Characteristics and Applications of Collocational Strength Calculation Formulas: A Case Study in International Chinese Language Education (Post-print)

Authors: Zhang Yongwei, Liang Jingzhi

Date: 2025-06-24T00:00:00+00:00

Abstract

[Purpose/Significance] This study analyzes the characteristics and performance differences of collocation strength calculation formulas in automatic extraction of Chinese window collocations and dependency collocations, aiming to provide references for Chinese collocation research and international Chinese education. [Method/Process] Seven typical collocation strength calculation formulas were selected to extract window collocations and dependency collocations for 60 typical words from an authentic corpus. After inviting experts for scoring validation, the performance of different formulas was analyzed. [Results/Conclusion] For international Chinese education, the formulas Dice coefficient, MI3, and log-likelihood ratio performed well in collocation extraction, while mutual information and collocation word frequency performed poorly. The precision of dependency collocation extraction was generally higher than that of window collocation, and using MI3 and Dice coefficient could achieve the highest recall rate, though it remained difficult to reach 100%. The research results provide a basis for the selection of collocation strength calculation formulas and the development of collocation extraction tools.

Full Text

Features and Applications of Collocation Strength Calculation Formulas: Taking International Chinese Language Education as an Example

Zhang Yongwei[1], **Liang Jingzhi**[2]
[1]Corpus and Computational Linguistics Research Center, Institute of Linguistics, Chinese Academy of Social Sciences, Beijing 100732, China

[2]School of Global Education & Development, University of Chinese Academy of Social Sciences, Beijing 100102, China

Abstract:
[**Purpose/Significance**] This study aims to analyze the characteristics and performance differences of collocation strength calculation formulas in the automatic extraction of Chinese window-based collocations and dependency-based collocations, providing references for Chinese collocation studies and international Chinese language education. [**Method/Process**] Seven typical collocation strength calculation formulas were selected to extract window-based collocations and dependency-based collocations for 60 representative words from authentic corpora. Following expert scoring validation, the performance of different formulas was analyzed. [**Result/Conclusion**] Regarding international Chinese language education, formulas such as Dice coefficient, MI^3, and log-likelihood ratio performed well in collocation extraction, while mutual information and collocate frequency showed poor performance. The precision of dependency-based collocation extraction was generally higher than that of window-based collocation extraction. Furthermore, the simultaneous use of MI^3 and Dice coefficient achieved the highest recall rates but still could not reach 100%. These findings provide a basis for selecting collocation strength calculation formulas and developing collocation extraction tools.

Keywords: Window-based collocation; Dependency-based collocation; Collocation strength calculation formula; Corpus

---

Collocation is a repeatedly occurring word combination characterized by arbitrariness, structural properties, and domain relevance, reflecting habitual expressions in language [1-2]. One can understand the meaning of a word through the other words that habitually co-occur with it [3]. As language is the carrier of culture, collocation embodies the organizational patterns of language and contains rich cultural information. In international Chinese language education, systematic collocation teaching can enhance the quality of Chinese vocabulary instruction.

A collocation consists of a node word (the focus of teaching and research) and a collocate (the co-occurring words that facilitate learning and understanding of the node word). When extracting collocations, collocation strength is used to measure the tightness of the relationship between the node word and collocate —the greater the strength, the tighter the relationship. Collocation strength calculation formulas (also called association measures) are mathematical formulas used to calculate the magnitude of collocation strength. Many corpus analysis tools provide automatic collocation retrieval functions, which represent one of the core features of such tools [4].

Collocations can be categorized as expert collocations (manually extracted) and automatic collocations (computer-extracted). Expert collocation extraction relies on expert knowledge and experience, yielding high-quality results but suf-

fering from subjectivity and being time-consuming. Automatic collocation extraction offers objectivity and efficiency, free from expert subjectivity, but faces issues such as insufficient coverage and lower quality. How to objectively and efficiently extract high-quality collocations that gradually approach expert collocations has long been a key research topic in corpus linguistics, with collocation strength calculation formulas being one of the most central issues, aiming to simulate expert judgment of collocations using statistical methods.

