the Vietnamese lexicon through statistical analysis, providing a valuable foundation for further research in lexical semantics, electronic dictionaries, part-of-speech tagging tools, and natural language processing applications.

Keywords: Computational linguistics, letter distribution, part-of-speech distribution, polysemy coefficient, Vietnamese dictionary.

INTRODUCTION

A dictionary is an essential resource for systematising lexical information, supporting both linguistic research and language teaching (Carstens, 1995; Cote González & Tejedor Martínez, 2011; Ezeh et al., 2022). Each dictionary entry not only provides information on phonetics, grammar, and semantics but also reflects a language's cultural and historical characteristics. Among the lexicographic works in Vietnam, the *Vietnamese Dictionary* compiled by the Institute of Linguistics and edited by Hoàng Phê (Hoàng, 2020), is a widely used resource in Vietnamese language research and education. However, previous uses of this dictionary have mostly been manual, without taking advantage of computational linguistic tools for systematic analysis.

Computational linguistics is an interdisciplinary field that combines computer science and linguistics, offering a new approach to dictionary analysis (Church & Liberman, 2021; Yang, 2023). This approach integrates automatic word form recognition, frequency analysis, and corpus-based research to enhance the development of linguistic resources (Hausser, 2001). With the support of modern tools, large-scale linguistic data can be processed with high precision, broadening the scope of observation and allowing for data-driven conclusions. However, in Vietnam, few studies have utilised natural language processing (NLP) techniques to systematically analyse the characteristics of Vietnamese vocabulary (Huynh et al., 2022; Thin et al., 2023, 2024). To address this gap, this study employed computational linguistic methods to analyse the *Vietnamese Dictionary* using the *Python* programming language and tools such as *NumPy* and *pandas* for statistical analysis, and *matplotlib* for data visualisation. Specifically, this research investigated the following questions:

1. What is the frequency distribution of letters and initial letters in Vietnamese lexical entries?
2. How are parts of speech distributed in the Vietnamese Dictionary?
3. What is the average multi-POS coefficient in the Vietnamese Dictionary, and which part-of-speech combinations are the most and least common in Vietnamese lexical entries?
4. What is the average polysemy coefficient of parts of speech in the Vietnamese Dictionary, and how is polysemy distributed within each part of speech?

By addressing these questions, this research not only helps identify distinctive linguistic features of Vietnamese but also contributes to the digitisation of Vietnamese linguistic data, particularly dictionary information, within the context of globalisation and international integration. The findings will provide insights into the Vietnamese lexicon and highlight the potential of computational linguistics in advancing lexicographic studies.

MATERIALS AND METHODS

The research process consists of four main stages: (1) collecting linguistic data and converting it into a tabular format, (2) preprocessing and cleaning the data, (3) creating the survey dataset, and (4) conducting statistical analysis and visualising the results (see Figure 1).

Data Collection and Tabular Conversion

The primary data source used in this study is the *Vietnamese Dictionary*, edited by Hoàng Phê and issued in the 2020 reprinted edition (Hoàng, 2020). The structure of the *Vietnamese Dictionary* includes various fields, such as *word (mục từ), Hanzi (Hán tự), part of speech (từ loại), definition (định nghĩa), usage example (ví dụ sử dụng), reference words (từ liên kết), synonyms (từ đồng nghĩa),* and *antonyms (từ trái nghĩa).* Although the dictionary was originally published in print, it was later digitised and organised into a structured, field-based format suitable for computational analysis. In the digitised version, each lexical entry is encoded as an XML structure with explicit tags for its lexical components (see Figure 2). A sample of the digitised *Vietnamese Dictionary* in XML format is available from CLC (2026).

The part-of-speech system is encoded using the following abbreviations: *noun (d: danh từ), verb (đg: động từ), adjective (t: tính từ), pronoun (đ: đại từ), adjunct (p: phụ từ), conjunction (k: kết từ), particle (tr: trợ từ),* and *interjection (c: cảm từ)*. Finally, all entries were converted into

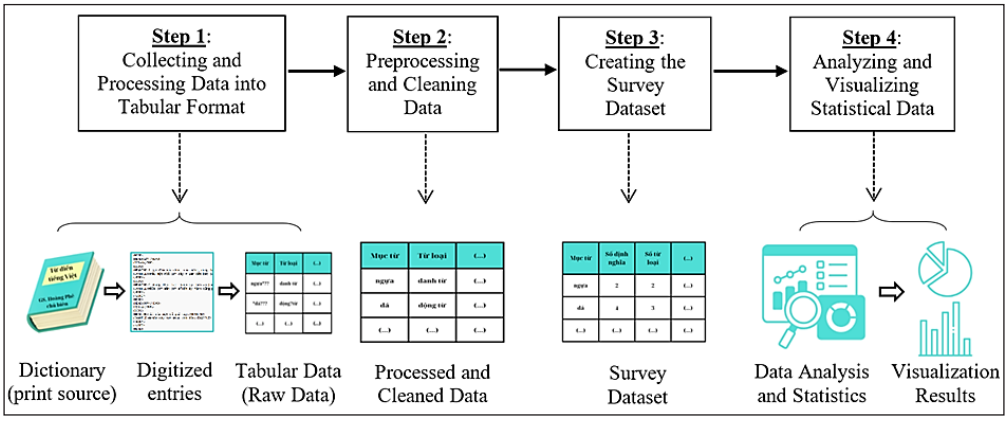

Figure 1. Overview diagram of the processing workflow

```
<WORD>
<HEAD>sách</HEAD>
<HANZI>冊</HANZI>
<POS>d</POS>
<BODY>
<MEANING>tập hợp một số lượng nhất định những tờ giấy có chữ in, đóng gộp lại thành quyển</MEANING>
<EXAMPLE>#sách tham khảo #đọc sách tới khuya #hiệu sách</EXAMPLE>
</BODY>
</WORD>
```

Figure 2. Sample digitised dictionary entry with tagged fields
Source: Hoàng (2020) and CLC (2026)

a tabular dataset (*csv* or *xlsx*) to support automated processing and quantitative analysis.

