Construction of Cilin Lexical Database: Exploring Metaphor and Metonymy in Classical Chinese Vocabulary

Shueh-Ying Liao

Construction of Cilin Lexical Database: Exploring Metaphor and Metonymy in Classical Chinese Vocabulary

LIAO, Shueh-Ying 廖學盈
PUD-PS
ENS Paris-Saclay[1]

Abstract

The *Cílín Diǎnyì* 詞林典腋 is a vocabulary index of classical Chinese poetry, often combined with other similar literary materials into volumes, and it traditionally falls under the category of reference books in the field of traditional sinology. The format and content of *Cílín Diǎnyì* were standardized during the Qing Dynasty (1636-1912). Today, it is commonly used as a writing tool and receives limited attention in literary research. However, the examples within *Cílín Diǎnyì* represent, in essence, typical Chinese word constructions formed through the creative practices of numerous poets and the careful selection and classification by countless editors, all guided by the aesthetic principles of traditional metrics. As a result, *Cílín Diǎnyì* serves as an entry point for studying the mechanisms of classical vocabulary generation and a rare historical linguistic resource for Chinese semantics. However, accessing *Cílín Diǎnyì* for research requires a basic understanding of poetic metrics, allowing scholars to systematically identify examples based on principles of tonal patterns, semantic similarities, and distinctions, whether they are individual words, word groups, phrases, or fine sentences. To facilitate greater engagement of scholars in the study of *Cílín Diǎnyì*, the creation of a digital database for *Cílín Diǎnyì* is becoming increasingly necessary. This study conducted an exploration of *Cílín Diǎnyì*, resulting in the identification of three distinct usage requirements: 1) distinguishing Chinese characters within the structures of word examples that demarcate semantic categories, 2) recognizing Chinese characters with strong lexical derivation capabilities in the semantic network, and 3) identifying the typical distribution of each Chinese character within various word structures. In addition to highlighting the fundamental functionalities required for future *Cílín Diǎnyì* databases, this research also uncovered metaphors and metonymies

[1] LIAO, Shueh-Ying, Ph.D. from the École Pratique des Hautes Études (EPHE), currently works as a data research engineer at the Plateforme Universitaire de Données Paris-Saclay (PUD-PS) of the École normale supérieure Paris-Saclay (ENS Paris-Saclay). He also serves as an associated researcher for GEO (Groupe d'études orientales, slaves et néo-helléniques - UR 1340) at the University of Strasbourg. Email: shueh-ying.liao@ens-paris-saclay.fr

generated through changes in the positions of Chinese characters within word structures during the exploration process. These findings are beneficial for scholars as they continue to uncover historical word construction patterns preserved in *Cílín Diǎnyì* and the concealed literary aesthetic principles within these word construction patterns.

Keywords : literary database, *Cílín Diǎnyì*, Chinese word construction, metonymy, metaphor

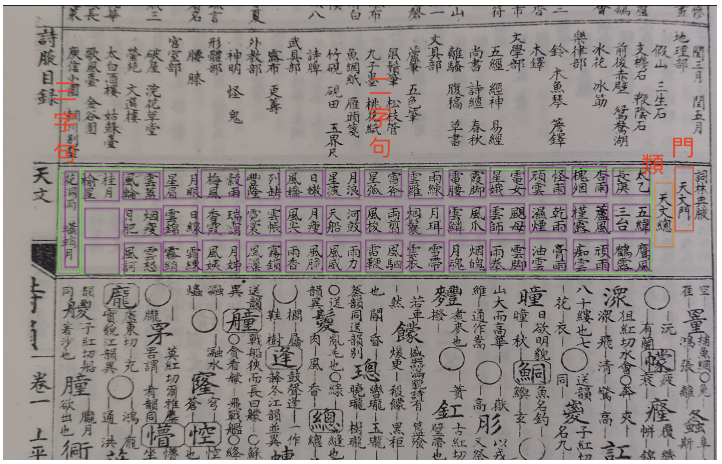
I. Introduction

The *Cílín Diǎnyì* 詞林典腋 is often regarded as a teaching resource in classical literature or as a reference material for writing. Within the realm of traditional sinology, it falls under the category of reference books (*lèishū* 類書). *Cílín Diǎnyì* categorizes words, word groups, phrases, or fine sentences based on themes, juxtaposing them in pairs for easy reference. This systematic arrangement aids scholars in constructing words and forming sentences based on established patterns. The examples found within *Cílín Diǎnyì* have undergone repeated validation through the creative practices of countless poets, making them quintessential to classical Chinese poetry themes and word constructions. However, perhaps due to its primarily utilitarian nature, *Cílín Diǎnyì* is often relegated to the role of an instructional tool and receives limited attention from literary researchers. Its wealth of lexical constructions is often regarded merely as a collection of diverse, old examples. This study aims to prepare a project for the construction of a relational database for *Cílín Diǎnyì*, with the goal of uncovering the inherent word construction patterns within *Cílín Diǎnyì* and the concealed literary aesthetic principles embedded within these patterns.

1. Structure of *Cílín Diǎnyì*

Cílín Diǎnyì encompasses a total of 31 thematic « categories » (*mén* 門) and 625 « subcategories » (*lèi* 類), with word examples organized from shorter to longer character sequences. These examples are arranged in word pairs or sentence pairs, juxtaposed side by side, considering tonal patterns and bearing semantic similarities. Such characteristics facilitate the examination of the correctness of each example sentence pair, word pair, and individual word.