In the field of international Chinese language education, non-native learners lack rich linguistic background knowledge, making vocabulary learning and usage particularly challenging. Effective word collocations help learners more comprehensively understand and master the meanings and usage of Chinese vocabulary. However, existing collocation extraction tools typically provide multiple collocation strength calculation formulas without offering guidance on formula selection, leaving users struggling to make appropriate choices. Therefore, this study examines three key questions in automatic collocation extraction: (1) What are the characteristics of commonly used collocation strength calculation formulas? (2) How do different formulas relate to each other—which show high similarity and which show significant differences? (3) How should one select collocation strength calculation formulas in the context of international Chinese language education? The findings will help international Chinese educators and learners choose appropriate formulas based on actual needs, improving collocation extraction efficiency and accuracy, and enhancing vocabulary teaching and learning quality.

The paper is organized as follows: Section 1 reviews related research on automatic collocation extraction, Section 2 details the automatic collocation extraction and expert validation experiments, Section 3 analyzes automatic collocation results against expert scoring, and the final section presents conclusions.

1.1 Overview of Automatic Collocation Extraction Methods

There are four methods for automatic extraction of binary word collocations: window-based methods extract co-occurring words within a specified distance from the target word; grammar-based methods use syntactic parsing to extract words with specific grammatical relationships to the target word; semantics-based methods employ semantic information, synonym substitution, and translation consistency to determine semantic associations between words and filter candidate collocations; and classification-based methods integrate features from the above three approaches, combining machine learning algorithms to classify candidate collocations and determine whether they constitute collocations with the node word [5]. When extracting large numbers of collocations, experts must further identify typical collocations. To improve expert identification efficiency, collocation strength calculation and ranking/filtering mechanisms can be introduced to provide a basis for expert judgment.

Window-based and dependency-based methods extract what are respectively

called window collocations and dependency collocations, which have been extensively studied. Many corpus analysis tools support automatic extraction of both types, including CQPWeb, Sketch Engine, English Corpora, WordSmith, AntConc, the Dependency Collocation Search System (DCS) [6], and Hanyu Zhuyan for window collocations; and CCA Chinese Collocation Assistant [7] and DCS for dependency collocations.

1.2 Overview of Collocation Strength Calculation Formulas

Collocation strength calculation formulas directly affect automatic collocation extraction effectiveness. When direct extraction yields numerous collocations but only a few typical ones are needed for teaching and learning, collocation strength calculation becomes particularly important. The degree to which automatic collocations can replace expert collocations serves as a crucial metric for evaluating formula effectiveness.

Wermter and Hahn [8] categorized collocation strength calculation formulas into frequency-based, entropy-based, and statistics-based methods. Many collocation extraction tools provide multiple formulas. For instance, CQPWeb supports nine formulas: mutual information (MI) [9], log-likelihood ratio (LLR) [10], MI^3 [11], t-score [12], z-score [13], Dice ratio [14], log ratio, conservative LR, and rank frequency. The DCS system supports nine formulas: pointwise mutual information, square mutual information (SMI) [15], t-score, log ratio, log-likelihood ratio, Dice coefficient (Dice' s coefficient, i.e., Dice ratio), relative frequency, co-occurrence frequency, and collocate frequency [6].

Previous research on collocation strength calculation formulas typically extracts all binary pairs from corpora first, then applies one or multiple formulas to quantify and evaluate these pairs, focusing on high-scoring collocations. These studies fall into two categories: those using characters as the basic unit, extracting character binary pairs from corpora to analyze formula characteristics by observing word formation [16-17]; and those using words as the basic unit, extracting word binary pairs to evaluate their association degree and identify word collocations [15,18-21]. Additionally, there are few comparative studies on collocation strength calculation methods based on fixed node words. For example, Liang Jingzhi [22] compared ten collocation strength calculation formulas, but only selected 20 collocates per node word—a relatively small number—without exploring the effects of using multiple formulas simultaneously. Notably, collocations serve diverse purposes with different evaluation criteria, necessitating more targeted research on formula characteristics for different applications.