Data Preprocessing and Cleaning

Data preprocessing is performed to ensure the accuracy of statistical analysis. The preprocessing steps include:

Step 1. Field Selection for Analysis

This step involves filtering the dataset to retain only the fields required for the statistical analyses within the scope of this paper, specifically the *headword* and *part-of-speech (POS)* information. Other fields are not deleted from the original dataset but are excluded from the analytical subset to maintain focus and ensure the accuracy of the reported results.

Step 2. Exclusion of Invalid or Out-of-Scope Entries

Entries that are not suitable for analysis in this study are excluded from the analytical subset. These include:

- Linguistic units larger than words such as idioms or proverbs (e.g., the idiom *sách gối đầu giường*)
- Non-lexical items such as symbols or special characters
- Entries with missing essential fields (e.g., missing *headword* or *definition*)

Since the present study focuses exclusively on the word as the basic unit of analysis, such entries fall outside the scope of processing and are therefore excluded from further analysis.

Step 3. Filtering Non-native Vietnamese Words

Applying the phonetic structure rules of Vietnamese syllables (Đoàn, 2007; Mai et al., 2008) (Table 1) along with the algorithm for verifying Vietnamese syllables based on their structural rules (Định & Lương, 2004) to identify and retain native Vietnamese words. The entire dataset was then manually reviewed to ensure accuracy.

Step 4. Standardising Lexical Morphology

This process includes unifying cases where *i* and *y* are interchangeable by Standardising them to *i* and adjusting the placement of diacritical marks to align with standard conventions.

Table 1
Vietnamese syllable structure

Onset (Âm đầu)	Tone (Thanh điệu)		
	Rhyme (Vần)		
	Glide (Âm đệm)	Nucleus (Âm chính)	Coda (Âm cuối)

Source: Đoàn (2007) and Mai et al. (2008)

Step 5. Storing Clean Data

Once the cleaning process is complete, the refined data is stored in *csv* or *xlsx* format for further analysis. A sample dataset in Excel format can be accessed via CLC (2026).

Survey Dataset Creation

The research team constructs a survey dataset consisting of a list of lexical entries, focusing on analysing letter distribution, part-of-speech characteristics, and polysemy levels.

Statistical Data Analysis and Visualisation

The data is analysed using *Python* (McKinney, 2017), with support from the *NumPy* and *pandas* libraries for efficient data processing and computation, as well as *matplotlib* for result visualisation.

Frequency Analysis

Lexical frequency distribution (Everitt & Skrondal, 2010) is calculated as the percentage ratio of an element's occurrences to the total number of elements in the dataset:

$$f(x) = \frac{N(x)}{N_{total}} \quad [1]$$

Where:

- f(x) is the frequency of the analysed element
- N(x) is the number of occurrences of the analysed element
- N_{total} is the total number of elements in the dataset

Part-of-Speech Distribution Analysis

In this statistical analysis, alongside the frequency statistics method, the research team also incorporated the combinatorial statistics method. Since each lexical entry may belong to one or more parts of speech, all parts of speech for each lexical entry were grouped into a single set, with duplicate combinations removed. Each unique part-of-speech combination was then listed and arranged according to its frequency of occurrence.

Multi-POS Coefficient Analysis

In order to examine the multi-POS coefficient (Everitt & Skrondal, 2010), the study determined the number of different parts of speech that each lexical entry possesses. The coefficient is calculated using the following formula:

$$Coeff = \frac{\sum_x^N pos(x)}{N} \quad [2]$$

Where:

- *Coeff* is the multi-POS coefficient
- *x* represents a lexical entry
- *pos(x)* is the number of parts of speech associated with a lexical entry
- *N* is the total number of lexical entries in the dataset

Polysemy Coefficient Analysis

To analyse the polysemy coefficient (Everitt & Skrondal, 2010), the study filtered data based on parts of speech, then determined the number of meanings each lexical entry has and calculated it using the following formula:

$$Coeff_{pos} = \frac{\sum_{x}^{N_{pos}} meaning_{pos}(x)}{N_{pos}} \quad [3]$$

Where:

- $Coeff_{pos}$ is the polysemy coefficient based on part of speech
- *Pos* is the part of speech being analysed
- N_{pos} is the set of lexical entries categorized under the given part of speech
- $meaning_{pos}(x)$ is the number of definitions of a lexical entry for the specified part of speech (pos)

RESULTS AND DISCUSSION

Based on the applied methods, the data were collected, processed, and analysed, providing the following statistical results on lexical entries.

Distribution of Letters in the Vietnamese Dictionary

Frequency Distribution of Letters in Lexical Entries

The frequency analysis of letters across all lexical entries in the *Vietnamese Dictionary* (Hoàng, 2020) reveals that the most frequently occurring letters include: *n* (29,356 times), *h* (24,262 times), *a* (18,456 times), *i* (17,391 times), *t* (17,333 times), *g* (13,729 times), *c* (13,085 times), and *u* (12,402 times) (Figure 3). Collectively, these eight letters account for approximately 60.6% of all letter occurrences in the dataset.