Figure 1: Cílín Diǎnyì compiled and supplemented by TANG Wenlu 湯文璐 *in the Qing Dynasty, included in the Augmented Compilation of Rhymes and Verses (Zēngbǔ Shīyùn Hébì* 增補詩韻合璧*)*



The compilation logic can be summarized as follows:

<table>
<tr><td colspan="4"></td><td rowspan="2">門</td></tr>
<tr><td colspan="3"></td><td rowspan="6">類</td></tr>
<tr><td colspan="3">Word Examples</td></tr>
<tr><td colspan="2">Sentence Pairs</td><td colspan="2">Word Pairs</td></tr>
<tr><td rowspan="2">Entry
(Lower Pair - Starting Sentence)
11</td><td rowspan="2">Entry
(Upper Pair - Starting Sentence)
9</td><td>Entry (Lower Pair) 2</td><td>Entry (Upper Pair) 1</td></tr>
<tr><td>Entry (Lower Pair) 4</td><td>Entry (Upper Pair) 3</td></tr>
<tr><td rowspan="2">Entry
(Lower Pair - Corresponding Sentence)
12</td><td rowspan="2">Entry
(Upper Pair - Corresponding Sentence)
10</td><td>Entry (Lower Pair) 6</td><td>Entry (Upper Pair) 5</td></tr>
<tr><td>Entry (Lower Pair) 8</td><td>Entry (Upper Pair) 7</td></tr>
</table>

In accordance with this, when examining the original text, « 太乙|長庚 » constitutes the first word pair with right-left parallelism, and « 五緯|三台 » forms the second word pair, both adhering to this pattern. When a right-left word pair parallelism results in the separation of sentence pairs or word pairs across two pages, it is modified to upper and lower correspondences. For example, the first instance of a three-character word pair in the leftmost vertical row of the column is considered an upper and lower correspondence. Sentence pairs are arranged below word pairs and continue downward. In the absence of sentence breaks, parallelism may appear, and sentence segmentation is determined based on meaning or tonal patterns.

For this study, the electronic version of *Cílín Diǎnyì* provided by Taiwanese linguist HSU André Chang-Mo (許長謨) was used for word and sentence pairs and word construction annotations. The initial observations, experiments, and analysis were conducted using two-character words as the basis.

2. Origin of the *Cílín Diǎnyì* Electronic Version

The electronic version of the *Cílín Diǎnyì* originates from the research conducted by Professor HSU in recent years and has found applications in language education and cultural promotion. Professor HSU, along with his research team, which included graduate students, undertook tasks such as inputting entries, identifying example pairs, annotating word constructions, and verifying éditions. They transformed the traditional printed manuscript into a digital text file, accompanied by essential

reference annotations. Additionally, a spreadsheet-style version was produced, offering basic search and filtering functions. While Professor HSU has published certain content from this electronic version in books and research reports, the complete dataset has not been publicly available until now. This electronic version, which has undergone years of revision and compilation by Professor HSU, primarily, but not exclusively, references the following historic editions of *Cílín Diǎnyì*:

editor	title	acronym	purpose
湯祥瑟	詩韻合璧	合	used for cross-referencing
湯文璐	詩韻全璧	全	used for cross-referencing
余照	增廣詩韻集成	集	used for cross-referencing
盧元駿	詩詞曲韻總檢	檢	served as the manuscript reference source

The electronic version of the *Cílín Diǎnyì* has gradually taken on the structure of a relational database by incorporating several crucial metadata elements. Through the calculation and encoding of entry lengths, this dataset can provide information on syllable patterns within word examples. In the case of word pairs, ranging from two-character words to nine-character words, numerical codes are used, and the upper and lower components are separated by a « / ». For example, « upper pair / lower pair » represents the structure of word pairs:

length	example	struct_1	n°	struct_2	struct_3	cross-reference
3	魚信斷/雁書遲		4908	SP	MH	《全》雁書還《檢》《集》雁聲遲

For sentence pairs, which also vary from two-character to nine-character words, the initial characters of French numerals are used to encode the « starting sentence » and « corresponding sentence » in accordance with the character length of the entries.[2] Upper and lower components are separated by « // », while « starting sentences » and « corresponding sentences » are separated by « / ». For instance, in the database, « QQ » denotes a sentence pair with the structure « four-character word / four-character word // four-character word / four-character word »:

length	example	struct_1	n°	struct_2	struct_3	cross-reference
QQ	角觝場中/擊丸蹴踘//踏歌聲裏/問柳尋花		1026			

[2] D (deux: 2), T (trois: 3), Q (quatre: 4), C (cinq: 5), S (six: 6), P (sept: 7), H (huit: 8), N (neuf: 9). The exception is for « 7 », which is represented by the letter « P » to avoid confusion with « 6 », which uses the letter « S ».

Or, for example, « QS » represents a sentence pair structure of « four-character word / six-character word // four-character word / six-character word »:

length	example	struct_1	n°	struct_2	struct_3	cross-reference
QS	吟嘯成群/感李陵于塞上//應接不暇/勞子敬于山陰		1254			

Following this logic, « QCC » signifies « four-character word / five-character word / five-character word // four-character word / five-character word / five-character word », and « QQQQ » designates « four-character word / four-character word / four-character word / four-character word // four-character word / four-character word / four-character word / four-character word ». This labeling and separation system is relatively rigorous and easily distinguishable.[3]

Professor HSU's team's work encompasses systems for annotating syllable patterns, arranging example pair structures, numbering example sequences, font verification for historical and modern éditions, and segmenting word meanings. These interpretive data provide a broad humanistic perspective for digital analysis of the *Cílín Diǎnyì*, laying a practical scientific foundation for the development of the *Cílín Diǎnyì* digital database. This research is conducted on this basis, utilizing text mining in the *Cílín Diǎnyì*. Through the author's analytical process and summary of results, remote discussions with Professor HSU are ongoing to explore structured and modeled possibilities for the *Cílín Diǎnyì*, evaluating the potential application of this literary and lexical data in studies of word construction, word meaning, and literary criticism.

II. Exploration Strategy

In the construction of the database, emphasis is placed not only on data completeness but also on data correlation and modeling, ensuring that real-world linguistic evidence is well reflected in the structure of the database. Therefore, if we wish for the database to assist in the exploration of word construction patterns and aesthetic principles, rather than being a mere repository of textual data, we need to interpret and elaborate on the relationships within the original data and establish a model for queries and responses.

The word construction patterns and aesthetic principles investigated in this study are not explicitly recorded within the Chinese characters themselves. However, by objectifying the compilation and

[3] For the sentences pairs, we found TS, TH, TC, SS, SQ, SP, SC, QS, QQ, QP, QH, QC, PS, PQ, HQ, HC, CT, CS, CQ, CC, SQQ, QQQQ, and QCC.

usage logic of word examples and clearly annotating the role and function of each character within this logic, it becomes possible to reconstruct the formalities of word construction and aesthetic standards. Hence, we have selected the following aspects for investigation:

1. The quantity and proportion of word construction types.
2. The differentiation power of word construction morphemes.
3. The network of nodes for word construction morphemes.
4. The function and position of word construction morphemes.

These investigations aim to reveal the underlying connections between vocabulary, structure, and categories.