2.1.2 Calculation Formulas

Liang Jingzhi [22] analyzed nine common collocation extraction tools and found they collectively support 30 collocation strength calculation formulas. Among these, mutual information, log-likelihood ratio, MI^3, t-score, z-score, and Dice coefficient (including its variants) are supported by at least six tools, making

them the most widely used. This study selected calculation formulas based on the principle of combining representativeness with diversity, considering both widespread support and different formula types. Therefore, from entropy-based methods we selected the most widely supported mutual information and MI^3; from statistics-based methods we selected t-score, log-likelihood ratio, and Dice coefficient. Although frequency-based methods are not universally supported, collocate frequency and co-occurrence frequency are commonly used to measure collocate typicality, so we included these two frequency-based methods. The final seven representative formulas are shown in Table 1 .

Table 1 Details of Collocation Strength Calculation Formulas
[Table content with formulas would appear here]

Note: Formula and symbol definitions reference Sketch Engine online documentation . Where f_A is node word frequency, f_B is collocate frequency, f_{AB} is co-occurrence frequency, N is corpus size, and xlx(N) is fl n (f).

2.1.3 Node Word Selection

To ensure representativeness, node words were selected using the following criteria: (1) To directly serve international Chinese language education practice, all candidate words were drawn from the *International Chinese Language Education Chinese Proficiency Grading Standards* vocabulary list. To avoid word length effects on subsequent expert scoring, only two-character words were selected. (2) To ensure sufficient usage value, only high-frequency words were selected. (3) To avoid ambiguity issues in collocation judgment caused by polysemy, only monosemous words were selected. (4) Based on modern Chinese part-of-speech system characteristics, 20 representative words were selected from each of the three main categories—nouns, verbs, and adjectives—as node words.

In the actual experiment, word frequency data were obtained through statistical analysis of the segmented and annotated CCL Modern Chinese Corpus, while word sense counts were determined based on the *Modern Chinese Dictionary* (7th edition). The final 60 node words are detailed in Table 2 .

Table 2 List of Node Words
Nouns: Government, department, product, president, policy, bank, expert, method, price, event, reason, countryside, approach, opportunity, work, hospital, athlete, industry, eyes, mother
Verbs: Believe, know, hold, improve, continue, include, realize, increase, obtain, achieve, cause, produce, announce, implement, expand, trust, reduce, see, consider, leave
Adjectives: Important, obvious, famous, extensive, excellent, complex, evident, thorough, significant, enthusiastic, unique, accurate, detailed, outstanding, pleasant, difficult, lovely, careful, interesting, precious

Note: The average frequency of the 60 node words is 61,578.78, with "precious" having the lowest frequency (7,148) and "government" the highest (253,572).

The experiment used seven formulas to extract the 50 highest-scoring dependency collocations and 50 window collocations for each of the 60 node words, yielding 42,000 collocation entries. To ensure collocations had sufficient usage and to highlight formula characteristics, the raw co-occurrence frequency threshold was set at $ $2.

For window collocation extraction, collocate part-of-speech was distinguished with a window size of 5. To facilitate expert validation while simplifying window collocation information, the experiment recorded whether collocates appeared to the left or right of the node word but not their specific positions within the window. For dependency collocation extraction, both collocate part-of-speech and dependency relations were distinguished. Using Dice coefficient as an example, the 10 highest-scoring window collocations and 10 dependency collocations for "opportunity" are shown in Tables 3 and 4 respectively.

Table 3 High-Strength Window Collocations for "机会" (Dice Coefficient)
[Table content would appear here]

Note: Collocate form and part-of-speech are separated by "/" . Left frequency indicates occurrences to the left of the node word; right frequency indicates occurrences to the right. Left + right frequency = co-occurrence frequency.

Table 4 High-Strength Dependency Collocations for "机会" (Dice Coefficient)
[Table content would appear here]

Note: Collocate form, part-of-speech, and dependency relation are separated by "/" .

2.2 Expert Validation

Six master' s students in international Chinese language education served as experts to score automatically extracted collocations using a 5-point scale: 5 points for definite collocations requiring teaching; 4 points for probable collocations but not necessarily requiring teaching; 3 points for uncertain collocations that could be omitted from teaching; 2 points for probable non-collocations with no teaching value; and 1 point for definite non-collocations that would burden teaching. "Requiring teaching" means the collocation meets international Chinese education needs, is common, and helps learners understand the node word. "Not necessarily requiring teaching" means that while it may be a collocation, mastery is not essential for learners or can be acquired through other means. "Could be omitted" means the collocation' s validity is questionable and its teaching value is low, so excluding it would not impact learners.