The high frequency of letters such as *n* and *h* is largely due to their broad grapheme distribution and frequent participation in multi-letter graphemes (i.e., some graphemes are composed of two or more letters), including *nh /ɲ, ŋ/, ng /ŋ/, ngh /ŋ/, th /tʰ/, ph /f/, kh /χ/*, and *gh /ɣ/*. These letters can be found in initial or final positions, and each occurrence of a multi-letter grapheme contributes to the counts of its component letters, increasing overall frequency. For instance, *n* appears at the beginning (*này, năm, nếu*), at the end (*bàn, nhân, cơn*), and in multi-letter graphemes like *ng* and *ngh (same phoneme /ŋ/)*. Similarly, *h* occurs in words such as *hành, hoa, học, hỏi, hiểu,* and frequently in multi-letter graphemes such as *nh* in *nhà*, *th* in *thơ*, *ph* in *phở*, and *kh* in *khỏe*. This observation aligns with the findings of Đinh and Đỗ (2015), who noted that *n* and *h* occur in multiple initial consonant positions *(n-, nh-, ng-, ngh-)* and final positions *(-n, -ng, -nh)*, resulting in their high frequency of occurrence.

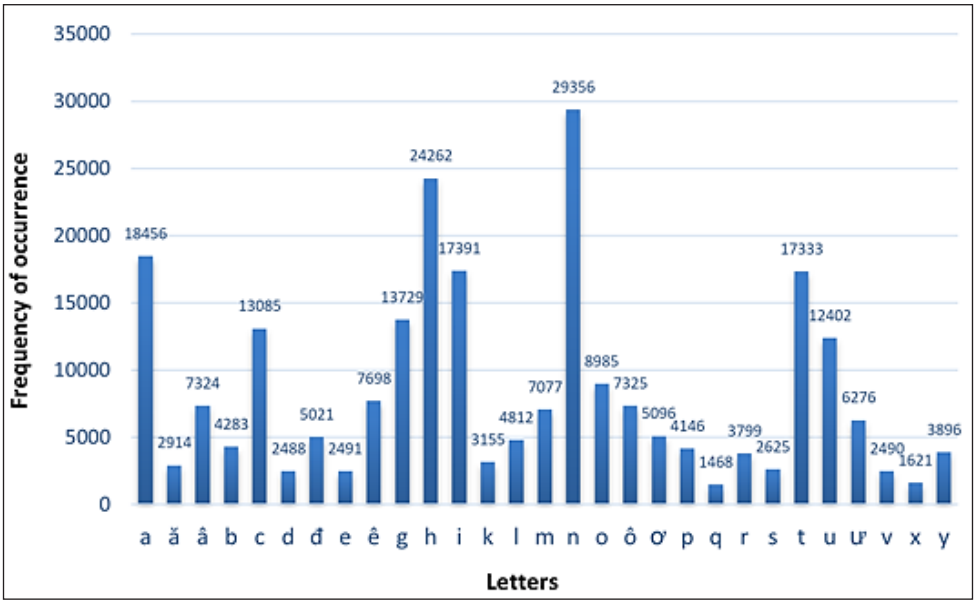

Figure 3. Frequency distribution of letters in Lexical Entries

Other high-frequency letters, including *t*, *g*, and *c*, also show high counts due to their occurrence in various word positions and common multi-letter graphemes. For instance, *t* appears at the beginning (*ta, tôi, tên*) and at the end of words (*bắt, đặt, viết*). The letter *g* is rarely found at word endings; its high frequency mainly comes from initial positions *(gạo, gà, gan, gặp)* and multi-letter graphemes like *ng* and *ngh (ngày, người, nghèo)*. The letter *c* occurs both at the beginning *(cá, cây, cỏ, con, cơm, cười)* and at the end of words *(bác, bọc, nhạc, lục, tóc)*, and frequently in multi-letter graphemes such as *ch (cháu, chị, chó, chuối)*.

The letters *a*, *i*, and *u* are widely distributed because they appear in many common words and frequently participate in multi-letter graphemes. For instance, *a* occurs in words like *anh, ai, áo, ba, bà, cha, cá, nhà*; *i* appears in *in, ít, tin, sinh*; and *u* is found in *út, thu, cung, lũ, trúc*. They also frequently form multi-letter graphemes such as *ai, ui,* and *iu*. This finding aligns with the analysis of Đinh and Đỗ (2015), who showed that *a, i,* and *u* can function as either nuclear vowels or final vowel elements, which contributes to their higher frequency of occurrence.

Conversely, lower-frequency letters such as *q (1,468 times), x (1,621 times), d (2,488 times), v (2,490 times), e (2,491 times), s (2,625 times), ă (2,914 times), k (3,155 times),* and *r (3,799 times)* are constrained by orthographic rules and positional restrictions. For instance, *q* appears almost exclusively in the multi-letter grapheme *qu* (e.g., *quan, quân, quốc, quý*), whereas *k* occurs mainly before *i* or *e* (e.g., *kiến, kẹo*). The letter *ă* appears in relatively few words compared with *a* (e.g., *ăn, mặc, bắt, trắng*), which makes

it much less frequent than *a*. Similarly, *e* is often replaced by *ê* in many common words (*bên, lên, tên, mền*), which limits the number of words in which *e* itself occurs. The letters *x, s, d, v,* and *r* mostly appear at the beginning of words (e.g., *xâm; sáng, sông; dạy; vào, vui; rau, rừng*) and rarely participate in multi-letter graphemes. These orthographic and positional constraints collectively account for the lower frequency of these letters compared with more widely distributed ones.

Frequency Distribution of the Initial Letter in Lexical Entries

In addition to analysing the frequency of letters across all lexical entries, the study also examines the frequency of the first letter in each entry (Figure 4). The results indicate that the most common initial letters are *t* (6,353 words), *c* (4,314 words), *n* (3,005 words), *đ* (2,564 words), *b* (2,451 words), *l* (2,193 words), and *h* (2,127 words). From a grapheme perspective, these figures indicate that dictionary headwords are unevenly distributed across initial letters, with a small number of initials occurring much more frequently than others. However, it should be noted that the statistics in Figure 4 consider only the first letter of each headword according to the Vietnamese alphabetical order. Therefore, the data reflect the distribution of initial letters used for dictionary indexing, rather than the actual distribution of multi-letter graphemes at the beginning of words. The high frequency of certain initials is partly due to the fact that common multi-letter graphemes are grouped under a single initial. For instance, headwords beginning with *th-* or *tr-* are counted under *t*; those beginning with *ch-* are counted under *c*; *gi-* and *gh-* are counted under *g*; *kh-* under *k*; *nh-, ng-*, and *ngh-* under *n*; and *ph-* under *p*.