III. Exploration Methods

The specific exploration methods are as follows: 1) Word construction is categorized using HSU's classification method[4], and the quantity and proportion of word constructions are tallied. 2) Chinese characters within word constructions are divided into central and non-central characters. The T-SNE (t-Distributed Stochastic Neighbor Embedding) method is then used to test the positions of characters that maximize the differentiation of categories in each word construction.[5] 3) Each pair of examples is considered as directional edges, with the characters contained in the example pair regarded as nodes, forming a node network. The number of connections in and out of each character node and their hierarchy is calculated. 4) The typical functions and positions of each character within the word construction are computed.

[4] Hsu, C. -M. 許長謨 (2010). 詞彙與語義之義證. In 漢語語言結構義證 (p. 125). 里仁書局, 臺北. Categories are as follows: 1) Simple Words: 聯綿 (LM, lianmian/binome) and 狀聲 (ON, onomatopoeia); 2) Compound Words: 並列 (CO, coordination), 偏正 (MH, modifier-head), 主謂 (SP, subject-predicate), 動賓 (VO, verb-object), and 動補 (VC, verb-complement); 3) Complex Words: 重疊 (DP, duplication) and 衍生 (AF, affix/derivation) 4) Location (CL, location), Proper Noun (PN, proper noun), Verb-Verb (VV, verb-verb), and Pivotal/Embedded (BY, baoyun/pivotal/embedded).

[5] In the survey conducted, character positions within words were used to distinguish categories. This approach emphasizes the significance of character order in classifying words and reveals structural patterns and relationships within the dataset. t-SNE was employed to visualize how positional differences impact word distribution in a lower-dimensional space, providing valuable insights for linguistic analysis and category characterization based on character positions. This methodology uncovers hidden linguistic patterns, offering a deeper understanding of lexical organization.

Each entry in *Cílín Diǎnyì* consists of two to nine Chinese characters, with most forming a single morpheme. Some entries have characters that cannot be separated and are composed of multiple characters, collectively forming a single morpheme. The labeling of characters follows these guidelines:

When extracting each character from the example : 1) Phonetic information is annotated, with tone distinctions converted to tonal patterns. If there are multiple pronunciations, they are separated by the « | » symbol. 2) The three-dimensional position of the character within the example (sentence pair or word pair) is marked. This includes the length of characters in the example, the upper-lower sequence of the entry (0 for upper, 1 for lower), and the character's position within the example (starting with 0 for the first character, 1 for the second, and so on). 3) The category of the example in which the character is located is noted. 4. The HSU's word construction model for the pair of entries in the example are labeled.

漢字	聲母	韻母	平仄	開合	等呼	上下對序	詞內字序	例對字長	詞門	詞類	詞例	詞構
太	透	泰	仄	開	一	0	0	4	【天文門】01TW	【天文總】01TW00	太乙|長庚	MH

An alternative approach is to employ a vocabulary management method using a relational database, with Chinese characters as the axis, creating a master table. Other information is organized in separate tables, with numerical references linking them to the master table. This approach can effectively maintain vocabulary consistency and optimize search efficiency.

漢字	聲母	韻母	平仄	開合	等呼	上下對序	詞內字序	例對字長	詞門	詞類	詞例	詞構
1	5	2422	1	0	1	0	0	4	1	1	1	4

By managing in this manner, it becomes possible to differentiate characters with the same font that, in fact, originate from different contextual backgrounds: 1) « Quantitative Features » can be used to investigate the grammatical functions of Chinese characters, such as the length of characters in examples, upper-lower sequence, and character order within words. 2) « Qualitative Features » can be employed to explore the semantic attributes of Chinese characters, including semantic category, semantic subcategory, word examples, word constructions, or phonetic patterns such as initials, finals, tonal patterns, and combinations of open and closed syllables. Currently, this study only examines cases involving two-character words.

IV. Investigation Results

1. Variety of Word Examples

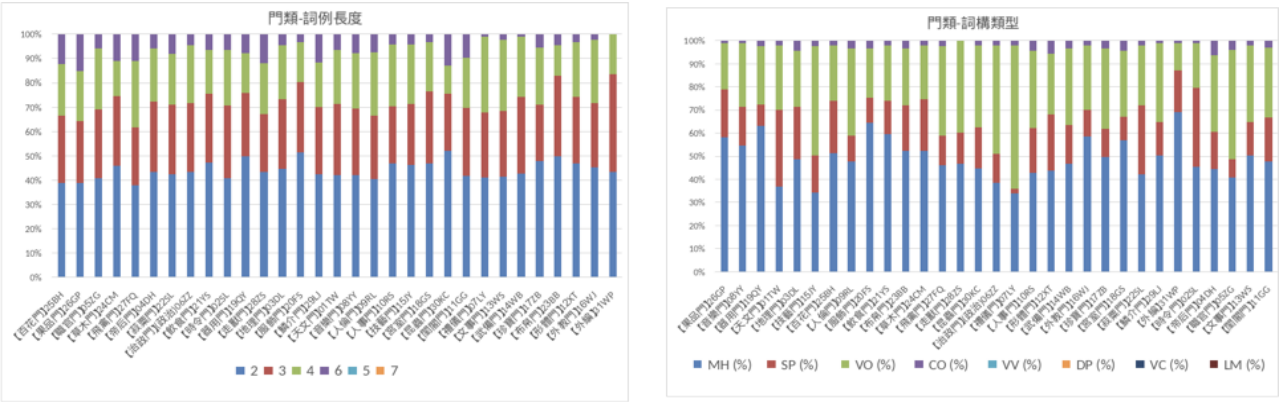
When categorizing word construction types based on the total number of examples, they are ranked in the following order: modifier-head, verb-object, subject-predicate, and coordination. The remaining word construction types have very few examples. The categories of word constructions under *Qìyòngmén* 器用門 (*Craftsmanship Category*), *Dìlǐmén* 地理門 (*Geography Category*), and *Guǒpǐnmén* 果品門 (*Fruit Category*) show the highest diversity in word construction types.