During scoring, collocations were evaluated independently without influence from other collocations or high-scoring quotas. Collocations had to meet international Chinese education needs, with collocates being common and helpful for learning the node word. For example, "mining management department" as a collocation for "department" and "reduce 38.04 million" for "reduce" were both scored low as they do not facilitate node word learning.

Among the 42,000 extracted collocation entries, 10,901 remained after deduplication, averaging 181.68 collocations per node word. Expert scoring statistics are shown in Table 5 .

Table 5 Expert Scoring Details
[Table content would appear here]

The study conducted reliability testing on the six experts' scores, yielding a Cronbach' s Alpha coefficient of 0.918, far exceeding the acceptable standard of 0.7, indicating high consistency and reliability suitable for subsequent analysis. Table 5 shows an inverted pyramid distribution in automatically extracted collocation quality, with low-quality collocations comprising the majority and high-quality collocations being scarce. Overall quality was low: 43.280% scored below 1 point, with a cumulative 72.204% below 2 points, meaning most automatically extracted collocations are unsuitable or of limited value for teaching. In contrast, high-quality collocations ($ $4 points) accounted for only 6.174%, with just 0.734% deemed essential for teaching, reflecting the scarcity of high-quality collocations. The average distribution further confirms this, with only 1.33 collocates per node word considered essential for teaching by all experts (score=5).

Given this scarcity, collocations scoring $ $4 points were treated as expert collocations, ensuring reasonable quantity while maintaining quality. The experiment assumed expert collocations were included in the automatically extracted set, which forms the premise for subsequent comparative analysis. Additionally, these results highlight the importance of precise collocation screening in international Chinese vocabulary teaching, with high-scoring expert collocations providing a reference for analyzing formula characteristics and evaluating automatic extraction performance.

3.1 Frequency Characteristics Analysis

This study analyzed the mean frequency of high-frequency collocates, mean collocation frequency, and their ratio, where a smaller ratio indicates greater dependency of collocate usage on the node word. Frequency information for window and dependency collocations is shown in Tables 6 and 7 respectively.

Table 6 High-Frequency Collocation Frequency Information (Window Collocations)
[Table content would appear here]

Table 7 High-Frequency Collocation Frequency Information (Dependency Collocations)
[Table content would appear here]

Tables 6 and 7 reveal the following frequency characteristics for each formula:

(1) **Mutual Information Formula**: For window collocations, both mean collocate frequency (3.02) and mean collocation frequency (3.23) are very

small, with a ratio of 0.94 , indicating a tendency to select low-frequency collocates whose usage heavily depends on the node word. For dependency collocations, both values increase significantly (66.70 and 16.07 respectively), with a ratio of 4.15, showing that dependency relations lead to selection of higher-frequency collocates. However, compared to other formulas, mutual information yields the smallest ratio for both collocation types, demonstrating its tendency to select collocates whose usage severely depends on the node word. This characteristic gives mutual information unique advantages in extracting rare but potentially significant collocations.

(2) **MI^3, T-score, and Log-Likelihood Ratio Formulas**: These three formulas maintain large collocation frequencies and ratios for both window and dependency collocations, showing good balance in extracting high-frequency, strongly associated collocations. They tend to select high-frequency collocates while ensuring high co-occurrence frequency, effectively reflecting both general usage patterns and tight word relationships.

(3) **Dice Coefficient Formula**: This yields relatively lower mean collocate and collocation frequencies, indicating a preference for relatively stable but not necessarily high-frequency collocations, capturing linguistic phenomena that traditional high-frequency methods might miss. Its frequency ratio falls between mutual information and other formulas for both types, reflecting a balanced strategy that considers both collocate independence and node-collocate co-occurrence frequency.

(4) **Collocate Frequency and Collocation Frequency Formulas**: Contrary to mutual information, collocate frequency selects the highest-frequency words, while collocation frequency selects words with highest co-occurrence with the node word. These formulas show the largest frequency ratios, indicating their extracted collocates are less dependent on the node word.