The study further disaggregates several high-frequency initial-letter categories into their corresponding word-initial multi-letter graphemes (i.e., sequences of two or more letters appearing at the beginning of a word) (Table 2). In the *c* category (4,314 entries), the two word-initial units *c-* and *ch-* are relatively balanced, with *c-* accounting for 2,436 entries (56.47%) and *ch-* for 1,878 entries (43.53%). The *g* category (1,334 entries) shows a markedly uneven distribution: *gi-* dominates with 818 entries (61.32%), while *g-* accounts for 437 entries (32.76%) and *gh-* is rare with only 79 entries (5.92%). In the *k* category (1,630 entries), *kh-* represents the majority with 918 entries (56.32%), compared with 712 entries for *k-* (43.68%). The *n* category (3,005 entries) consists of four word-initial units. The single-letter unit *n-* accounts for 911 entries (30.32%). Multi-letter graphemes make up the majority with 2,094 entries (69.68%), including *nh-* with 1,027 entries (34.17%), *ng-* with 850 entries (28.29%), and *ngh-* with 217 entries (7.22%). A particularly skewed distribution is observed in the *p* category (1,083 entries), in which *p-* appears in only 4 entries (0.37%) and *ph-* accounts for almost all cases (1,079 entries, 99.63%). Finally, in the *t* category (6,353 entries), *t-* is most frequent with 2,888 entries (45.46%), followed by *th-* (2,250 entries, 35.42%) and *tr-* (1,215 entries, 19.12%).

In the lower-frequency portion of the distribution (Figure 4), vowel letters occur much less frequently as the initial letter of words. Letters such as *ê* (40 words), *ơ (*40 words)*, e* (52 words), and *i* (70 words) rarely appear in the initial position, indicating that vowel-initial headwords constitute only a small proportion of the dictionary. This distribution reflects the general pattern in Vietnamese orthography, in which an initial consonant letter or consonant cluster is predominant (Đinh & Đỗ, 2015), whereas vowel-initial headwords occur less frequently.

In addition, vowel letters with diacritics (e.g. *ê, ơ*) appear less frequently at the beginning of words than more common vowels, which further reduces their frequency in the initial position. Most words beginning with these vowels are either monosyllabic or onomatopoeic, such as *ơ (kìa), ê (ê a), e (em, eo),* and *im (lặng)*.

Distribution of Part of Speech in the Vietnamese Dictionary

The distribution of part of speech in the *Vietnamese Dictionary* reflects the characteristics of an isolating language, where syntactic relationships rely primarily on word order and context rather than morphological inflection. The statistical results show a total of 49,438 definitions classified into major parts of speech (Table 3), in which nouns account for the highest proportion at 44.7% (22,098 definitions), followed by verbs at 31.58% (15,613 definitions) and adjectives at 21.22% (10,488 definitions). However, it should be noted that the relatively high proportion of nouns is not unique to Vietnamese; it is also a common outcome of general-purpose dictionary compilation, which aims to provide broad coverage of nameable entities and concepts (e.g., objects, institutions, proper names, etc.) that are typically listed

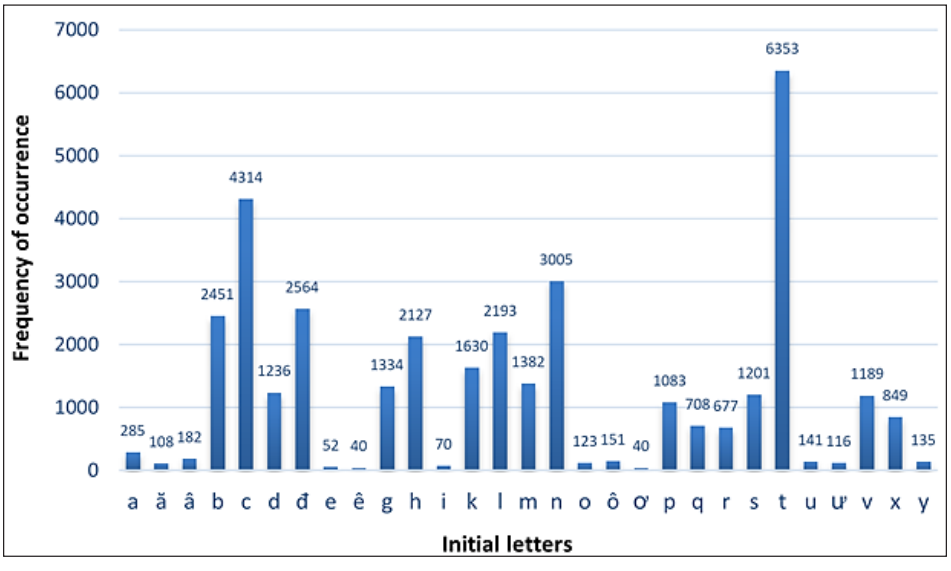

Figure 4. Frequency distribution of the initial letters in Lexical Entries

Table 2
Distribution of Vietnamese word-initial multi-letter Graphemes (ch, gh/gi, kh, ng/ngh/nh, ph, th/tr)