Table 1: Word Construction Quantity in Each Category (Two-Character Words)

門類	MH	SP	VO	CO	DP	VC	LM	PN	VV	ON	詞構種類
【器用門】19QY	291	12	90	14	4	7	2	8	1	0	9
【地理門】03DL	193	56	90	19	4	2	2	1	0	0	8
【果品門】26GP	174	27	41	4	2	6	1	1	0	0	8
【天文門】01TW	130	71	88	9	6	2	1	0	0	0	7
【音樂門】08YY	67	8	31	2	1	0	1	2	0	0	7
【服飾門】20FS	209	8	51	16	1	12	0	11	0	0	7
【草木門】24CM	243	53	91	11	7	1	2	0	0	0	7
【百花門】25BH	312	59	102	20	5	4	0	4	0	0	7
【飛禽門】27FQ	137	22	95	7	4	8	0	0	0	1	7
【走獸門】28ZS	84	10	65	1	1	2	1	0	0	0	7
【昆蟲門】30KC	66	16	48	4	2	2	1	0	0	0	7
【人倫門】09RL	93	11	80	10	2	13	0	0	0	0	6
【形體門】12XT	68	24	40	12	3	11	0	0	0	0	6
【武備門】14WB	44	9	35	5	0	1	0	1	0	0	6
【技藝門】15JY	55	10	70	5	2	6	0	0	0	0	6
【珍寶門】17ZB	40	3	26	4	0	3	0	2	0	0	6
【宮室門】18GS	166	10	67	12	0	4	0	1	0	0	6
【飲食門】21YS	178	13	63	8	2	0	0	4	0	0	6
【布帛門】23BB	45	7	17	4	0	0	1	2	0	0	6
【外編】31WP	90	13	20	4	4	10	0	0	0	0	6
【時令門】02SL	145	53	66	3	0	1	0	0	0	0	5
【帝后門】04DH	43	8	21	8	1	0	0	0	0	0	5
【職官門】05ZG	68	6	47	9	0	0	0	1	0	0	5
【治政門】06ZZ	29	6	30	2	0	1	0	0	0	0	5
【禮儀門】07LY	19	1	29	2	0	0	0	0	1	0	5
【人事門】10RS	105	29	86	15	0	2	0	0	0	0	5
【文事門】13WS	76	14	32	4	0	2	0	0	0	0	5
【外教門】16WJ	66	7	30	3	0	0	0	2	0	0	5
【菽粟門】22SL	72	19	33	4	4	0	0	0	0	0	5
【鱗介門】29LJ	57	4	35	1	0	1	0	0	0	0	5
【闈閣門】11GG	37	8	27	3	0	0	0	0	0	0	4
Total Quantity	3402	597	1646	225	55	101	12	40	2	1	
Total Percentage (%)	55.94	9.82	27.07	3.7	0.9	1.66	0.2	0.66	0.03	0.02	

Bǎihuāmén 百花門 (*Flower Category*), *Guǒpǐnmén* 果品門 and *Zhíguānmén* 職官門 (*Official Category*) show the most diverse word example lengths. In each category, two-character, three-character, and four-character word examples are predominant, with a fair representation of six-character examples. Five-character, seven-character, and other parallel structures are also present. In terms of word constructions, apart from *Jìyìmén* 技藝門 (*Art Category*), *Lǐyímén* 禮儀門 (*Etiquette Category*), and *Zhíguānmén* 職官門, which contain a substantial number of verb-object structures in their examples, the remaining categories are primarily composed of modifier-head structures in their word examples.

Figure 2: Diversity of Word Examples



The table for the constituent components of each category is as follows :

Table 2: Word Construction Ratios in Each Category (Two-Character Words)

門類	MH (%)	SP (%)	VO (%)	CO (%)	DP (%)	VC (%)	LM (%)	PN (%)	VV (%)	ON (%)
【天文門】01TW	42.35	23.13	28.66	2.93	1.95	0.65	0.33	0	0	0
【時令門】02SL	54.1	19.78	24.63	1.12	0	0.37	0	0	0	0
【地理門】03DL	52.59	15.26	24.52	5.18	1.09	0.54	0.54	0.27	0	0
【帝后門】04DH	53.09	9.88	25.93	9.88	1.23	0	0	0	0	0
【職官門】05ZG	51.91	4.58	35.88	6.87	0	0	0	0.76	0	0
【治政門】06ZZ	42.65	8.82	44.12	2.94	0	1.47	0	0	0	0
【禮儀門】07LY	36.54	1.92	55.77	3.85	0	0	0	0	1.92	0
【音樂門】08YY	59.82	7.14	27.68	1.79	0.89	0	0.89	1.79	0	0
【人倫門】09RL	44.5	5.26	38.28	4.78	0.96	6.22	0	0	0	0
【人事門】10RS	44.3	12.24	36.29	6.33	0	0.84	0	0	0	0
【閨閣門】11GG	49.33	10.67	36	4	0	0	0	0	0	0
【形體門】12XT	43.04	15.19	25.32	7.59	1.9	6.96	0	0	0	0
【文事門】13WS	59.38	10.94	25	3.12	0	1.56	0	0	0	0
【武備門】14WB	46.32	9.47	36.84	5.26	0	1.05	0	1.05	0	0
【技藝門】15JY	37.16	6.76	47.3	3.38	1.35	4.05	0	0	0	0
【外教門】16WJ	61.11	6.48	27.78	2.78	0	0	0	1.85	0	0

Table 2: Word Construction Ratios in Each Category (Two-Character Words)

門類	MH (%)	SP (%)	VO (%)	CO (%)	DP (%)	VC (%)	LM (%)	PN (%)	VV (%)	ON (%)
【珍寶門】17ZB	51.28	3.85	33.33	5.13	0	3.85	0	2.56	0	0
【宮室門】18GS	63.85	3.85	25.77	4.62	0	1.54	0	0.38	0	0
【器用門】19QY	67.83	2.8	20.98	3.26	0.93	1.63	0.47	1.86	0.23	0
【服飾門】20FS	67.86	2.6	16.56	5.19	0.32	3.9	0	3.57	0	0
【飲食門】21YS	66.42	4.85	23.51	2.99	0.75	0	0	1.49	0	0
【菽粟門】22SL	54.55	14.39	25	3.03	3.03	0	0	0	0	0
【布帛門】23BB	59.21	9.21	22.37	5.26	0	0	1.32	2.63	0	0
【草木門】24CM	59.56	12.99	22.3	2.7	1.72	0.25	0.49	0	0	0
【百花門】25BH	61.66	11.66	20.16	3.95	0.99	0.79	0	0.79	0	0
【果品門】26GP	67.97	10.55	16.02	1.56	0.78	2.34	0.39	0.39	0	0
【飛禽門】27FQ	50	8.03	34.67	2.55	1.46	2.92	0	0	0	0.36
【走獸門】28ZS	51.22	6.1	39.63	0.61	0.61	1.22	0.61	0	0	0
【鱗介門】29LJ	58.16	4.08	35.71	1.02	0	1.02	0	0	0	0
【昆蟲門】30KC	47.48	11.51	34.53	2.88	1.44	1.44	0.72	0	0	0
【外編】31WP	63.83	9.22	14.18	2.84	2.84	7.09	0	0	0	0