Comparing window and dependency collocations reveals that dependency collocations have generally lower mean frequencies but higher ratios, suggesting that while less frequent overall, they better capture specific relationships between words.

3.2.1 Precision of Different Calculation Formulas

Precision (P) and recall (R) were used to evaluate extraction quality, with P@n and R@n representing precision and recall for the top n collocations respectively, calculated using formulas (1) and (2).

[Formulas (1) and (2) would appear here]

In these formulas, "correct collocations" refer to those with average expert scores $ $4. Precision measures extraction accuracy—the proportion of correct collocations among extracted ones. Recall measures extraction completeness—the

proportion of expert collocations successfully extracted. Calculating precision and recall at various n values provides comprehensive performance reflection.

When collocation number n increased from 5 to 50 in steps of 5, precision for different formulas is shown in Figure 2 [Figure 2: see original paper].

Figure 2 Precision of Collocation Extraction
[Figure would appear here]

Figure 2 shows significant precision differences among formulas. As collocation number increases, most formulas show declining precision, but with varying degrees and patterns. Dice coefficient performs best overall, especially with smaller collocation numbers. MI^3 and log-likelihood ratio show similar declining trends, with MI^3 slightly better. T-score and collocation frequency formulas have relatively low but stable precision (9.67%-14.13%). Mutual information and collocate frequency perform worst, far below practical requirements. Notably, mutual information' s precision can be improved by raising the minimum frequency threshold.

For dependency collocation extraction, all formulas show similar precision trends, but with significantly higher values overall, demonstrating the effectiveness of dependency relations in improving precision. Specifically, Dice coefficient performs best for small collocation numbers (5-20). As numbers increase, MI^3 and log-likelihood ratio maintain high precision in medium ranges (25-40), joining Dice coefficient as the top three formulas. T-score and collocation frequency, though slightly inferior, remain stable and outperform their window collocation counterparts. Mutual information and collocate frequency improve with increasing numbers but remain the lowest.

From a precision perspective, Dice coefficient, MI^3, and log-likelihood ratio are recommended; collocate frequency and mutual information with low-frequency thresholds are not recommended.

3.2.2 Recall of Different Calculation Formulas

Recall rates for different formulas when n increased from 5 to 50 are shown in Figure 3 [Figure 3: see original paper].

Figure 3 Recall of Collocation Extraction
[Figure would appear here]

Figure 3 shows that Dice coefficient, MI^3, and log-likelihood ratio excel in both collocation types, with recall improving significantly as collocation number increases. For window collocations, Dice coefficient performs best; for dependency collocations, MI^3, log-likelihood ratio, and Dice coefficient perform similarly well. T-score and collocation frequency show moderate performance with steady but less impressive improvement. Mutual information and collocate frequency perform poorly for both types, with low recall even as numbers increase. Most formulas show slowing improvement after reaching 20-30 collocations.

Among all 50 extracted collocations, the highest recall achieved by a single formula is 95.12% (Dice coefficient) for window collocations and 85.54% (MI^3) for dependency collocations. From a recall perspective, Dice coefficient, MI^3, and log-likelihood ratio remain recommended; collocate frequency and low-threshold mutual information are not recommended. Comparing Figures 2 and 3 reveals significant performance differences among formulas.

3.3 Correlation Analysis of Calculation Formulas

This study calculated consistency between collocations extracted by different formulas as a correlation measure. Correlation heatmaps (Figures 4 [Figure 4: see original paper] and 5 [Figure 5: see original paper]) visualize these relationships, with cell color intensity and numerical values (0.00-1.00) indicating correlation strength. To analyze the impact of collocation number, heatmaps were generated based on top 25 (n=25) and top 50 (n=50) collocations.

Figure 4 Correlation Heatmap of Calculation Formulas (Window Collocations)
[Figure would appear here]

Figure 4 shows light-colored regions for mutual information, indicating extremely low correlation with other formulas. Collocate frequency also shows mostly light colors, with only a moderately deep cell with collocation frequency, showing medium correlation. The central dark region comprising MI^3, T-score, log-likelihood ratio, and Dice coefficient indicates high correlations among them, particularly between T-score and collocation frequency, and between MI^3 and log-likelihood ratio. Dice coefficient shows moderate correlation, primarily with MI^3 and log-likelihood ratio. Overall, the color pattern remains consistent between n=25 and n=50, suggesting collocation number has limited impact on formula correlations for window collocations.