Alphabetic initial letter	Total count	Word-initial Multi-Letter Graphemes	Count	Percentage
c	**4314**	c	2436	56.47%
		ch	1878	43.53%
g	**1334**	g	437	32.76%
		gh	79	5.92%
		gi	818	61.32%
k	**1630**	k	712	43.68%
		kh	918	56.32%
n	**3005**	n	911	30.32%
		ng	850	28.29%
		ngh	217	7.22%
		nh	1027	34.17%
p	**1083**	p	4	0.37%
		ph	1079	99.63%
t	**6353**	t	2888	45.46%
		th	2250	35.42%
		tr	1215	19.12%

and indexed as nouns. Comparable evidence can be observed in the Kaikki-derived multilingual lexicon (machine-readable dictionaries extracted from Wiktionary using Wiktextract) (Ylönen, 2022), where nouns constitute the largest category in several languages, including English (934,609 noun senses), Russian (235,389 noun senses), Chinese (106,048 noun senses), Japanese (78,948 noun senses), Indonesian (32,723 noun senses), and Vietnamese (20,763).

Table 3
Number of definitions for each part of speech in the Vietnamese Dictionary

Part of Speech	Number of Definitions	Distribution Percentage (%)
Noun *(Danh từ)*	22,098	44.70%
Verb *(Động từ)*	15,613	31.58%
Adjective *(Tính từ)*	10,488	21.22%
Adjunct *(Phụ từ)*	587	1.19%
Conjunction *(Kết từ)*	233	0.47%
Pronoun *(Đại từ)*	180	0.36%
Particle *(Trợ từ)*	138	0.28%
Interjection *(Cảm từ)*	101	0.20%

Turning to the remaining major classes in Hoàng (2020), verbs make up 31.58% (15,613 definitions), demonstrating a rich verbal system. However, Vietnamese verbs do not inflect for tense, aspect, or mood as in inflectional languages. Instead, verb meanings depend on context, word order, and auxiliary words, such as *đã đi* (past) or *sẽ đi* (future). Adjectives account for a lower proportion at 21.22% (10,488 definitions) and are often accompanied by degree markers such as *rất* (very), *khá* (quite), or *hơi* (slightly), as in *rất đẹp* (very beautiful), *khá cao* (quite tall), *hơi lạnh* (slightly cold).

Auxiliary parts of speech, including adjuncts (1.19%), conjunctions (0.47%), pronouns (0.36%), particles (0.28%), and interjections (0.2%), occur at a lower frequency but play an essential role in syntax organisation, meaning expression, and communication. *Adjuncts* modify the degree, time, or direction of verbs and adjectives, such as *đã, đang, sẽ* indicating verb tense, or *rất, khá, hơi* modifying the intensity of adjectives. Conjunctions *(và, nhưng, bởi vì)* function as sentence connectors. Pronouns, including personal pronouns *(tôi, mình, chúng tôi)*, demonstrative pronouns *(này, kia)*, and indefinite pronouns *(ai, người ta)*, refer to participants or entities, show social relationships such as hierarchy, familiarity, and politeness, and maintain cohesion and coherence in discourse. *Particles (mà, nhé, chứ)* add modal meanings, emphasis, or speaker attitudes. Interjections *(ôi, chao, ồ)* express emotions.

Multi-POS Coefficient in the Vietnamese Dictionary

Statistical analysis of the *Vietnamese Dictionary* indicates a total of *35,739* distinct lexical entries. An examination of the multi-POS coefficient reveals that, on average, each entry has 1.11 parts of speech. Specifically, 32,204 entries (90.11%) belong to only one part of speech; 3,160 entries (8.84%) function as two parts of speech; 340 entries (0.95%) have three parts of speech; 32 entries (0.09%) exhibit four parts of speech; and only 3 entries (0.01%) possess five parts of speech (Table 4 and Figure 5).

The data indicate that although the Vietnamese lexicon allows for part-of-speech conversion, its flexibility remains

Table 4
Number of entries in the multi-POS distribution of the Vietnamese Dictionary

Number of parts of speech	Number of Lexical Entries
1	32,204
2	3,160
3	340
4	32
5	3
Total Lexical Entries	35,739
Multi-POS Coefficient	1.110