2. Category Differentiation

Each category is treated as a collected sample, with the word examples within the category serving as features. Subsequently, the characters contained in the examples are divided into two groups: the previous character in one group and the subsequent character in another. Based on TF-IDF (Term Frequency-Inverse Document Frequency) normalized word frequency, the distances between samples are calculated using the t-SNE (t-Distributed Stochastic Neighbor Embedding) algorithm, and then projected into a two-dimensional plane. Finally, the concentration and dispersion of sample data points in the analyzed graph are observed.

The higher the Silhouette Score, indicating more concentration of data points within the same category, and the lower the Davies-Bouldin Index, signifying greater dispersion of data points between different categories, the more balanced the differentiation of characters' positions. This is reflected in a higher Quality Ratio defined in this work.

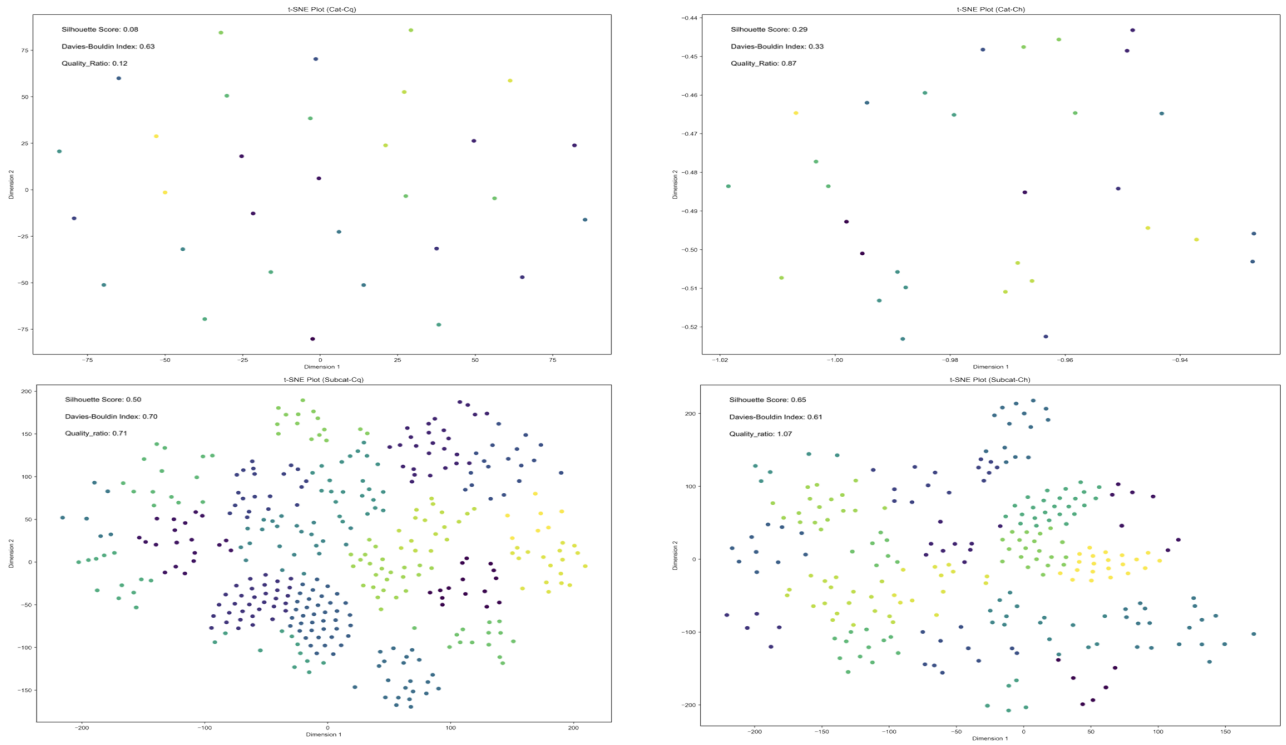
In the following, « C » represents chinese character, lowercase « q » (*qián* 前) represents the previous character, and lowercase « h » (*hòu* 後) represents the subsequent character. The key positions that differentiate characters based on « category » and « subcategory » for various word constructions are as follows:

Table 3: Testing of Chinese Character Category Differentiation Positions

Type	Pos.	Silhouette Score (SS, range from -1 to 1)			Davies-Bouldin Index (DBI, range from 0 to 1)			Quality Ratio (SS/DBI)		
		Cat	Sub-cat	Diff.	Cat	Sub-cat	Diff.	Cat	Sub-cat	Diff.
MH	Cq	0,08	0,5	0,42	0,63	0,7	0,07	0,12	0,71	0,59
	Ch	0,29	0,65	0,36	0,33	0,61	0,28	0,87	1,07	0,2
SP	Cq	0,1	0,6	0,50	0,51	0,67	0,16	0,19	0,9	0,71
	Ch	0,47	0,73	0,26	0,13	0,59	0,46	3,7	1,24	-2,46
VO	Cq	0,28	0,74	0,46	0,32	0,52	0,2	0,86	1,41	0,55
	Ch	0,05	0,63	0,58	0,52	0,68	0,16	0,1	0,92	0,82
CO	Cq	0,28	0,89	0,61	0,44	0,45	0,01	0,63	1,98	1,35
	Ch	0,31	0,86	0,55	0,38	0,53	0,15	0,84	1,62	0,78

In summary, for both « category » and « subcategory », the key positions for differentiation are as follows: the subsequent character is key in subject-**predicate** (S**P**) and modifier-**head** (M**H**); the previous character is key in **verb**-object (**V**O). In coordination (CO) structures, the subsequent character distinguishes « category », while the previous character distinguishes « subcategory ». When shifting analysis from « category » to « subcategory » within the same word construction and at the same character position, notable differences in clustering strength (Silhouette Score) are observed. For instance, in modifier-head (MH), the clustering strength for « modifier (Cq) » increased by 42%, whereas « head (Ch) » only increased by 36% :

Figure 4. t-SNE Test for Modifier-Head Structure



In summary, characters originally considered key positions in « category » tend to weaken their clustering function when transitioning to « subcategory » within the same word construction, and these characters are integrated into « subcategory » by characters that were originally non-key positions. For example, subject in subject-predicate, modifier in modifier-head, object in verb-object, and the subsequent character in coordination, while they cannot govern « category », can integrate « subcategory ». Additionally, the Davies-Bouldin Index values, which represent the interval between Chinese characters' clustering, also show corresponding changes during the transition to a different « subcategory ». It is important to note that the testing results are most reliable for modifier-head (55.94%) and verb-object (27.07%) based on the ratio of subcategory word construction quantity to the total quantity.