Figure 5 Correlation Heatmap of Calculation Formulas (Dependency Collocations)
[Figure would appear here]

Figure 5 shows that for dependency collocations, mutual information and collocate frequency have extremely low correlations with other formulas, while MI^3, T-score, log-likelihood ratio, and Dice coefficient show high inter-correlations. Compared to window collocations, collocate frequency' s correlations decrease while Dice coefficient' s correlations increase, indicating that collocation type significantly affects formula relationships. Formula selection should therefore consider specific research purposes and collocation types.

3.4 Recall Using Two Formulas Simultaneously

Single-formula recall rates show maximum values of 95.12% (window) and 85.54% (dependency) at n=50. In practice, to quickly build comprehensive collocation databases, multiple formulas are often used simultaneously. This study analyzed recall rates when using two formulas together.

Given collocate frequency' s extremely low recall, it was excluded from analysis. The remaining six formulas were paired, with extracted collocations merged equally. Recall rates for window and dependency collocations when each formula extracted 5-50 collocations are shown in Figure 6 [Figure 6: see original paper].

Figure 6 Recall of Collocation Extraction (Using Two Formulas)
[Figure would appear here]

Note: When both formulas extracted the same correct collocation, it was counted only once.

Figure 6 shows all curves trending upward for both types, indicating improved recall with increased collocation numbers. For window collocations, the MI^3&Dice combination shows the steepest curve, reaching 98.22% recall, while for dependency collocations, MI^3&Dice also performs best at 91.41%—improvements of 3.10% and 5.87% over single-formula maximums. This demonstrates window collocations' advantage in recall. Larger gaps between curves for window collocations indicate formula selection has greater impact, while dependency collocation curves are more concentrated, showing less sensitivity to formula choice.

For window collocations, combinations including Dice coefficient generally achieve higher recall. For dependency collocations, combinations including MI^3 perform better. Using both simultaneously yields the highest recall for each type. Some combinations show large performance differences between types—for example, T-score&Dice works well for window but ranks lower for dependency collocations. Therefore, formula combination selection must consider collocation type. Figure 6 also shows that even when extracting 50 collocations per formula, no combination reaches 100% recall, suggesting that increasing numbers or using formula combinations may not fully solve the recall problem. In practice, merging window and dependency results or exploring more efficient formula combinations may be necessary for comprehensive extraction.

This study reveals the characteristics and performance differences of various collocation strength calculation formulas in extracting window and dependency collocations. Through correlation heatmaps, it visually demonstrates inter-formula relationships. The analysis of single-formula precision/recall and two-formula recall provides systematic performance comparisons.

These findings are significant not only for Chinese corpus linguistics but also for collocation teaching and research in international Chinese education. By understanding formula characteristics and performance differences, researchers and educators can more effectively select and apply formulas, improving extraction effectiveness and vocabulary teaching outcomes. The study also provides theoretical foundations for developing more efficient and accurate Chinese collocation extraction tools.

Future work includes: (1) Expanding formula analysis scope; (2) Using larger-scale, more diverse corpora; (3) Broadening word selection across more parts of speech and frequency ranges; (4) Exploring formula improvements based on

current analysis.