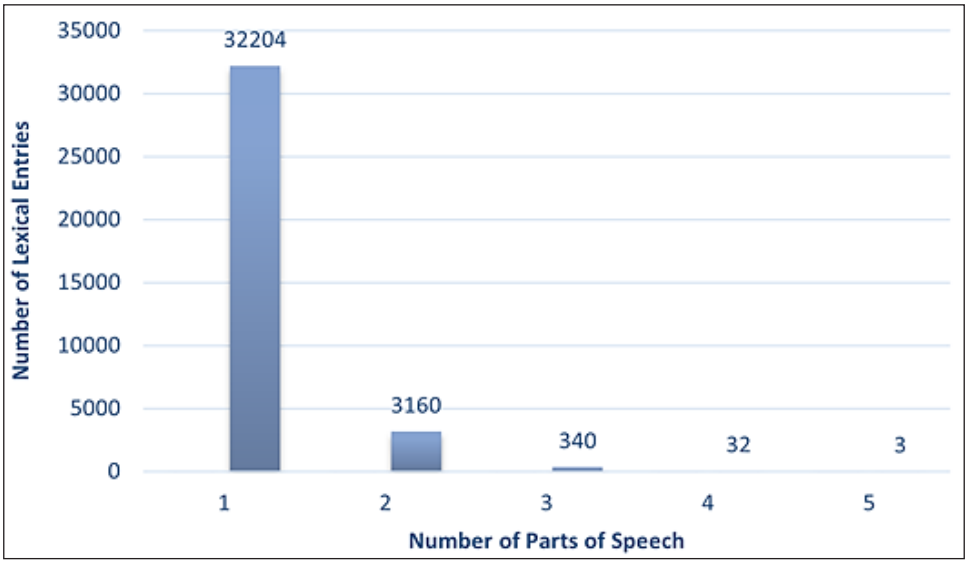

Figure 5. Distribution of Multi-POS in the Vietnamese Dictionary

relatively moderate. Vietnamese grammar primarily relies on context and word combinations to determine parts of speech. The limited number of lexical entries with three or more parts of speech reflects the stability of the word classification system in Vietnamese, which aligns with the characteristics of an isolating language, where grammatical relationships depend on word order rather than inflection. Part-of-speech conversion in Vietnamese mainly occurs between nouns, verbs, and adjectives. For example, the verb *ăn* in the sentence *Tôi ăn cơm* (I have a meal) can be nominalised as *bữa ăn* in *Một bữa ăn ngon* (A delicious meal), or the adjective *đẹp* in *Cô ấy đẹp* (She is beautiful) can function as a noun in *cái đẹp* (beauty). In contrast, English exhibits a higher degree of part-of-speech conversion. For instance, *run* can function as a verb

(I run every day), a noun (go for a run), or an adjective (a run-down building).

Additionally, a statistical analysis of part-of-speech combinations was conducted. The chart in Figure 6 shows that the group of words functioning as both nouns and verbs *(d, đg)* accounts for the highest proportion at 38.78%, indicating that many verbs in Vietnamese can be nominalised, such as *học (*study)→ *việc học* (the act of studying). Following this, the noun-adjective group *(d, t)* constitutes 22.74%, demonstrating the nominalisation potential of adjectives, such as *tốt* (good) → *điều tốt* (a good thing). The verb-adjective group *(đg, t)* makes up 21.33%, illustrating that some adjectives can function as verbs to express actions. For instance, the adjective *lạnh* (cold) can act as a verb in the sentence *Nước lạnh đi* (The water gets colder). Only 5.8% of lexical entries assume all three roles (*d, đg, t*) - noun, verb, and adjective. Other less common combinations, including those involving primary parts of speech with auxiliary parts of speech, account for a total of 11.35%.

A detailed analysis (Table 5) reveals several notable characteristics of part-of-speech combinations in Vietnamese. First, rare combinations tend to appear only once or twice across the dataset, such as *(đ, t), (p, k, tr)*, or *(c, đg, tr, d)*. This pattern suggests that not all parts of speech in Vietnamese can combine freely. In particular, auxiliary categories, including *adjunct (p)*, *conjunction (k)*, and *interjection (c)*, exhibit very low combination frequency. The primary reason for this limitation lies in their syntactic properties. *Adjuncts (p)* primarily accompany verbs or adjectives to modify meaning, rather than forming independent lexical units. *Conjunctions (k)* function as connectors, linking clauses or phrases but lacking independent lexical

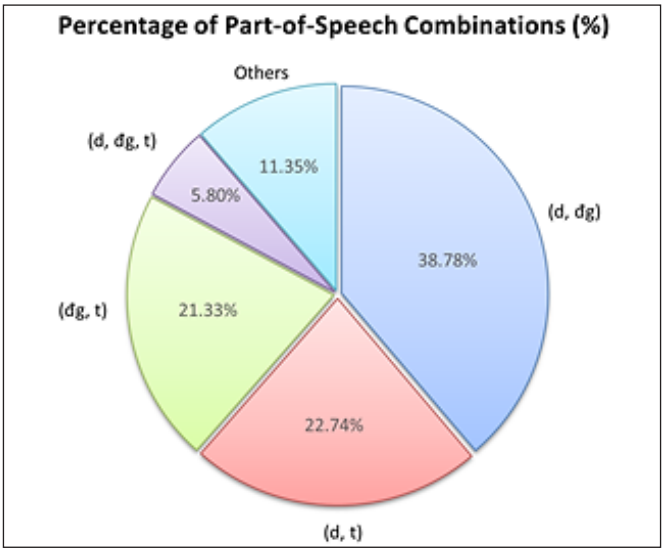

Figure 6. Distribution of part-of-speech combinations

Table 5
Distribution of part-of-speech combinations in the Vietnamese Dictionary

No.	Part-of-Speech Combination	Quantity	Percentage (%)	No.	Part-of-Speech Combination	Quantity	Percentage (%)
1	(d, đg)	1,371	38.78	38	(đg, t, tr)	4	0.11
2	(d, t)	804	22.74	39	(đ, đg, t)	3	0.08
3	(đg, t)	754	21.33	40	(c, đg, d)	3	0.08
4	(p, t)	51	1.44	41	(d, đ, tr)	2	0.06
5	(p, đg)	49	1.39	42	(c, đ, tr)	2	0.06
6	(d, p)	26	0.74	43	(d, p, tr)	2	0.06
7	(k, đg)	17	0.48	44	(d, k, tr)	2	0.06
8	(d, đ)	14	0.4	45	(d, p, đ)	2	0.06
9	(d, k)	11	0.31	46	(d, k, t)	2	0.06
10	(c, đg)	11	0.31	47	(c, t, d)	2	0.06
11	(đg, tr)	8	0.23	48	(đ, t, tr)	1	0.03
12	(d, tr)	7	0.2	49	(c, tr, d)	1	0.03
13	(k, t)	5	0.14	50	(d, đ, k)	1	0.03
14	(đ, đg)	5	0.14	51	(p, k, tr)	1	0.03
15	(đ, tr)	4	0.11	52	(c, đ, d)	1	0.03
16	(c, t)	4	0.11	53	(d, p, k)	1	0.03
17	(p, tr)	4	0.11	54	(c, đg, t)	1	0.03
18	(t, tr)	4	0.11	55	(p, k, đg)	1	0.03
19	(c, d)	3	0.08	56	(p, k, t)	1	0.03
20	(c, tr)	2	0.06	57	(c, p, đg)	1	0.03
21	(p, k)	2	0.06	58	(d, p, đg, t)	4	0.11
22	(đ, t)	1	0.03	59	(d, đg, k, t)	4	0.11
23	(c, đ)	1	0.03	60	(d, đ, đg, tr)	2	0.06
24	(đ, p)	1	0.03	61	(c, đg, tr, d)	2	0.06
25	(k, tr)	1	0.03	62	(d, p, đg, đ)	2	0.06
26	(d, đg, t)	205	5.8	63	(d, đg, k, tr)	2	0.06
27	(d, p, t)	25	0.71	64	(p, đg, t, tr)	2	0.06
28	(d, p, đg)	20	0.57	65	(d, p, k, đg)	2	0.06
29	(p, đg, t)	14	0.4	66	(đ, p, t, tr)	1	0.03
30	(d, k, đg)	9	0.25	67	(d, k, t, tr)	1	0.03
31	(d, đ, đg)	5	0.14	68	(d, p, t, tr)	1	0.03
32	(đg, k, t)	5	0.14	69	(d, đg, t, tr)	1	0.03
33	(d, đg, tr)	4	0.11	70	(d, đ, k, đg)	1	0.03
34	(p, t, tr)	4	0.11	71	(đ, p, đg, tr)	1	0.03
35	(d, t, tr)	4	0.11	72	(c, p, đg, t)	1	0.03
36	(d, đ, t)	4	0.11	73	(đg, k, t, tr)	1	0.03
37	(đg, k, tr)	4	0.11	74	(d, p, đg, tr)	1	0.03