3. Chinese Characters Semantic Network

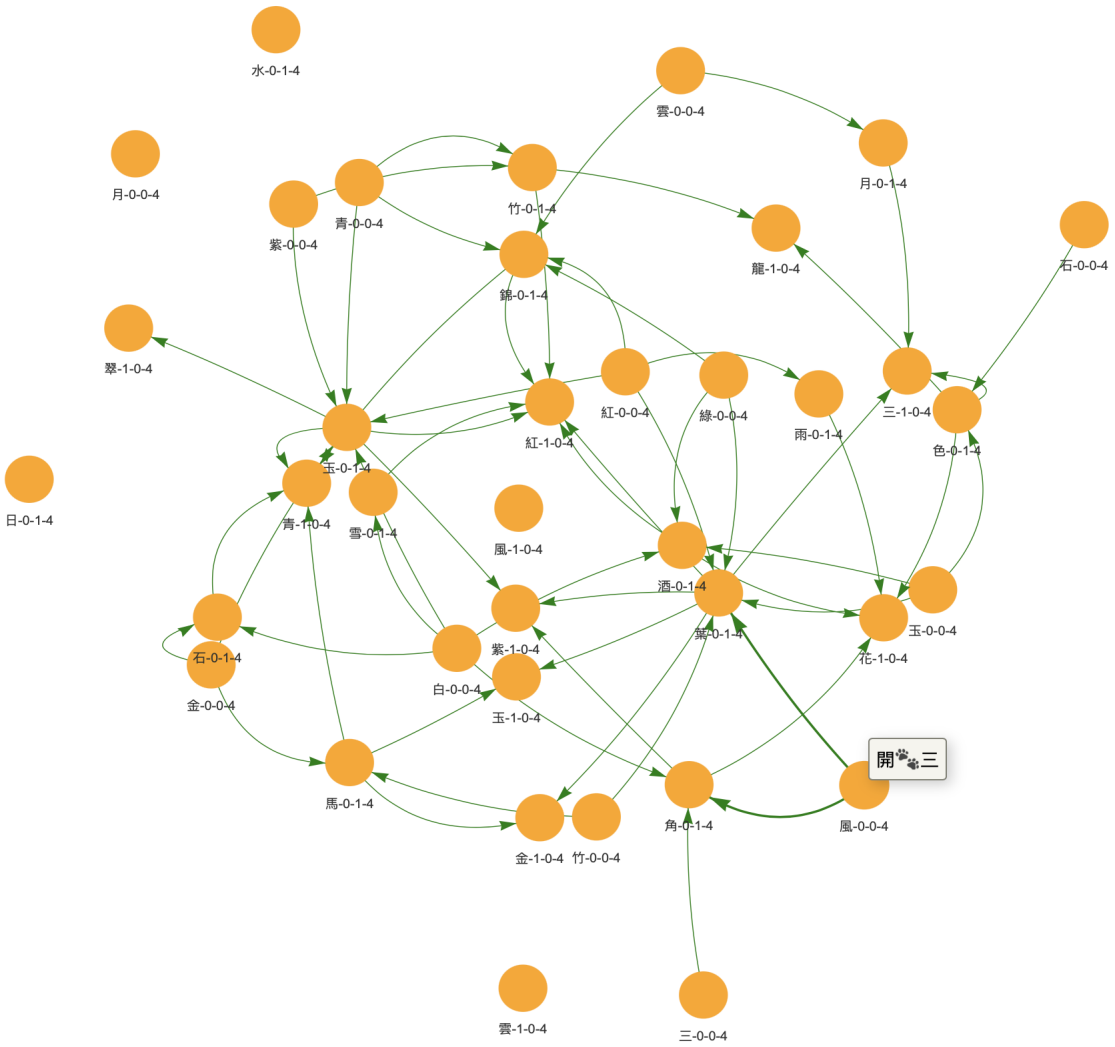
In this section, we explore the creation of a network with Chinese characters as nodes. Each Chinese character, accompanied by its quantitative attributes, such as its position in a word pair (upper/lower), its position in the pair's sequence (previous/subsequent), and the number of characters in the pair (4, 6, 8…18), serves as a « node » in the network. The connections, or « directed edges », between nodes are determined by the order of characters in word examples. For example, in the word pair « 太乙|長庚 », the directed edges connect characters in this sequence: « 太-乙-長-庚 ».

Positional information for each character is recorded to keep track of where they appear within different word pairs or sentence pairs. This information ensures that each character's connections are unique to their specific context. We calculate the number of nodes that can function as previous character(s) and those that can serve as subsequent character(s). Additionally, we assess the level at which they can extend to form complete words. This analysis helps us understand the connection capabilities of each node within the network.

The top ten characters with the most generative capacity and their respective word construction positions are as follows: « 三-1-0-4 »[6], « 雨-0-1-4 », « 月-0-0-4 », « 風-1-0-4 », « 雲-1-0-4 », « 風-0-0-4 », « 雲-0-0-4 », « 月-0-1-4 », « 石-0-1-4 », and « 玉-0-0-4 ». These characters play a significant role in forming various word structures within the network.

[6] The « 三 » represents Chinese characters, « 1 » indicates being in the lower pair the example, « 0 » signifies being the previous character in the pair below the example, and « 4 » represents a two-character pair with a total word length of four Chinese characters. All other nodes are interpreted in a similar manner.

Figure 5. Chinese Characters with More than Thirty Layers of Sequential Node Connections



4. Distribution of Chinese Characters' Functions and Positions within Word Constructions

Building upon the analyses conducted earlier, we've identified the crucial positions that differentiate categories within each word construction, such as the subsequent characters in modifier-head structures. Additionally, based on the second analysis, we've pinpointed potential nodes that facilitate the generation of additional word entries within each word construction.[7] For example, nodes like « 風-0-0-4 », « 雲-0-0-4 », or « 月-0-1-4 » within the node network.