References

[1] Sinclair J. Corpus Concordance Collocation[M]. Oxford: Oxford University Press, 1991.
[2] Sun Maosong, Huang Changning, Fang Jie. A Quantitative Analysis of Chinese Collocations[J]. Chinese Language, 1997(1): 29-38.
[3] Firth J R. Papers in Linguistics 1934-1951[M]. Oxford: Oxford University Press, 1957.
[4] Zhang Yongwei, Wu Bingxin. Review of Core Functions of Fourth-Generation Web-Based Corpus Analysis Tools[J]. Contemporary Linguistics, 2023, 25(4).
[5] Wong K F, Li W, Xu R, et al. Introduction to Chinese natural language processing[M]. San Rafael: Morgan & Claypool Publishers, 2009.
[6] Zhang Yongwei, Ma Qiongying. Research on a Dependency Collocation Retrieval System for Chinese Dictionary Compilation[J]. Lexicographical Studies, 2022(4): 30-40, 125.
[7] Hu Renfen, Xiao Hang. Research on Construction and Application of Chinese Collocation Knowledge Base for Second Language Teaching[J]. Applied Linguistics, 2019(1).
[8] Wermter J, Hahn U. Collocation extraction based on modifiability statistics[C]//Proceedings of the 20th International Conference on Computational Linguistics. 2004: 980-986.
[9] Church K, Hanks P. Word association norms, mutual information, and lexicography[J]. Computational Linguistics, 1990, 16(1): 22-29.
[10] Dunning T E. Accurate methods for the statistics of surprise and coincidence[J]. Computational Linguistics, 1993, 19(1): 61-74.
[11] Oakes M P. Statistics for Corpus Linguistics[M]. Edinburgh: Edinburgh University Press, 1998.
[12] Church K, Gale W A, Hanks P, et al. Using statistics in lexical analysis[M]//Zernik U, ed. Lexical acquisition: exploiting on-line resources to build a lexicon. New York: Psychology Press, 1991: 115-164.
[13] Berry-Rogghe G. The computation of collocations and their relevance in lexical studies[M]//Aitken A, Bailey R, Hamilton-Smith N. The Computer and Literary Studies. Edinburgh: Edinburgh University Press, 1973: 103-112.
[14] Dice L R. Measures of the amount of ecologic association between species[J]. Ecology, 1945, 26(3): 297-302.
[15] Zhang H, Zhang Y, Yu J. Collocation extraction using square mutual information approaches[J]. International Journal of Knowledge and Language Processing, 2011, 2(1): 53-58.
[16] Sproat R, Shih C. A statistical method for finding word boundaries in Chinese text[J]. Computer Processing of Chinese & Oriental Languages, 1990, 4(4): 336-351.
[17] Luo Shengfen, Sun Maosong. Experimental Research on Automatic Chinese Word Extraction Based on Internal Binding Strength of Character

Strings[J]. Journal of Chinese Information Processing, 2003(3): 9-14.
[18] Sun Jian, Wang Wei, Zhong Yixin. A Statistical Method for Discovering Common Word Collocations[J]. Journal of the China Society for Scientific and Technical Information, 2002(1): 12-16.
[19] Wang Daliang, Zhang Dezheng, Tu Xuyan, et al. A Collocation Extraction Method Based on Relative Conditional Entropy[J]. Journal of Beijing University of Posts and Telecommunications, 2007, 30(6): 40-45.
[20] Qian Y. Dynamism of collocation in L2 English writing: a bigram-based study[J]. International Review of Applied Linguistics in Language Teaching, 2022, 60(2): 339-360.
[21] Su Q, Gu C, Liu P. Association measures for collocation extraction: automatic evaluation on a large-scale corpus[J]. International Journal of Corpus Linguistics, 2024, 29(1): 59-86.
[22] Liang Jingzhi. Characteristics of Collocation Strength Calculation Formulas and Their Implications for International Chinese Language Education[D]. Beijing: University of Chinese Academy of Social Sciences, 2024.

Notes:

http://ccl.pku.edu.cn:8080/ccl_{corpus}
https://brat.nlplab.org/index.html
http://ltp.ai/docs/appendix.html
https://www.sketchengine.eu/documentation/statistics-used-in-sketchengine/

A collocation consists of a node word and collocate. Besides the node word, collocates can co-occur with other words, so collocate frequency is usually not less than collocation frequency. However, in window collocations, when a collocate appears between two identical node words, collocate frequency may be less than collocation frequency, resulting in a ratio <1. For example, when extracting collocations from "···important trade port, which is of important significance for the development of Swahili culture···" with "important" as node word and "Swahili" as collocate, "Swahili" frequency is 1 while its collocation frequency with "important" is 2.

The tendency of mutual information to select low-frequency collocations in this experiment is related to the minimum frequency threshold of 2. Different threshold settings significantly affect mutual information' s extraction results for window collocations.

Note: Figure translations are in progress. See original paper for figures.