Table 5 (*continued*)

No.	Part-of-Speech Combination	Quantity	Percentage (%)	No.	Part-of-Speech Combination	Quantity	Percentage (%)
75	(c, p, tr, đ)	1	0.03	78	(đg, t, p, tr, d)	1	0.03
76	(p, k, t, tr)	1	0.03	79	(k, đg, p, tr, đ)	1	0.03
77	(đg, p, k, t)	1	0.03	80	(k, đg, t, p, tr)	1	0.03

Note. Abbreviations: d = noun; đg = verb; t = adjective; đ = pronoun; p = adjunct; k = conjunction; tr = particle; c = interjection

meaning outside their relational role. Interjections *(c)* are expressive words that mainly convey emotions, reactions, or exclamations, rather than participating in grammatical structures or undergoing part-of-speech conversion. As a result, combinations such as *(p, k, tr), (c, đ, d)*, and *(c, p, đg, t)* appear only once (0.03%), illustrating the restrictions on their ability to function across multiple parts of speech.

A notable pattern in the dataset is the sharp decline in frequency as the number of parts of speech in a combination increases. For instance, the three-part combination *(d, đg, t)* appears relatively frequently (5.8%), but when an additional category is added *(d, đg, t, tr)*, its occurrence drops dramatically to just one instance (0.03%). Further analysis reveals that five-part-of-speech combinations are extremely rare in the *Vietnamese Dictionary*. Only three words exhibit such classifications: *có (đg, t, p, tr, d)*, *qua (k, đg, p, tr, đ)*, and *rồi (k, đg, t, p, tr)*. The rarity of these combinations suggests that Vietnamese does not favour excessive grammatical complexity. This trend can be explained by the minimalist structure of Vietnamese grammar. As an isolating language, Vietnamese relies on word order rather than inflectional morphology. Nouns, verbs, and adjectives serve as the core elements of sentence structure, while other parts of speech (such as adjuncts, particles, conjunctions, and interjections) function mainly as modifiers rather than forming complex multi-category roles. The distribution of part-of-speech combinations reflects linguistic habits in Vietnamese, emphasising structural simplicity in word classification.

Polysemy Coefficient in the Vietnamese Dictionary

The polysemy coefficient is an important indicator that reflects the degree of polysemy in Vietnamese. Since computers cannot distinguish between polysemy and homography, these two types of relationships are grouped. Therefore, in this paper, the term *polysemy* refers to both. Research results show that, on average, each lexical entry in the *Vietnamese dictionary* has 1.245 meanings, indicating a moderate level of polysemy in Vietnamese lexical entries (Table 6). Among the different parts of speech, particles (1.394) exhibit the highest polysemy coefficient due to their dependency on context. Pronouns (1.353)

Table 6
Polysemy coefficient by part of speech

Part of Speech	Number of Lexical Entries	Polysemy Coefficient
Noun *(Danh từ)*	17,831	1.239
Verb *(Động từ)*	12,088	1.291
Adjective *(Tính từ)*	8,749	1.199
Pronoun *(Đại từ)*	133	1.353
Particle *(Trợ từ*)	99	1.394
Interjection *(Cảm từ)*	92	1.098
Conjunction *(Kết từ)*	186	1.253
Adjunct *(Phụ từ)*	509	1.153
Entire Dictionary	49,438	1.245

also demonstrate high polysemy compared to other parts of speech, reflecting their contextual flexibility and socially influenced usage.

For instance, the pronoun *nó* can refer to a person, an animal, or an object, depending on context and social relationships, contributing to its semantic ambiguity. Verbs (1.291) rank third in terms of polysemy, highlighting their semantic adaptability in a non-inflectional language. In Vietnamese, meaning shifts in verbs primarily occur through syntactic combinations rather than morphological changes. For example, the verb *chạy* can convey different meanings: denoting physical movement, as in *Anh ấy chạy rất nhanh* (He runs very fast), expressing urgency in task completion, as in *chạy dự án* (rush a project), or describing machine functionality, as in *Máy chạy tốt* (The machine runs well). Conjunctions (1.253) exhibit polysemy, reflecting their role in connecting clauses within varied semantic and logical relationships. Nouns (1.239) demonstrate a moderate level of polysemy, primarily through metaphorical and metonymic extension. Adjectives (1.199) have a relatively lower polysemy coefficient, suggesting that descriptive words tend to retain more stable meanings than nouns and verbs. Adjuncts (1.153) have a low polysemy coefficient and mainly function as modifiers, providing additional detail or emphasis rather than undergoing semantic shifts. Interjections (1.098) show the lowest polysemy coefficient. As they express emotions or reactions, their meanings remain fixed and do not change much across different contexts.