For Chinese characters with stronger connectivity, we delve deeper to examine their distribution ratios across various positions within each word construction, corresponding categories, and

[7] The node graph can be queried and filtered based on nodes, connections, categories, positions, or phonetic conditions.

semantic classes. This allows us to understand the typical distribution and category correspondence for each character within word constructions. For instance, an investigation into the distribution of the character « 風 » across different word constructions reveals that « 風 » can be used across various word constructions, particularly in modifier-head (MH), verb-object (VO), and subject-predicate (SV) structures. In modifier-head structures, it predominantly occupies the lower part and frequently serves as the subsequent character in categories with strong differentiation. In verb-object structures, it also tends to be in the lower part and often serves as the subsequent character in categories with weaker differentiation. In subject-predicate structures, it mainly appears as the previous character and does not play a role as the subsequent character, with weaker category differentiation. In summary, despite its widespread usage across various word constructions, « 風 » primarily assumes a non-central role within the word construction:

條件:('風', 'VO', 1, 1, 4), 數量: 33, 比率: 18.97%, 詞條:逐雨|追風.....
條件:('風', 'VO', 0, 1, 4), 數量: 13, 比率: 7.47%, 詞條:因風|伴月 占風|觀日.....
條件:('風', 'VC', 1, 0, 4), 數量: 1, 比率: 0.57%, 詞條:雨折|風掀
條件:('風', 'SP', 1, 0, 4), 數量: 27, 比率: 15.52%, 詞條:雨剪|風梭 日嫩|風嬌.....
條件:('風', 'SP', 0, 0, 4), 數量: 16, 比率: 9.20%, 詞條:風脆|雨香 風勁|雲濃.....
條件:('風', 'MH', 1, 1, 4), 數量: 32, 比率: 18.39%, 詞條:穀雨|梅風 十雨|五風.....
條件:('風', 'MH', 1, 0, 4), 數量: 21, 比率: 12.07%, 詞條:月姐|風姨 清白|風流.....
條件:('風', 'MH', 0, 0, 4), 數量: 18, 比率: 10.34%, 詞條:風腳|雨拳 風駟|雷鞭.....
條件:('風', 'MH', 0, 1, 4), 數量: 9, 比率: 5.17%, 詞條:鷹風|鶴雲 濕風|陰雲.....
條件:('風', 'CO', 1, 0, 4), 數量: 2, 比率: 1.15%, 詞條:儒雅|風神 富貴|風流
條件:('風', 'CO', 0, 0, 4), 數量: 2, 比率: 1.15%, 詞條:風流|搖曳 風渙|隨雷

For example, let's consider the character « 龜 » as the subject of investigation, limiting the focus to four-character word examples:

條件:('龜', 'VO', 1, 1, 4), 數量: 1, 比率: 5.88%, 例對:棲鵠|麗龜
條件:('龜', 'SP', 1, 0, 4), 數量: 1, 比率: 5.88%, 例對:魚戲|龜遊
條件:('龜', 'MH', 1, 0, 4), 數量: 6, 比率: 35.29%, 例對:鼇石|龜山 椿算|龜齡 鶴眼|龜文 鳳彩|龜文 貝錦|龜紗 馬蹄|龜背
條件:('龜', 'MH', 0, 0, 4), 數量: 5, 比率: 29.41%, 例對:龜齡|馬齒 龜甲|蛛絲 龜殼|繭頭 龜甲|鶴文 龜甲|龍文
條件:('龜', 'MH', 1, 1, 4), 數量: 2, 比率: 11.76%, 例對:砥柱|寶龜 銅鶴|神龜
條件:('龜', 'MH', 0, 1, 4), 數量: 1, 比率: 5.88%, 例對:洛龜|岡鳳
條件:('龜', 'CO', 1, 0, 4), 數量: 1, 比率: 5.88%, 例對:熊虎|龜蛇

It is evident that the character « 龜 » most frequently appears as a modifier in the « four-character modifier-head pairs » in the lower part (35.29%) and as a modifier in the upper part (29.41%). It is relatively less common for « 龜 » to function as the headword in modifier-head pairs, with occurrences of 11.76% and 5.88%, respectively. Additionally, « 龜 » is observed in structures containing verbs, specifically in subject-predicate and verb-object structures, appearing only in the lower part of these structures and serving as the subject or object. Earlier, using the T-SNE method,

it was verified that subject-predicate and verb-object structures, respectively with the predicate and the verb as key characters, have a higher discriminative power in terms of categories. Consequently, in structures containing verbs, « 龜 » does not show strong category-differentiating power.

In summary, « 龜 » has not functioned as a verb and is mostly employed as a modifier in modifier-head structures. While the central role in these structures is not its strength, it is not insignificant either. Therefore, innovating by changing the position of « 龜 » in modifier-head structures is not easy. Conversely, in subject-predicate or verb-object structures, successfully altering « 龜 » grammatical roles can yield unique effects. For example, using « 龜 » as a verb implies « to hide », resulting in the derived word « 龜著 » (*to hide*) in the affix/derivation structure. An example usage might be: *When he got scared, he would run back home and hide (*他一害怕就跑回家龜著*)*.

This investigation clarifies the similar and distinct conditions required for generating new words and creating new meanings. Phonetics can also be used as search criteria to explore new words or meanings generated from similar phonetic structures, investigating analogies or logical connections between new words or meanings and phonetic patterns.

V. Conclusion and Discussions

This article extensively explores the intricate relationship between Chinese lexical constructions and themes, emphasizing the dynamic nature of language. It particularly highlights how Chinese characters, as morphemic components, show variations in their positions, functions, meanings, and poetic nuances within different word constructions.

In conclusion, the results of this investigation reveal that when isolated from their word constructions, Chinese characters represent mere morphemes or partial components of morphemes and do not inherently define their parts of speech.

Analyzing the distribution of the same character on either side of a central word's core is essentially an analysis of its ability to create metonymic transformations within morphemes or morphemic components. For instance, in modifier-head structures, characters that can be interchanged (i.e., functionally interchangeable) to form words of different categories provide an example of such

associations[8]. This kind of association stems from everyday experiences and corresponds to a poet's ability to employ metonymies to *touch* (*chùwù* 觸物) things. Analyzing within word constructions, when the same position, morpheme, or morphemic component can be replaced while maintaining the same word category, it delves into the metaphorical structure of a word. The example pairs in the word forest highlight the metaphoric logic behind words within the same theme. Such analogies arise from rational analysis and correspond to the poet's ability to *comprehend* (*yuánlǎn* 圓覽) in poetic metaphors. Viewing a poet's approach to word formation from this perspective helps understand how in poetry, *despite being from different backgrounds* (*wù sūi hú yuè* 物雖胡越), *things can eventually harmonize* (*hé zé gān dǎn* 合則肝膽)[9].

Under the constraints of classical metrics, classical Chinese poetry is actually required to be created at least on the level below phrases. Therefore, it's challenging for poetry to have a strict sense of sentence structure: we simply use the concept of everyday sentences to interpret the formal sentences in poetry as a temporary measure. In fact, a complete sentence in poetry, a paragraph defined by a self-contained meaning and distinct from other meanings, often spans several formal sentences. Moreover, throughout the entire poem, it's as if the entire poem is a single word, and all the Chinese characters are morphemes, like a series of language gestures to pass over different

[8] By altering the node settings and removing directional information from the nodes while filtering for the situation where the subsequent character in the upper pair can be exchanged with the previous character in the lower pair, you can identify all the Chinese characters whose upper and lower character positions can be interchanged : 三-重, 雲-鶴, 別-鶴, 梅-欺, 槐-綠, 市-槐, 清-香, 月-落, 明-月, 月-桂, 夜-月, 繅-風, 吹-風, 搖-風, 從-風, 星-流, 星-繁, 星-編, 停-雲, 雲-飛, 暖-風, 綺-霞, 日-觀, 滴-雨, 煙-磨, 洞-雲, 天-行, 管-絃, 弓-張, 賓-鴻, 紫-蓋, 飛-高, 花-黃, 銀-黃, 合-璧, 振-玉, 玉-碎, 屑-玉, 玉-縷, 玉-白, 烏-金, 紫-金, 楓-江, 燕-玉, 飛-龍, 山-飛, 飛-鸞, 破-鏡, 珠-連, 垂-珠, 寶-珠, 國-香, 雪-香, 珠-綴, 散-珠, 文-龍, 織-蒲, 翔-鸞, 盤-龍, 作-解, 凝-烟, 簪-花, 白-花, 吟-秋, 木-落, 沙-眠, 濃-陰, 紫-苔, 戀-蝶, 戲-蝶, 盤-鴉, 水-面, 梅-熟, 化-女, 化-雀, 紅-蓼, 殘-紅, 紅-醉, 帶-露, 三-春, 祖-竹, 嘯-竹, 乳-花, 柳-舒, 柳-細, 柳-穿, 紅-輕, 流-翠, 丹-書, 埋-雪, 草-黃, 新-詩, 冰-彈, 綠-葉, 紅-苞, 寒-燈, 舞-鶴, 停-針, 白-薤, 立-竹, 塢-竹, 毛-竹, 徂-暑, 交-衣, 薦-酒, 吟-蛩, 密-葉, 子-白, 駕-鶴, 怨-鶴, 炭-獸, 奔-崖, 中-虛, 馬-馳, 衣-錦, 孫-桐, 化-魚, 狐-疑, 丹-心, 池-鳳, 棲-鳳, 調-鼎, 調-鹽, 勁-節, 批-鱗, 細-腰, 螭-蟠, 身-輕, 心-虛, 翅-薄.

[9] These four terms are used by Liu Xie [劉勰] (465-521) in his famous work Wénxīn Diāolóng 文心雕龍 to create a concluding envoi at the end of the Bǐ-Xīng 比興 chapter, which explains how the metaphoric and metonymic forms encompass semantic categories.

semantic categories within a poetic body[10]. The word construction pattern is thus the key to access a poem's vision.[11]

Since poetic parallelism also relies on the compositional structure of classical verse, the question arises whether the combination of phonetics and the metaphorical structure of example pairs can generate analogical relationships. Can the pronunciation's place and manner of articulation of phonemes be related to the metonymic mechanism of morphemes or morphemic components within word constructions to produce such wide-ranging yet nuanced connections? As we continue to build the Cilin Lexical Database based on the reconstructed model of classical morphemic relationships, these questions, accompanied by the analysis process, will gradually lead us to systematic explanations.

Building on our investigation, word embedding practices in Chinese studies prompt a crucial reevaluation. It's evident that our understanding of Chinese word meanings calls for a shift from surface-level word embedding to a deeper exploration of morphemes in text mining, including the consideration of their positional meanings. Current word embedding techniques limit our understanding to the surface, hindering an in-depth grasp of the complexities in the Chinese language. Annotating morphemes—examining phonemes, lexemes, and their positions and positional meanings—promises a more profound comprehension, overcoming the constraints of word-centric methods.

[10] Zheng, Y.-Y. [鄭毓瑜]. (2017). 姿與言: 詩國革命新論. 麥田出版, 臺北. The author emphasizes the categorical association of words organized like functional organisms within one poetic body, including the universe, nature, people, the environment (p.67), and all physical and physiological activities of a poet when they pronounce words (pp.45-46).

[11] In the abstract of Liu, C.-L. [劉昭麟], Chang, W.-T., Chu, C.-T., & Zheng, T.-Y. (2023), Machine learning and data analysis for word segmentation of classical Chinese poems: Illustrations with Tang and Song examples, published in *Digital Scholarship in the Humanities*, the authors discuss the word's crucial role in accessing a Chinese classical poem.

References

1. Liu, C.-L. [劉昭麟], Chang, W.-T., Chu, C.-T., & Zheng, T.-Y. (2023). Machine learning and data analysis for word segmentation of classical Chinese poems: Illustrations with Tang and Song examples. *Digital Scholarship in the Humanities*. DOI : https://doi.org/10.1093/llc/fqad073

2. Tang, W.-L. [湯文璐] (1983). 增廣詩韻全璧 (民國六年上海廣益書局本). 華正書局. 臺北.

3. Hsu C.-M. [許長謨] (2010). 漢語語言結構義證: 理論與教學應用. 里仁書局, 臺北.

4. Zheng, Y.-Y. [鄭毓瑜] (2017). 姿與言: 詩國革命新論. 麥田出版, 臺北.