CONCLUSION

This study analysed lexical entries in the *Vietnamese Dictionary* (Hoàng, 2020) from a computational linguistics perspective, focusing on letter and part-of-speech distribution, multi-POS coefficient, polysemy coefficient analysis to explore the organisational patterns of Vietnamese lexicon. The results indicate that Vietnamese frequently uses certain initial consonants such as *n, h, t, g, c* and vowels like *a, i, u,* reflecting the structural characteristics of syllables and the flexibility

of phoneme combinations. In terms of part-of-speech distribution, *nouns*, *verbs*, and *adjectives* account for the largest proportions among word classes. The multi-POS coefficient suggests that part-of-speech conversion occurs at a relatively low rate, primarily between *nouns*, *verbs*, and *adjectives*. Additionally, polysemy analysis identifies particles and pronouns as the most polysemous parts of speech, indicating their greater semantic flexibility in various linguistic contexts. These findings contribute to a clearer understanding of the structural characteristics of Vietnamese lexicon and provide valuable data for applied linguistics research and language resource digitisation. However, this study is limited by its reliance on a single dictionary source. Future research could expand by incorporating more diverse linguistic data sources and applying modern language models to analyse lexical entries.

ACKNOWLEDGEMENT

The authors would like to express sincere gratitude to the Computational Linguistics Center for providing essential resources that supported this research. The authors acknowledge the editorial team who processed this research paper. The authors received no financial support for the research, authorship, or publication of this article.

REFERENCES

Carstens, A. (1995). Language teaching and dictionary use: An overview. *Lexikos, 5*(1). https://doi.org/10.5788/5-1-1059

Church, K., & Liberman, M. (2021). The future of computational linguistics: On beyond alchemy. *Frontiers in Artificial Intelligence, 4*. https://doi.org/10.3389/frai.2021.625341

Computational Linguistics Center (CLC). (2026). *Dictionaries: CLC_VietDict – Vietnamese dictionary (download samples in Excel and XML)*. Retrieved January 4, 2026, from https://www.clc.hcmus.edu.vn/?page_id=469&lang=en

Cote González, M., & Tejedor Martínez, C. (2011). The effect of dictionary training in the teaching of English as a foreign language. *Alicante Journal of English Studies / Revista Alicantina de Estudios Ingleses, 24*, 31–52. https://doi.org/10.14198/raei.2011.24.02

Đinh, Đ., & Đỗ, Đ. H. (2015, October). Chữ quốc ngữ hiện nay qua các con số thống kê [Current national Vietnamese language through statistics]. In *Hội thảo cấp quốc gia về chữ quốc ngữ: Sự hình thành, phát triển và những đóng góp vào văn hoá Việt Nam* (National workshop on Vietnamese language: Formation, development and contributions to Vietnamese culture).

Định, N. G., & Lương, T. T. (2004). Thuật toán kiểm tra âm tiết tiếng Việt dựa trên luật cấu tạo âm tiết [An algorithm for Vietnamese syllable checking based on syllable-structure rules]. *Tạp chí khoa học, Đại học Huế, 25*.

Đoàn, T. T. (2007). *Ngữ âm tiếng Việt* (4th ed.). Nhà xuất bản Đại học Quốc gia Hà Nội.

Everitt, B., & Skrondal, A. (2010). *The Cambridge dictionary of statistics*. Cambridge University Press.

Ezeh, N. G., Anyanwu, E. C., & Onunkwo, C. M. (2022). Dictionary as an effective resource in teaching and learning of English as a second language: Complementing instructions. *English Language Teaching, 15*(4), Article 108. https://doi.org/10.5539/elt.v15n4p108

Hausser, R. (2001). *Foundations of computational linguistics: Human-computer communication in natural language*. Springer.

Hoàng, P. (2020). *Từ điển tiếng Việt* (9th ed.). Nhà xuất bản Đà Nẵng – Trung tâm Từ điển học.

Huynh, T. V., Nguyen, K. V., & Nguyen, N. L. T. (2022). ViNLI: A Vietnamese corpus for studies on open-domain natural language inference. In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 3858–3872). International Committee on Computational Linguistics. https://aclanthology.org/2022.coling-1.339

Mai, N. C., Vũ, Đ. N., & Hoàng, T. P. (2008). *Cơ sở ngôn ngữ và tiếng Việt* (9th ed.). Nhà xuất bản Giáo dục.

McKinney, W. (2017). *Python for data analysis: Data wrangling with pandas, NumPy, and IPython* (2nd ed.). O'Reilly Media.

Thin, D. V., Hao, D. N., & Nguyen, N. L. T. (2023). Vietnamese sentiment analysis: An overview and comparative study of fine-tuning pretrained language models. *ACM Transactions on Asian and Low-Resource Language Information Processing, 22*(6), 1-27. https://doi.org/10.1145/3589131

Thin, D. V., Hao, D. N., & Nguyen, N. L. T. (2024). A study of Vietnamese sentiment classification with ensemble pre-trained language models. *Vietnam Journal of Computer Science, 11*(1), 137-165. https://doi.org/10.1142/s2196888823500173

Yang, F. (2023). A computational linguistic approach to English lexicography. *Transactions on Computer Science and Intelligent Systems Research, 2*, 39-44. https://doi.org/10.62051/wepk6t89

Ylönen, T. (2022). Wiktextract: Wiktionary as machine-readable structured data. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)* (pp. 1317–1325).